Final Version of our investigations into:

Report on dangers and opportunities posed by large search engines, particularly Google

September 30, 2007

H. Maurer, Co-author, editor and responsible for the project, Institute for Information Systems and Computer Media, Graz University of Technology

Co-authors in alphabetical order:

Dipl. Ing. Dr. Tilo Balke, L3S Hannover

Professor Dr. Frank Kappe, TU Graz

Dipl. Ing. Narayanan Kulathuramaiyer, TU Graz

Priv. Dozent Dr. Stefan Weber, Uni Salzburg

Dipl. Ing. Bilal Zaka, TU Graz

We want to acknowledge input from Professor Dr. Nick Scerbakov, Univ. Doz. Dr. Denis Helic, both TU Graz, and we want to thank Mrs. M. Lampl for her infinite patience with various versions of this work and her formatting it into a somewhat coherent document.

We want to particularly thank the Austrian Federal Ministry of Transport, Innvoation and Technology (Bundesministerium für Verkehr, Innovation und Technologie) for the financial support of this report.

Table of Contents

Overview

The aim of our investigation was to discuss exactly what is formulated in the title. This will of course constitute a main part of this write-up. However, in the process of investigations it also became clear that the focus has to be extended, not to just cover Google and search engines in an isolated fashion, but to also cover other Web 2.0 related phenomena, particularly Wikipedia, Blogs, and other related community efforts.

It was the purpose of our investigation to demonstrate:

- Plagiarism and IPR violation are serious concerns in academia and in the commercial world
- Current techniques to fight both are rudimentary, yet could be improved by a concentrated initiative
- One reason why the fight is difficult is the dominance of Google as THE major search engine and that Google is unwilling to cooperate
- The monopolistic behaviour of Google is also threatening how we see the world, how we as individuals are seen (complete loss of privacy) and is threatening even world economy (!)

In our proposal we did present a list of typical sections that would be covered at varying depth, with the possible replacement of one or the other by items that would emerge as still more important.

The preliminary intended and approved list was:

Section 1: To concentrate on Google as virtual monopoly, and Google's reported support of Wikipedia. To find experimental evidence of this support or show that the reports are not more than rumours.

Section 2: To address the copy-past syndrome with socio-cultural consequences associated with it.

Section 3: To deal with plagiarism and IPR violations as two intertwined topics: how they affect various players (teachers and pupils in school; academia; corporations; governmental studies, etc.). To establish that not enough is done concerning these issues, partially due to just plain ignorance. We will propose some ways to alleviate the problem.

Section 4: To discuss the usual tools to fight plagiarism and their shortcomings.

Section 5: To propose ways to overcome most of above problems according to proposals by Maurer/Zaka. To examples, but to make it clear that do this more seriously a pilot project is necessary beyond this particular study.

Section 6: To briefly analyze various views of plagiarism as it is quite different in different fields (journalism, engineering, architecture, painting, …) and to present a concept that avoids plagiarism from the very beginning.

Section 7: To point out the many other dangers of Google or Google-like undertakings: opportunistic ranking, analysis of data as window into commercial future.

Section 8: To outline the need of new international laws.

Section 9: To mention the feeble European attempts to fight Google, despite Google's growing power.

Section 10. To argue that there is no way to catch up with Google in a frontal attack.

Section 11: To argue that fighting large search engines and plagiarism slice-by-slice by using dedicated servers combined by one hub could eventually decrease the importance of other global search engines.

Section 12: To argue that global search engines are an area that cannot be left to the free market, but require some government control or at least non-profit institutions. We will mention other areas where similar if not as glaring phenomena are visible.

Section 13: We will mention in passing the potential role of virtual worlds, such as the currently over-hyped system "second life".

Section 14: To elaborate and try out a model for knowledge workers that does not require special search engines, with a description of a simple demonstrator.

Section 15 (Not originally part of the proposal): To propose concrete actions and to describe an Austrian effort that could, with moderate support, minimize the role of Google for Austria.

Section 16: References (Not originally part of the proposal)

In what follows, we will stick to Sections 1 -14 plus the new Sections 15 and 16 as listed, plus a few Appendices.

We believe that the importance has shifted considerably since the approval of the project. We thus will emphasize some aspects much more than ever planned, and treat others in a shorter fashion. We believe and hope that this is also seen as unexpected benefit by BMVIT.

This report is structured as follows:

After an Executive Summary that will highlight why the topic is of such paramount importance we explain in an introduction possible optimal ways how to study the report and its appendices. We can report with some pride that many of the ideas have been accepted by the international scene at conferences and by journals as of such crucial importance that a number of papers (constituting the appendices and elaborating the various sections) have been considered high quality material for publication.

We want to thank the Austrian Federal Ministry of Transport, Innovation and Technology (BMVIT) for making this study possible. We would be delighted if the study can be distributed widely to European decision makers, as some of the issues involved do indeed involve all of Europe, if not the world.

Executive Summary: The Power of Google and other Search engines

For everyone looking at the situation it must be clear that Google has amassed power in an unprecedented way that is endangering our society.

Here is a brief summary:

Google as search engine is dominating (Convincing evidence on this is easily available and presented in Section 1). That on its own is dangerous, but could possibly be accepted as “there is no real way out”, although this is not true, either. (We would rather see a number of big search engines run by some official non-profit organisations than a single one run by a private, profit driven company.) However, in conjunction with the fact that Google is operating many other services, and probably silently cooperating with still further players, this is unacceptable.

The reasons are basically:

– Google is massively invading privacy. It knows more than any other organisation about people, companies and organisations than any institution in history before, and is not restricted by national data protection laws.

– Thus, Google has turned into the largest and most powerful detective agency the world has ever known. I do not contend that Google has started to use this potential, but as commercial company it is FORCED to use this potential in the future, if it promises big revenue. If government x or company y is requesting support from Google for information on whatever for a large sum, Google will have to comply or else is violating its responsibilities towards its stockholders.

– Google is influencing economy by the way advertisements are ranked right now: the more a company pays, the more often will the add be visible. Google answers that result from queries are also already ranked when searches are conducted (we give strong evidence for this in Section 1): Indeed we believe it cannot avoid ranking companies higher in the future who pay for such improved ranking: Google is responsible to stockholders to increase the company’s value. Google is of course doing this already for ads.

– Since most material that is written today is based on Google and Wikipedia, if those two do not reflect reality, the picture we are getting through “googeling reality” as Stephan Weber calls it, is not reality, but the Google-Wikipedia version of reality. There are strong indications that Google and Wikipedia cooperate: some sample statistics show that random selected entries in Wikipedia are consistently rated higher in Google than in other search engines.

– That biased contributions can be slipped into Wikipedia if enough money is invested is well established.

– Google can use its almost universal knowledge of what is happening in the world to play the stock market without risk: in certain areas Google KNOWS what will happen, and does not have to rely on educated guesses as other players in stock market have to. This is endangering trading on markets: by game theory, trading is based on the fact that nobody has complete information (i.e. will win sometimes, but also loose sometimes). Any entity that never looses rattles the basic foundations of stock exchanges!

– It has to be recognized that Google is not an isolated phenomenon: no society can leave certain basic services (elementary schooling, basic traffic infrastructure, rules on admission of medication,…) to the free market. It has to be recognized that Internet and the WWW also need such regulations, and if international regulations that are strong enough cannot be passed, then as only saving step an anti-Trust suite against Google has to be initiated, splitting the giant in still large companies, each able so survive, but with strict “walls” between them.

- It has to be recognized that Google is very secretive about how it ranks, how it collects data and what other plans it has. It is clear from actions in the past (as will be discussed in this report) that Google could dominate the plagiarism detection and IPR violation detection market, but chooses not to do so. It is clear that it has strong commercial reasons to act as it does.

- Google’s open aim is to “know everything there is to know on Earth”. It cannot be tolerated that a private company has that much power: it can extort, control, and dominate the world at will.

I thus call for immediate action, and some possibilities are spelt out in this report.

One word of warning is appropriate: This report originated from a deep concern about plagiarism using Google, about the Google Copy Paste syndrome, as one of the authors has called it. Consequently, this aspect is covered more deeply than some others in the main body of the paper, some economical and political issues are moved into independent appendices.

If the danger of Google is your main concern, skip now to Sections 7 and 8, otherwise read the introduction which basically explains how this voluminous document should be read depending on your preferences.

Hermann Maurer, Graz/Austria, September 30, 2007
hmaurer@iicm.edu   www.iicm.edu/maurer

Introduction : A guide how to read this document

As it has hopefully become clear in the executive summary, this report started with an important but still limited topic: fighting plagiarism and IPR violations. It lead rapidly into the study of threats posed by new methods of data-mining, employed by many organisations due to the fact that no international laws regulate such activities. The archetypical company that has turned this approach into a big success for itself but a veritable threat for mankind is Google.

As we discuss how to possibly minimize the power of Google and similar activities we implicitly show that there is also room for European endeavours that would strengthen the role of Europe and its economy.

The leader of this project team H. Maurer from Graz University of Technology was thus forced to put together a team consisting of both researchers and implementers (who would try out some ideas) in “no time” and to compile a comprehensive report on the current situation and possible remedies in much less than a year, although the undertaking at issue would have needed much more time, resources and manpower.

Thus we are proud to present a thorough evaluation including concrete recommendations in this report and its appendices, yet we realize that the necessary parallelism of work has created some redundancies.

However, we feel that these redundancies are actually helpful in the sense that this way the 15 Sections and the six Appendices we present can be read fairly independently. However, to give a bit of a guidelines we recommend to read the report, depending on what your main interests are, in different ways, and we are trying to suggest some as follows:

If you are mainly interested in how Google, and surprisingly Wikipedia (two organisations that work much closer together than they are willing to admit) are changing the way we do research and learn (and hence influence the fabric of society), and that somehow this trend should be reverted, then Sections 1-5 are the ones you should definitely start with.

If you are more interested in the concept of plagiarism and how it applies to IPR violations, we recommend to start with Appendix 1, with a more specific analysis in Appendix 2, and a totally different and new approach that we are still going to pursue vigorously in Section 6.

Those more concerned about the influence that Google (and data mining) has on privacy but also other issues may want to start with Appendices 4, 5 and 6.

Readers who want to learn why we consider Google a serious threat to economy, much beyond invasion of privacy, elimination of intermediaries, threatening traditional ways of advertising, etc. should start with Section 7, and have an extended look a the situation in Section 8.

We feel it is not sufficient to recognize the danger that Google poses, but we also need alternatives. We regret that Europe is not even taking up the challenge the search engine Google poses, let alone all the other threats, as we discuss in Section 9. We do propose specific measures how to minimize the impact of Google as search engine (and how to by-pass Google’s unwillingness to help with plagiarism detection) in Sections 10 and 11. We explain why data-mining is an area of vital interest to all of humanity, like the provision of water, elementary schooling, etc. in Section 12 and hence should be recognized as such by governments on all levels.

We take a quick look at other new developments on the Internet (particularly Second Life) and try to make a guess how they might correlate with other aspects of data-mining in Section 13.

We show in Section 14 that the tools we have proposed (e.g. the collaborative/ syndicated tools of Section 10) can also be used to make knowledge workers more efficient, an important aspect in the fight for economic prosperity in Europe.

In Section 15 we briefly discuss one approach that we are currently testing. If carried out with some support by government and sponsors on a European level as proposed in Section 11 it would not only reduce the power of Google as search engine dramatically, but create new jobs and economic value.

It remains to say that Appendix 3 has been put in as a kind of afterthought, to present some issues that we did not cover elsewhere, particularly the relation between plagiarism detection and IPR violation.

It is worth mentioning that we have collected much more data in the form of informal interviews or E-Mails than we dare to present since it would not stand up in court as acceptable evidence, if we are sued by a large data-mining company. However, this evidence that we cannot present here has increased our concern that Google is well on the way of tying to dominate the world, as one of the IT journalists of Austria (Reischl) has very courageously put it.

We hope that this collection of probes, tests and research papers that we have compiled under enormous time pressure due to what we see as very real danger will help to move some decision makers to take steps that reduce some of the dangers that we do point out in this report, and help to create new job opportunities for Europe.

Section 1: Data Knowledge in the Google Galaxy- and Empirical Evidence of the Google-Wikipedia Connection

(Note: Sections 1-5 are basically material produced by S. Weber after discussions with H. Maurer, with H. Maurer doing the final editing)

In the beginning of the third millennium, we are faced with the historically unique situation that a privately owned American corporation determines the way we search and find information – on a global scale. In former times, the self-organisation of storage and selection of our common knowledge base was primarily the task of the scientific system – especially of the librarians of the Gutenberg Galaxy. In the last 15 years, this has changed dramatically: We are moving with enormous speed from the Gutenberg to the Google Galaxy. Fuelled by techno enthusiasm, nearly all social forces – political, scientific, artistic, and economic ones – have contributed to the current situation.

Just think of the affirmative optimism we all felt in the nineties: With Altavista, we started to search for key terms of our interest – and the results were really astonishing: The notion of "networking" quickly gained a completely new dimension. All of a sudden, it was possible to collect information we otherwise und before had no means to gather.

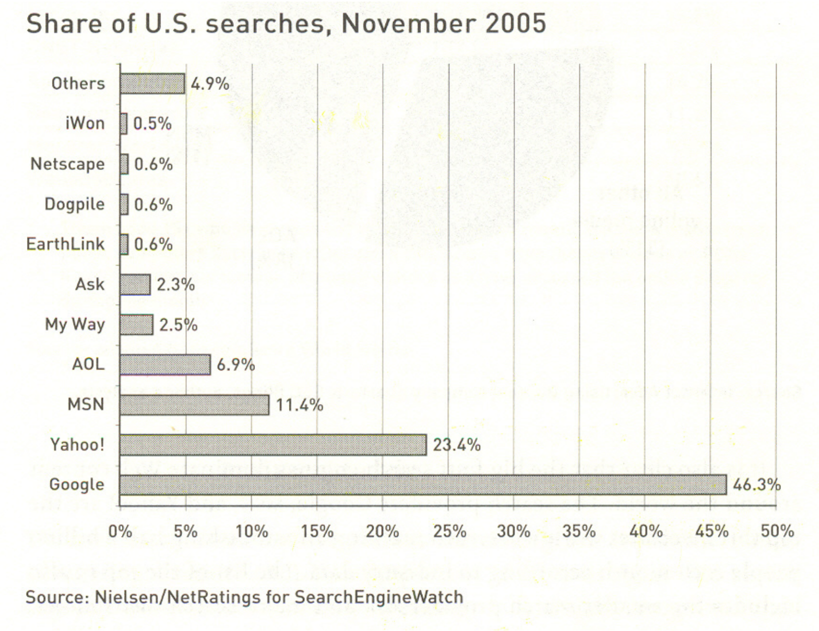
An example: In 1996, one of the authors of this report was deeply into autopoietic systems theory of German sociologist Niklas Luhmann. Typing the term "systems theory" into Altavista did not only lead to several research groups directly dealing with Niklas Luhmann's autopoietic theory at that time, but also to the work of Berlin sociologist Rodrigo Jokisch who developed his own sociological "theory of distinctions" as a critique as well as an extension of Luhmann's theory. In a web commentary found with Altavista, the author read for the first time about Jokisch's book "Logik der Distinktionen" which was to be released. For the first time the net (especially the search engine) did a connection on a syntactic level that made sense in the semantic and pragmatic dimension as well: Data turned into knowledge.

Seen from today, it is easy to reconstruct this situation and also to re-interpret why we all were fascinated by this new technology. One of the first scientific metaphors used for search engines from that period was the so called "meta medium". Search engines were in the beginning of theory-building described as "meta media" [Winkler 1997] and compared for example to magazines covering the TV programme: They were seen as media dealing with other media (or media content), interpreted as media "above" media, as second order media (and remember in this context sites like thebighub.com: a meta search engine, a meta medium for many other meta media).

The self-organisation of the (economic and technological) forces of the web has led to a very ambivalent situation today: As Google came up in 1998, step-by-step Altavista lost its position as leading search engine defining the standards of what we seek and find, and Google became increasingly the new number one. None the less media science still describes the web search situation with old metaphors from the print era: Instead of speaking of a (neutral!) "meta medium", today the metaphor of Google as the new "gate keeper" of the web is widely spread [for example Machill & Beiler 2007]. Note that the "gate keeper" is always something dialectical: It enables (some information to come in) and it prevents (other information to reach the user). More on search engines see the excellent book [Witten 2006] and on the history of Google see [Batelle 2005] and [Vise 2005].

To demonstrate the over-all importance of Google, just have a look at some data:

Figure 1: Google clearly leading the ranking of US searches in 2005

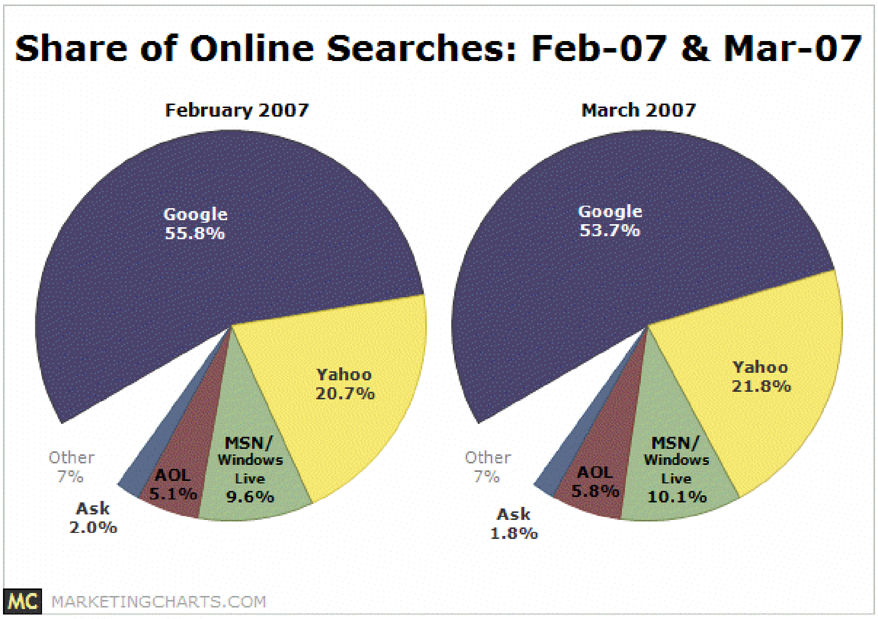


Source: Nielsen/NetRatings for SearchEngineWatch

[Machill & Beiler 2007, 107]

In this figure, you clearly see the dominating role of Google amongst the searches in the US in November 2005 (according to Nielsen/ NetRatings): Nearly every second search in the US was done with Google. If one compares autumn 2005 data with spring 2007, one can see that Google has gained once more (from 46,3% to about 55%) and the other search engines lost:

Figure 2: More than every second online search is done with Google



Source: Nielsen/Netratings MegaView Search

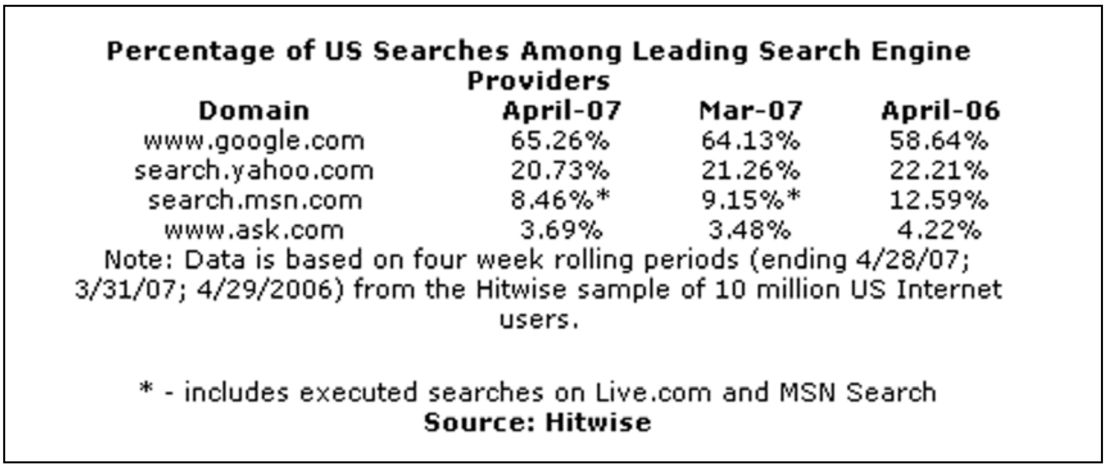
[http://www.marketingcharts.com/interactive/share-of-online-searches-by-engine-mar-2007-294]

The world-wide success story of Google is unique in the history of media technologies: Not only that Google in fact had developed the best search algorithm (the so-called "PageRank algorithm" after its inventor Larry Page – and not after the page in the sense of a site), there also happened a strange socio-cultural process which is also known as "Matthew effect" or labelled as "memetic" spreading: As more and more people started to use the new search engine, even more and more people started to use it. The Google revolution was (a) emergent und (b) contingent (in the sense that nobody could forecast it in the mid-nineties). In the eye of the users, it seemed to be an evolutionary development, but as seen in the context of the historical evolution of searching, archiving and finding information, we were witnesses of an incomparable revolution. And the revolution still goes on.

What has happened? Sergey Brin, one of the two founders of Google, had the idea that information on the web could be ordered in a hierarchy by the so-called "link popularity" (this means: the more often a link directs to a specific page, the higher this page is ranked within the search engine results). Other factors like the size of a page, the number of changes and its up-to-dateness, the key texts in headlines and the words of hyperlinked anchor texts were integrated into the algorithm. Step-by-step, Google improved an information hierarchisation process which most net users trust blindly today. The result was perplexing: "If Google doesn't find it, it doesn't exist at all" quickly became the hidden presupposition in our brains.

The importance of Google increased further and further as not only the search algorithms constantly evolved, but also as the socio-cultural "Matthew effect" remained unbroken. We propose an experimental situation in which a group of people is asked to search a term on the net. We let the test persons start with an unsuspicious site (e. g. www.orf.at in Austria, the web site of the nationwide broadcast cooperation). Our hypothesis is that about 80 or even 90 percent of the people will start their search with Google (maybe in an experimental setting also Yahoo or Microsoft Live searchers will tend to use Google). There is much evidence for this virtual monopoly of Google. Just look at the following table:

Table 1: Percentage of US searches among leading search engine providers – Google rated even higher



Percentage of US Searches Among Leading Search Engine Providers

Domain	April-07	Mar-07	April-06
www.google.com	65.26%	64.13%	58.64%
search.yahoo.com	20.73%	21.26%	22.21%
search.msn.com	8.46%*	9.15%*	12.59%
www.ask.com	3.69%	3.48%	4.22%

Note: Data is based on four week rolling periods (ending 4/28/07; 3/31/07; 4/29/2006) from the Hitwise sample of 10 million US Internet users.

* - includes executed searches on Live.com and MSN Search

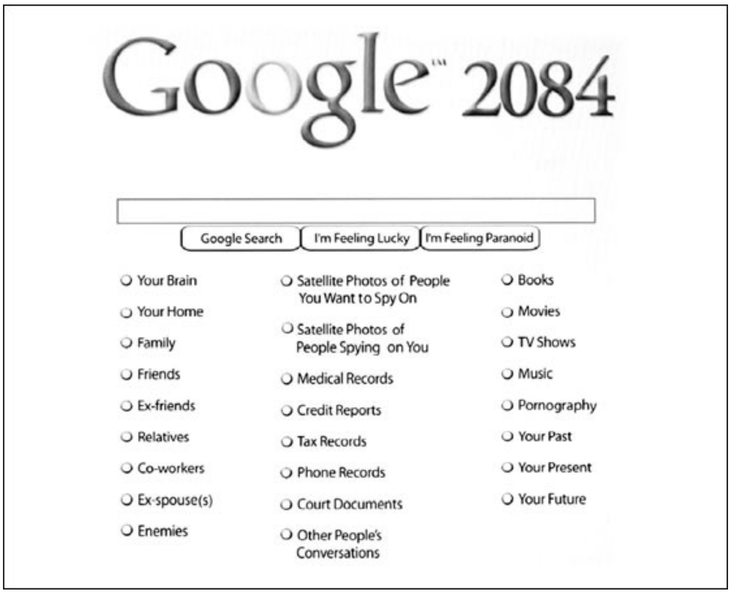
Source: Hitwise

[http://www.hitwise.com/press-center/hitwiseHS2004/search-engines-april-2007.php]

One can clearly see that the success story of Google did not stop within the last twelve months. On the contrary, from April 2006 to April 2007 the percentage of Google searches amongst all US searches rose from 58.6 to 65.3 (which means that Google ranks hit-wise even higher than Nielsen/ NetRatings). Yahoo and especially MSN did lose (in the moment it is questionable if Microsoft's "new" search engine Live can stop this development in the long run). Please note that the top-three search engines make up 94.45 percent of the whole “search-cake”. This number shows dramatically what a new search engine would have to do if it wants to get a piece from that cake. To enforce that users leave Google and begin to trust another search engine would need an amount of money as well as technological competence and marketing measures beyond all thought. And not to forget the "butterfly effect" in combination with the "Matthew effect" in the socio-cultural context: People must start to use the new search engine because other people already do so whom they trust (two-step-flow or second-order effect). Let us call the new search engine wooble.com. Then opinion leaders worldwide have to spread: "Did you already use Wooble? It shows you better results than Google." In the moment, this process is completely unrealistic.

The overwhelming virtual monopoly of Google already led to the equation "Google = Internet". Will Google also develop a web browser and thus swallow Internet Explorer of Microsoft (as the Explorer swallowed Netscape Navigator before)? Will Google develop more and more office software tools (first attempts can be tried out in the "Google Labs" zone, for example the "Text&Tabellen" programme or the by now quite wide-spread Gmail free mail software)? Will Google turn from the virtual monopoly on the web to the Internet itself? In a blog one could read recently that Google is not really a competitor anymore, but already the environment [cited after Kulathuramaiyer & Balke 2006, 1737]. Will we one day be faced with the situation that our new all-inclusive medium (as prime address in all concerns worldwide), the Internet and a private owned corporation with headquarters near San Francisco will fall together, will form a unity?

Figure 3: Vision of Google in 2084

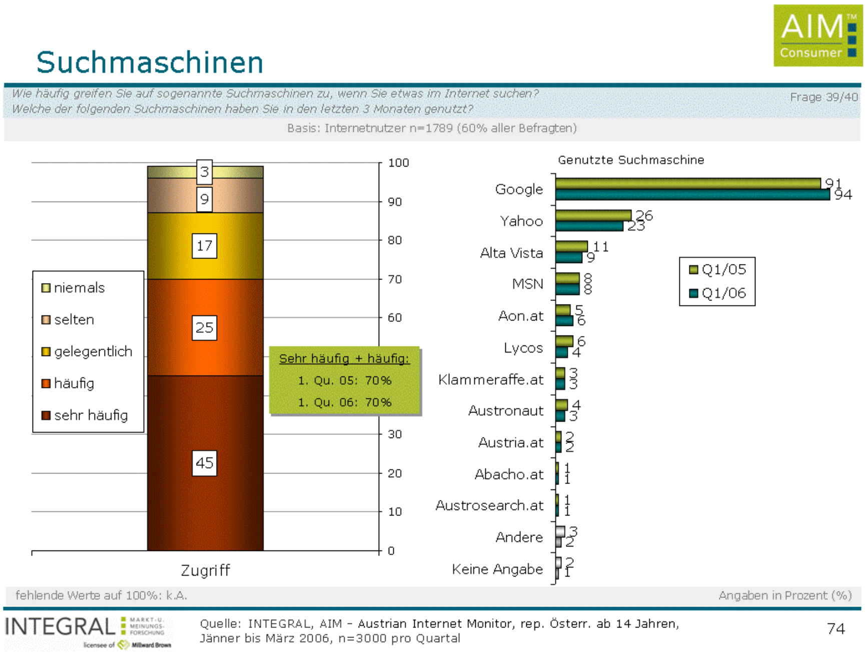


[From "New York Times", 10 October 2005, http://www.nytimes.com/imagepages/2005/10/10/opinion/1010opart.html]

Hyper-Google = the net itself then will know everything about us: Not only the way we currently feel, the things we bought (and would like to buy) and searched for, but also our future (see figure 3). In fact this vision is not very comfortable, and this is the main reason why strategies are discussed to stop Google, mainly through legal limitations or otherwise restrictions from the government or the scientific system [Machill & Beiler 2007, see also Maurer 2007a and 2007c].

As shown in the figures above, in the moment at least about two thirds of all US web inquiries are executed via Google. If you look at a specific European country, e. g. Austria, the situation is not different at all.

Figure 4: Users' search engine preferences in Austria 2006 compared to 2005



Suchmaschinen

Wie häufig greifen Sie auf sogenannte Suchmaschinen zu, wenn Sie etwas im Internet suchen?
Welche der folgenden Suchmaschinen haben Sie in den letzten 3 Monaten genutzt?

Frage 39/40

Basis: Internetnutzer n=1789 (60% aller Befragten)

Sehr häufig + häufig:
1. Qu. 05: 70%
1. Qu. 06: 70%

fehlende Werte auf 100%: k.A.

Angaben in Prozent (%)

Quelle: INTEGRAL, AIM - Austrian Internet Monitor, rep. Österr. ab 14 Jahren,
Jänner bis März 2006, n=3000 pro Quartal

74

[Austrian Internet Monitor, http://www.integral.co.at, thanks for the slide to Sandra Cerny]

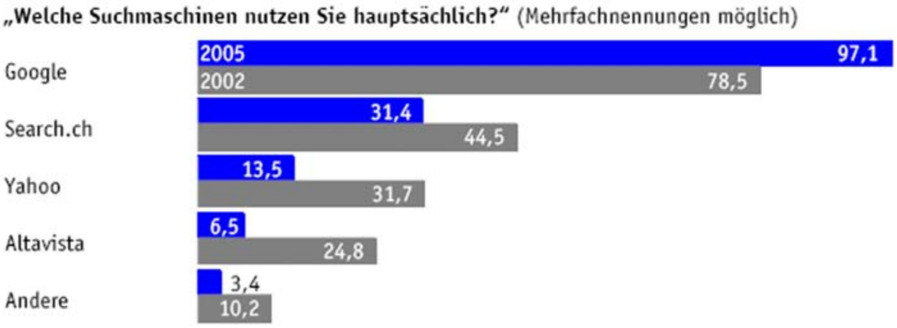
94 percent of the Austrian Internet users said that they used Google at least once in the last three months. The number has increased from 91 to 94 percent between March 2005 and March 2006. So if users are asked (instead of examining actual searches amongst users), the dominance of Google gets even clearer: In the moment there seems to be no way to compete with Google any more – although we know that web users can globally change their behaviour and preferences very quickly in an indeterminable way ("swarm behaviour").

Finally, to present empirical evidence for the dominance of Google, we can not only look at actual searches or for example at the users of a specific nation, but we can also have a look at a specific group of people, for example at journalists, scientists or students: Surprisingly enough, we nearly have no hard facts about the googling behaviour of pupils and students. We know that pupils more and more tend to download ready-made texts from the net, for example from Wikipedia or from paper mills. We have no large-scale statistics on the Google Copy Paste behaviour of students as described in some alarming case studies by Stefan Weber [Weber 2007a].

But we do have some empirical evidence on the googling behaviour of another group of text-producing people: of journalists. More and more, initiatives to maintain journalistic quality standards complain that also journalistic stories are increasingly the result of a mere "googlisation of reality". One drastic example is described by the German journalist Jochen Wegner [Wegner 2005]: A colleague of him did a longer report on a small village in the north of Germany. He reported about a good restaurant with traditional cooking, a region-typical choir doing a rehearsal in the village church and about a friendly farmer selling fresh agricultural products. If you type the name of the small village into Google, the first three hits are the web sites of the restaurant, the farmer and the choir. And if you compare the complete story of the journalist with the texts on the web sites found by Google, you will see: As a journalist of the 21st century, you don't have to be at a place to write a "personal" story about it. Google can do this for you.

Two recent empirical studies tried to enlighten the googling behaviour of journalists in German-speaking countries: The Swiss study "Journalisten im Internet 2005" and the Austrian study "So arbeiten Österreichs Journalisten für Zeitungen und Zeitschriften" in the year 2006. In both studies, the overwhelming dominance of Google was evident: Between 2002 and 2005, the number of Swiss journalists who do primarily Google research grew from 78.5 to 97.1 (!) percent. Therefore one can say that nearly every Swiss journalist at least also used Google in 2005 – and this won't have changed much until today.

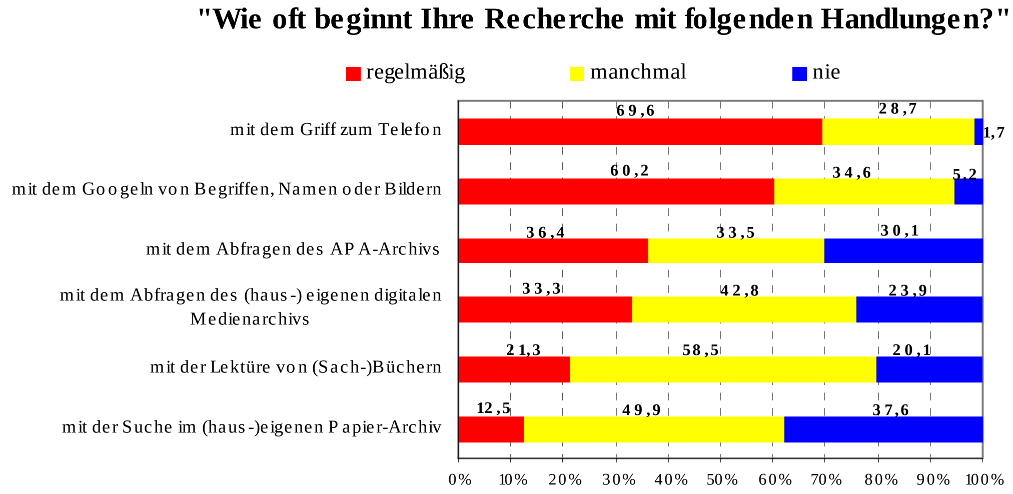
Figure 5: Search engine preferences of Swiss journalists 2005 compared to 2002



[Keel & Bernet 2005, 11]

Also, in Austria 94.8 percent of the print journalists asked in a survey admitted that they start their research for a story at least sometimes with the googling of keywords, names, or images. 60 percent of the journalists google continuously.

Figure 6: Googling as new starting point of research for journalists in Austria 2005



Methode: Online-Befragung; Erhebungszeitraum: 16.11.2005 bis 5.12.2005; n=296 Printjournalisten

[Weber 2006a, 16]

All the data published here could be reported almost world-wide in the same way, with a few exceptions. Google for example is not strong in South Korea, where a publishing group with "Ohmynews" is dominating the scene. Anyway, without being pathetic we have to state that Google rules the world of information and knowledge. If the world has turned into an information society or even into a knowledge society (according to philosopher Konrad Paul Liessmann the notion of a "knowledge society" is only a myth), than googling is the new primary cultural technique to gather information (or better: to gather data from the web). The Google interface and the (hidden) way

Google ranks the information found is the core of the information society in the moment. Never before in history was this organized by a private enterprise.

In the current academic world we do not only observe the so-called "Google Copy Paste Syndrome" (GCPS) as new questionable cultural technique of students [Weber 2007a, Kulathuramaiyer & Maurer 2007], but also an interesting by-product of the GCP-technique: the obvious Google-Wikipedia connection (GWC). As seen from net user behaviour, this means when typing a specific technical term into Google, one will tend to click on the Wikipedia link because one will probably trust Wikipedia more than other sources (because of the collective generation of knowledge in the Wikipedia leading to a specific form of consensus theory of truth, one might probably feel that information on this net encyclopaedia is less biased than information from various other sites. This notion is much defended in [Surowiecki 2004] but equally well attacked in the “must-read” book [Keen 2007].

In this context we also recommend an experimental setting in which we tell students to search for a list of technical terms on the net and copy & paste the definitions obtained – the hypothesis is that a great majority will consult Wikipedia and copy & paste definitions from that site.

Our everyday experience with the googling of technical terms led us to the intuitive assumption that Wikipedia entries are significantly more often ranked under the top three or even on the first place of the Google search results than other links. One reason could be the Google search algorithm which probably ranks sites higher if they are big and continuously updated. In a Google watchblog a blogger recently wrote that the "new" Google search algorithm ranks sites higher the more they are not only externally, but also internally linked and the more often the site has changed (see http://www.jmboard.com/gw/2007/04/28/neues-patent-zum-zeitlichen-ranking-von-webseiten). Of course this would be a good explanation why Wikipedia entries are very often ranked very high in the Google matches list. But the ranking of Wikipedia entries could also have another more unpleasant reason: It is possible that Google ranks some sites intentionally higher than others. This would mean that information is not only put into a hierarchical order by an "abstract" mathematical algorithm after indexing the text galaxy of the net, but also by human forces tagging some sites with higher points than others (which will lead to a higher ranking, again by an algorithm). This is not only a speculation. Wikipedia itself reported in 2005 a strange fact about a cooperation between the net encyclopaedia and Yahoo:

"An agreement with Yahoo in late Spring 2004 added them as one of the search engines widgets offered for web/Wikipedia searching when our internal search goes down, and we also provided them with a feed of our articles (something we needed to do anyway). In exchange, we were apparently linked more prominently, or more effectively, in their searches, and have access to stats on how many click-throughs we get from them. It's not clear how we were linked more prominently, but click-throughs increased by a factor of 3 in the months after this agreement (and then levelled off)." [http://meta.wikimedia.org/wiki/Wikimedia_partners_and_hosts; boldface introduced by the authors of this report]

So Wikipedia admits that after an agreement with Yahoo they observed that Wikipedia results began to climb up on the list of search matches. The question remains: Due to a – coincidentally – change or improvement of the Yahoo search algorithm or due to some kind of "human intervention"? This case addresses the attention to a blind spot of large search engines: Is everything done by computer programmes, or is there some kind of intentional intervention into the search results? Remember also that the Chinese version of the Google web site has shown that Google is able to and also will bias the information one will retrieve.

Also a German scientist wrote about a cooperation between Yahoo and Wikipedia as well as between Google and Wikipedia:

"Some months ago the 'Wikimedia Foundation' has signed an agreement with 'Yahoo's' rival 'Google' that guarantees that – as well as with 'Yahoo' – a matching Wikipedia article is ranked all above on the list of search results."
[Lorenz 2006, 86 f.; translation by the authors of this report]

The same scientist also speculated: "How does Wikipedia finance that? Or does Google donate that service?" [Lorenz 2006, 87, footnote 14; translation by the authors of this report] In a personal eMail correspondence, the author informed us that the speaker of Wikipedia Germany strictly denied the existence of such an agreement.

The topic of probably intentionally biasing search results is central for the credibility and reliability of large search engines. As shown above, in the current information or knowledge society search engines – and especially Google – determine what we search and find. They are not only "meta media", they are not only "gate keepers" of information. In fact they are much more: Especially Google has become the main interface for our whole reality. This ranges from the search of technical terms to news searches on a specific topic. If we speak about the interface for our whole reality, we mean that in an epistemological kind of way. Of course many documents in the web (especially in the hidden "deep web", in sites which can only be accessed with passwords, in paid content sites etc.) won't be found by Google. And many documents (not only older ones!) are still only available offline. To be precise: With the Google interface the user gets the impression that the search results imply a kind of totality. In fact, one only sees a small part of what one could see if one also integrates other research tools. (Just type the word "media reception" into Google. The first results won't give you any impression on the status quo of this research field. In fact especially with scientific technical terms one often gets the impression that the ranking of the results is completely arbitrary.)

For this report, we did the following experiment: We randomly chose 100 German and 100 English keywords from the A-Z index of the two Wikipedia versions and typed these keywords into four large search engines. We noted the place of the Wikipedia link in the specific ranking list. This experiment for the first time shows how our knowledge is organised in rather different ways by different research tools on the net.

Detailed description of the experiment: We typed 100 randomly chosen keywords with already existing text contributions (and some few forward-only keywords) from http://de.wikipedia.org/wiki/Hauptseite into:

- http://www.google.de
- http://de.yahoo.com
- http://de.altavista.com and
- http://www.live.com/?mkt=de-de

and noted the ranking..

We decided to use these four search engines because A9.com does not work with its own search engine, but shows the web search results of live.com, the products search results of amazon.com itself and additionally results of answers.com and others. And also AOL research is powered by Google – so we would have compared Google with Google or Microsoft Live with Microsoft Live.

For the duplication of the experiment in English language, we randomly took 100 keywords from the English Wikipedia site http://en.wikipedia.org/wiki/Main_Page (also from the A-Z index) and typed them into:

- http://www.google.com
- http://www.yahoo.com
- http://www.altavista.com
- http://www.live.com

and again noted the ranking.

And here are the results based on 100 randomly chosen keywords:

Figure 7: Percentage of Wikipedia entries ranked as first results

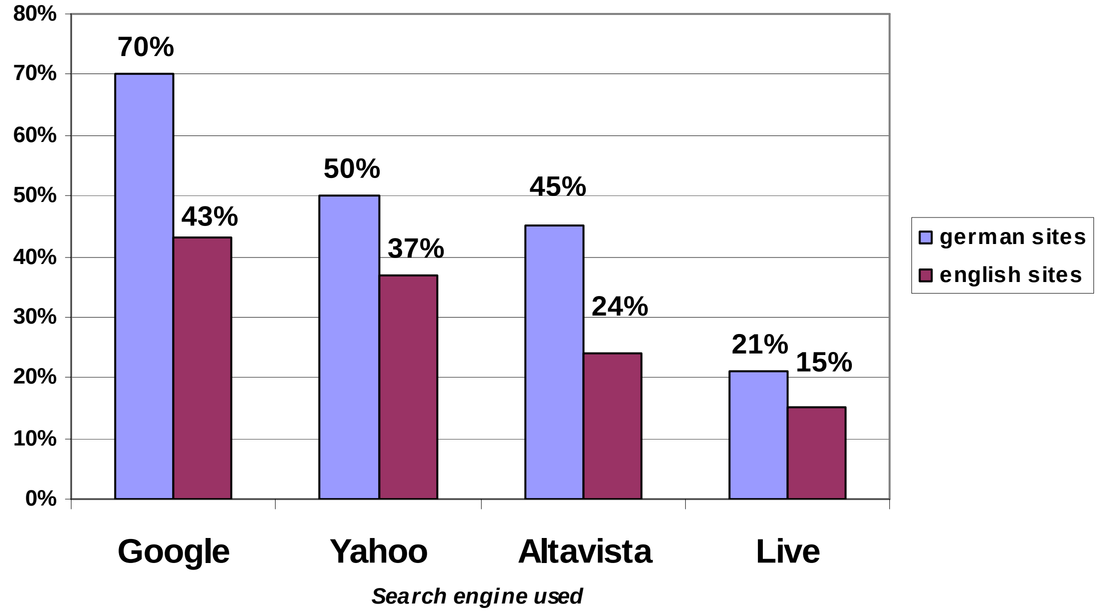


Figure 8: Percentage of Wikipedia entries within the first three results

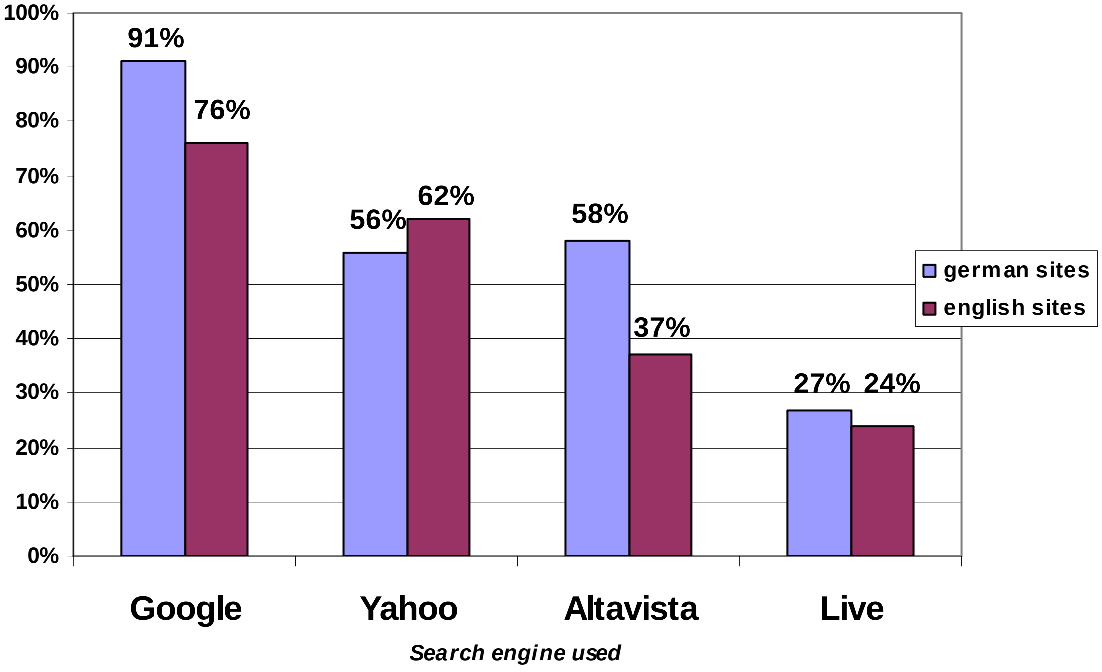
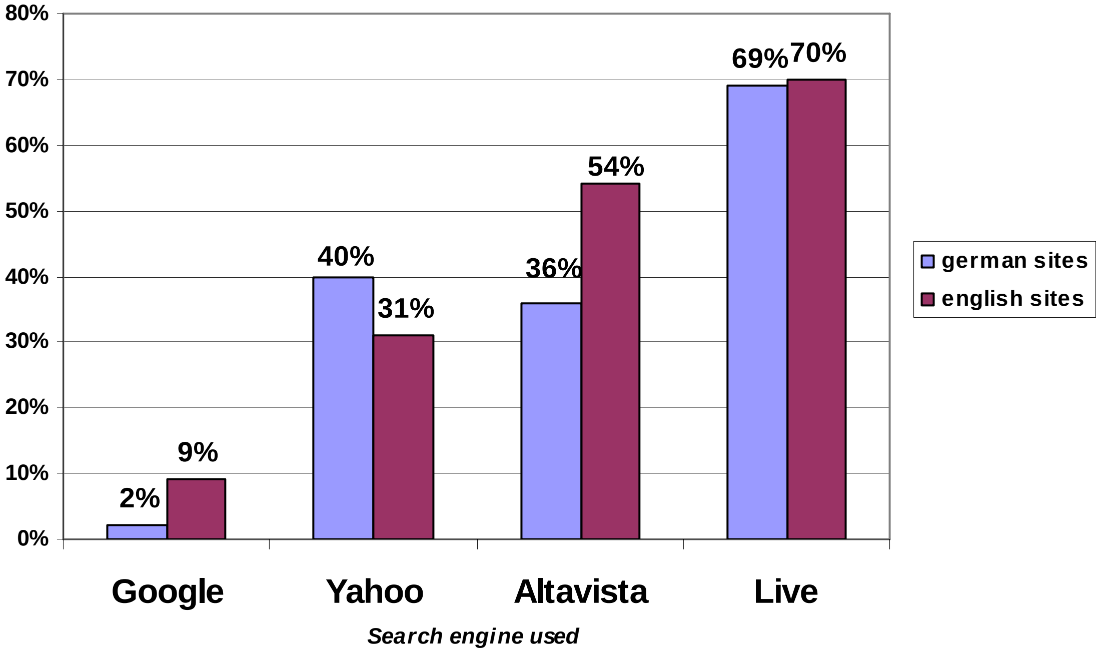


Figure 9: Percentage of Wikipedia entries beyond top ten results



The results prove two clear tendencies:

1) Google is clearly privileging Wikipedia sites in its ranking – followed by Yahoo. Microsoft's search engine Live only rarely ranks Wikipedia entries very high.
2) The German Google version privileges the German Wikipedia site significantly more than the international Google site privileges the English Wikipedia site.

Of course we can exclude the assumption that the German and the international Google sites operate with completely different algorithms. So we are able to draw two logical conclusions from the astonishing result mentioned under 2):

1) Context factors influence the ranking: One explanation would be the fact that the German Wikipedia is much more internally linked than the English mother site (for which we have no evidence). Another (weak) explanation could be the fact that the German Wikipedia in German-speaking countries is more popular than the English site in English-speaking countries (which is true) and that this fact indirectly leads to a better ranking because more links lead to the German Wikipedia (which is questionable).
2) The other conclusion is scary: Google does in a strange and unknown way privilege Wikipedia entries – followed by Yahoo; and Google does this intentionally more with the German version.

We recommend immersing deeper into this apparent correlations or strange coincidences. For this report, we only did a pilot. To do comparative research in search engine results ranking in fact would be a new and crucial field of research which in the moment still is a big desideratum.

The apparent Google-Wikipedia connection (GWC) is also problematic from an epistemological point of view: When people google key terms, they need no brain effort to do research: everybody can type a word or a phrase into a search engine (in former times, one needed basic knowledge about the organisation of a library and the way a keyword catalogue operates, and one needed to work with the so-called "snowball system" to find new fitting literature in the reference lists of already found literature). So there is a clear shift in the field of research towards a research without brains. But there also is another shift in the way encyclopaedic knowledge is used: In former times facts from a print encyclopaedia maximally marked the starting point of a research (who ever cited the Encyclopaedia Britannica or the Brockhaus verbatim?). Today one must observe countless students copying passages from Wikipedia. Thus a term paper can be produced within a few minutes. Students lose the key abilities of searching, finding, reading, interpreting, writing and presenting a scientific paper with own ideas and arguments, developed after a critical close reading process of original texts. Instead of that they use Google, Copy & Paste and PowerPoint. Their brains are now contaminated by fragmented Google search terms and the bullet points of PowerPoint. For a critique on PowerPoint see also [Tufte 2006].

The online encyclopaedia Wikipedia is problematic not only because of vandalism or fabrication of data. It is also problematic because of the often unknown origin of the basic texts than adapted by the authors' collective of net users. In some reported cases already the very first version of a Wikipedia entry was plagiarised, often copied nearly verbatim without the use of much "brain power" from an older print source. Some of these cases are precisely described in [Weber 2005b] and [Weber 2007a, 27 ff.]. We have to state that there is a systematic source problem in Wikipedia, because the problem of plagiarism was ignored too long, and is still being ignored.

For example, just type the word "Autologisierung" into the German Wikipedia site. You will find an entry which was published some times ago by an unknown person. But the entry is rather brainless plagiarism of the subchapter "Autologisierung" in a scientific contribution by one of the authors of this report (Stefan Weber) which appeared 1999 in a print anthology. Nearly nobody ever used that word since than, and in fact there is absolutely no reason why it is a keyword in the German Wikipedia :-).

These are only some examples or case studies of an evolving text culture without brains also on the Wikipedia. We should also not overlook that American Wikipedia and Google critic Daniel Brandt reported about 142 cases of plagiarism on the Wikipedia [N. N. 2006b].

The knowledge reliability of Wikipedia remains a big problem for our common knowledge culture. Doing unreflecting cut and paste from Wikipedia seems to be a new cultural technique which has to be observed with care. If Google privileges Wikipedia and thus makes the way to do copy & paste even more straight, the problem is a double one: a search monopoly of a private corporation and the informational uncertainty of a collective encyclopaedia which also marks a new knowledge monopoly.

We will return to some of the above issues in Section 7. For references see Section 16.

Section 2: Google not only as main door to reality, but also to the Google Copy Paste Syndrome: A new cultural technique and its socio-cultural implications

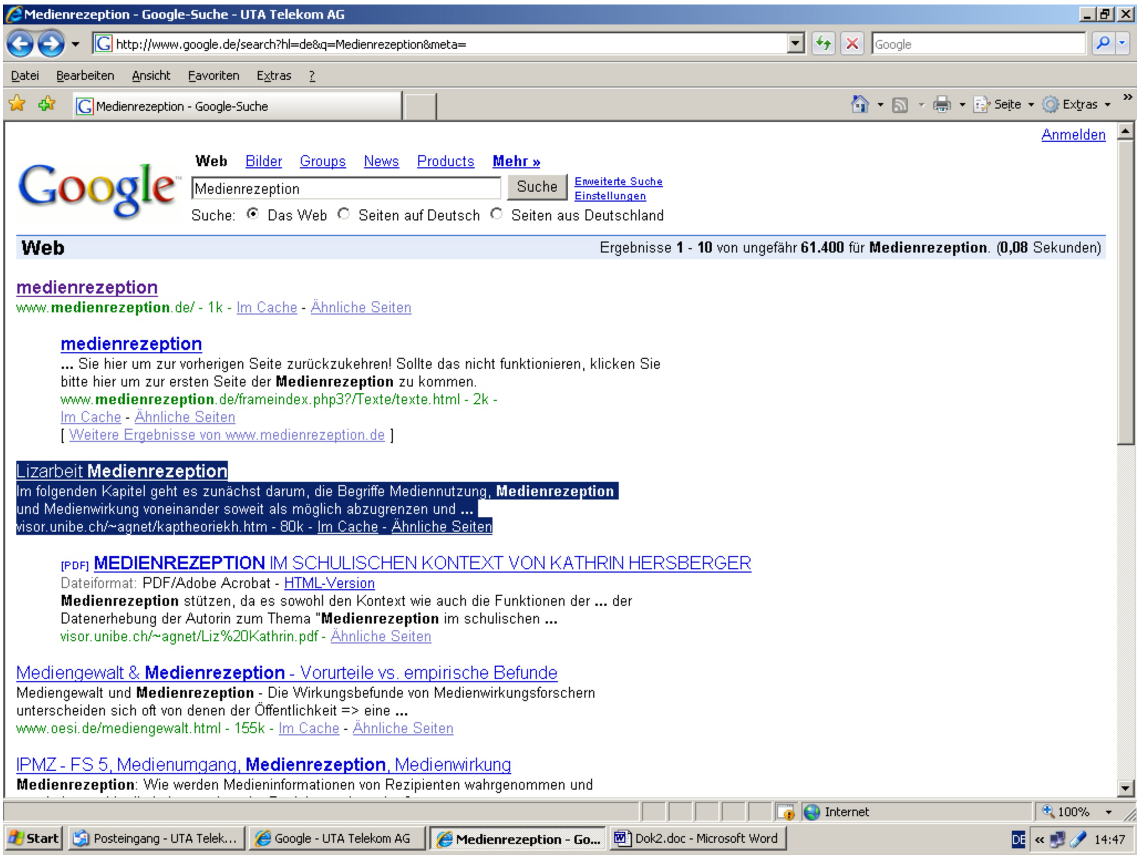
(Note: Sections 1-5 are basically material produced by S. Weber after discussions with H. Maurer, with H. Maurer doing the final editing)

As shown in Section 1, at the moment Google determines the way we search and find information in the net to a degree that media critics, scientists and politicians cannot longer remain silent about: they should actively raise their voice. Section 1 dealt with the epistemological revolution in the net age: The ranking algorithm of a private owned corporation listed on the stock exchange dictates which information we receive and which information is neglected or even intentionally suppressed. As elaborated in Section 1, this is a unique situation in the history of mankind.

But there is also an important consequence affecting our whole knowledge production and reception system on a more socio-cultural level: After googling a technical or common term, a name or a phrase or whatever, especially the younger generation tends to operate with the found text segments in a categorically different way than the generation of the Gutenberg Galaxy did: While the print generation tended towards structuring a topic and writing a text by themselves, the new "Generation Google" or "Generation Wikipedia" is rather working like "Google Jockeys" [N. N. 2006a] or "Text Jockeys" [Weber 2007f]: They approach text segments in a totally different manner then the print-socialised generation. Text segments found on the web are often appropriated despite their clearly claimed authorship or despite their clearly communicated copyright restrictions because they are seen as "free" and/or "highly reliable". One often hears persons accused of net plagiarism justifying themselves: "But it's already written on the web – why should I put it in new words anyway?". See the quite disenchanting example in [Weber 2007a, 4] The new generation of text jockeys starting with googling key terms or phrases tends to cut and paste found information of the search engine's result list directly into their document and claim a new authorship from now on. Of course plagiarism also occurred in the print era (as will be shown later on), but the Google Copy Paste technique is something categorically new in the history of the relationship between text and author.

We will give the following clear and unmasking example as an introduction: Type the German word "Medienrezeption" (media reception) into Google. You will see that the third entry of the search results list is a Swiss "Lizentiatsarbeit" (a kind of equivalent to a master thesis) about media reception in the context of school kids. Absurdly enough, this found site has nearly nothing to do with the current status quo of media reception research. We have no idea how many external links must lead to this completely arbitrary result that it went up to rank 3 on the Google list (we checked it for the last time on 20 May 2007; but the link has already been on the top three of the list of Google results months – and probably years – ago).

Figure 10: Searching "Medienrezeption" with Google



Medienrezeption

Suche: ● Das Web ○ Seiten auf Deutsch ○ Seiten aus Deutschland

Web Ergebnisse **1** - **10** von ungefähr **61.400** für **Medienrezeption**. (**0,08** Sekunden)

medienrezeption
www.**medienrezeption**.de/ - 1k - Im Cache - Ähnliche Seiten

medienrezeption
... Sie hier um zur vorherigen Seite zurückzukehren! Sollte das nicht funktionieren, klicken Sie bitte hier um zur ersten Seite der **Medienrezeption** zu kommen.
www.**medienrezeption**.de/frameindex.php3?/Texte/texte.html - 2k -
Im Cache - Ähnliche Seiten
[Weitere Ergebnisse von www.medienrezeption.de]

Lizarbeit **Medienrezeption**
Im folgenden Kapitel geht es zunächst darum, die Begriffe Mediennutzung, **Medienrezeption** und Medienwirkung voneinander soweit als möglich abzugrenzen und ...
visor.unibe.ch/~agnet/kaptheoriekh.htm - 80k - Im Cache - Ähnliche Seiten

[PDF] **MEDIENREZEPTION** IM SCHULISCHEN KONTEXT VON KATHRIN HERSBERGER
Dateiformat: PDF/Adobe Acrobat - HTML-Version
Medienrezeption stützen, da es sowohl den Kontext wie auch die Funktionen der ... der Datenerhebung der Autorin zum Thema "**Medienrezeption** im schulischen ...
visor.unibe.ch/~agnet/Liz%20Kathrin.pdf - Ähnliche Seiten

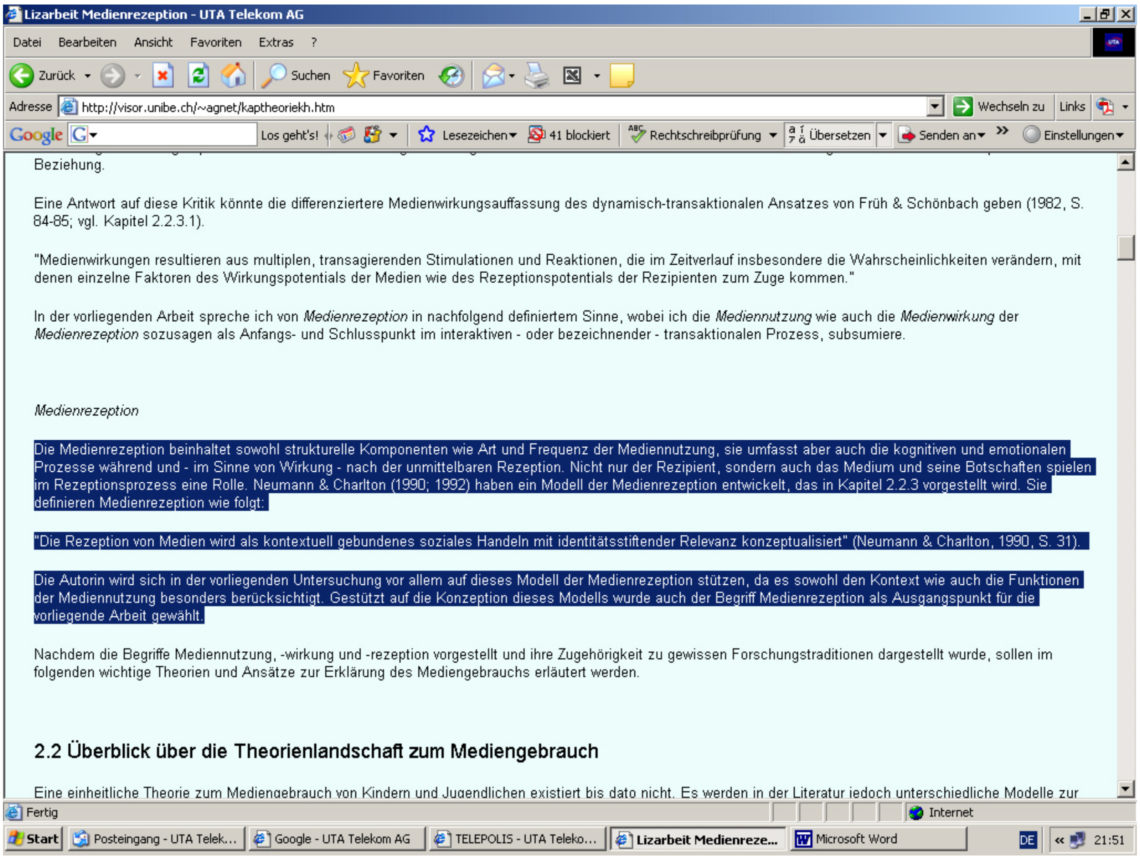
Mediengewalt & **Medienrezeption** - Vorurteile vs. empirische Befunde
Mediengewalt und **Medienrezeption** - Die Wirkungsbefunde von Medienwirkungsforschern unterscheiden sich oft von denen der Öffentlichkeit => eine ...
www.oesi.de/mediengewalt.html - 155k - Im Cache - Ähnliche Seiten

IPMZ - FS 5, Medienumgang, **Medienrezeption**, Medienwirkung
Medienrezeption: Wie werden Medieninformationen von Rezipienten wahrgenommen und

[Screenshot, 17 May 2007]

If one clicks on the link, one will find the following text segment (which was – as we suppose – written by the author of the Swiss master thesis and not copied from elsewhere):

Figure 11: Document found on third place of the Google results



Beziehung.

Eine Antwort auf diese Kritik könnte die differenziertere Medienwirkungsauffassung des dynamisch-transaktionalen Ansatzes von Früh & Schönbach geben (1982, S. 84-85; vgl. Kapitel 2.2.3.1).

"Medienwirkungen resultieren aus multiplen, transagierenden Stimulationen und Reaktionen, die im Zeitverlauf insbesondere die Wahrscheinlichkeiten verändern, mit denen einzelne Faktoren des Wirkungspotentials der Medien wie des Rezeptionspotentials der Rezipienten zum Zuge kommen."

In der vorliegenden Arbeit spreche ich von *Medienrezeption* in nachfolgend definiertem Sinne, wobei ich die *Mediennutzung* wie auch die *Medienwirkung* der *Medienrezeption* sozusagen als Anfangs- und Schlusspunkt im interaktiven - oder bezeichnender - transaktionalen Prozess, subsumiere.

Medienrezeption

Die Medienrezeption beinhaltet sowohl strukturelle Komponenten wie Art und Frequenz der Mediennutzung, sie umfasst aber auch die kognitiven und emotionalen Prozesse während und - im Sinne von Wirkung - nach der unmittelbaren Rezeption. Nicht nur der Rezipient, sondern auch das Medium und seine Botschaften spielen im Rezeptionsprozess eine Rolle. Neumann & Charlton (1990; 1992) haben ein Modell der Medienrezeption entwickelt, das in Kapitel 2.2.3 vorgestellt wird. Sie definieren Medienrezeption wie folgt:

"Die Rezeption von Medien wird als kontextuell gebundenes soziales Handeln mit identitätsstiftender Relevanz konzeptualisiert" (Neumann & Charlton, 1990, S. 31).

Die Autorin wird sich in der vorliegenden Untersuchung vor allem auf dieses Modell der Medienrezeption stützen, da es sowohl den Kontext wie auch die Funktionen der Mediennutzung besonders berücksichtigt. Gestützt auf die Konzeption dieses Modells wurde auch der Begriff Medienrezeption als Ausgangspunkt für die vorliegende Arbeit gewählt.

Nachdem die Begriffe Mediennutzung, -wirkung und -rezeption vorgestellt und ihre Zugehörigkeit zu gewissen Forschungstraditionen dargestellt wurde, sollen im folgenden wichtige Theorien und Ansätze zur Erklärung des Mediengebrauchs erläutert werden.

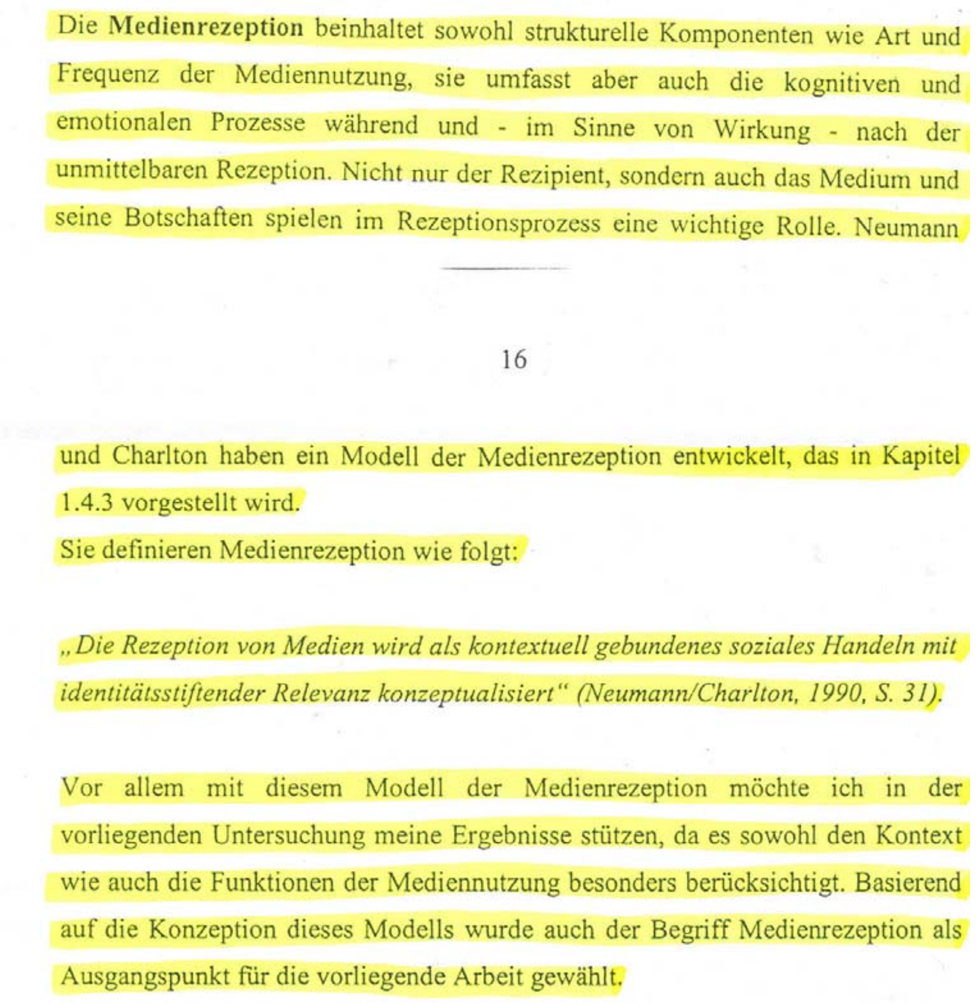
2.2 Überblick über die Theorienlandschaft zum Mediengebrauch

Eine einheitliche Theorie zum Mediengebrauch von Kindern und Jugendlichen existiert bis dato nicht. Es werden in der Literatur jedoch unterschiedliche Modelle zur

[http://visor.unibe.ch/~agnet/kaptheoriekh.htm, visited 29/5/07]

The marked text segment went straight into the diploma thesis of an Austrian student – of course without any reference to the original text or context found on the web.

Figure 12: This document used for plagiarism



Die **Medienrezeption** beinhaltet sowohl strukturelle Komponenten wie Art und Frequenz der Mediennutzung, sie umfasst aber auch die kognitiven und emotionalen Prozesse während und - im Sinne von Wirkung - nach der unmittelbaren Rezeption. Nicht nur der Rezipient, sondern auch das Medium und seine Botschaften spielen im Rezeptionsprozess eine wichtige Rolle. Neumann

16

und Charlton haben ein Modell der Medienrezeption entwickelt, das in Kapitel 1.4.3 vorgestellt wird.
Sie definieren Medienrezeption wie folgt:

„Die Rezeption von Medien wird als kontextuell gebundenes soziales Handeln mit identitätsstiftender Relevanz konzeptualisiert" (Neumann/Charlton, 1990, S. 31).

Vor allem mit diesem Modell der Medienrezeption möchte ich in der vorliegenden Untersuchung meine Ergebnisse stützen, da es sowohl den Kontext wie auch die Funktionen der Mediennutzung besonders berücksichtigt. Basierend auf die Konzeption dieses Modells wurde auch der Begriff Medienrezeption als Ausgangspunkt für die vorliegende Arbeit gewählt.

[Scan from N. N., Wickie und die starken Männer – TV-Kult mit Subtext. Diploma thesis, 2004, p. 15 f.]

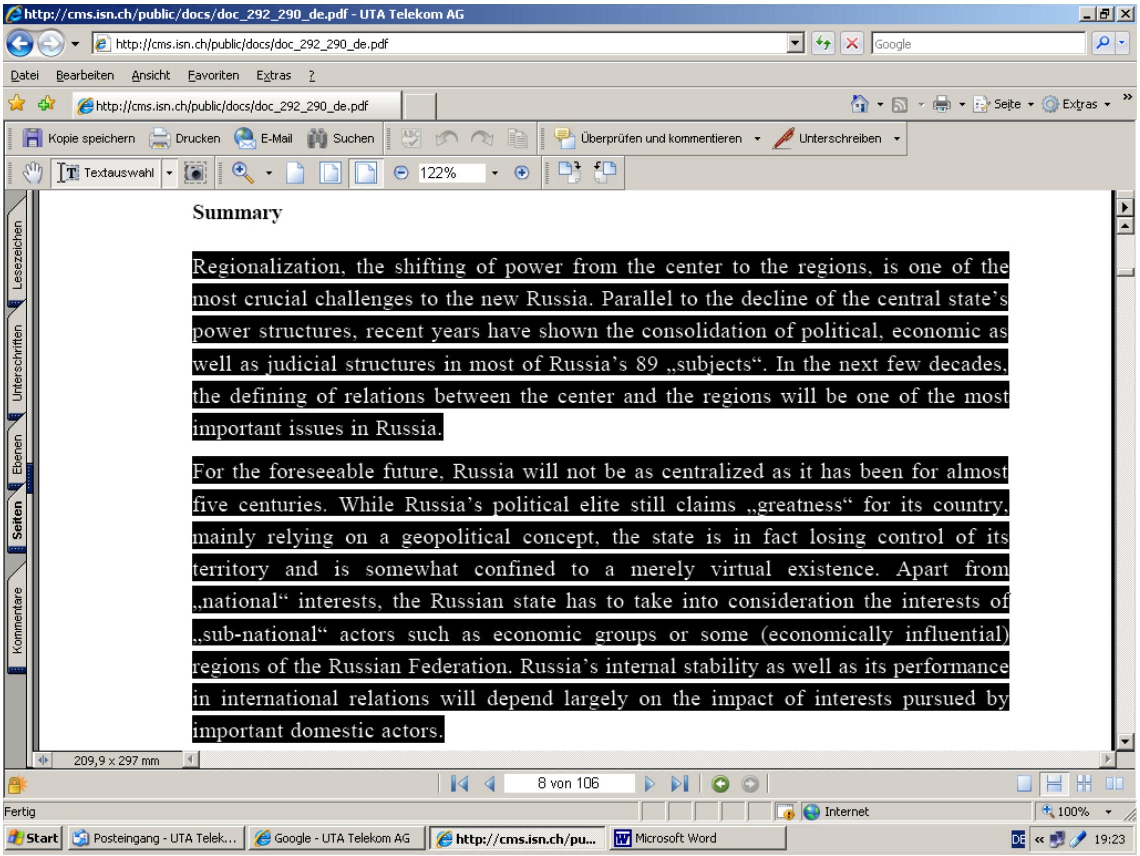
Please note: This text segment would contain "quote plagiarism" also if the original text from the net was quoted "properly". In humanities you are only allowed to reproduce a quote within another quote when you are unable to obtain the original. With the Neumann/Charlton quote this would not be the case.

It can be excluded that plagiarism went the other way round or that both documents have a common third and unknown origin. The Swiss master thesis ranked by Google on third position dated 1998 (see http://visor.unibe.ch/~agnet/Liz%20Kathrin.pdf), the Austrian master thesis was approved in 2004. Also see [Weber 2007a, 74 ff] for a discussion of the example. And it turned out that the Austrian author of the diploma thesis – at that time a scientific assistant at the department of media and communication research at Alpen Adria University Klagenfurt – plagiarised about 40 percent of her whole text (for an online documentation see http://www.wickieplagiat.ja-nee.de). The author of the plagiarised master thesis has been dismissed from her job at the university in August 2006.

Sadly enough this was no singular case of cut and paste plagiarism breaking with all known and well-established rules of academic honesty and integrity. One of the authors of this study, Stefan Weber, has meanwhile collected 48 cases of plagiarism which occurred mainly on Austrian (and some German) universities, on the web and in journalism between 2002 and 2007. The spectrum ranges from small term papers completely copied & pasted from one single web source (you need about ten minutes do to this – including layout work) to post-doctoral dissertations and encyclopaedias of renowned professors emeriti. The majority of the cases is connected with some kind of net plagiarism. Two further examples will be discussed in the following.

An Austrian political scientist and expert on asymmetric warfare (!) has copied at least 100 pages – and probably much more – from the web into his doctoral thesis in 2002, as always without giving any credit. The "funny" thing is that he even nearly verbatim appropriated the summary and the conclusions of another paper written four years before. The original paper was published as a PDF online, it can be found here: http://cms.isn.ch/public/docs/doc_292_290_de.pdf. This document was written by scientists on the technical university of Zurich in 1998 and counts 106 pages. One can find the summary and the conclusions of this document (and nearly all following chapters) mainly verbatim in a Viennese doctoral thesis from the year 2002. Just compare the following two screenshots:

Figure 13: Original text as PDF file on the web



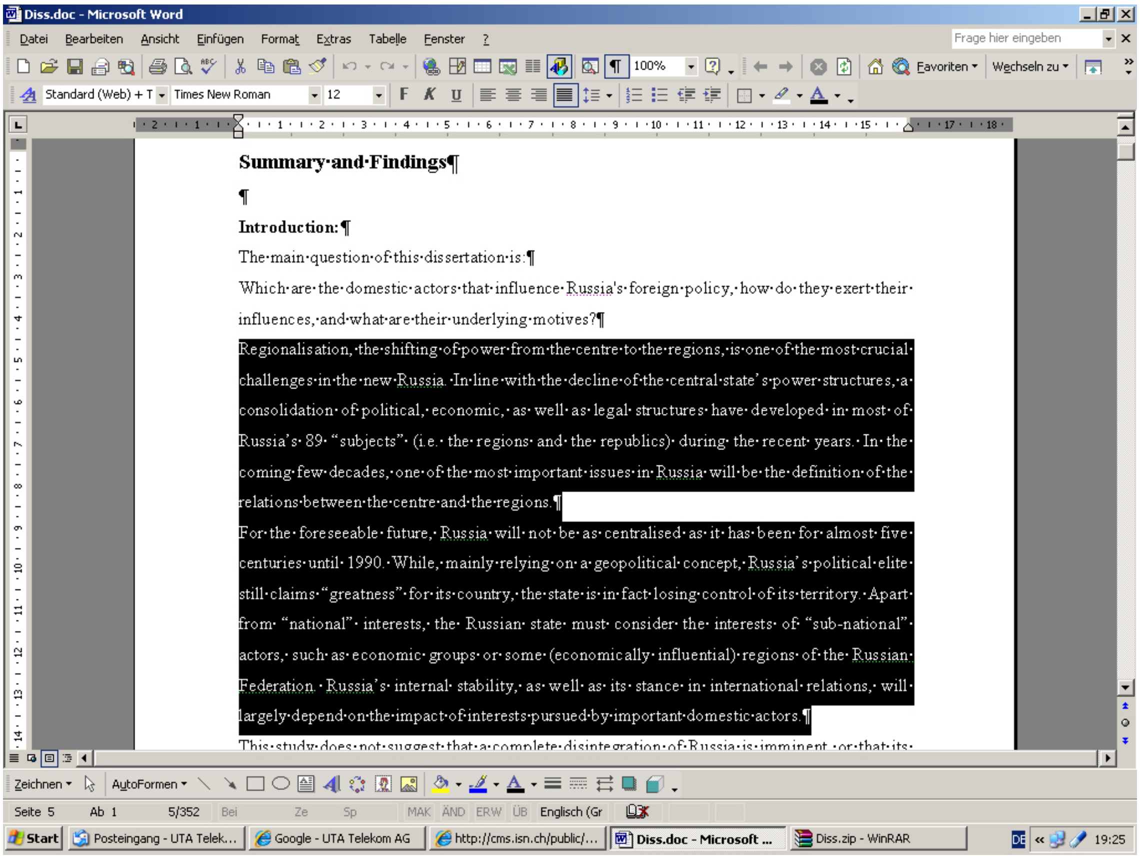
Summary

Regionalization, the shifting of power from the center to the regions, is one of the most crucial challenges to the new Russia. Parallel to the decline of the central state's power structures, recent years have shown the consolidation of political, economic as well as judicial structures in most of Russia's 89 „subjects". In the next few decades, the defining of relations between the center and the regions will be one of the most important issues in Russia.

For the foreseeable future, Russia will not be as centralized as it has been for almost five centuries. While Russia's political elite still claims „greatness" for its country, mainly relying on a geopolitical concept, the state is in fact losing control of its territory and is somewhat confined to a merely virtual existence. Apart from „national" interests, the Russian state has to take into consideration the interests of „sub-national" actors such as economic groups or some (economically influential) regions of the Russian Federation. Russia's internal stability as well as its performance in international relations will depend largely on the impact of interests pursued by important domestic actors.

[http://cms.isn.ch/public/docs/doc_292_290_de.pdf, p. 5, original text dating 1998, visited 29/5/07]

Figure 14: Plagiarised version of the same text


Diss.doc - Microsoft Word
Datei Bearbeiten Ansicht Einfügen Format Extras Tabelle Fenster ?
Frage hier eingeben
Standard (Web) + T
Times New Roman
12
100%
Favoriten
Wechseln zu
Zeichnen
AutoFormen
Seite 5 Ab 1 5/352 Bei Ze Sp MAK ÄND ERW ÜB Englisch (Gr
Start
Posteingang - UTA Telek...
Google - UTA Telekom AG
http://cms.isn.ch/public/...
Diss.doc - Microsoft ...
Diss.zip - WinRAR
DE
19:25

Summary and Findings

Introduction:

The main question of this dissertation is:

Which are the domestic actors that influence Russia's foreign policy, how do they exert their influences, and what are their underlying motives?

Regionalisation, the shifting of power from the centre to the regions, is one of the most crucial challenges in the new Russia. In line with the decline of the central state's power structures, a consolidation of political, economic, as well as legal structures have developed in most of Russia's 89 "subjects" (i.e. the regions and the republics) during the recent years. In the coming few decades, one of the most important issues in Russia will be the definition of the relations between the centre and the regions.

For the foreseeable future, Russia will not be as centralised as it has been for almost five centuries until 1990. While, mainly relying on a geopolitical concept, Russia's political elite still claims "greatness" for its country, the state is in fact losing control of its territory. Apart from "national" interests, the Russian state must consider the interests of "sub-national" actors, such as economic groups or some (economically influential) regions of the Russian Federation. Russia's internal stability, as well as its stance in international relations, will largely depend on the impact of interests pursued by important domestic actors.

This study does not suggest that a complete disintegration of Russia is imminent , or that its

[Screenshot from dissertation of N. N., text dating 2002]

Start to compare the two texts with "Regionalisation, the shifting of power" in the doctoral thesis and note that the plagiarising author made very small changes ("Regionalization" turned into "Regionalisation"; "Parallel to the decline" turned into "In line with the decline" and so on). Whenever one can identify a clearly copied document with very small changes (1 or 2 words replaced with synonyms or different spelling per sentence), this is a strong indicator for not only a "mistake of the computer" or a "software problem".

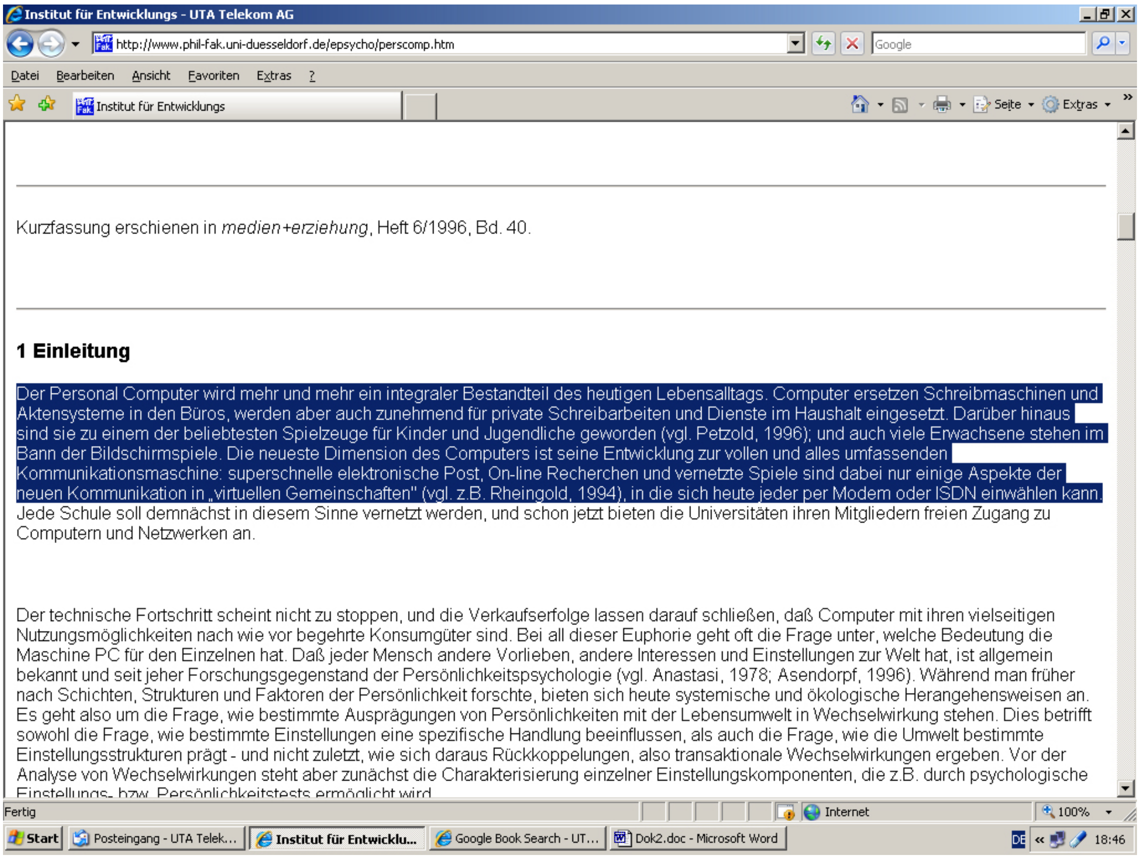
The third example of net based copy & paste "writing" is the most rigorous one known to us until now: The author – a former student of psychology, coincidentally again at Klagenfurt university – compiled the prose of her doctoral thesis from probably more than one hundred un-cited text fragments from various online sources (including non-scientific ones). Hard to believe, the first 20 pages of the dissertation were not much more than the addition of text chunks found on the following web sites (and of course nothing was referred):

- http://www.phil-fak.uni-duesseldorf.de/epsycho/perscomp.htm
- http://www.diplomarbeiten24.de/vorschau/10895.html
- http://www.bildungsserver.de/zeigen.html?seite=1049
- http://www.foepaed.net/rosenberger/comp-arb.pdf
- http://www.behinderung.org/definit.htm
- http://www.dieuniversitaet-online.at/dossiers/beitrag/news/behinderung-integration-universitat/83/neste/1.html
- http://www.behinderung.org/definit.htm
- http://info.uibk.ac.at/c/c6/bidok/texte/steingruber-recht.html
- http://info.uibk.ac.at/c/c6/bidok/texte/finding-sehbehindert.html

- http://www.arbeitundbehinderung.at/ge/content.asp?CID=10003%2C10035
- http://www.behinderung.org/definit.htm
- http://ec.europa.eu/employment_social/missoc/2003/012003/au_de.pdf
- http://www.grin.com/de/preview/23847.html
- http://www.bmsg.gv.at/cms/site/attachments/5/3/2/CH0055/CMS1057914735913/behindertenbericht310703b1.pdf
- http://www.parlinkom.gv.at/portal/page?_pageid=908,221627&_dad=portal&_schema=PORTAL
- http://www.bpb.de/publikationen/ASCNEC,0,0,Behindertenrecht_und_Behindertenpolitik_in_der_Europ%E4ischen_Union.html
- http://ec.europa.eu/employment_social/disability/eubar_res_de.pdf
- http://ec.europa.eu/public_opinion/archives/ebs/ebs_149_de.pdf

This case was also documented in [Weber 2006b]. After media coverage of this drastic example of plagiarism, the university of Klagenfurt decided to control all dissertations and master thesis approved in the last five years for suspicious plagiarism cases with the software Docol©c (see chapter 4). An explicit plagiarism warning was published on the web site of the university including an extended definition of plagiarism (http://www.uni-klu.ac.at/main/inhalt/843.htm). To find copied text chunks in the highly questionable dissertation, it is sufficient to start with the very first words of the preface. The author didn't write in the traditional way any more, rather nearly everything was copied from the web.

Figure 15: Original from the web...


Institut für Entwicklungs - UTA Telekom AG
http://www.phil-fak.uni-duesseldorf.de/epsycho/perscomp.htm
Google
Datei
Bearbeiten
Ansicht
Favoriten
Extras
?
Institut für Entwicklungs
Seite
Extras
Fertig
Internet
100%
Start
Posteingang - UTA Telek...
Institut für Entwicklu...
Google Book Search - UT...
Dok2.doc - Microsoft Word
DE
18:46

Kurzfassung erschienen in *medien+erziehung*, Heft 6/1996, Bd. 40.

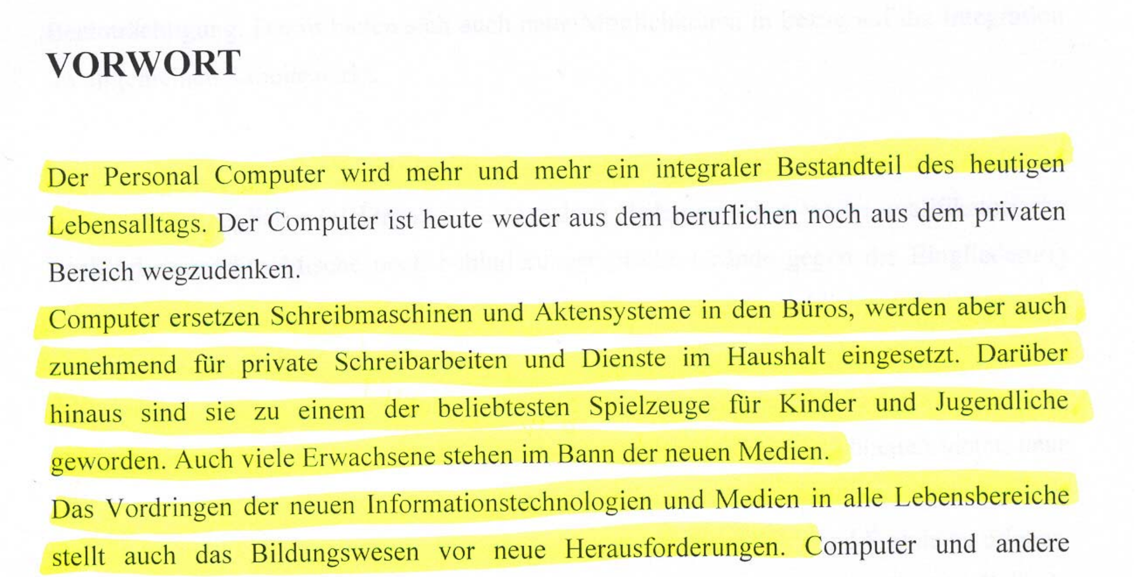
1 Einleitung

Der Personal Computer wird mehr und mehr ein integraler Bestandteil des heutigen Lebensalltags. Computer ersetzen Schreibmaschinen und Aktensysteme in den Büros, werden aber auch zunehmend für private Schreibarbeiten und Dienste im Haushalt eingesetzt. Darüber hinaus sind sie zu einem der beliebtesten Spielzeuge für Kinder und Jugendliche geworden (vgl. Petzold, 1996); und auch viele Erwachsene stehen im Bann der Bildschirmspiele. Die neueste Dimension des Computers ist seine Entwicklung zur vollen und alles umfassenden Kommunikationsmaschine: superschnelle elektronische Post, On-line Recherchen und vernetzte Spiele sind dabei nur einige Aspekte der neuen Kommunikation in „virtuellen Gemeinschaften" (vgl. z.B. Rheingold, 1994), in die sich heute jeder per Modem oder ISDN einwählen kann. Jede Schule soll demnächst in diesem Sinne vernetzt werden, und schon jetzt bieten die Universitäten ihren Mitgliedern freien Zugang zu Computern und Netzwerken an.

Der technische Fortschritt scheint nicht zu stoppen, und die Verkaufserfolge lassen darauf schließen, daß Computer mit ihren vielseitigen Nutzungsmöglichkeiten nach wie vor begehrte Konsumgüter sind. Bei all dieser Euphorie geht oft die Frage unter, welche Bedeutung die Maschine PC für den Einzelnen hat. Daß jeder Mensch andere Vorlieben, andere Interessen und Einstellungen zur Welt hat, ist allgemein bekannt und seit jeher Forschungsgegenstand der Persönlichkeitspsychologie (vgl. Anastasi, 1978; Asendorpf, 1996). Während man früher nach Schichten, Strukturen und Faktoren der Persönlichkeit forschte, bieten sich heute systemische und ökologische Herangehensweisen an. Es geht also um die Frage, wie bestimmte Ausprägungen von Persönlichkeiten mit der Lebensumwelt in Wechselwirkung stehen. Dies betrifft sowohl die Frage, wie bestimmte Einstellungen eine spezifische Handlung beeinflussen, als auch die Frage, wie die Umwelt bestimmte Einstellungsstrukturen prägt - und nicht zuletzt, wie sich daraus Rückkoppelungen, also transaktionale Wechselwirkungen ergeben. Vor der Analyse von Wechselwirkungen steht aber zunächst die Charakterisierung einzelner Einstellungskomponenten, die z.B. durch psychologische Einstellungs- bzw. Persönlichkeitstests ermöglicht wird.

[http://www.phil-fak.uni-duesseldorf.de/epsycho/perscomp.htm, original text dating 1996, visited 29/5/07]

Figure 16: ... again used for plagiarism



VORWORT

Der Personal Computer wird mehr und mehr ein integraler Bestandteil des heutigen Lebensalltags. Der Computer ist heute weder aus dem beruflichen noch aus dem privaten Bereich wegzudenken.

Computer ersetzen Schreibmaschinen und Aktensysteme in den Büros, werden aber auch zunehmend für private Schreibarbeiten und Dienste im Haushalt eingesetzt. Darüber hinaus sind sie zu einem der beliebtesten Spielzeuge für Kinder und Jugendliche geworden. Auch viele Erwachsene stehen im Bann der neuen Medien.

Das Vordringen der neuen Informationstechnologien und Medien in alle Lebensbereiche stellt auch das Bildungswesen vor neue Herausforderungen. Computer und andere

[Scan from dissertation of N. N., 2004, p. 7]

Of course we have to mention that Google is not only the problem (as first part of the fatal triad Google, Copy, and Paste), but also one possible solution: All the above documented cases of plagiarism were also detected by googling phrases from the suspicious texts. One can reconstruct this detection easily by just googling a few words from the plagiarised text material (for example "Der Personal Computer wird mehr und mehr" – in most cases, a few words are absolutely sufficient!). But this does not mean that the problem is solved by Google – not at all! We always have to remember that first came the misuse, and then we have the possibility to reveal some plagiarised work, but by far not all plagiarised work. Just one example: Order a diploma thesis from http://www.diplom.de and pay about 80 Euros for the whole text burnt on a CD. If you use that text for plagiarism, Google won't help at all to detect the betrayal.

Whenever media cover big cases of plagiarism, journalists ask the few experts in the field: How widespread is this behaviour on universities in the moment? Responsible persons of university managements often speak of very few cases of problematic plagiarism. However, some statistics report that about 30 percent or more students (partly) plagiarised their work at least once a time. Other reports say that about also 30 percent of all submitted works contain plagiarised material.[1]

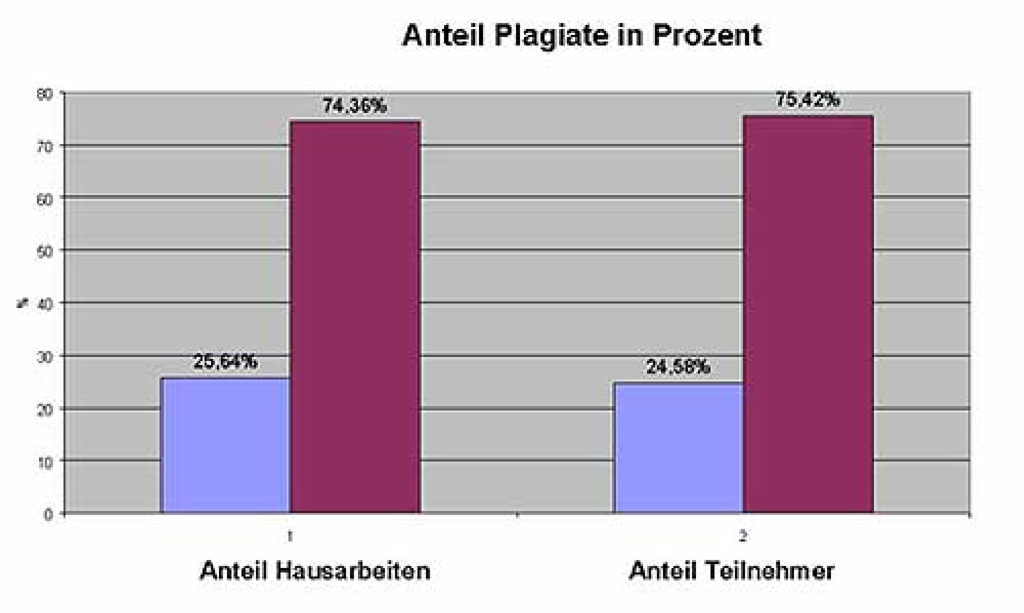
However, we will discuss in Section 6 that plagiarism is seen very different in different fields. Most of the careful examination done by one of the authors of this report concentrated on material whose intrinsic value is text-based, quite different from other areas in which the intrinsic value is not the wording, but the new idea, the new formula, new drawing, new computer program. Thus, plagiarism , and particularly the GCP syndrome, is more of a problem in text-only related areas, and less so in engineering disciplines, architecture, etc.

Cut and paste plagiarism became a topic for German universities and for the German media after the University of Bielefeld (D) published that in 2001/2002 in a sociological seminar of Professor

[1] "Erste Hinweise von Universitätsprofessoren aus dem In- und Ausland lassen jedoch vermuten, daß die Erstellung von Plagiaten mithilfe des Internets eine deutlich steigende Tendenz aufweist. So ist zum Beispiel an der University of California (Berkeley/USA) für einen Zeitraum von drei Jahren (Stichjahr: 1997) eine Zunahme von Täuschungsversuchen um 744 Prozent beobachtet worden."
(Zitat aus Fußnote (1) von: http://www.hochschulverband.de/presse/plagiate.pdf)

Wolfgang Krohn about 25 percent of the submitted papers and about 25 percent of the participants were involved in some kind of plagiarism. No detailed studies on the percentage of increase through the internet are available to us, yet the papers quoted in the Introduction in Appendix 1 all mention a significant increase.

Figure 17: Percentage of plagiarism on a German university in 2001/02 (share of cases of plagiarism in blue)



[http://www.uni-bielefeld.de/Benutzer/MitarbeiterInnen/Plagiate/iug2001.html, visited 20/5/07]

Amongst many surveys on plagiarism worldwide since then two studies delivered highly reliable results: The survey of Donald L. McCabe, executed for the "Center for Academic Integrity" (CAI) at Duke University in the USA between 2002 and 2005 (N>72.950 students and N>9.000 staff members), and a survey executed by OpinionpanelResearch for "Times Higher Education Supplement", executed amongst students in Great Britain in March 2006 (N=1.022 students). Both studies revealed nearly the same fact that about one third of the students already were involved in some kind of plagiarism. Have a look at the data in detail [for a summary also see Weber 2007a, 51 ff.]:

Table 2: Plagiarism in the US

Survey of Donald L. McCabe, US (N>72.950 students; N>9.000 staff members):

Cheating on written assignments:	**Undergraduates***	Graduate Students*	Faculty**
"Paraphrasing/copying few sentences from written source without footnoting it"	38 %	25 %	80 %
"Paraphrasing/copying few sentences from Internet source without footnoting it "	36 %	24 %	69 %
"Copying material almost word for word from a written source without citation"	7 %	4 %	59 %

* Values represent % of students who have engaged in the behaviour at least once in the past year.

** Values represent % of faculty who have observed the behaviour in a course at least once in the last three years.

[Donald L. McCabe, "Cheating among college and university students: A North American perspective", http://www.ojs.unisa.edu.au/journals/index.php/IJEI/article/ViewFile/14/9, 2005, p. 6]

Table 3: Plagiarism in GB

Survey of Opinionpanel, GB (N=1.022 students):

Action:	**"Since starting university, which of the following have you ever done?"**
"Copied ideas from a book on my subject"	37 %
"Copied text word-for-word from a book on my subject (excluding quoting)"	3 %
"Copied ideas from online information"	35 %
"Copied text word-for-word from online information (excluding quoting)"	3 %

[OpinionpanelResearch, "The Student Panel", paper from Times Higher Education Supplement, personal copy, received July 2006, p. 4]

Similar smaller studies were done in Germany and in Austria and led to similar results (for example an Austrian students' online platform did a small survey in June 2006 asking students "Did you ever use texts without citations?", and 31 percent answered with "yes", see Weber 2007a, 55). A recent study carried out as a master thesis on the university of Leipzig (D) revealed that even 90 percent (!) of the students in principle would plagiarise if there is an occasion to do so. The survey was executed online using the randomized-response-technique (RRT) to ensure that the students were willing to fill out very confidential questions with true answers with a higher probability. In sum nearly all studies say the following:

1) There is a very high willingness to plagiarise by the current generation of students. The data collected so far are indicators for an increasing culture of hypocrisy and betrayal at universities.
2) About 3 to 7 percent of the students admit some kind of "hardcore plagiarism" also when they were asked in a scientific context (probably in difference to what they would say to peer groups; one author of this report was faced with more than one student telling that he or she was proud of his or her betrayal).
3) About 30 percent admit some kind of "sloppy referencing" or problematic paraphrasing. In many of these cases, we do not know if we should talk of plagiarism or of a single "forgotten" footnote – which in fact must be decided on each singular case.

If one bears in mind that there is always a discrepancy between what test persons say about their behaviour in a scientific context and what they actually do, one will soon realise that the current problem of plagiarism and Google-induced copy & paste culture at universities is without any doubt a big one. We think that therefore the problem – interpreted in the context of a general shift of cultural techniques – should move into the centre of the agenda in the academic world. On the other hand, the solution is not so much a-posteriory plagiarism check, but first educating what does constitute a serious plagiarism case (copying a few words without footnote is not acceptable, yet was more considered a small offense compared to the very stringent rules that now start to deal with plagiarism), and second the educational system should make sure that plagiarism cannot occur, or cannot occur easily. We return to this in the lengthy section 14 by introducing two new concepts.

Meanwhile, the work definitions of what constitutes plagiarism get more and more draconic. We are right now on the way that also the so called "sloppy referencing" is not tolerated any longer. Just look at the following widespread US definition of plagiarism:

"All of the following are considered plagiarism:

- turning in someone else's work as your own
- copying words or ideas from someone else without giving credit
- failing to put a quotation in quotation marks
- giving incorrect information about the source of a quotation
- changing words but copying the sentence structure of a source without giving credit
- copying so many words or ideas from a source that it makes up the majority of your work, whether you give credit or not"

[http://turnitin.com/research_site/e_what_is_plagiarism.html, visited 20/5/07]

This list in fact means that also a wrong footnote ("giving incorrect information about the source of a quotation") or a text comprising quote after quote (the "majority of your work", that means at least 50 percent must be one's genuine prose!) can constitute plagiarism. In Europe only some universities have adopted strong definitions of plagiarism so far. Most institutions differ between sloppy citation and "real" plagiarism – which of course gives responsible persons at universities the possibility to play down intentional plagiarism as sloppiness.

However, we should also mention that the above very rigorous definition of what plagiarism is comes from Turnitin, a company whose business is to sell plagairsim detection software. That such company tries to define plagiarism down to a very fine level of granularity must not come as surprise!

At the university of Salzburg 2006 a leading professor refused any consequences for a student who plagiarised at least 47 pages of his diploma thesis verbatim from the web by simple copy & paste – some typing errors and misspellings from the original source remained uncorrected. The responsible person called the copy & paste work of the student "sloppy citation".

After some serious cases of plagiarism at the Alpen Adria University Klagenfurt a working group published in 2007 a new and stronger definition of what constitutes plagiarism:

"Plagiat ist die unrechtmäßige Aneignung von geistigem Eigentum oder Erkenntnissen anderer und ihre Verwendung zum eigenen Vorteil. Die häufigsten Formen des Plagiats in wissenschaftlichen Arbeiten sind:
1) Die wörtliche Übernahme einer oder mehrerer Textpassagen ohne entsprechende Quellenangabe (Textplagiat).
2) Die Wiedergabe bzw. Paraphrasierung eines Gedankengangs, wobei Wörter und der Satzbau des Originals so verändert werden, dass der Ursprung des Gedankens verwischt wird (Ideenplagiat).
3) Die Übersetzung von Ideen und Textpassagen aus einem fremdsprachigen Werk, wiederum ohne Quellenangabe.
4) Die Übernahme von Metaphern, Idiomen oder eleganten sprachlichen Schöpfungen ohne Quellenangabe.
5) Die Verwendung von Zitaten, die man in einem Werk der Sekundärliteratur angetroffen hat, zur Stützung eines eigenen Arguments, wobei zwar die Zitate selbst dokumentiert werden, nicht aber die verwendete Sekundärliteratur (Zitatsplagiat)."
[http://www.uni-klu.ac.at/main/inhalt/843.htm, visited 20/5/07]

For the first time at least in Austria also idea plagiarism and even "quote plagiarism" were included (the latter means that you cite a quote you have read elsewhere without checking the original, for example you cite Michel Foucault from the web or from the secondary literature and make a footnote to his original book which you have never read).

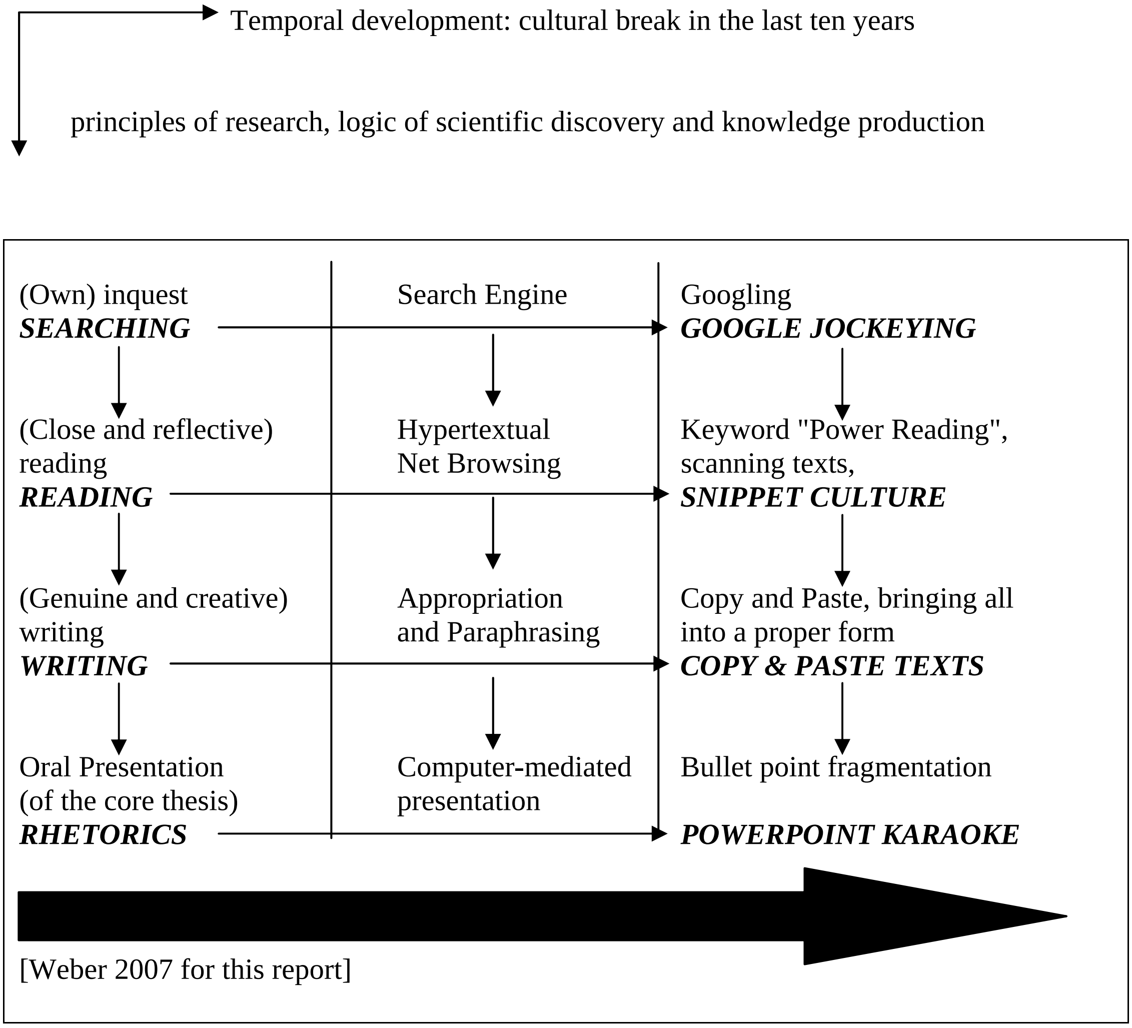
In the ideal case a scientific work whose value is based mainly on the textual component contains three text categories:

1) Genuine prose written by yourself or by the group of authors listed on the paper. This should be the largest part of the text, because science always has to do with innovation and with your own or the authors' critical reflections of the scientific status quo as reported in the literature and/or of current developments.
2) To prevent that scientific texts are only personal comments on a given topic and to contextualise own reflections into as many others' works and ideas working on the same specific field as possible, **direct quotes** should be used to reproduce inventive positions and thesis articulated already before the new work was written word-for-word (without changing anything – the real 1:1 reproduction is here essential for the scientific reference system).
3) If you refer to an idea, a genuine concept or also an empirical date you have drawn from elsewhere, you have to refer to the original work by [see Weber] or – as used in German humanities – by [Vgl.] (= compare to). This is usually called **indirect referring**.

There is empirical evidence that the triad of genuine prose, direct quotes and indirect referring is collapsing in the moment. The "Generation Google" or "Generation Wikipedia" tends to produce texts in a radically different way: They do not start writing in an inductive manner, but they deductively come from the texts they have already marked and cut from the web. They re-arrange these texts, write some connecting sentences between the already existing text chunks (what formerly was the genuine prose!) and bring the whole text into a proper form. Appropriation, paraphrasing (which means simple synonym replacement today!) and Shake & Paste are the new cultural techniques.

But is this still science? Are we moving towards a text culture without brains? What is happening in the moment? Cognitive capacities get free because searching, reading and writing can be delegated to Google, Copy, and Paste – but for what purpose? There is even no need for rhetoric any more, the bullet points of PowerPoint are sufficient in most contexts (which, in fact, is another big problem of current computer-based learning and teaching). The following graphic shows these changes.

Table 4: From human brain involvement to a text culture without brains?



Temporal development: cultural break in the last ten years

principles of research, logic of scientific discovery and knowledge production

(Own) inquest ***SEARCHING***	Search Engine	Googling ***GOOGLE JOCKEYING***
(Close and reflective) reading ***READING***	Hypertextual Net Browsing	Keyword "Power Reading", scanning texts, ***SNIPPET CULTURE***
(Genuine and creative) writing ***WRITING***	Appropriation and Paraphrasing	Copy and Paste, bringing all into a proper form ***COPY & PASTE TEXTS***
Oral Presentation (of the core thesis) ***RHETORICS***	Computer-mediated presentation	Bullet point fragmentation ***POWERPOINT KARAOKE***

[Weber 2007 for this report]

The table features some problematic aspects of an evolving text culture without brains [also see Weber 2007a, Kulathuramaiyer & Maurer 2007]. The optimistic way of reading this development is full of (naive?) hope that the cognitive capacities which are freed by Google, Copy & Paste, and PowerPoint will be occupied with other (more?) useful things. A pessimistic way of reading of the tendencies listed above is the diagnosis of a crisis of humanities as such [also see Weber 2006b]: Cases of plagiarism give proof of an increasing redundancy of the knowledge production of humanities. For example: If there already exist hundreds of diploma thesis on the history of the Internet, the students needn't rewrite this again, it is sufficient to produce a collage of found text chunks from the Internet itself (where you will find more than enough "raw material" on the history of the Internet – just try it out). If we take a closer look, we see that this discussion runs into the wrong direction. We should debate on the redundancy of specific topics and not of cultural science as such – it is the non-creativity of the lecturers and professors which is responsible for the current situation and not "the" science.

We do not have a real answer to all that happens around us as far as the flood of information is concerned. There are reports that information is doubling each year. On the other hand, knowledge double sonly every 5 to 12 years (depending on area and whom you believe). This difference points out clearly that more is written all the time about the same knowledge, creating by necessity a kind of plagiarism of some ideas, at least!

In sum, we observe the spreading of a "culture of mediocrity" [Kulathuramaiyer & Maurer 2007] and a "culture of hypocrisy" in the current academic world. "Mediocrity" also refers to the trends of...

- rhetorical bullshitting substituting controversial discussions based on critical reflections;
- effects dominating over real contents;
- affirmative research PR and techno PR instead of facing the real problems of scientific knowledge production (plagiarism, fraud, and manipulation or fabrication of data];
- and also to the trend of doing more and more trivial research in a micro scale which only has the function to confirm given common sense assumptions ("Mickey Mouse Research", see Weber 2007a, 148 ff.).

Let us return to the next scene of the revolution, and therefore again to Google. Google does not only mark the starting point of the Copy Paste Syndrome, it will change or already has begun to change our interactions with texts on a second arena: http://books.google.com. It is no doubt that the digitalisation of 15 millions or more printed books is the next big revolution (and other related initiatives like amazon's "Search Inside"). Information will be accessible for everybody on the web – but for what price, and information in which dose and in which context? Critical voices should not be overheard. Michael Gorman, president of the American Library Association, was cited in "Nature" with the following warning words:

"But Gorman is worried that over-reliance on digital texts could change the way people read — and not for the better. He calls it the 'atomization of knowledge'. Google searches retrieve snippets and Gorman worries that people who confine their reading to these short paragraphs could miss out on the deeper understanding that can be conveyed by longer, narrative prose. Dillon agrees that people use e-books in the same way that they use web pages: dipping in and out of the content." [Bubnoff 2005, 552]

There are some empirical hints that the reading ability and the ability of the younger generation to understand complex texts is diminishing. The fact that some books like the Harry Potter series are bestsellers hides the fact that many readers are indeed adults and not children! With the possibility to surf through (often parts of!) books and to do keyword search online, the cultural technique of reading a whole text as an entity defined by its author could become obsolete. It is a bit of a paradox: In the scanning streets of Asia, Google and Amazon are digitalising millions of books. But the way how Google Book Search will operate could make an end to all books as textual entities. Thus, the ideas of a core thesis, of a central argument with its branches, of the unfolding of a sophisticated theory or the complex description of empirical research settings and results could turn into anachronisms soon. Instantaneous application knowledge found within a second by key word search is the new thing.

In the galaxy of the ubiquitous media flow, will we still have time for reading a whole book? The way Google Book Search could probably change (or put an end to) the production of scientific texts in the academic world once more: A master thesis with at least 100 pages, a dissertation with 200 or 300 pages and a post-doctoral dissertation with even more pages – this could soon be history. Will then also change the ways we define plagiarism? In the end: Are people engaged in fighting plagiarism anachronists by themselves? In the era of copyleft licenses, creative commons and scanning millions of printed books – in many cases without explicit permission of the publishing companies – everything changes so quick that a critical reflection is essential (for optimistic descriptions of the current media situation and for "friendly" future scenarios see for example Johnson 2005, Johnson 1999, Gleich 2002, and Pfeifer 2007; for a critique especially on Google see Jeanneney 2006, Maurer 2007a, b, and c, and Weber 2007a; for a general apocalyptic version of the web development and especially the Web 2.0 see Keen 2007; a relatively neutral discussion of major developments can be found in Witten & Gori & Numerico 2007, here especially Chapter 2 on the digitalisation of printed books).

For references see Section 16.

Section 3: The new paradigm of plagiarism – and the changing concept of intellectual property

(Note: Sections 1-5 are basically material produced by S. Weber after discussions with H. Maurer, with H. Maurer doing the final editing)

Plagiarism today transgresses the narrow boundaries of science. It has become a major problem for all social systems: for journalism and economics, for the educational system as well as – as we will show – for religion (!).

In the recent years of theory building, sociology tended to describe society by one single macro trend. But this in fact happened about 40 or even more times, so that the contemporary sociological discourse knows many "semantics of society": Just think of "risk society", "society of communication" or "society of simulation". For this report we suggest to add the term "copying society" to the meanwhile long list of self descriptions of society. A copying society is a society in which the distinction between an original and a copy is always crucial and problematic simultaneously, and it is also a society in which we can observe many social processes in the micro, meso, and macro scale with the distinction of "original/copy" (if you think of things or objects) or "genuinely doing/plagiarising" (if you think of actions or processes).

Plagiarism is, as mentioned, by far not an exclusive problem of science, and it is by far not an exclusive problem of text culture. The following examples will give an impression of that. Plagiarism and a general crisis of the notion of "intellectual property" are intertwined phenomena. Many discussions are about playing down severe cases of plagiarism by arguing with the freedom of text circulation on the web. There is a big misunderstanding at the moment that for example copyleft licenses imply the rejection of any concept of "plagiarised work". Also people who fight against plagiarism are often accused to maintain a conservative text ideology or at least not to be up-to-date with the current net developments [in this style see for example IG Kultur 2007]. The concepts of authorship and intellectual property are regarded as chains which should be overcome by a free cybernetic circulation of text fragments in the Internet – without any claim of authorship, without any legal constraints.

In this confusing situation it is not easy to argue that people who mix the plagiarism debate with the copyleft discussion make a categorical mistake: Everybody should opt for plagiarism-free work – regardless if under a copyright or a copyleft license. Also within the framework of a copyleft license, the texts, ideas, or images published should not imply intellectual theft. The plagiarism debate is about the genesis of an intellectual product, the copyright/copyleft debate about its further distribution. Nevertheless there are some major frictions which don't make the situation easier, especially for the younger Google Copy Paste generation used to the net. Just have a closer look at a paragraph of the "GNU Free Documentation License", an older license for example still prominently used by Wikipedia. By this license the copying of texts, images, and otherwise information is allowed under the condition that the copier publishes the copied version under the same license (alone this would be impossible in the classical academic publishing system!) and is mentioning the source. Furthermore it is even allowed to modify the original version and publish it again – under the same condition of adopting the GNU license:

"You may copy and distribute a Modified Version of the Document [...], provided that you release the Modified Version under precisely this License, with the Modified Version filling the role of the Document, thus licensing distribution and modification of the Modified Version to whoever possesses a copy of it."
[http://en.wikipedia.org/wiki/Wikipedia:Text_of_the_GNU_Free_Documentation_License, in original http://www.gnu.org/copyleft/fdl.html, both visited 20/5/07]

Recent publications on free licenses on the web are very euphoric about these developments and nearly totally ignore the problem of plagiarism [see for example Dobusch & Forsterleitner 2007; for a criticism on that book see Weber 2007c]. The current ideology is a bit like: If you want to be hip and give your work a progressive touch, publish your material under a specific Creative Commons license of your taste. Again we have to state that there is little intellectual reflection of this. But if you push a Creative Commons license to its extremes, it could mean the following: Transliterate an interview and publish it under a Creative Commons license. When a modified version is allowed, one can take that interview and change some central sayings. If the interviewed person complains that he or she never said what is published now, tell him or her about your specific license and that you have mentioned the original publication in a correct manner.

So at the moment there is really a kind of "reference war" between researchers concerned about (often Google-induced, as already shown) Cut and Paste plagiarism on the web and people thinking euphorically about free licenses. In this report we argue that free licenses on the web (GNU, Creative Commons, and others) should be seen sceptically – especially in the context of scientific or artistic production which means in all social systems concerned with creative processes and intellectual products.

To show how widespread in all social systems plagiarism already is, just let us give the following examples. We start with the "chronological" aspect of a questionable socialisation into plagiarism und continue to discuss case studies of plagiarism in various social systems.

- The new cultural technique of verbatim appropriation of found text segments can start in childhood if the child already has access to a cellular phone or to a computer: Youngsters for example tend to spread identically worded messages by SMS (this is already quite well documented by empirical media psychology), they also tend to appropriate contingent formulas for the headlines of their nickpages, for "personal" slogans and the reflection of basic attitudes towards life. Sentences like "ï ÑëËð Ÿøµ ±Ø £ïƒË! Wï±hØµ± ÿØµ ±hË wØ®£ð ï$ bØ®ïÑg!" or longer phrases as for example...

"×._YôÛ ÂrÊ ôNê SpÊcÎâL PêRsÔn Ôf 1o0o0o_.× ×._BûT yÔu ÂrÊ ThÊ ÔnÊ BâBê – WhÔ î LôVê Sô DÂmN MûCh_.× ×._ThÂnK YôÛ BîG HêÂrT FôR ThÊ MôMêNts YôÛ LîStn tÔ mÊ_.× ×._Dô YôÛ KnÔw.. ThÂt YôÛ ÂrÊ SôMêThÎnG SpÊcÎâL?_.× ×._LôVê YôÛ_.×"
[Text examples taken from Weber 2007a, 122 ff.]

... give proof of a quite dramatic change in text culture: Text fragments circulate without the need to involve one's brain very much. Text chunks spread in a "memetic" way, the idea of "authorship" or a "real" and "authentic" message doesn't need to come up any more. The new viral text galaxy of redundant "Weblish" or Leetspeak formulas marks the first step towards a text culture without brains, towards a new cognitive distance between what your brain is occupied with and the text transmitted.

For a generation which downloads songs, video files, ring tones, and cellular phone logos from the Internet, also the downloading and appropriation of text seems to be something completely normal. It is a big deficit of nearly all current empirical studies dealing with young people and their relationship to new media that the target group was only asked about the downloading of images, videos, and music, but not about the downloading of texts.

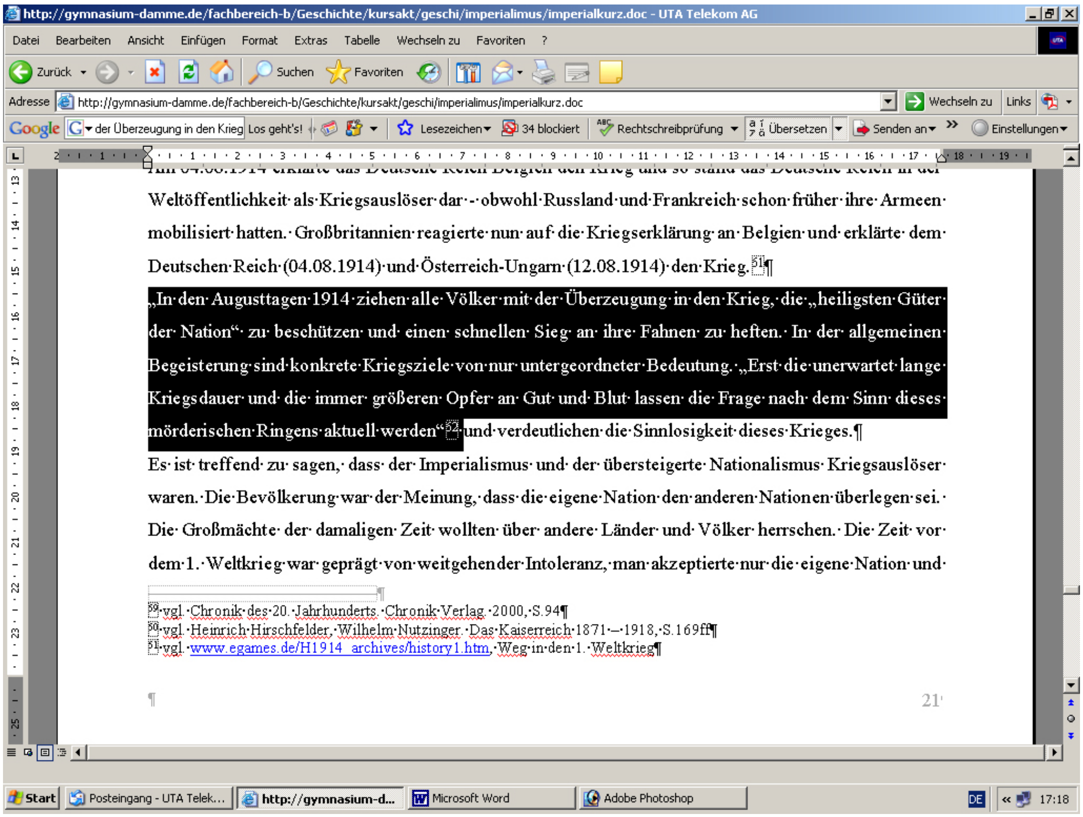
- The copy & paste culture introduced by Leetspeak formulas is continued with the (mis)use of Wikipedia by many pupils: In the moment there are no empirical investigations on the abuse of Wikipedia texts, but many singular reports and case studies give a hint of a dramatic change in knowledge production: For doing a school presentation on the topic of "Trafalgar Square", it often seems to be sufficient to copy & paste the whole text from the Wikipedia entry. Thus, pupils learn quickly how knowledge is produced in the beginning of the third millennium: Google the key term, then click on the Wikipedia link (which is very often ranked on top position or at least amongst the top three, as proven in Chapter 1), then copy & paste this text into your document (if you are "creative", use Google image search and produce a cool PowerPoint presentation with some images or even short movies not already found on the Wikipedia). An Austrian teacher reported in this context that he warned a pupil not to do copy and paste from Wikipedia. The pupil answered that he doesn't understand this warning: If it's already written in the Wikipedia, it can – and virtually must – be used for a presentation.

- When pupils do written assignments, they do not need to write texts on their own any more. They can go to paper mills especially addressed to pupils with countless ready-made school assignments, as for example http://www.schoolunity.de. This web site already welcomes the willing plagiarist with an ethically highly problematic statement:

"No feeling like working on yourself? Just search amongst thousands of ready-made written assignments, presentations, specialised texts or biographies and print whatever you need."
[From http://www.schoolunity.de – translation by the authors of this report]

Of course there is no standardised plagiarism check on this site. The only indicator for the quality of a paper is the mark the "author" got in his or her school. We suppose that many papers on such paper mills are plagiarised – which means when one uses them without quotation he or she plagiarises plagiarism (and with correct citation you still cite plagiarism). The problem of second order plagiarism comes into view: Compare the following two screenshots and you will clearly see that the assignment on http://www.schoolunity.de is already plagiarised from an older written assignment, done for another school. But also in this "original" work the original paragraph from an older book is referenced so sloppy and in such a grubby way that it is impossible to clear the references any more without consulting the original book source [for a documentation and discussion of this example also see Weber 2007a, 59 ff.]. Of course such paper mills always suggest that there is no need to go back to the original source.

Figure 18: Example of a written assignment for school (with sloppy citation)


http://gymnasium-damme.de/fachbereich-b/Geschichte/kursakt/geschi/imperialimus/imperialkurz.doc - UTA Telekom AG
Datei Bearbeiten Ansicht Einfügen Format Extras Tabelle Wechseln zu Favoriten ?
Zurück Suchen Favoriten
Adresse http://gymnasium-damme.de/fachbereich-b/Geschichte/kursakt/geschi/imperialimus/imperialkurz.doc
Wechseln zu Links
Google der Überzeugung in den Krieg Los geht's! Lesezeichen 34 blockiert Rechtschreibprüfung Übersetzen Senden an Einstellungen
Start Posteingang - UTA Telek... http://gymnasium-d... Microsoft Word Adobe Photoshop
DE 17:18

Am 04.08.1914 erklärte das Deutsche Reich Belgien den Krieg und so stand das Deutsche Reich in der Weltöffentlichkeit als Kriegsauslöser dar - obwohl Russland und Frankreich schon früher ihre Armeen mobilisiert hatten. Großbritannien reagierte nun auf die Kriegserklärung an Belgien und erklärte dem Deutschen Reich (04.08.1914) und Österreich-Ungarn (12.08.1914) den Krieg.[51]

„In den Augusttagen 1914 ziehen alle Völker mit der Überzeugung in den Krieg, die „heiligsten Güter der Nation“ zu beschützen und einen schnellen Sieg an ihre Fahnen zu heften. In der allgemeinen Begeisterung sind konkrete Kriegsziele von nur untergeordneter Bedeutung. „Erst die unerwartet lange Kriegsdauer und die immer größeren Opfer an Gut und Blut lassen die Frage nach dem Sinn dieses mörderischen Ringens aktuell werden“[52] und verdeutlichen die Sinnlosigkeit dieses Krieges.

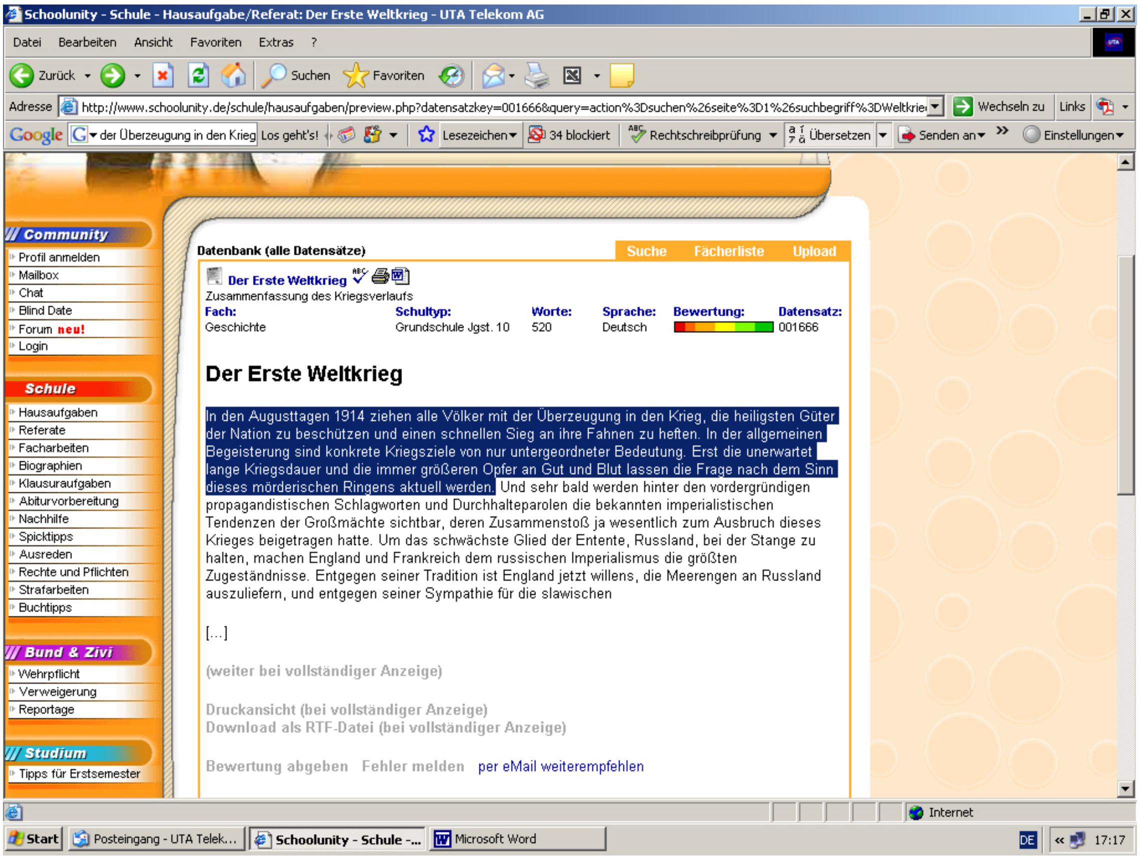
Es ist treffend zu sagen, dass der Imperialismus und der übersteigerte Nationalismus Kriegsauslöser waren. Die Bevölkerung war der Meinung, dass die eigene Nation den anderen Nationen überlegen sei. Die Großmächte der damaligen Zeit wollten über andere Länder und Völker herrschen. Die Zeit vor dem 1. Weltkrieg war geprägt von weitgehender Intoleranz, man akzeptierte nur die eigene Nation und

[49] vgl. Chronik des 20. Jahrhunderts. Chronik Verlag. 2000, S.94
[50] vgl. Heinrich Hirschfelder, Wilhelm Nutzinger. Das Kaiserreich 1871 – 1918, S.169ff
[51] vgl. www.egames.de/H1914_archives/history1.htm, Weg in den 1. Weltkrieg

21

[http://gymnasium-damme.de/fachbereich-b/Geschichte/kursakt/geschi/imperialimus/imperialkurz.doc, p. 21, visited 20/5/07]

Figure 19: Plagiarised version of this assignment in a paper mill for school kids



Datenbank (alle Datensätze)

Der Erste Weltkrieg
Zusammenfassung des Kriegsverlaufs

Fach:	Schultyp:	Worte:	Sprache:	Bewertung:	Datensatz:
Geschichte	Grundschule Jgst. 10	520	Deutsch	🟥🟧🟨🟩	001666

Der Erste Weltkrieg

In den Augusttagen 1914 ziehen alle Völker mit der Überzeugung in den Krieg, die heiligsten Güter der Nation zu beschützen und einen schnellen Sieg an ihre Fahnen zu heften. In der allgemeinen Begeisterung sind konkrete Kriegsziele von nur untergeordneter Bedeutung. Erst die unerwartet lange Kriegsdauer und die immer größeren Opfer an Gut und Blut lassen die Frage nach dem Sinn dieses mörderischen Ringens aktuell werden. Und sehr bald werden hinter den vordergründigen propagandistischen Schlagworten und Durchhalteparolen die bekannten imperialistischen Tendenzen der Großmächte sichtbar, deren Zusammenstoß ja wesentlich zum Ausbruch dieses Krieges beigetragen hatte. Um das schwächste Glied der Entente, Russland, bei der Stange zu halten, machen England und Frankreich dem russischen Imperialismus die größten Zugeständnisse. Entgegen seiner Tradition ist England jetzt willens, die Meerengen an Russland auszuliefern, und entgegen seiner Sympathie für die slawischen

[...]

(weiter bei vollständiger Anzeige)

Druckansicht (bei vollständiger Anzeige)
Download als RTF-Datei (bei vollständiger Anzeige)

Bewertung abgeben Fehler melden per eMail weiterempfehlen

[http://www.schoolunity.de/schule/hausaufgaben/preview.php?datensatzkey=001666&query=action%, visited 20/5/07]

- When pupils once socialised with the Google Copy Paste technique later come to university, they are confronted with strict scientific citation guidelines whose deeper sense they do not fully understand any more. In their eyes quotation seems to be something annoying, sometimes even something ridiculous (once I heard a student say "Why do I have to cite it, it's clear that I have it from the web!"). Amongst the current generation there is no or only very little feeling for the need of reflecting a source or re-checking that which seems to be evident because it's written on the net. In one case one author of this report proofed that a term paper from a German student published on http://www.hausarbeiten.de (on the Austrian philosopher and psychologist Ernst von Glasersfeld) was fully copied from a years older web source. But the author earned money with plagiarism, and the text could be plagiarised by other students willing to pay 4.99 Euro and thus committing second order plagiarism [the case is documented in Weber 2007e].

- Unfortunately plagiarism is not limited to pupils and students. Stefan Weber collected many cases of plagiarism by teachers and professors. One case dealt with a stolen PowerPoint presentation of a whole term lecture [Weber 2007a, 64], another one with dozens of uncited and paraphrased pages in a post-doctoral dissertation. Another professor copied more than 100 pages of his former colleague into his post-doctoral dissertation. Plagiarism by teachers and professors is always a big problem because the question arises: Which citation ethics, which reference culture do they teach in their courses, and which degree of misuse do they tolerate?

- Plagiarism in the Web 2.0: There are many reported cases of plagiarism in Wikipedia [N. N. 2006b], in other Wikipedia clones on the net as well as in weblogs [some cases are discussed in Weber 2007b]. One problem with Wikipedia is the fact that the origin of a text is often uncertain [for a critique see also Weber 2005b and 2007a, 27ff.]. Nobody can control if an original text – later on subject of various changes/adaptations by the net community – is plagiarised or not. For example the text could come from a book source which is not cited properly [also Weber 2005b and 2007a showing concrete examples]. In some way Web 2.0 applications like Wikis and Weblogs produce a second order knowledge galaxy often originating in print sources, but very often with an unclear reference system. This gap makes data, information, and knowledge drawn from the web often problematic. Again the logic of RSS Feeds (e. g. the automatic republishing of news snippets) from the net and the logic of exclusive publishing from the print era collide. Some bloggers also confuse the (legal) RSS feed possibility with the illegal appropriation of a text from another site: Syndication and plagiarism usually are not the same. But nevertheless a constructive dialogue between these two paradigms is absolutely necessary and should be put on top place of the agenda in media science as well as in copyright law.

Table 5: Copyright and copyleft paradigm in comparison

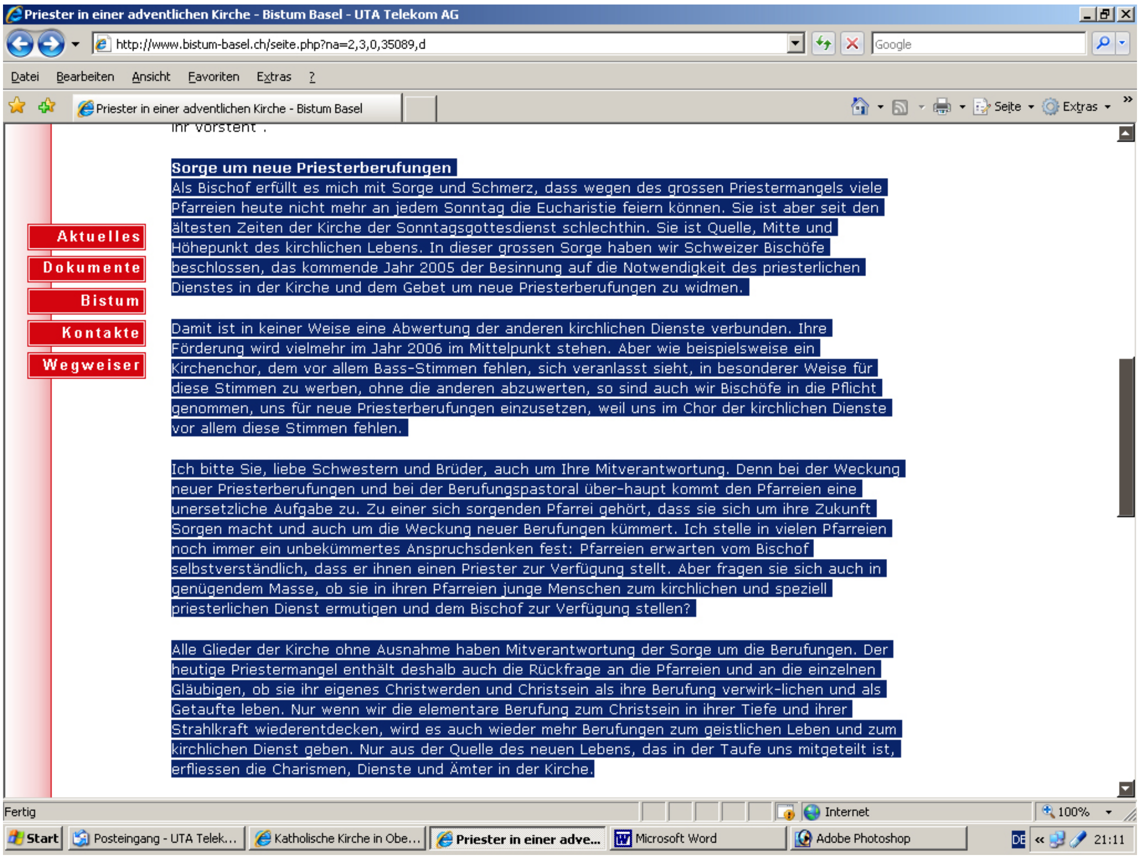
PRINT LOGIC **Copyright paradigm**	**NET (OR WEB 2.0) LOGIC** **Copyleft paradigm**
One author or a group of authors	Author(s) not necessarily mentioned or nickname(s), avatars,...
Publishing companies as publishers	Free Licenses
Control over distribution by publishers	RSS Feeds, free flow of information

[Weber 2007 for this report]

- As already mentioned, plagiarism transcends scientific knowledge production and the educational system. It also affects nearly every other social system. Plagiarism also transcends computer-based or net-based information, plagiarism was also a (often neglected) concern in the print galaxy. Several cases of book plagiarism have also been detected by Stefan Weber (as described with one case study in Section 4).

- In the following we would like to show one interesting example of plagiarism in religion: In the year 2007 an Austrian bishop plagiarised a sermon for Lenten season from another bishop from Switzerland originally dating from 2004. (Please note that the following two screenshots stem from the web sites of two different bishops – there are of course no quotes or references:)

Figure 20: Original sermon of a Swiss bishop on the web


Priester in einer adventlichen Kirche - Bistum Basel - UTA Telekom AG
http://www.bistum-basel.ch/seite.php?na=2,3,0,35089,d
Google
Datei
Bearbeiten
Ansicht
Favoriten
Extras
?
Priester in einer adventlichen Kirche - Bistum Basel
Seite
Extras
Aktuelles
Dokumente
Bistum
Kontakte
Wegweiser
Fertig
Internet
100%
Start
Posteingang - UTA Telek...
Katholische Kirche in Obe...
Priester in einer adve...
Microsoft Word
Adobe Photoshop
DE
21:11

ihr vorsteht .

Sorge um neue Priesterberufungen

Als Bischof erfüllt es mich mit Sorge und Schmerz, dass wegen des grossen Priestermangels viele Pfarreien heute nicht mehr an jedem Sonntag die Eucharistie feiern können. Sie ist aber seit den ältesten Zeiten der Kirche der Sonntagsgottesdienst schlechthin. Sie ist Quelle, Mitte und Höhepunkt des kirchlichen Lebens. In dieser grossen Sorge haben wir Schweizer Bischöfe beschlossen, das kommende Jahr 2005 der Besinnung auf die Notwendigkeit des priesterlichen Dienstes in der Kirche und dem Gebet um neue Priesterberufungen zu widmen.

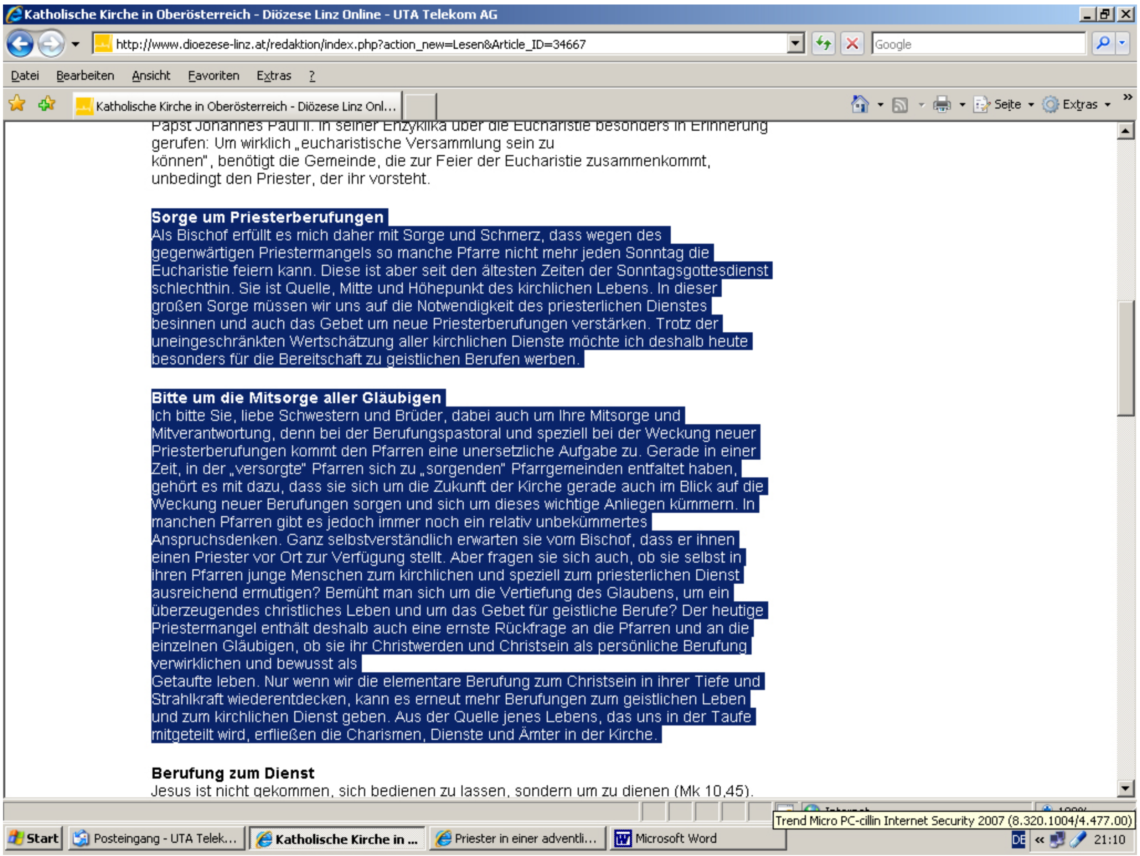
Damit ist in keiner Weise eine Abwertung der anderen kirchlichen Dienste verbunden. Ihre Förderung wird vielmehr im Jahr 2006 im Mittelpunkt stehen. Aber wie beispielsweise ein Kirchenchor, dem vor allem Bass-Stimmen fehlen, sich veranlasst sieht, in besonderer Weise für diese Stimmen zu werben, ohne die anderen abzuwerten, so sind auch wir Bischöfe in die Pflicht genommen, uns für neue Priesterberufungen einzusetzen, weil uns im Chor der kirchlichen Dienste vor allem diese Stimmen fehlen.

Ich bitte Sie, liebe Schwestern und Brüder, auch um Ihre Mitverantwortung. Denn bei der Weckung neuer Priesterberufungen und bei der Berufungspastoral über-haupt kommt den Pfarreien eine unersetzliche Aufgabe zu. Zu einer sich sorgenden Pfarrei gehört, dass sie sich um ihre Zukunft Sorgen macht und auch um die Weckung neuer Berufungen kümmert. Ich stelle in vielen Pfarreien noch immer ein unbekümmertes Anspruchsdenken fest: Pfarreien erwarten vom Bischof selbstverständlich, dass er ihnen einen Priester zur Verfügung stellt. Aber fragen sie sich auch in genügendem Masse, ob sie in ihren Pfarreien junge Menschen zum kirchlichen und speziell priesterlichen Dienst ermutigen und dem Bischof zur Verfügung stellen?

Alle Glieder der Kirche ohne Ausnahme haben Mitverantwortung der Sorge um die Berufungen. Der heutige Priestermangel enthält deshalb auch die Rückfrage an die Pfarreien und an die einzelnen Gläubigen, ob sie ihr eigenes Christwerden und Christsein als ihre Berufung verwirk-lichen und als Getaufte leben. Nur wenn wir die elementare Berufung zum Christsein in ihrer Tiefe und ihrer Strahlkraft wiederentdecken, wird es auch wieder mehr Berufungen zum geistlichen Leben und zum kirchlichen Dienst geben. Nur aus der Quelle des neuen Lebens, das in der Taufe uns mitgeteilt ist, erfliessen die Charismen, Dienste und Ämter in der Kirche.

[http://www.bistum-basel.ch/seite.php?na=2,3,0,35089,d, visited 20/5/07]

Figure 21: Plagiarised version of this sermon by an Austrian bishop


Katholische Kirche in Oberösterreich - Diözese Linz Online - UTA Telekom AG
http://www.dioezese-linz.at/redaktion/index.php?action_new=Lesen&Article_ID=34667
Google
Datei
Bearbeiten
Ansicht
Favoriten
Extras
?
Katholische Kirche in Oberösterreich - Diözese Linz Onl...
Seite
Extras
Trend Micro PC-cillin Internet Security 2007 (8.320.1004/4.477.00)
Start
Posteingang - UTA Telek...
Katholische Kirche in ...
Priester in einer adventli...
Microsoft Word
DE
21:10

Papst Johannes Paul II. in seiner Enzyklika über die Eucharistie besonders in Erinnerung gerufen: Um wirklich „eucharistische Versammlung sein zu können", benötigt die Gemeinde, die zur Feier der Eucharistie zusammenkommt, unbedingt den Priester, der ihr vorsteht.

Sorge um Priesterberufungen
Als Bischof erfüllt es mich daher mit Sorge und Schmerz, dass wegen des gegenwärtigen Priestermangels so manche Pfarre nicht mehr jeden Sonntag die Eucharistie feiern kann. Diese ist aber seit den ältesten Zeiten der Sonntagsgottesdienst schlechthin. Sie ist Quelle, Mitte und Höhepunkt des kirchlichen Lebens. In dieser großen Sorge müssen wir uns auf die Notwendigkeit des priesterlichen Dienstes besinnen und auch das Gebet um neue Priesterberufungen verstärken. Trotz der uneingeschränkten Wertschätzung aller kirchlichen Dienste möchte ich deshalb heute besonders für die Bereitschaft zu geistlichen Berufen werben.

Bitte um die Mitsorge aller Gläubigen
Ich bitte Sie, liebe Schwestern und Brüder, dabei auch um Ihre Mitsorge und Mitverantwortung, denn bei der Berufungspastoral und speziell bei der Weckung neuer Priesterberufungen kommt den Pfarren eine unersetzliche Aufgabe zu. Gerade in einer Zeit, in der „versorgte" Pfarren sich zu „sorgenden" Pfarrgemeinden entfaltet haben, gehört es mit dazu, dass sie sich um die Zukunft der Kirche gerade auch im Blick auf die Weckung neuer Berufungen sorgen und sich um dieses wichtige Anliegen kümmern. In manchen Pfarren gibt es jedoch immer noch ein relativ unbekümmertes Anspruchsdenken. Ganz selbstverständlich erwarten sie vom Bischof, dass er ihnen einen Priester vor Ort zur Verfügung stellt. Aber fragen sie sich auch, ob sie selbst in ihren Pfarren junge Menschen zum kirchlichen und speziell zum priesterlichen Dienst ausreichend ermutigen? Bemüht man sich um die Vertiefung des Glaubens, um ein überzeugendes christliches Leben und um das Gebet für geistliche Berufe? Der heutige Priestermangel enthält deshalb auch eine ernste Rückfrage an die Pfarren und an die einzelnen Gläubigen, ob sie ihr Christwerden und Christsein als persönliche Berufung verwirklichen und bewusst als
Getaufte leben. Nur wenn wir die elementare Berufung zum Christsein in ihrer Tiefe und Strahlkraft wiederentdecken, kann es erneut mehr Berufungen zum geistlichen Leben und zum kirchlichen Dienst geben. Aus der Quelle jenes Lebens, das uns in der Taufe mitgeteilt wird, erfließen die Charismen, Dienste und Ämter in der Kirche.

Berufung zum Dienst
Jesus ist nicht gekommen, sich bedienen zu lassen, sondern um zu dienen (Mk 10,45).

[http://www.dioezese-linz.at/redaktion/index.php?action_new=
Lesen&Article_ID=34667, visited 20/5/07]

When accused of plagiarism by an Austrian radio station, the bishop said the text was a common work of both bishops and that he forgot the footnotes because he had to leave for a funeral to Rome.

- Plagiarism in journalism is also an often neglected problem [see Fedler 2006]. On the web site http://www.regrettheerror.com Craig Silverman wrote that plagiarism was the biggest problem for journalism in the year 2006. Some cases are reported in which even journalists only did copy & paste from Wikipedia [Weber 2007a, 35]. In another documented case a journalist just took over a PR release from a commercial TV station and sold it to a renowned online magazine as his own story. While in the nineties some severe cases of fabrication shocked media ethics, cases of cut and paste plagiarism are a relatively new way of misconduct in journalism often not seen by the public. Especially in this field we recommend to do much more empirical research.

- "Famous" plagiarism cases in the context of politics were sometimes reported worldwide in the media – usually without any consequences for the plagiarists: President Putin was accused to have stolen text in his dissertation from an older American book; the British government was accused to have simply copied text for a decisive Iraq dossier from a ten years older dissertation, and so on.

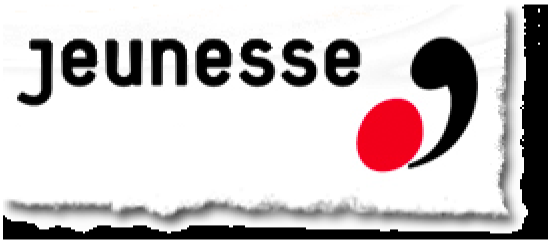
- Also in fiction more than a handful of big plagiarism accusations and also proved cases of plagiarism drew much public attention in the last years.

- In arts plagiarism again transcends the domain of texts and also affects stolen melodies from pop hits or plagiarised narratives of cinema movies.

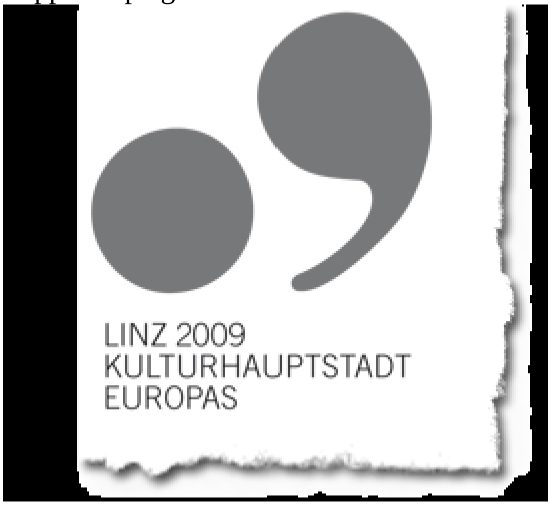
- Plagiarism accusations and cases are also reported from all fields of the so-called "creative industries". – The following example shows a case of a supposed logo theft.

Figures 22 and 23: Accusation of logo plagiarism

The original:



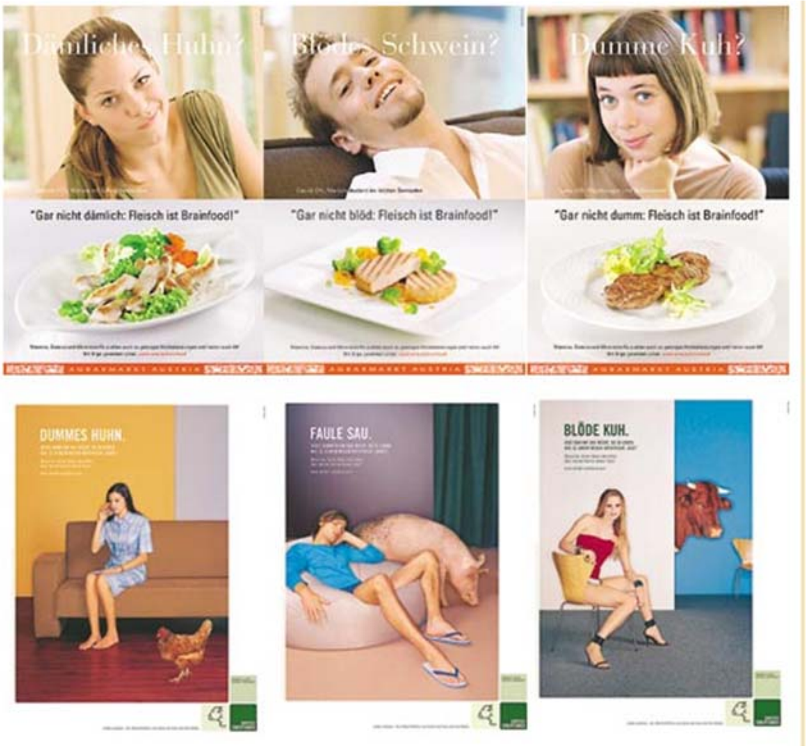
Supposed plagiarism:



[Both images from http://www.qxm.de/gestaltung/20060626-163801/plagiat-oder-zufall?com=1, visited 20/5/07]

At last we have to mention that plagiarism of ideas is always hard to identify. In the following example the PR campaign below (dating from 2006) was accused to be plagiarism of the campaign above from 2002 (for a non-expert observer this might be quite hard to reproduce).

Figure 24: Accusation of idea plagiarism in a PR campaign


Dämliches Huhn?
Blödes Schwein?
Dumme Kuh?
"Gar nicht dämlich: Fleisch ist Brainfood!"
"Gar nicht blöd: Fleisch ist Brainfood!"
"Gar nicht dumm: Fleisch ist Brainfood!"
AGRARMARKT AUSTRIA
DUMMES HUHN.
FAULE SAU.
BLÖDE KUH.

[Comparison from "medianet" print edition, 16 January 2007, p. 13]

Not only texts, but also mathematical formulas and illustrating figures can be plagiarised in scientific and otherwise works. In the following (also a good example for plagiarism from an older print source!) you see a page from a plagiarised Austrian post-doctoral dissertation compared to the original.

The original page from a concise dictionary (anthology) dating 1979:

Figure 25: Original scientific text from 1979

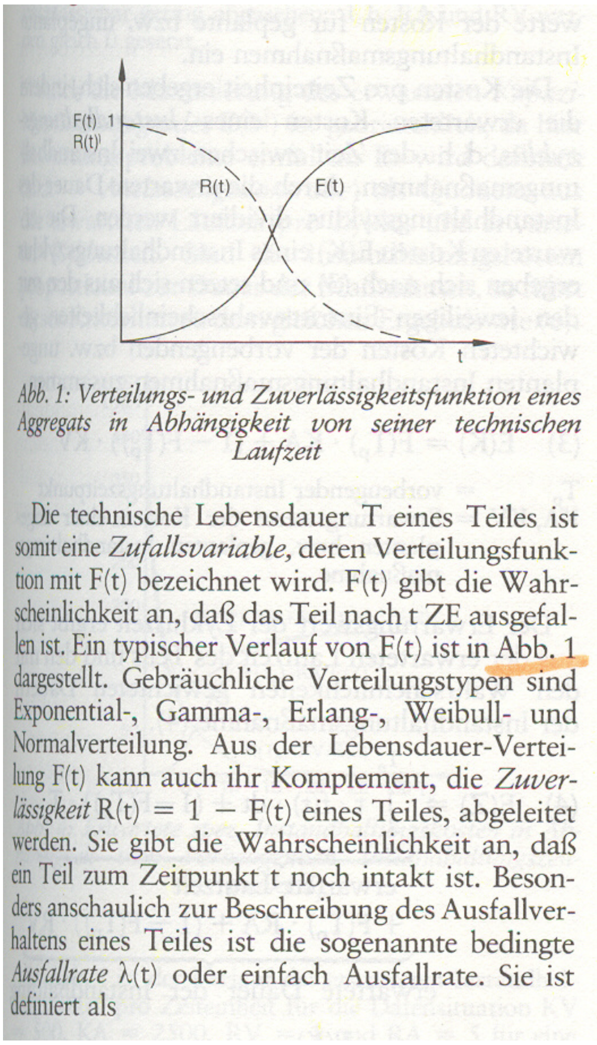

F(t) 1
R(t)
R(t)
F(t)
t

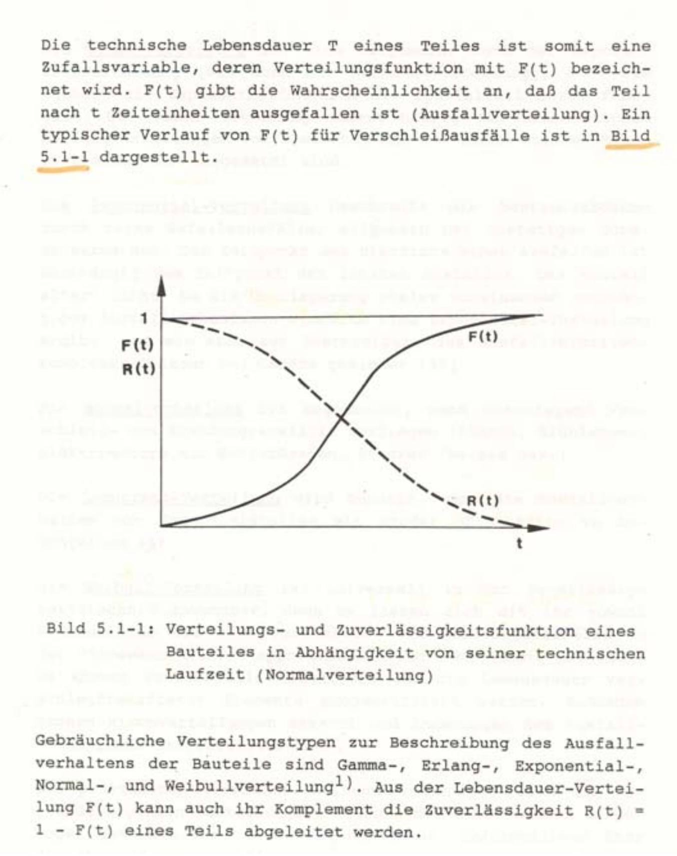
Abb. 1: Verteilungs- und Zuverlässigkeitsfunktion eines Aggregats in Abhängigkeit von seiner technischen Laufzeit

Die technische Lebensdauer T eines Teiles ist somit eine *Zufallsvariable,* deren Verteilungsfunktion mit F(t) bezeichnet wird. F(t) gibt die Wahrscheinlichkeit an, daß das Teil nach t ZE ausgefallen ist. Ein typischer Verlauf von F(t) ist in Abb. 1 dargestellt. Gebräuchliche Verteilungstypen sind Exponential-, Gamma-, Erlang-, Weibull- und Normalverteilung. Aus der Lebensdauer-Verteilung F(t) kann auch ihr Komplement, die *Zuverlässigkeit* R(t) = 1 − F(t) eines Teiles, abgeleitet werden. Sie gibt die Wahrscheinlichkeit an, daß ein Teil zum Zeitpunkt t noch intakt ist. Besonders anschaulich zur Beschreibung des Ausfallverhaltens eines Teiles ist die sogenannte bedingte *Ausfallrate* λ(t) oder einfach Ausfallrate. Sie ist definiert als

[Scan from Werner Kern (ed.). Handwörterbuch der Produktionswirtschaft. Stuttgart: Poeschel, 1979, column 825]

The plagiarised page in the post-doctoral dissertation from 1988:

Figure 26: Plagiarised scientific text from 1988



Die technische Lebensdauer T eines Teiles ist somit eine Zufallsvariable, deren Verteilungsfunktion mit F(t) bezeichnet wird. F(t) gibt die Wahrscheinlichkeit an, daß das Teil nach t Zeiteinheiten ausgefallen ist (Ausfallverteilung). Ein typischer Verlauf von F(t) für Verschleißausfälle ist in Bild 5.1-1 dargestellt.

1
F(t)
R(t)
F(t)
R(t)
t

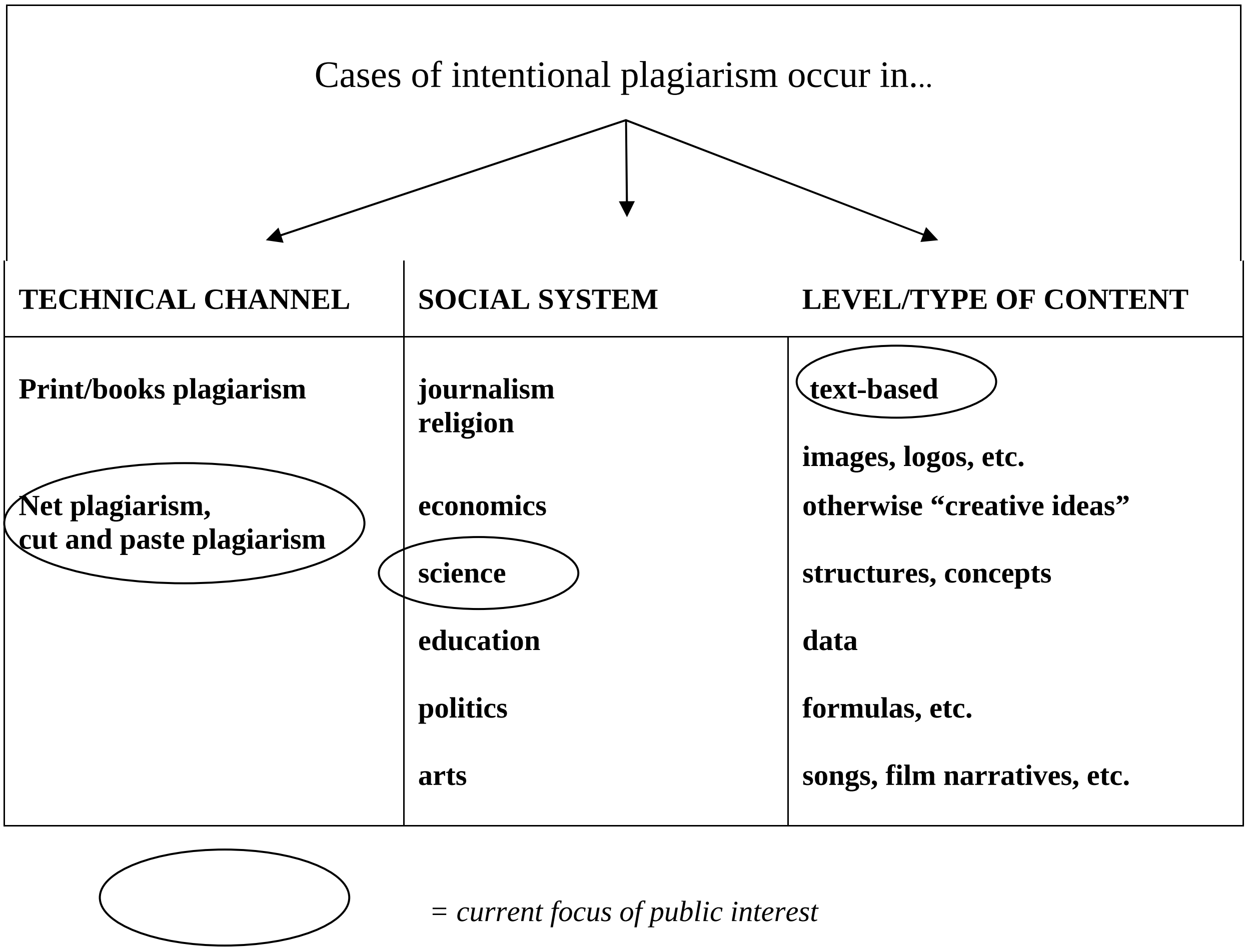
Bild 5.1-1: Verteilungs- und Zuverlässigkeitsfunktion eines Bauteiles in Abhängigkeit von seiner technischen Laufzeit (Normalverteilung)

Gebräuchliche Verteilungstypen zur Beschreibung des Ausfallverhaltens der Bauteile sind Gamma-, Erlang-, Exponential-, Normal-, und Weibullverteilung[1]. Aus der Lebensdauer-Verteilung F(t) kann auch ihr Komplement die Zuverlässigkeit R(t) = 1 - F(t) eines Teils abgeleitet werden.

[Scan from N. N., Anlagenwirtschaft unter besonderer Akzentuierung des Managements der Instandhaltung. Post-doctoral dissertation, 1988, p. 91]

Thinking of all faces of plagiarism mentioned above, we come to the following diagram:

Table 6: **The various faces of plagiarism**



Cases of intentional plagiarism occur in...

TECHNICAL CHANNEL	**SOCIAL SYSTEM**	**LEVEL/TYPE OF CONTENT**
Print/books plagiarism	**journalism** **religion**	**text-based**
		images, logos, etc.
Net plagiarism, **cut and paste plagiarism**	**economics**	**otherwise “creative ideas”**
	science	**structures, concepts**
	education	**data**
	politics	**formulas, etc.**
	arts	**songs, film narratives, etc.**

= *current focus of public interest*

[Weber 2007 for this report]

One big problem is the fact that the majority of the social systems concerned has no institutionalised routines for dealing with accusations and actual proved cases of plagiarism (we do not speak of the legal dimension here, but of the ethical aspect and the aspect of the eminently important presupposition that we always have to trust and rely upon the knowledge-producing social systems!). There is still a relatively widespread mentality to treat cases of intentional intellectual theft as harmless peccadillos. Often plagiarists as well as responsible persons tend to play down the problem: They speak of a "computer mistake", a "problem with the floppy disk" (when the plagiarism occurred in the eighties or nineties) or of some other kind of sloppiness. In many cases people are more worried about the image of an institution and the prevention of bad media coverage than on what actually happens in the concrete knowledge production: the image comes before the content.

In the next Section we would like to discuss some strategies on how to overcome plagiarism and why they all are insufficient so far: The plagiarist always seems to be far ahead of the plagiarism searcher or "hunter" – be it a person or a computer software.

For references see Section 16.

Section 4: Human and computer-based ways to detect plagiarism – and the urgent need for a new tool

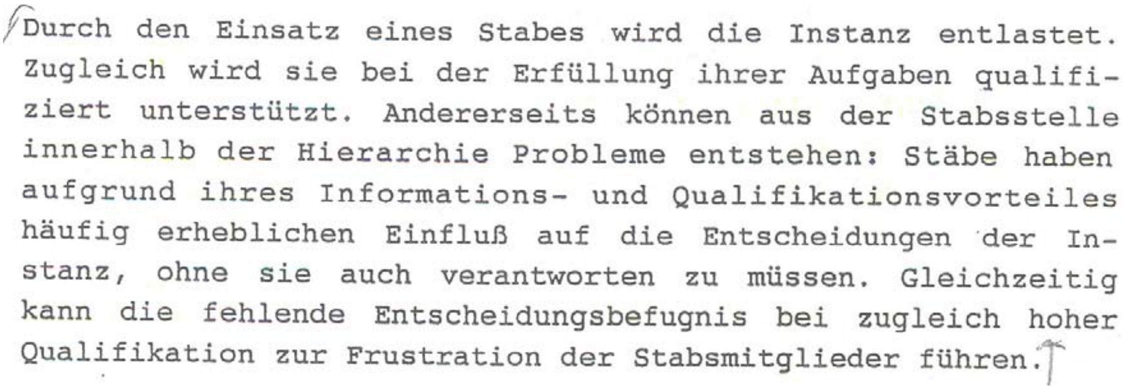
(Note: Sections 1-5 are basically material produced by S. Weber after discussions with H. Maurer, with H. Maurer doing the final editing)

In the current cheating culture the plagiarists often are superior to the originators or the evaluating teaching staff. Of course this is also a question of generations: The younger people – pupils and students – are more familiar with the latest technological developments than the older generation. In fact, we suppose that a "skills gap" between pupils and students on the one side and the faculty on the other side started about the year of 2000. This was the time when students realised that they could cut and paste texts from the web very easily – and that the chance to be detected was very small. Of course, this can be interpreted and is a sign of progress in the younger generation, yet in this case the progress is exploited in an undesirable way. Meanwhile many professors and lecturers know that they have to google when they once read suspicious text fragments, e. g. written in a highly elaborated prose with tacit knowledge the student simply couldn't have. Usually, such texts are very easy to detect, and it is a shame for some universities that some clearly plagiarised texts remained undetected and were even evaluated with good or very good marks [see the documented cases in Weber 2007a]. But we have to notice that even today, Google – as possible first step into plagiarism – is often helpless when plagiarists use advanced techniques of text theft. Some strategies are mentioned in the following:

- A student willing to plagiarise efficiently can order some current master or doctoral thesis in the context he or she should write on by interlending. For example if you study in Berlin, order some master thesis from Munich or Zurich. The web can help the plagiarist: Exclude the possibility that the professors which have judged the other thesis know your own professor closer, exclude that they could be in nearer contact (you can do that by just googling their names with a plus). Exclude that the borrowed master thesis are already as full text versions on the web (again use Google for that check!). Then make a "cross-over", a post-modern mash-up of the ordered thesis. If you have excluded all these risk factors, the probability that your hybrid work will ever be detected is rather small.
- A student willing to plagiarise efficiently can order some current master or doctoral thesis by interlending from foreign countries and translate them or let them translate. Again, proceed as described above. And again, you can succeed with this method!
- Another method is to use a given text and to replace about every third or fourth noun or verb with a synonym. The disadvantage is that you have to use your brain a little for this method, the big advantage is that current anti-plagiarism software won't detect your betrayal (as we will show below).
- Probably still the best method is to plagiarise a print source which isn't already digitized. The problem is that one day it could be digitised. In the following, we will discuss such an example:

In 1988 an Austrian professor wrote in his post-doctoral dissertation the following "genuine" text (we suppose the text was genuinely written because there are no quotes, references or footnotes around the text):

Figure 27: A suspicious text segment

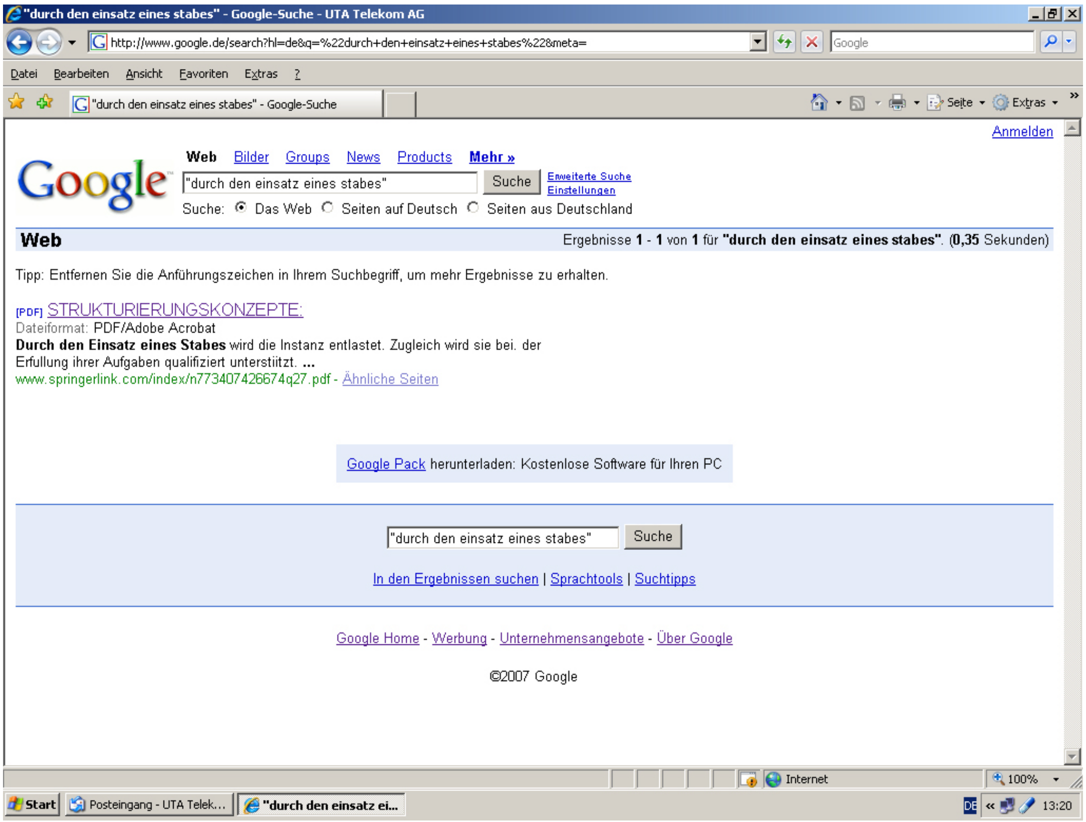


Durch den Einsatz eines Stabes wird die Instanz entlastet. Zugleich wird sie bei der Erfüllung ihrer Aufgaben qualifiziert unterstützt. Andererseits können aus der Stabsstelle innerhalb der Hierarchie Probleme entstehen: Stäbe haben aufgrund ihres Informations- und Qualifikationsvorteiles häufig erheblichen Einfluß auf die Entscheidungen der Instanz, ohne sie auch verantworten zu müssen. Gleichzeitig kann die fehlende Entscheidungsbefugnis bei zugleich hoher Qualifikation zur Frustration der Stabsmitglieder führen.

[Scan from N. N., Anlagenwirtschaft unter besonderer Akzentuierung des Managements der Instandhaltung. Post-doctoral dissertation, 1988, p. 135]

If you want to do a manual Google check, it is sufficient to select just a few words (usually you needn't search for special elaborated prose parts). In this case it is enough to type "Durch den Einsatz eines Stabes" (the first five words of the paragraph) into Google. Surprisingly, Google only finds one match. With this simple but effective method, it has become possible to detect plagiarism originating from the print era. In this case, the author plagiarised a book first released in 1986 in the year of 1988. Meanwhile the book was digitized and made available online via springerlink.com. Of course you have to pay to obtain the full text of the book chapter. But nevertheless already the fragment found by Google gives enough evidence of plagiarism:

Figure 28: The original document found by Google



"durch den einsatz eines stabes" - Google-Suche - UTA Telekom AG
http://www.google.de/search?hl=de&q=%22durch+den+einsatz+eines+stabes%22&meta=
Google
Datei Bearbeiten Ansicht Favoriten Extras ?
"durch den einsatz eines stabes" - Google-Suche
Seite Extras
Anmelden
Web Bilder Groups News Products Mehr »
Google
Suche
Erweiterte Suche
Einstellungen

"durch den einsatz eines stabes"

Suche: ● Das Web ○ Seiten auf Deutsch ○ Seiten aus Deutschland

Web Ergebnisse **1** - **1** von **1** für **"durch den einsatz eines stabes"**. (**0,35** Sekunden)

Tipp: Entfernen Sie die Anführungszeichen in Ihrem Suchbegriff, um mehr Ergebnisse zu erhalten.

[PDF] STRUKTURIERUNGSKONZEPTE:
Dateiformat: PDF/Adobe Acrobat
Durch den Einsatz eines Stabes wird die Instanz entlastet. Zugleich wird sie bei. der Erfullung ihrer Aufgaben qualifiziert unterstiitzt. **...**
www.springerlink.com/index/n773407426674q27.pdf - Ähnliche Seiten

Google Pack herunterladen: Kostenlose Software für Ihren PC

"durch den einsatz eines stabes"

Suche
In den Ergebnissen suchen | Sprachtools | Suchtipps
Google Home - Werbung - Unternehmensangebote - Über Google
©2007 Google
Internet
100%
Start
Posteingang - UTA Telek...
"durch den einsatz ei...
DE
13:20

[Screenshot, 18 May 2007]

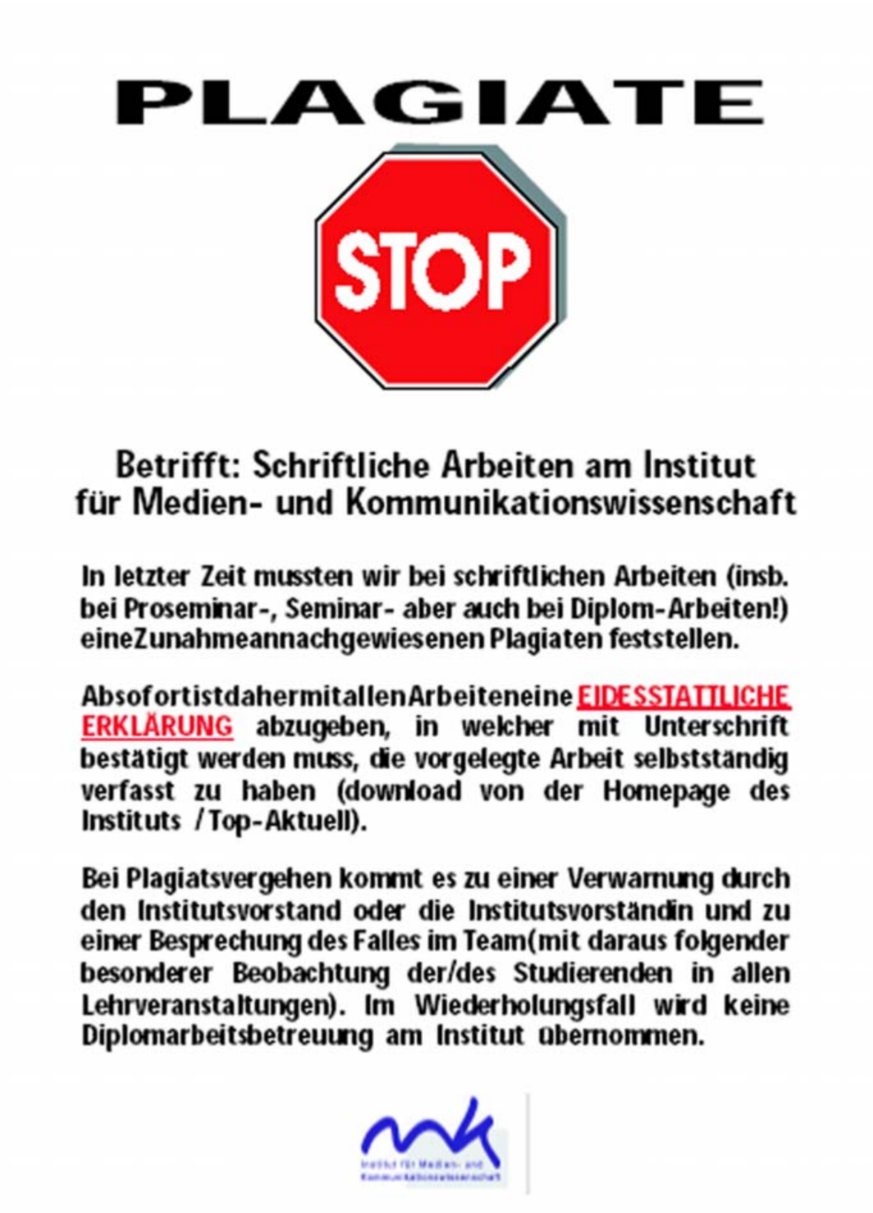
So we come to a central question of our report: If Google's indexing of web sites is as effective as just shown in this example (please note that a document in the paid content area was found!), why didn't Google already develop its own anti-plagiarism or text comparison/text similarity tool? – Steven Johnson already imagined such an "intelligent" text tool in the year of 1999 [Johnson 1999, 157 ff.] – this tool should be able to find texts by tagging keywords "automatically" on a semantic level and also be able to compare them with each other.

As we know, "Google Labs" is the place of permanent try-outs of everything by Google. "Google Books" enables the user to browse through millions of books and scan textual snippets (and not to read whole books!). "Google Scholar" enables the user to look up how a term or person is mentioned or cited in scientific texts. There are various other applications in an experimental state dealing with text – but no text comparison or anti-plagiarism tool as far as one can see. Is Google not interested in this topic? Does Google not see the relevancy in current copying society? Or does Google intentionally look away to protect the interests of plagiarists or even of the whole cut and paste culture (in a way, Google itself does cut and paste, for example automatically with "Google News" when fragments of web sites of other news media are fed into their site). Maybe there are ideological reservations. The problem is that Google is not very transparent about such questions. One author of this study mailed to the German Google spokesman two times about text comparison and anti-plagiarism tools, but there absolutely came no response. It is also possible that Google is following the development of the market for anti-plagiarism tools closely and will move in for a “kill” if it turns out to be economically interesting, see Appendix 2.

As long as this situation won't change, we can only use Google and our brains to detect plagiarism and hope that always more and more texts will be digitized. – Before we have a look upon various anti-plagiarism tools already on the market, let us differ between plagiarism prevention in advance and plagiarism detection thereafter. In an idealistic situation, intentional plagiarism couldn't occur at all: If all students are really interested in what they search, read and write and if they realise the deeper sense of scientific knowledge production, no fake culture could be introduced. Of course, we don't live in this idealistic context, the world has changed dramatically in the last ten years – due to technological, economical, and political transformations. In the educational system, in some way title marketing replaced the classical concepts of education and enlightenment: Many students do not want to accumulate and critically reflect knowledge any more, instead of that they are in search for the quickest way towards an academic degree. Therefore often the fight against plagiarism on the level of its prevention seems to be a lost game.

Two examples of more or less successful "sensitisation strategies" are given. Nearly each academic institute meanwhile warns students not to plagiarise. But in the moment it's not sure if flyers as the following (an example from an Austrian university) are really able to eliminate a kind of "betrayal energy" amongst the students:

Figure 29: Plagiarism warning flyer from the University of Klagenfurt, Austria



PLAGIATE

STOP

Betrifft: Schriftliche Arbeiten am Institut für Medien- und Kommunikationswissenschaft

In letzter Zeit mussten wir bei schriftlichen Arbeiten (insb. bei Proseminar-, Seminar- aber auch bei Diplom-Arbeiten!) eineZunahmeannachgewiesenen Plagiaten feststellen.

AbsofortistdahermitallenArbeiteneine EIDESSTATTLICHE ERKLÄRUNG abzugeben, in welcher mit Unterschrift bestätigt werden muss, die vorgelegte Arbeit selbstständig verfasst zu haben (download von der Homepage des Instituts /Top-Aktuell).

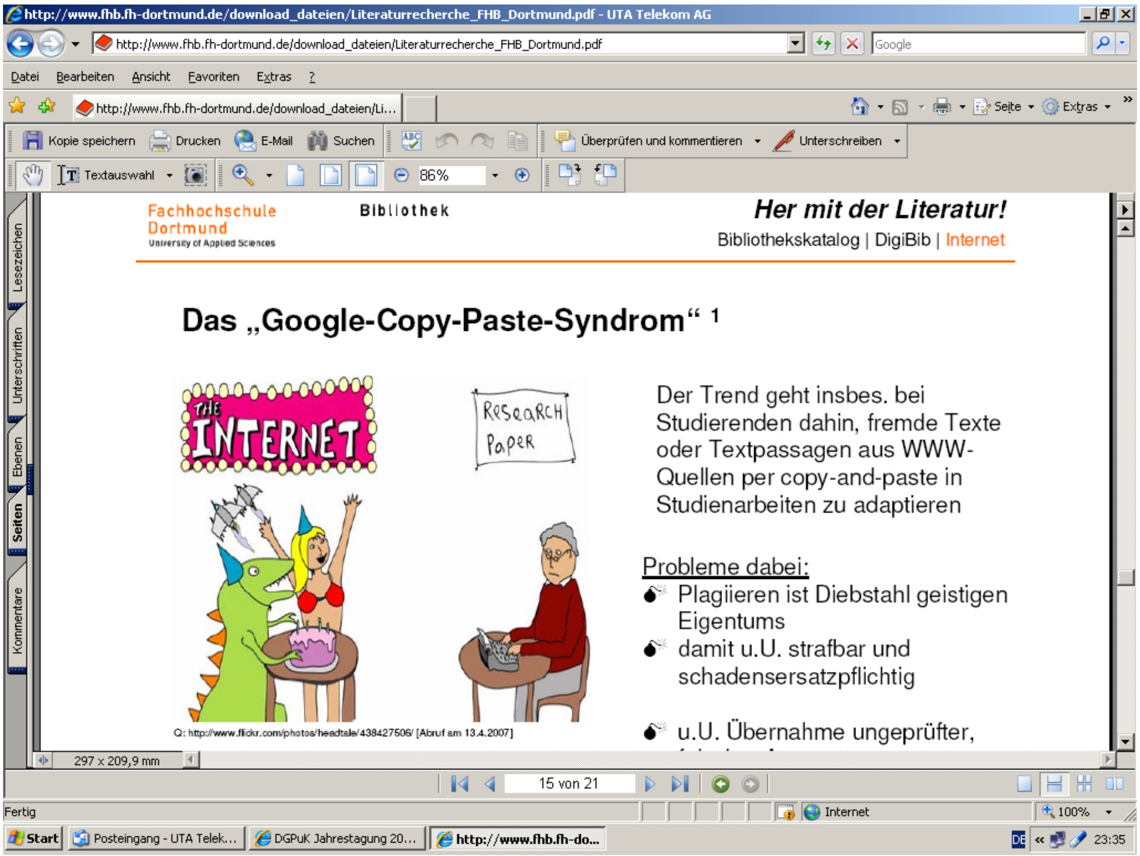
Bei Plagiatsvergehen kommt es zu einer Verwarnung durch den Institutsvorstand oder die Institutsvorständin und zu einer Besprechung des Falles im Team(mit daraus folgender besonderer Beobachtung der/des Studierenden in allen Lehrveranstaltungen). Im Wiederholungsfall wird keine Diplomarbeitsbetreuung am Institut übernommen.

mk

[http://www.uni-klu.ac.at/mk0/studium/pdf/plagiat01.pdf, visited 22/5/07]

Another example is taken from a university in Germany. In a PDF available online, some problematic aspects of web references are discussed and also the Google Copy Paste Syndrome (GCPS) is explicitly mentioned:

Figure 30: The Google Copy Paste Syndrome mentioned in a brochure from a university in Dortmund, Germany



http://www.fhb.fh-dortmund.de/download_dateien/Literaturrecherche_FHB_Dortmund.pdf - UTA Telekom AG
http://www.fhb.fh-dortmund.de/download_dateien/Literaturrecherche_FHB_Dortmund.pdf
Google
Datei Bearbeiten Ansicht Favoriten Extras ?
http://www.fhb.fh-dortmund.de/download_dateien/Li...
Seite
Extras
Kopie speichern
Drucken
E-Mail
Suchen
Überprüfen und kommentieren
Unterschreiben
Textauswahl
86%
Lesezeichen
Unterschriften
Ebenen
Seiten
Kommentare
297 x 209,9 mm
15 von 21
Fertig
Internet
100%
Start
Posteingang - UTA Telek...
DGPuK Jahrestagung 20...
http://www.fhb.fh-do...
DE
23:35
THE INTERNET
ReseaRCH PaPeR

Fachhochschule Dortmund
University of Applied Sciences

Bibliothek

Her mit der Literatur!

Bibliothekskatalog | DigiBib | Internet

Das „Google-Copy-Paste-Syndrom“ 1

Der Trend geht insbes. bei Studierenden dahin, fremde Texte oder Textpassagen aus WWW-Quellen per copy-and-paste in Studienarbeiten zu adaptieren

Probleme dabei:

- Plagiieren ist Diebstahl geistigen Eigentums
- damit u.U. strafbar und schadensersatzpflichtig
- u.U. Übernahme ungeprüfter,

Q: http://www.flickr.com/photos/headtale/438427506/ [Abruf am 13.4.2007]

[http://www.fhb.fh-dortmund.de/download_dateien/Literaturrecherche_FHB_Dortmund.pdf, p. 15, visited 22/5/07]

An overview of how academic institutions worldwide react after proven cases of plagiarism is given in [Maurer & Kappe & Zaka 2006, 1052 ff.; Note: This paper which was written as part of this report is added as Section 17: Appendix 1].

Please note that there is a broad spectrum of possible consequences depending on the gravity of plagiarism – from an obligatory seminar in science ethics to a temporal relegation from university. Generally one can observe that universities in the US and in GB fight plagiarism still more effectively than most universities in (the rest of) Europe. Especially some reported cases from Austria and Germany show that plagiarism is still played down or even tolerated when it occurred within the faculty. But also for plagiarising professors one could imagine some adequate consequences. In critically commenting a current German case of a plagiarising professor in law, the main German "plagiarism hunter" Debora Weber-Wulff wrote in her blog:

"Here we have a clear determination of plagiarism on the part of a professor. Surely something must happen? I can imagine all sorts of things: research funding moratorium; taking away his research assistant the next time it is free; making him take a course in ethics; making the whole department – which appears to have a culture in which such behavior is acceptable – take an ethics course; assigning hours of public service such as doing something about the cataloging backlog in the library or whatever; requesting that he donate the proceeds from the book to financing a course on avoiding plagiarism for students. Surely there must be something which is the equivalent of failing a student for a course in which they submit a plagiarism." [http://copy-shake-paste.blogspot.com, visited 22/5/07]

In the following, we will concentrate on digital measures to fight against plagiarism, which means on strategies that are applied after plagiarism prevention (failed). But we should not forget that also effective digital tools have an important preventive didactic function!

In the moment, plagiarism software systems have problems with at least four varieties of plagiarism:

- plagiarism of print-only sources (with the exception that some print sources could already formerly have been checked by a software which compiles all checked documents in own databases provided that students accepted that storage – than that kind of print plagiarism can be detected). Some systems also have reported problems with online documents only available on pay sites or hidden in the deep web.
- plagiarism of sources which are only available in image formats. If you scan a page not already digitized and accessible online and then convert it with OCR software (optical character recognition) into a text format, all current plagiarism software systems won't detect it.
- plagiarism of passages written in foreign languages. In the moment, no plagiarism software on the market features an integrated automatic translation function. There are experimental attempts to develop a software which is able to translate text originating from foreign languages into a basic English form with normalised words [see Maurer & Kappe & Zaka 2006, 1079 f.; **Note: This paper which was written as part of this report is added as Appendix 1.**]
- plagiarism of original texts in which many words are replaced by synonyms.

To demonstrate that, note the following synonym experiment [for a similar example also see Maurer & Kappe & Zaka 2006, 1067 ff.; **Note: This paper which was written as part of this report is added as Appendix 1.**]

.

We take the following two segments copied verbatim from the web:

1) While it has many nicknames, information-age slang is commonly referred to as leetspeek, or leet for short. Leet (a vernacular form of "elite") is a specific type of computer slang where a user replaces regular letters with other keyboard characters.

2) Emoticon ist ein Kunstwort, das sich aus Emotion und Icon zusammensetzt. Bezeichnet wird damit eine Zeichenfolge (aus normalen Satzzeichen), die einen Smiley nachbildet, um in der schriftlichen elektronischen Kommunikation Stimmungs- und Gefühlszustände auszudrücken.

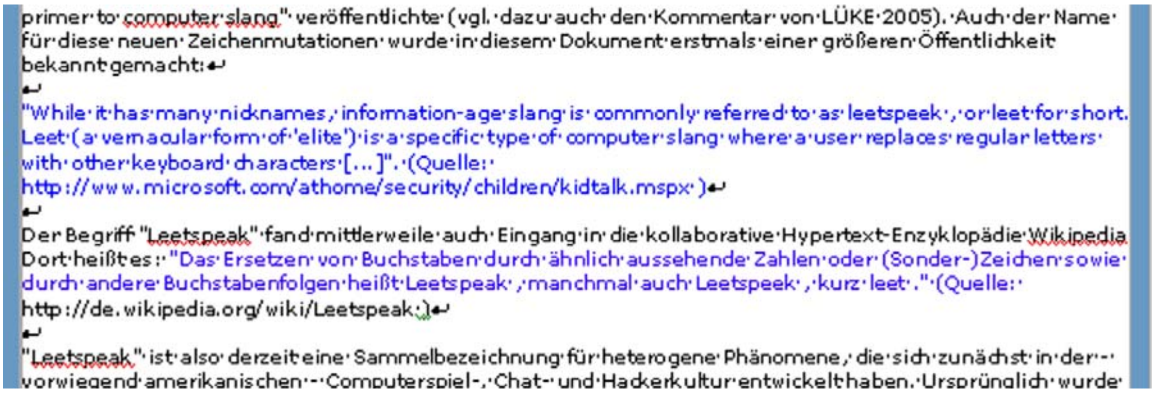
In both text segments, we replace a number of words by synonyms:

*1) While it has many **cognomens**, information-age **vernacular** is **typically** referred to as leetspeek, or leet **in brief**. Leet (a **slang** form of "elite") is a **special** type of computer slang where a user **substitutes common** letters with other **key pad** characters.*

*2) Emoticon ist ein **künstliches Wort**, das sich aus Emotion **(Gefühl)** und Icon **(Bildzeichen)** zusammensetzt. **Gemeint ist** damit eine **Zeichenkette** (aus normalen Satzzeichen), die einen Smiley **imitiert**, um in der schriftlichen [...] Kommunikation **am PC** Stimmungs- **bzw.** Gefühlszustände **zu artikulieren**.*

If you check the first copied passage "borrowed" from the web with a currently used anti-plagiarism tool, the text segment (in our example below a part of another bigger text) is found on the web and therefore in the report marked in blue – see the upper paragraph of the following screenshot. But note that in this case no plagiarism occurred because of correct citation in quotation marks – this is necessary interpretation work of a human!

Figure 31: Mydropbox finding copied text from the web



primer to computer slang" veröffentlichte (vgl. dazu auch den Kommentar von LÜKE 2005). Auch der Name für diese neuen Zeichenmutationen wurde in diesem Dokument erstmals einer größeren Öffentlichkeit bekannt gemacht:

"While it has many nicknames, information-age slang is commonly referred to as leetspeek , or leet for short. Leet (a vernacular form of 'elite') is a specific type of computer slang where a user replaces regular letters with other keyboard characters [...]". (Quelle: http://www.microsoft.com/athome/security/children/kidtalk.mspx)

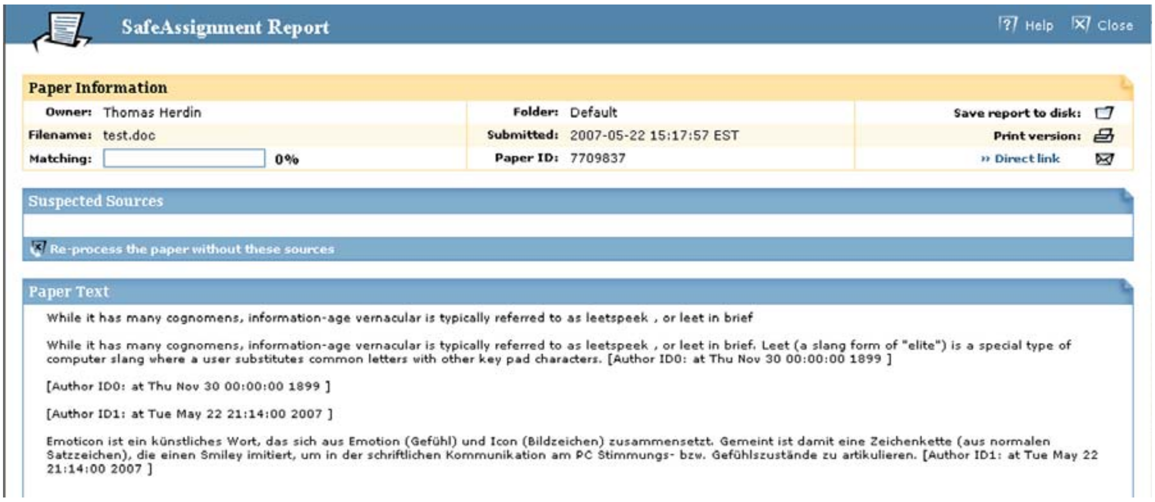
Der Begriff "Leetspeak" fand mittlerweile auch Eingang in die kollaborative Hypertext-Enzyklopädie Wikipedia Dort heißt es : "Das Ersetzen von Buchstaben durch ähnlich aussehende Zahlen oder (Sonder-)Zeichen sowie durch andere Buchstabenfolgen heißt Leetspeak , manchmal auch Leetspeek , kurz leet ." (Quelle: http://de.wikipedia.org/wiki/Leetspeak)

"Leetspeak" ist also derzeit eine Sammelbezeichnung für heterogene Phänomene, die sich zunächst in der -- vorwiegend amerikanischen -- Computerspiel-, Chat- und Hackerkultur entwickelt haben. Ursprünglich wurde

[Screenshot, 14 May 2007]

But if you feed the same system with the first and the second altered passage as rewritten above, the software won't detect anything and reports 0 percent of matches!

Figure 32: Mydropbox failing to detect the same paragraph with synonyms



Paper Information

Owner:	Thomas Herdin	Folder:	Default	Save report to disk:	
Filename:	test.doc	Submitted:	2007-05-22 15:17:57 EST	Print version:	
Matching:	0%	Paper ID:	7709837	» Direct link	

Suspected Sources

Paper Text

While it has many cognomens, information-age vernacular is typically referred to as leetspeek , or leet in brief

While it has many cognomens, information-age vernacular is typically referred to as leetspeek , or leet in brief. Leet (a slang form of "elite") is a special type of computer slang where a user substitutes common letters with other key pad characters. [Author ID0: at Thu Nov 30 00:00:00 1899]

[Author ID0: at Thu Nov 30 00:00:00 1899]

[Author ID1: at Tue May 22 21:14:00 2007]

Emoticon ist ein künstliches Wort, das sich aus Emotion (Gefühl) und Icon (Bildzeichen) zusammensetzt. Gemeint ist damit eine Zeichenkette (aus normalen Satzzeichen), die einen Smiley imitiert, um in der schriftlichen Kommunikation am PC Stimmungs- bzw. Gefühlszustände zu artikulieren. [Author ID1: at Tue May 22 21:14:00 2007]

[Screenshot, 22 May 2007]

This is a proof that current plagiarism software fails to detect synonym plagiarism. In the context of synonym and translation plagiarism, three things should be noticed:

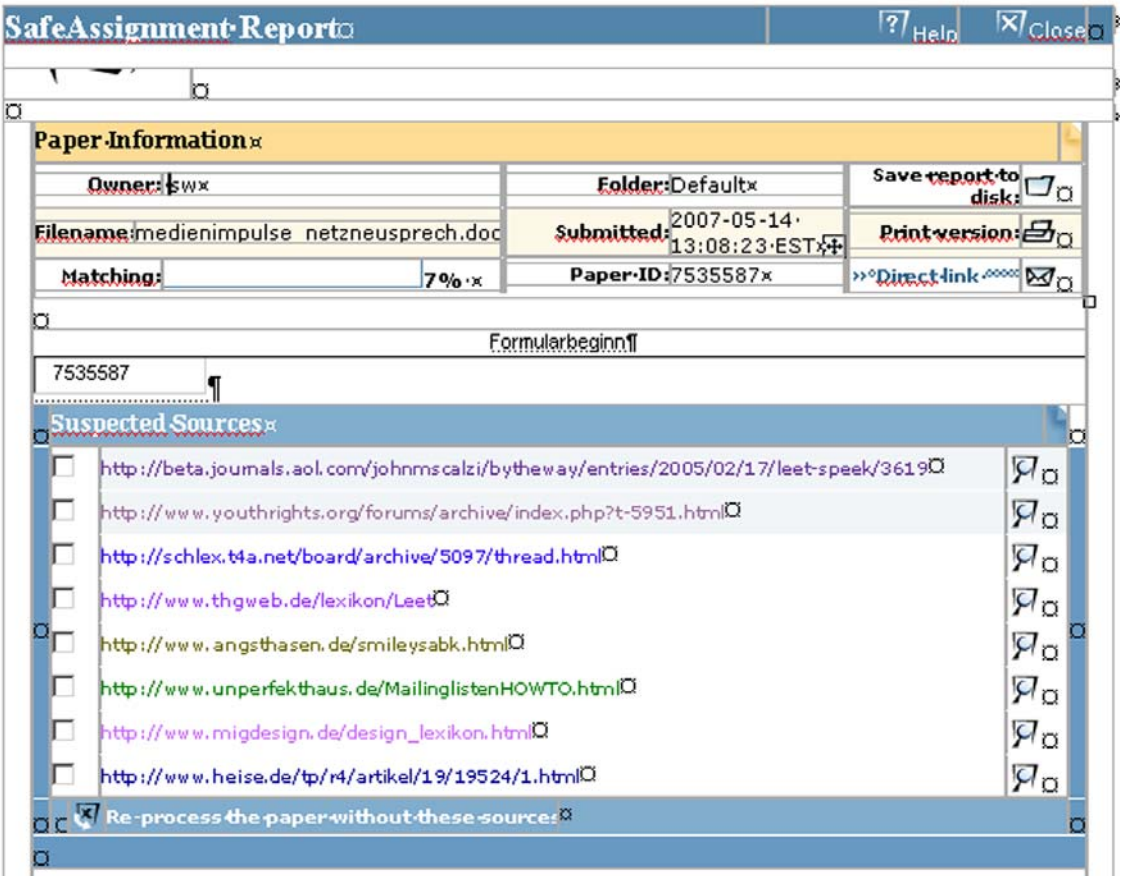
- When a plagiarist replaces single words in sentences by synonyms, the total number of words per sentence won't change a lot. In our examples above, the numbers changed from
  - 17, 23 to 17, 24 in the first paragraph, and from
  - 11, 22 to 14, 24 in the second paragraph.

  This means that a mathematical analysis (maybe with stochastics) of a similar sequence of numbers of words per sentence could indicate plagiarism (of course only when a reference text corpus is given).
- When a plagiarist replaces words in sentences by synonyms, he or she normally must leave all direct quotes and bibliographical references unchanged. The analysis of the citation structure of a document (comparable to the function "related documents" in some scientific databases) therefore allows indicating plagiarism also when single words of the genuine prose were replaced.

- Please note that also in the case of translation plagiarism direct quotes and bibliographical titles on the reference list might remain in their original language. The German professor for informatics Wolfgang Coy already mentioned in 2002 the problem of translation plagiarism and this approach as a possibility to combat it: "In such a case a critical search engine check of the cited literature may help an investigating teacher or reviewer." [Coy 2002, 142]

For this report, we made an unveiling test with one anti-plagiarism tool currently used by many universities worldwide. We submitted a text which is completely available online – as a free PDF which is also found easily by search engines (as a keyword check can show). The URL is http://www.schule.at/dl/Weber-_-Phaenomen_Cyber-Neusprech.pdf. We took a MS Word version of this file and sent it through the plagiarism detection software. The report showed 7 percent similarity with found documents on the web. Of course, 100 percent similarity should have been shown [for comparable results in a broader experiment see also Maurer & Zaka 2007; **Note: This paper which was written as part of this report is added as Appendix 2**.]

Figure 33: Mydropbox reporting 7 percent similarity for a text which is completely online



Paper Information

Owner: swx	Folder: Default	
Filename: medienimpulse_netzneusprech.doc	Submitted: 2007-05-14 13:08:23 EST	
Matching: 7%	Paper ID: 7535587	

7535587

Suspected Sources

□ http://beta.journals.aol.com/johnmscalzi/bytheway/entries/2005/02/17/leet-speek/3619
□ http://www.youthrights.org/forums/archive/index.php?t-5951.html
□ http://schlex.t4a.net/board/archive/5097/thread.html
□ http://www.thgweb.de/lexikon/Leet
□ http://www.angsthasen.de/smileysabk.html
□ http://www.unperfekthaus.de/MailinglistenHOWTO.html
□ http://www.migdesign.de/design_lexikon.html
□ http://www.heise.de/tp/r4/artikel/19/19524/1.html

[Screenshot, 14 May 2007]

The software detected various correctly cited quotes from Wikipedia (for example special definitions of terms like "Erikativ" which couldn't be found in print literature at the time the paper was written – in summer 2005). So the fact that the software finds similarities doesn't mean that the segments were plagiarised – it is also possible that text copied from the web was cited correctly. In this case, everything reported by the software was quoted correctly, but the software didn't find the whole original paper on the web.

German plagiarism expert Debora Weber-Wulff is generally sceptical about anti-plagiarism software not only due to the arguments and findings listed above. In 2004 she fabricated 10 intentionally (semi)plagiarised texts and fed a couple of software systems with them [as part of the first German online tutorial for plagiarism prevention and detection, see Weber-Wulff 2004]. Weber-Wulff tested PlagiarismFinder, turnitin, CopyCatchGold, Damocles, Eve2, MyDropBox, Copyscape, Docol©c, PlagiarismSleuth, Plagiarism Detection Tool, and Glatt Plagiarism Screen Tool [for details see http://plagiat.fhtw-berlin.de/ff/05hilfen/programme.html, visited 22/5/07]. The result was that *out of eleven tested programmes only four were classified as good*. Weber-Wulff's conclusion was that instead of using the majority of the anti-plagiarism programmes one can also throw a coin. She is currently working on new demo versions of plagiarised texts and will repeat her test in summer of 2007 [for another recent test with similar ambivalent results done by a German journalist see http://www.ftd.de/technik/medien_internet/188947.html, visited 22/5/07]

All efforts mentioned so far deal with the idea that there is a text suspicious of plagiarism and a given collection of reference texts (in most cases the indexed web sites of a search engine, be it Google – as in the case of Docol©c and some others which use the Google API – or an own specialised search engine – as for example in the case of turnitin or MyDropBox). All tools compare a given text to an extrinsic "docuverse". There is an alternative method which is relatively new: The so-called "intrinsic plagiarism detection" [Meyer zu Eissen & Stein 2006; Stein & Meyer zu Eissen 2006]. With this method – which is also called stylometry – breaks in the writing style can be quantified by special features – like the "Averaged Word Frequency Class" – which could give a hint for plagiarism. (Of course these tools only indicate breaks in style and thus can limit the search for plagiarism to certain areas, and nothing is said about where the plagiarised material is derived from). There are also interesting attempts in the field of reliable authorship verification, again a development Steven Johnson has welcome in 1999 when he reflected the upcoming revolution in text culture [Johnson 1999, 157 ff.].

We think that intrinsic plagiarism detection or stylometry and authorship verification will have a great future especially for cases of suspected IPR violation from older print texts that will never be digitised. When you examine texts without the possibility to compare them to an online text corpus (there are exceptions as shown above when material is published online later on), there are only intrinsic indicators left.

We made an experiment with a 1988 post-doctoral dissertation from which we already knew that chapter 2 was nearly fully plagiarised and asked Benno Stein from the university of Weimar to do a stylometric analysis for us. The result was that there is a number of other subchapters with style breaks (chapters 3.3, 4.3, 4.4.1, 5.3, 6.2.2.2.1, and 6.2.3). This result could encourage further examinations. A recently developed prototype of a stylometric plagiarism indication system can be found under http://www.picapica.net. In the context of what is developed in the next chapter, a collaboration with the stylometry group of Weimar University seems to be fruitful.

For references see Section 16.

Section 5: Design of a pilot project for advanced plagiarism detection

(Note: Sections 1-5 are basically material produced by S. Weber after discussions with H. Maurer, with H. Maurer doing the final editing)

In analogy to what already exists in the US with the "Center for Academic Integrity" (CAI) we propose such an institution not only responsible for honesty in academia, but in all social systems with a special focus on intellectual property: a *European Centre for Plagiarism and IPR violation detection* (in short ECPIRD, for concrete suggestions see Kulathuramaiyer & Maurer 2007) and more specifically Appendix 3. (This appendix was written as independent paper for this particular study.) This centre should deal with cases of intellectual property rights violation and plagiarism in

- science
- arts
- journalism
- economics
- education
- politics, and
- religion (as shown in table 7 in this report).

This broad spectre implies a new inter-disciplinary research field which we would like to call "copy and paste studies" throughout all social systems. Cases of intellectual property theft in all its varieties regarding their technical substrate, their content, and the level of plagiarism should be focussed:

Technical substrate:

- print-to-print-plagiarism
- print-to-online-plagiarism
- online-to-print-plagiarism
- online-to-online-plagiarism [see also Weber 2007a, 47]

Plagiarised content:

- texts
- structures, concepts, and ideas
- images, illustrations, and diagrams
- drawings
- music
- videos
- computer programmes, programme codes etc.

Level of plagiarism:

- 1 to 1 plagiarism, total copy
- Partial plagiarism
- Paraphrasing/kinds of mutating the original
- Structure plagiarism
- Idea plagiarism
- Conceptual plagiarism

This broad focus means that all conceivable kinds of plagiarism should be of interest for ECPIRD.

In this new research field to be founded at least

- informatics
- science of semantic technologies

- computer linguistics
- information retrieval science
- knowledge management science
- copyright law, and
- empirical media research

must work together trans-disciplinary in two main fields:

- the elaboration of sensitisation strategies for plagiarism prevention, starting with families and schools
- the development of viable new tools for plagiarism detection.

The centre should also be engaged in doing empirical research to allow insights in how big the problem for example in the academia or in the arts world today in fact is. For example a pan-European students and staff questionnaire should deal with topics like

- betrayal at tests and exams – also with the help of modern technologies (PDAs, cellular phones)
- betrayal in written assignments
- actual cases or willingness for translation plagiarism
- actual cases or willingness to fool anti-plagiarism systems, e. g. by plagiarism from non-digitised sources or by using advanced technical anti-anti-plagiarism methods

Similar research has been done by the CAI in the US [see http://www.academicintegrity.org and http://www.ojs.unisa.edu.au/journals/index.php/IJEI/article/ViewFile/14/9, both visited 22/5/07]. In Europe plagiarism and scientific misconduct research is not on top of the agenda until now although these phenomena are increasingly seen as big problems in nearly all social systems (as shown in this report).

The centre should not only do empirical research as outlined above, but also continuously test existing products on the market which promise plagiarism detection including a periodical analysis of their shortcomings. – A further focus should be a research field which Google itself recently described with "*Searchology*" (in the context of a presentation of their attempts towards "Universal Search", which means the integration of web, images, groups, and news search on one single ranking list). We are convinced that we shouldn't leave over the search engine research field to Google alone. Possible areas of research would be:

- experimental investigation of the linkage between "googlisation of reality" and disposition for doing copy & paste [as described in Weber 2007a]
- further investigation in specific sites privileged by large search engines and what this fact means for net users (for example the Google-Wikipedia connection as empirically indicated in Section 1)
- *eyetracking* of users of search engines: Poynter institute carried out some studies in online news reading behaviour a few years ago, this should be extended to online search engine use and information seek behaviour, for example with devices like the *EyeLink II* eyetracker working with the bright-pupil method and allowing to analyse fixations and saccades, for example on the list of the first top ten results of search engines.

In this context we propose to do much more research. We also recommend establishing a professorship on "*searchology & wikipedistics*". For example in German-speaking media science there is not very much research coverage of these latest developments in the moment. The socio-cultural reflection and empirical investigation of the current revolutions in information search and processing must better start today than tomorrow.

The main focus of the centre should be the concentration of forces to develop an *integral anti-plagiarism package*. In the following such an ideal tool for plagiarism detection is shortly described. We propose a *pilot project* to develop a prototype of such a machine which should consist of an effective combination of known and so far still experimental measures:

- The package should allow an *online reference text check* – if possible in cooperation with Google Web search and Google Book Search. We are fully aware of the fact that a cooperation with Google would mean "sleeping with the devil". But at least we should try to ask. If Google is unwilling to cooperate, the centre should search for alternative partners or build up specialised search engines for the social systems and subsystems mentioned above of its own. In this context we should also search for strategies to index the deep web and paid content sites as far as it makes sense.
- The package should try to solve the problem of handling *special characters* in specific languages, at least in the main languages spoken in Europe.
- The package should also try to solve the problem of *translation plagiarism* by featuring the option to convert each text into basic English form.
- The package then should allow statistical features for comparison like
  - averaged word lengths in defined text chunks (including standard deviation)
  - averaged number of words per sentence (including standard deviation)
  - nouns to verbs relation
  - and others
- The package should also allow *synonym checks* with advanced techniques on a semantic level by the usage of synonym classes.

The development of such an experimental integral anti-plagiarism package should go hand-in-hand with the establishment of topic-centred specialised search engines [as described in Maurer 2007a]. Indeed such an anti-plagiarism package and several specialised search engines could be two features of one single application (comparable to the interface of Docol©c).

An advanced integrative anti-plagiarism tool could function as follows (logical steps of an analysis of a submitted text-only file, also see the description in Maurer 2007a):

- Removal of fill words (like "a", "the", "at", "from", "of", "by"...)
- Replacing words by word stems with the help of specially developed or adapted *stemming algorithms* ("replacing" –> "replace", "took" –> "take", "machines" –> "machine"...)
- Replacing every word stem by its synonym class/its larger word class, therefore complex *synonym dictionaries* are necessary ("truck" –> "car", "azure" –> "blue"...)
- If the original text submitted is written in a foreign language, then automatic translation into English ("standard form")
- Comparing this text chain to a reference collection of texts (in cooperation with Google, Google Books, and/or other search engines, and/or other databases, including especially scientific ones)
- Detection of an overall quote of similarity

We agree with German plagiarism expert Debora Weber-Wulff that one can't solve social problems with software. But a new and revolutionary software could be a useful addition to your brain and Google. This is what ECPIRD should work on. We propose to develop a prototype of that software in a research consortium with at least the following institutions [also see Maurer 2007a and 2007c]:

- FAST, Oslo and Munich
- L3S, Hannover
- Hyperwave, Graz
- IICM and KNOW Center, Graz

- Bauhaus University of Weimar, Faculty of Media, Stylometry Research Group
- Fraunhofer Institute for Computer Graphics IGD, Darmstadt
- Middlesex University, Compton
- and other institutions to be invited

The fight against plagiarism is not about a digital arms race. It's about guaranteeing the honesty and reliability of information and knowledge production as well as mediation in society and preventing a future in which a culture of mediocrity and hypocrisy dominates over those who want to give their best with honest means.

However, as serious the situation concerning plagiarism is, and has been described to be, the real danger of Google and other data-mining operations goes far beyond the problem with plagiarism, as has already been pointed out in Section 2: Executive Summary. We return to this issue in other Sections, in particular in Section 7 and in Section 12.

For references see Section 16.

Section 6: Different kinds of plagiarism: How to avoid them, rather to detect them

In this study we have mainly described plagiarism in the context of textual material. It must be said at least once quite clearly that different areas do have different rules and desires. If an organisation is interested in a study on some topic, it usually does not care whether the study is plagiarized or not, as long as it will not face problems because of IPR violations. A Japanese colleague has surprised us with the statement that "something is not good, if is not plagiarized". And although he gave as example a material one: "that a Rolex has such a high reputation is only due to the fact that it has been copied (=plagiarized) many times" this may indeed also apply to intellectual artefacts. Architects have said more than once if they build a landmark and its special design is copied, this is nothing to be worried about, but is indeed a big complement. When I pointed out to an electronic engineer that the introduction and outlook section of a thesis of some of the students were completely plagiarized he shrugged his shoulders and could not care less: "My students get their grade on what they develop and on the documentation thereof, if some of the brimborium around this is copied I really don't care." We could continue listing many other cases from other disciplines, but will return now to the problem of textual plagiarism that has been and is the main focus of those parts of the study that have to do with plagiarism.

In particular we want to come to some conclusions based on the facts mentioned earlier that no tools (not even the distributed tools to be described in Section 14), are powerful enough to prevent plagiarism a posteriori. I.e. it does not make sense to look at documents, thesis, whatever,… after they have been written and to try to detect plagiarism cases: yes, some will be detected, but many will pass undetected, simply due to the problems that many documents are not available for plagiarism detection tools since they are hidden in the "deep web" (i.e. in protected databases), or because they do not exist in digitized form, or they have been obtained by translation from an obscure language or by systematically using synonyms, etc.

Thus, it makes much more sense to provide environments that make plagiarism difficult or impossible. We are proposing such "plagiarism preventing" environment ICARE as "E-Learning Ecosystem" where plagiarism is actually helpful during the creative phases, but does not result in actual plagiarism. We present the following material in the form of a finished paper on "Learning Ecosystems for dealing with the Copy-Paste Syndrome."

(Note: The rest of this Section is joint work between N. Kulathuramaiyer and H. Maurer)

Abstract : The fact that people of all walks of life are becoming more and more reliant on a wide-range of easily-available digital content is often called the Copy-Paste Syndrome. It implies the indiscriminate usage of material i.e. without checking for reliability or a concern for violations of intellectual property rights or plagiarism, a kind of cheating that has become uncomfortably widespread. A holistic approach is required to address this universal problem combining an institutional approach together with the application of viable technologies, rather than a-posteriory checks with software of doubtful reliability. This paper describes a learning ecosystem, ICARE, that addresses the Copy-Paste Syndrome by minimizing the possibility for unwanted copy-and-paste situations.

1. Introduction

The Web is experiencing a phenomenal growth, with the explosion of user-generated content. As tools get easier to use, and access become more widespread, it also becomes easier for networked learners to misuse the possibilities for plagiarism and IPR violation [Pannepacker, 2007]. It will also continue to become much simpler to acquire information from the web community as opposed to meeting up with co-learners and experts in the real world [Alexander, 2006]. The openness of the Web environment thus poses a number of challenges in monitoring and keeping track of the explorative expressions of learners.

The term copy-paste is used in this paper to refer to an emerging practice of fast and easy publication by millions of people. The 'Google Copy-Paste Syndrome' (GCPS), [Weber, 2006] describes a common activity of performing a fast, easy and usually "not diligently researched" copying of passages of text by people of all walks of life including scientists, journalists, academics and students. The GCPS has resulted in a proliferation of infringements such as plagiarism and IPR violations. Acquiring insights is performed by 'conveniently searching' the Web as opposed to a rigorous process of learning through scientific discovery. Information from Web sources such as Google and Wikipedia are often used without even considering the validity of the source. According to Weber, GCPS and Web mining can actually impede the inquiry-driven scientific process, as answers conveniently pop up, with minimal effort. This syndrome thus endangers original writing and thinking by de-emphasizing the need for deliberate and insightful reasoning. [Weber, 2006] This emerging phenomenon in turn encourages mediocrity in published works due to the lack of careful thought and understanding.

Due to this eminent danger in store, it is of utmost importance for us to explore innovative means of addressing these issues.

We will concentrate our attention on the important phenomenon of plagiarism and the Copy-Paste Syndrome (CPS). Current learning environments are often more concerned about identifying problem situations after they actually happen. We believe that the detection of plagiarism or copy-paste activities after some work is finished is neither reliable nor a good approach. Rather, there is a need to explore preventive ways of making sure that unwanted versions of copy-and-paste just cannot happen. A learning ecosystem coupled strongly with pedagogical aspects and techniques that control copy-and-paste situations throughout will best serve the emerging needs of educational institutions.

2. Dealing with plagiarism (and the Copy-Paste Syndrome)

Students are often expected to read the policy in the handbook and thereafter comply with a non-plagiarizing attitude. This approach is likely to be unsuccessful as the core problem lies in the students lack of understanding of the concept of plagiarism and, most of all, their inability to deal with it [Kennedy, 2004]. Students are also generally not aware of the full implication of the acts of copy-paste. They also do not value the importance of intellectual property or take pride in their ability to produce creative works [Kennedy, 2004]. As pointed out by [Duff et. al., 2006] there is the lack of appreciation of the Western system of scholarship, especially among new students and foreign students. There is thus a need to teach the skills required for paraphrasing, summarizing and referencing accurately [Kennedy, 2004].

There is a need to instill moral and ethical values in students regarding their education. Students will begin to understand the need to respect other people's copyright when they themselves are actively engaged in creating their own intellectual property [Midolo, Scott, 2003]. Best practices in teaching and learning and academic integrity can be further achieved if students are aware that their inputs have been valuable and considered carefully by instructors [Kennedy, 2004].

Another proposed approach to address Copy-Paste Syndrome is through the employment of well-structured and clearly articulated assessment tasks. Course designers will have to carefully design courses and course content to ensure that they do not indirectly encourage plagiarism. Factors that encourage plagiarism include the same questions being set repeatedly year to year, questions which cannot be understood clearly or when clear criteria are not specified [Kennedy, 2004].

There are a number of approaches that can be employed to reduce plagiarism as suggested by works in [Harris, 2004]. Instructors are encouraged to enforce the use of one or more sources not written within the past year. This approach effectively invalidates results of paper mills [Harris, 2004]. By enforcing the use of one or more specific articles or specific information, students can be encouraged to formulate their own thoughts. Another effective technique describe by Harris, is to enforce the production of assignments as a series of process steps as they lead to the final completion of projects. Student learning can then be continuously checked and assessed at each stage. The administration of personalized student tracking and assessment tends to overwhelm instructors. A careful selection of viable technologies is required to minimize the effort required. A technological platform can also be applied to guide students in using material from various sources in a constructive way and promote critical thinking.

3.   Typical Approach for Dealing with Plagiarism (and also Copy-Paste)

A typical approach used in dealing with plagiarism in educational institutions is to employ tools for plagiarism detection such as Turnitin or Mydropbox. However, a single tool by itself is not adequate for copy-paste detection. A suite of tools is required to detect plagiarism or copy-paste effectively to establish and substantiate the detection of plagiarism with as much evidence as possible. An overview of a broad range of tools required for fighting Plagiarism and IPR violation is presented in [Maurer et al, 2006]. A layered application of plagiarism detected has been further proposed by [Kulathuramaiyer, Maurer, 2007] to systematically perform elaborate mining by focusing on relevant subsets of documents. Table 1 describes the availability of multiple approaches for detecting the various aspects of plagiarism. Despite the availability of these tools and techniques, their usage has mainly been employed in the detection of plagiarism and Copy-Paste situations. We propose the application of these tools and techniques in preventing the Copy-Paste Syndrome

Table 1: Tools for Plagiarism Detection

Task	**Tool**
Manual Technique	Search Engines [Maurer et al, 2006]
Text-based Document Similarity Detection	Dedicated Software, Search and Web Databases [Maurer, Zaka, 2007]
Writing Style Detection	Stylometry software [Eissen, Stein, 2006]
Document Content Similarity	Semantic Analysis [Dreher, Williams, 2006], [Ong, Kulathuramaiyer, 2006], [Liu et. al., 2006]
Denial of Plagiarism	Cloze Procedure [Standing, Gorassini, 1986]
Content Translation	Normalized Representation [Maurer, Zaka, 2006]
Multi-site Plagiarism	Distributed Plagiarism [Kulathuramaiyer, Maurer, 2007]

4.   Comprehensively Addressing the Copy-Paste Syndrome

4.1 Rationale

In exploring a technological solution to comprehensively address the copy-paste syndrome, the first question clearly is: Will it be ever be possible to comprehensively address the copy-paste syndrome by software to check a paper submitted without any knowledge how the paper was compiled. Our answer is a clear “no”. We have pointed out the existence of paper mills [Paper Mills, 2006] which even prepare papers to order [Kulathuramaiyer, Maurer, 2007]. Documents may also contain large portions that are translations of some material in a not-so-common language making it nearly impossibly to find out if material is plagiarized. Furthermore, there are large collections of materials available in either closed databases or in not digitized form that are not available to any plagiarism checking software. As such a different approach is needed: we believe the key issue is to monitor the work of learners continuously.

We will discuss issues of an E-Learning ecosystem called ICARE. We are trying out a number of components of a learning ecosystem at Graz University of Technology. We refer to the proposed suite of software and content as ICARE, aimed at controlling copy-paste situations.

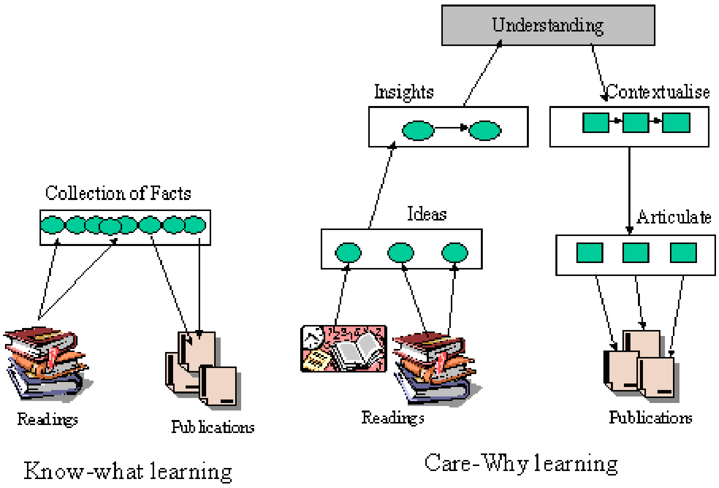
4.2 The main concept of ICARE[2]

ICARE stands for Identify-Correlate-Assimilate-Rationalize-Express. ICARE denotes the five steps involved in the cultivation of academic reading and writing. These steps can be elaborated as:

- Identification: Identify key points (relevant) while reading a text document
- Correlate: Associate reading with concepts in the mind map of a learner
- Assimilate: Associate concepts learnt with prior knowledge of learner
- Rationalize: Formulate ideas based on concepts arising from student learning
- Express: Express idea in learners own words

As opposed to the inadvertent (improper) copy-paste, ICARE enforces care on the part of the students' understanding of concepts, enabling them to apply learnt concepts in the appropriate manner. The proposed approach to copy-paste will thus be seen as focusing on deeper appreciation and understanding ('care-why learning') as opposed to a less-diligent focusing on facts ('know-what learning'). Figure 34ontrasts these two forms of learning. Learning should not be based on a mere a collection of facts, it should rather be viewed as a connection to a learner's thoughts [Sathya Sai Baba, 2001]. Support mechanisms are required to allow students to connect readings to the construction of knowledge. We believe that E-Learning systems should focus more on personal knowledge management activities and in fostering a deeper understanding. This will then effectively reduce the occurrence of improper copy-paste.

Figure 34 Types of Learning Modes



Practicing a constructive form of copy-paste supports a learner's ability of absorbing concepts, and consolidating and assimilating them before expressing ideas with a deeper understanding. The proposed ecosystem guides and allows students to become aware of the correct approach of reading, digesting and applying knowledge. At the same time, the platform fosters creativity in their associational and expressive ability. The proposed ecosystem allows an instructor to view and monitor the learning process of students, in observing and monitoring the rightful practice of 'copy-paste

[2] To be read as 'I Care'

skills'. At the same time creative expressions of students can be pinpointed, highlighted and recorded. We believe that helping and assessing students does need a certain amount of supervision that cannot be put into the undesirable "Big Brother" neighbourhood.

4.3 Towards a Holistic Learning Ecosystem

Although a variety of forms of E-learning have been explored, the predominant form of E-Learning employs an E-book paradigm. For the proposed ecosystem, however multiple learning paradigms need to be incorporated. It will also need to enable pedagogical aspects of learning via technology enhanced knowledge transfer [Helic, 2007].

Current E-learning systems tend to employ a blended learning environment which involves the combination of instructional modalities or the instructional methods via the combination of online or face-to-face instruction. [Graham, 2004]. Options currently available in such learning systems [Graham, 2004] include self-study options such as web-based courseware, simulations, systems and books together with live teaching options such as web-casting, live video, conference calls, and instructor-led training. Each of these are often treated as standalone training objects delivered either via face-to-face (F2F) or computer mediated learning (CML) instruction [Valiathan, 2002]. In this case each training object represents a particular modality of learning where CML training objects are seen as alternative learning modes to classroom-based F2F approaches. The main weakness of this approach is that it does not allow composing training objects that contain both aspects of F2F and CML.

The realization of ICARE requires an E-learning ecosystem that mixes F2F and CML within the context of a learning scenario that also minimizes the unwanted use of copy-and-paste by guiding the learner through the process. ICARE enables the complementary use of technology to harness the systematic development of both personal learning and collective intelligence. Table 2 describes the differences between blended learning in traditional E-Learning and the proposed learning ecosystem.

Table 2: Comparing the proposed Learning System Functionalities against a typical E-Learning system

Teaching-Learning Activity	**Typical E-Learning environment**	**Proposed E-Learning Ecosystem**
Announcements (Communicating timely messages to students)	Learning Management System or E-mail	Dynamically Activated from an Event database, RSS feeds
Overview session	Email, E-books	Reading Scenario (E-Room)
Self-paced learning	Web-based tutorial. E-books simulations	Learning Specifications, Project-Rooms, E-Books
Student Question Answering	email , Frequently Asked Questions	Active documents, schedule E-mentoring sessions
Assessment	Simulations, Online test, Submission system	Knowledge maps, Testing scenarios (can be personalized, collaborative, or peer-reviewed); Student Activity Logs and Reports
Collaborative Sessions	Discussion groups, Bulletin Boards, Chat	Brainstorming Scenario, Peer ranking
Feedback	Email	Examination Rooms
Continuous Assessment	Student Records	Student Portfolio, Learning Plans, Performance Monitoring Tool

5. *Realization of the ICARE Ecosystem*

5.1 Overall Design

ICARE will incorporate many of the experimental features in WBT-Master [WBT 2006] coupled with the knowledge management capabilities found in Hyperwave Information Server [Mödritscher et. al., 2005]. It will also be augmented with a copy-paste detection and administration suite together with specifically prepared E-learning modules to address both the issues mentioned earlier and anticipated future learning needs.

We propose additional functionalities to the E-Learning platform, (currently not available in any E-Learning system we know) for providing personalized guidance for administering academic reading and writing. Our previous works in the development of tools for fighting plagiarism and IPR violation has provided insights on the requirements of the proposed ecosystem. [Kulathuramaiyer, Maurer, 2007]

Table 3 summarizes the functions to be incorporated describing the technological requirements for the ICARE ecosystem. The ecosystem employs an effective administration of E-Learning together with powerful tools for guiding and managing student learning and interaction.

The various learning functions together with well-designed assessments are crucial. In the next section components of an experimental system will address these issues. Tracking and analysis will be required to keep track of a variety of student works such as term papers, projects, examinations, etc. Tracking of activities will also be important in providing insights on learning and knowledge creation activities. The copy-paste handling suite of tools and techniques are required to assist and support the learner in the mastery of the rightful copy-paste skills. Specifically developed E-Learning modules enable the learners to master the fundamentals of academic reading and writing and promote an understanding of academic publishing culture.

Table 3: ICARE Ecosystem: Needs versus Required Functionality

Needed Functionality	**Required Functionality**
Effective Administration of Learning	Ability to incorporate pedagogy in a learning environment combined with an ability to Structure Assessment; this includes the ability to discover and visualize student learning (knowledge maps) and integrate this with assessment, The management of capability-driven student learning, ability to manage and guide collaborative group-centered (project) work and Flexible design of assessment tasks to manage learning as a series of steps
Guided Learning Process	Controlled Environment for keeping track of learner activities, and Workflow management and compliance checking
Tracking Learners' Copy-paste activity	Integrated Copy-paste handling capability enabled by a suite of similarity checking software
Appreciation and Mastery of ICARE principles and Process	To incorporate -Learning Modules on: Western Scholarship Academic Reading and Writing Valuing Intellectual Property and Ethics

5.2 Key Components of ICARE Ecosystem

The following features from our past experimental developments in projects such as WBT-Master and Hyperwave will facilitate the realization of the learning ecosystem:

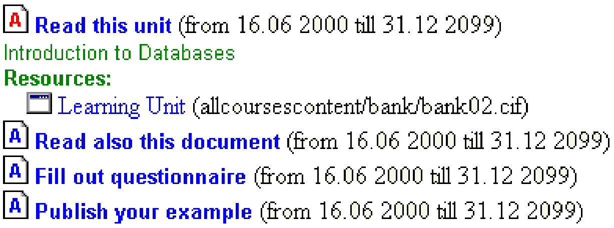
- Ability to define training scenarios as training objects or study rooms: A controlled environment can be established to track both the explorative and collaborative activities of students [Helic et. al., 2004a].
- Pedagogy driven learning: A teaching scenario or environment can be built where a tutor works with a group of learners in both synchronous and asynchronous mode, leading them to achieve a particular learning goal.
- Project-oriented learning: A controlled learning environment can be built to allow a group of learners working together on a project, e.g., a software engineering project [Helic et. al., 2003]
- Adaptive discovery of personalized background knowledge: A reading room paradigm can be created for enabling learners to chart their knowledge discovery process. This can be supported by the automated linking to related contents or background knowledge [Mödritscher et. al., 2005]
- Annotations: Annotations allow the attachment of text segments, system or media objects or an URL to a learning object or material [Korica et. al., 2005]. It is possible to annotate any kind of material such as papers, parts of a digital library, other user contributions, etc.
- Active Documents: The idea of active documents present an efficient way of students learning in a collaborative question-answering environment. Active documents present an innovative mean to demonstrate student learning and at the same time, an effective way for an instructor to direct knowledge discovery [Heinrich, Maurer, 2000].
- Visualisation as knowledge maps: The cluster of a document with documents containing similar concepts or ideas can be visualized via a knowledge map typical of knowledge management systems. A knowledge map with similar articles can be created and visualized [Helic et. al., 2004b]. "Knowledge cards" are used to describe particular concept (i.e. semantic entity). Knowledge cards may be combined into a semantic network. For example, the knowledge card "Student's Discovered concept" may be related as "is a part of" to the knowledge card "Course Domain Ontology".
- Workflow management and compliance checking capabilities: Learning can be visualized as a process flow of learning tasks. Non-compliance can then be automatically flagged by the system.

5.3 Controlled Environment For Pedagogy-driven E-Learning

ICARE provides a controlled environment in which the instructor is able to track the usage of reference materials by students. Such a controlled environment makes it much easier to curtail unethical practices and also promotes constructivist learning among students. Furthermore, user tracking and user activity-logging facilities can also be used to enforce learners to read certain parts of a document before being allowed to annotate an article or ask questions about some part of it [Helic et. al. 2004a].

An environment that closely monitors students' knowledge construction and collaborative activities can help the instructor to assess and guide students' ability to publish effectively. Process level support can be achieved via the workflow management and compliance checking capabilities of systems. The system can be trained to recognize non-conforming patterns to be able to flag instructors. Discovered patterns regarding a student's learning can then be captured and stored within a learner's profile. Knowledge profiling is supported in the acquisition, structuring, and reuse of extracted expert knowledge. By maintaining individual learner profiles, personalized learning can be supported. Personalized learning units then can be designed for each student as shown below:

Figure 35: Personalised Learning units

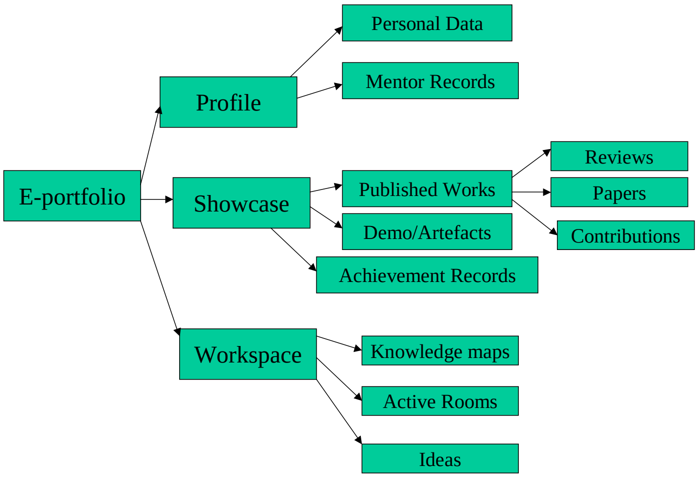


Read this unit (from 16.06 2000 till 31.12 2099)
Introduction to Databases
Resources:
Learning Unit (allcoursescontent/bank/bank02.cif)
Read also this document (from 16.06 2000 till 31.12 2099)
Fill out questionnaire (from 16.06 2000 till 31.12 2099)
Publish your example (from 16.06 2000 till 31.12 2099)

Interactive collaborative scenarios [Helic, 2007] are employed to administer and respond directly to individual student learning activities. For example, active documents can then be employed to keep track of learner interactions and learning support within the context where learning occurs.

An explicit and implicit profiling of students has to be applied to keep track of the learning process of students. E-Portfolios [Alexander, 2006] enable the recording of student participation and contribution to support the profiling of students. E-Portfolios are important in allowing students to start valuing their own contributions and also other student contributions. An example of a student portfolio structure is shown in Figure 36 As shown here, the ecosystem provides a workspace for students to continuously expand their knowledge base while taking responsibility for their own learning. Records of student achievement then immediately become available to mentors and instructors for personalized evaluation and guidance.

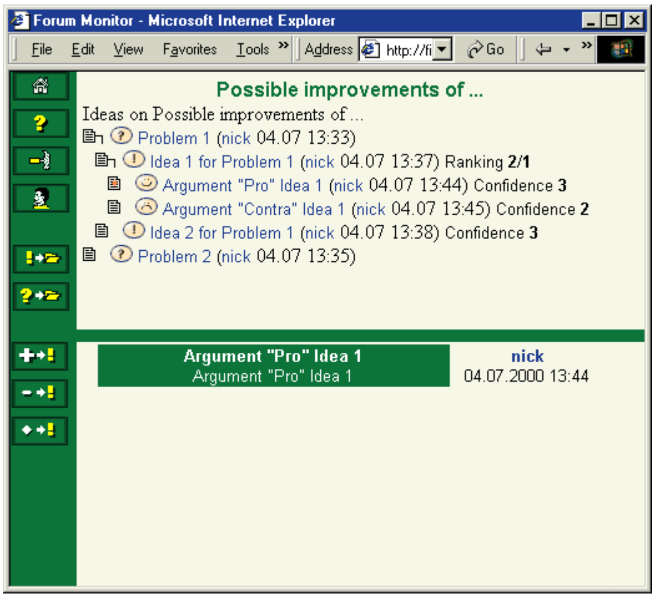
Figure 36: Student E-Portfolio



Incentive schemes can be tied to E-portfolios in order to acknowledge and highlight student achievement. Recommendation systems proposed to make explicit the valuations associated with each student’s contribution. Recommendation systems play an important role in the development of rational impartial judgment among students. A combination of human and automated ranking of important topics, ideas, suggestions and contributions can further be applied personalized interaction among students with similar background and interests.

A number of tools are available for creating an environment for students to collaborate among themselves and with their instructors and mentors. These include Peer-Evaluation support, Collaborative concept mapping, brainstorming and discussion forums. Brainstorming also incorporates mechanisms for specifying ranks and incorporating personal evaluation. (See Figure 37Annotations are again a key feature to represent and organize collective student learning. Annotations have also been proposed to represent links to Knowledge Cards to reflect the knowledge construction process of students.

Figure 37: Brainstorming support (extracted from WBT-Master Manual)



Possible improvements of ...

Ideas on Possible improvements of ...

- Problem 1 (nick 04.07 13:33)
  - Idea 1 for Problem 1 (nick 04.07 13:37) Ranking **2/1**
    - Argument "Pro" Idea 1 (nick 04.07 13:44) Confidence **3**
    - Argument "Contra" Idea 1 (nick 04.07 13:45) Confidence **2**
  - Idea 2 for Problem 1 (nick 04.07 13:38) Confidence **3**
- Problem 2 (nick 04.07 13:35)

Argument "Pro" Idea 1
Argument "Pro" Idea 1

nick
04.07.2000 13:44

Integrated visual tools will be applied in the management and display of information in illustrating student learning and mastery of concepts. They allow the instructor to impart particular skills, to refine processes used for a specific task, or to organize information into a structured form. They can also be used by students to express their understanding of concepts. Knowledge visualization tools will also be applied as a form of assessment of students' incremental knowledge gain over a period of time. Learners also need to be supported by means of personalized knowledge retrieval facilities. Such a tool will be effective in identifying potential infringements by students and can be used to aid students in the mastery of useful skills. The visualization capability for concept maps further allows the incremental visualization of concepts formulated by students.

Knowledge-cards (K-Cards) enable the specification of concepts of is-a and instance-of links for ICARE knowledge maps. The semantic relationships built upon K-Cards essentially define a semantic graph. The knowledge card mechanism is also used to automatically link to peer-learners and resource persons in collaborative mode. Two types of K-Card usage have been defined: personal knowledge card attached to each learner, and context based knowledge cards attached to assignments or scenarios. The use of K-Cards supports the creation of points of interests by students. A knowledge card can also be linked to other related readings that students may associate (if required). These K-cards will then allow students to link to a concept map, which will demonstrate the students' understanding process.

5.4 Incorporating the Ability to handle Copy-Paste into ICARE

ICARE benefits from the administration of academic reading procedures which can be integrated directly into the ICARE ecosystem. By enabling a business process model view of E-Learning, [Helic et. al., 2003] the learning process can be supported at each step.

E-learning modules on effective copy-paste would then be embedded to educate students on the rightful procedure of academic publishing (reading and writing). Apart from employing a plagiarism or copy-paste detection suite for summative assessment of a breach of conduct, we propose the formative application of such tools for self-plagiarism checking and in cultivating constructive 'copy-paste skills'. For example, existing document similarity detection (as used in plagiarism detection tools) can be applied in conjunction with a learning scenario paradigm for facilitating students to master academic publishing. By consolidating the results from similarity search engines on local databases as well as the Internet, a plagiarism detection tool can be applied to assist students to teach them how and when to cite another publication.

5.4.1 Copy Paste Handling Software Suite

The Copy-Paste Handling Software Suite incorporates self-plagiarism checking tools and techniques to help students in mastering copy-paste. Both simple and advanced forms of copy-paste checking are supported. We propose the use of the plagiarism detection tools and techniques to achieve this task (see table 1). This suite will be applied in two modes; closed world and open world modes. This will allow the operation of the copy-paste handling in both a supervised mode (assisted by an instructor) and an unsupervised mode (self learning).

In the closed world mode a student uses the copy-paste wizard as guide for the academic reading and writing process. This wizard is described in the next section. Here the text that students select for copy-paste will be used as a fingerprint and applied as query string to search the whole published text of the student for a weak or blatant copy-paste case. The similarity checking engine identifies the degree of similarity in determining the extent of paraphrasing (or the lack of it). The system is also able to check for compliance or negligence citation. A string similarity checking mechanism is applied for this purpose. In the case of identifying an improper copy-paste, the system presents its findings as an advice to students. The changes made by students are noted by the system and can be used in a mentoring session.

In the open world mode, students are not guided or restricted in terms of usage of specified references. Similarity detection is then applied to a larger collection of documents where it checks the web for all possible improper copy-paste actions performed by the students. Student's past years papers are also checked for similar text strings to determine improper copy-paste and lack of citation. The system produces statistical information for the instructor to assess the mastery level of students.

A number of learning scenarios can be built by a selective application of one or more copy-paste handling tools. As described here, these scenarios could either be applied in a supervised manner assisted by an instructor or a mentor or the unsupervised manner with system inputs.

During the mentoring process a manual selection approach for plagiarism detection may be employed checking with one or more search engines. This process can provide the system a set of constrained documents to be used for similarity checking. Specific tools to approve or disprove suspected plagiarism such as Cloze may also be applied when a dispute arises. A Cloze procedure [Maurer, Zaka, 2006] has been used to judge the originality of authorship of published works. As part of the copy-paste detection, alternative techniques such as stylometry can be applied to discover similar (or dramatically changing) stylistic patterns such as syntactic forms usage, text structure of published works and the usage of key terms to indicate that some copying may have taken place.

5.4.2 Copy-Paste (Academic Reading and Writing) wizard

This wizard has been proposed to enable learners to acquire the skills of academic reading and writing in a controlled environment. The wizard can be used by learners to perform the following:

- Highlight key points and annotate selected phrases, using the annotation feature of WBT-Master. A highlighting mechanism is supported to allow learners to highlight key points.
- Create a knowledge-card for the point discovered, label it and link it to known concepts (or form a new concept)

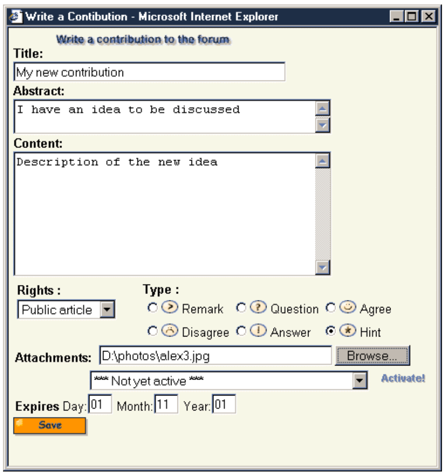
- Review the internal concept map and assimilate new ideas found in reading. This may range from concept links to concept map restructuring. This stage involves substantiating the body of knowledge cards with links and metadata
- Formulate an idea and add information to knowledge cards
- Express an idea and present it as a descriptive text

Annotations will be employed in linking original document to relevant information sources to perform the above steps. This enables the tracing of students' activities to check on process-flow of academic writing. Separate readings can be assigned to each student to track individual student activities and also to avoid plagiarism. At the same time a single document may also be used for an entire class or a smaller group of students. In this way a comparative analysis of students' learning can be visualized and studied. These documents can then be constructed as active documents that allows collaborative learning to take place, built upon students' comprehension and ability. As with our previous experiments on active documents, we know that when 500-1000 users have viewed a particular document, all possible questions that need experts become answered. [Dreher, Maurer, 2000]

The technological support to prevent blatant copying by students is realized by imposing the use of annotations (through specially designed interface templates) which overcomes the need to duplicate content. Figure 38 illustrates the interface that allows students to express ideas, opinions, contribution to collaborative sessions, ask questions, etc. Additionally, the copy-paste interface further displays the highlighted text, representing key points with a ranking of importance, paraphrased text, comments, etc. Students' published works will then be stored as annotations to the original text and visualized separately by the instructor for evaluation.

The use of annotations can be explored as a means of training students' use of the correct form of citations and referencing. By using annotations, a much simpler similarity checking system would suffice to overcome plagiarism to a large extent in ICARE. Annotations and its sophisticated communicational and collaborative features play an important role in the realization of a culture of Web-based reading and writing.

Figure 38: Interface for learners to annotate documents



Write a Contibution - Microsoft Internet Explorer

Write a contribution to the forum

Title: My new contribution

Abstract: I have an idea to be discussed

Content: Description of the new idea

Rights : Public article

Type : ○ Remark ○ Question ○ Agree ○ Disagree ○ Answer ● Hint

Attachments: D:\photos\alex3.jpg

*** Not yet active ***

Expires Day: 01 Month: 11 Year: 01

5.5 Design of Assessment

ICARE also includes mechanisms for the careful design and execution of assessments. The pedagogy driven learning together with the ability to define learning scenarios and rooms allow for highly personalized assessment design and curriculum development.

Beyond the features of the ICARE system as described, the ability to operate in the following modes is instrumental:

- Guided mode: interactive session (system auto-suggestions) with closed systems monitoring
- Self-learning mode: minimal non-interactive feedback, closed world systems monitoring but with feedback provided only on student request
- Diagnostic mode (formative): closed world systems monitoring but with no feedback, results are archived for self-driven assessment
- Evaluative mode (summative): open world mode, with text analysis performed (copy-paste analysis) and used as support for self-paced assessment
- Mentor-assisted mode: similar to diagnostic mode but with feedback sent to a mentor, who responds to students
- Peer-learning mode: open world learning mode, with the system tracking learner participation and contributions

These modes of operation can be realized as scenarios (training objects) in WBT-Master. This system also allows assessments to be broken up into smaller parts as a means of supporting continuous assessment, and in the monitoring of student learning process.

As an example of the application of ICARE in a classroom, we propose the following illustration:

1. Students in a class are first asked to collaboratively construct a collective concept map for a domain of study.
2. Individual students are then required to construct a personalized concept map representing their personal learning space
3. Subsequently, students are assigned selected reading material. An online copy of the reading material is placed in the reading room of each student (or a group of students)
4. Students are then required to identify key points by using the wizard in closed monitoring mode with all activities tracked by system. The highlighted text segments by students can be used to reflect their understanding. Both individual student learning and group learning can be highlighted.
5. The highlighted texts are then visualized for the instructor as annotations attached to the selected document. Statistical information is used to demonstrate student learning e.g. common mistakes made by student, misunderstanding of text, etc.
6. Instructors’ comments can either be placed in personal spaces of students or public spaces for the whole class
7. Students are then requested to paraphrase the texts selected in guided mode
8. A visualization of all student inputs is then made available for the instructor. Additional statistical information is presented to support student evaluation. Non-compliance in student learning workflows is visualized
9. The next step involves a peer-learning mode, where student are requested to discuss the points selected by their peers in the brainstorming room. All points being discussed are referenced and the system links them together for visualization. The instructor or facilitator then provides interactive feedback in the brainstorming room
10. Students are then required to update their personal concept maps, with the knowledge gained in 9.
11. Statistics of popular concepts in knowledge-map, popularly selected key points, list of questions posed during brainstorming or during any other phase in the exercise are all presented to the classroom.
12. As the final task, students are asked to collaboratively construct a single concept map while continuing with discussions in the brainstorming rooms. All concepts in the knowledge map are uniquely identifiable as they are implemented using knowledge cards. Thus, students are able to discuss the addition of particular concepts or places for links and types of links as well.

The above hypothetical assessment has been defined to illustrate the various functions for the explorative employment in a classroom. A typical classroom usage may only require a subset of the tasks listed. This clearly highlights the power and potential of the proposed ecosystem, to serve as basis for the design of future E-Learning systems.

6. Conclusion

We have adopted the stand that copy-paste need not be entirely considered a wrong-doing. Students would then need to be educated and guided on the constructive use of copy-paste skills as a learning mechanism. We have presented an academic ecosystem with technological support to comprehensively address the copy-paste-syndrome.

We proposed the use of an advanced E-Learning system, together will carefully planned student assessments and the close monitoring of student learning to address the problem. Plagiarism and copy-paste syndrome avoidance mechanisms and procedures are integrated into the ecosystem and applied throughout the program of study. E-learning modules together with a suite of copy-paste handling tools enable the formative development of ‘effective copy-paste skills’. A complete suite of copy-paste detection and avoidance tools will need to be established in all educational institutions.

By effectively addressing the copy-paste-syndrome many of the social problems that we are likely to face (arising from the degradation scientific quality and even possibly leading to quality of life) in future can be averted. Without the full institutional backing and commitment of academics however, a culture that withstands and compensates the prevalent copy-paste culture cannot be achieved.

References

[Alexander, 2006] Alexander, B., (2006) Web 2.0: A New Wave of Innovation for Teaching and Learning?, EDUCAUSE Review, Vol. 41, No. 2, 2006 pp. 32–44

[Bersin, 2004] Bersin, J., (2004) The Blended Learning Handbook: Best Practices, Proven Methodologies, and Lessons Learned (excerpt), Pfeiffer Wiley, San Francisco, USA http://media.wiley.com/product_data/excerpt/67/07879729/0787972967.pdf

[Downes, 2006] Downes, S., (2006) E-learning 2.0, ACM eLearn Magazine, 2006, http://www.elearnmag.org/subpage.cfm?section=articles&article=29-1

[Dreher, et. al., 1994] Dreher, H. V., Dreher, L. H., McKaw, K., (1994) The Active Writing Project - small movements in the real world, Proceedings of Asia Pacific Information Technology in Training and Education, Brisbane, Australia

[Dreher, Maurer, 2000] Dreher, H., and Maurer, H., (2000) Active Documents: Concept, Implementation and Applications, Journal of Universal Computer Science, Vol. 6, No. 12, 2000, pp. 1197-1202

[Dreher, et. al., 2004] Dreher, H., Krottmaier, H., Maurer, H., (2004) What we Expect from Digital Libraries JUCS, Journal of Universal Computer Science Vol. 10, No. 9, 2004, pp. 1110-1122,

[Dreher, Williams, 2006] Dreher, H., Williams, R., (2006) Assisted Query Formulation Using Normalised Word Vector and Dynamic Ontological Filtering, Proceedings of 7th International Conference of Flexible Query Answering Systems, (FQAS 2006), pp. 282 –294, Milan, Italy

[Duff, et. al., 2006] Duff, A. H., Rogers, D. P., Harris, M. B., (2006) International Engineering Students - Avoiding Plagiarism through Understanding the Western Academic Context of Scholarship, European Journal of Engineering Education, Vol. 31, No. 6, 2006, pp. 673

[Eissen, Stein, 2006] Eissen, S., Stein, B., (2006) Intrinsic Plagiarism Detection, Proceedings of the 28th European Conference on Information Retrieval, Lecture Notes in Computer Science, Vol. 39, No. 36, 2006, pp. 565-569 Springer Publishing Company

[Graham, 2004] Graham, C. R., (2004) Blended Learning Systems: Definition, Current Trends, and Future Directions, in Bonk, C. J., Graham, C. R., (Eds.) Handbook of Blended Learning: Global Perspectives, Local Designs, Pfeiffer Wiley, San Francisco, USA

[Harris, 2004] Harris, R., (2004) Anti-Plagiarism Strategies for Research Papers, Virtual Salt, November 17, 2004 http://www.virtualsalt.com/antiplag.htm

[Helic, 2007] Helic, D., (2007) Formal Representation of E-Learning Scenarios: A Methodology to Reconfigure e-Learning Systems, Journal of Universal Computer Science Vol. 13, No. 4, 2007, pp. 504-530

[Helic et. al., 2003] Helic, D., Krottmaier, H., Maurer, H., Scerbakov, N., (2003) Implementing Project-Based Learning in WBT Systems, Proceedings of E-Learn 2003, pp 2189-2196, AACE, Charlottesville, USA

[Helic et. al., 2004a] Helic, D., Maurer, H., Scerbakov, N., (2004a) Discussion Forums as Learning Resources in Web-Based Education, Advanced Technology for Learning, Vol. 1, No. 1, 2004, pp. 8-15

[Helic et. al., 2004b] Helic, D., Maurer, H., Scerbakov, N., (2004b) Knowledge Transfer Processes in a Modern WBT System, Journal of Network and Computer Applications, Vol. 27, No. 3, 2004, pp. 163-190

[Kennedy, 2004] Kennedy, I., (2004) An Assessment Strategy to Help Forestall Plagiarism Problems, Studies in Learning, Evaluation, Innovation and Development, Vol. 1, No. 2, 2004, pp. 1–8 http://sleid.cqu.edu.au/viewissue.php?id=5#Refereed_Articles

[Korica et. al., 2005] Korica, P., Maurer, H., Scerbakov, N., (2005) Extending Annotations to Make them Truly Valuable, Proceedings of E-Learn 2005, pp. 2149-2154, AACE, Vancouver, Canada

Krottmaier, H., Helic, D., (2002) More than Passive Reading: Interactive Features in Digital Libraries, Proceedings of E-Learn 2002, pp 1734-1737, AACE, Charlottesville, USA

[Kulathuramaiyer, Maurer, 2007] Kulathuramaiyer, N., Maurer, H., (2007) Why is Fighting Plagiarism and IPR Violation Suddenly of Paramount Importance? Proceedings of International Conference on Knowledge Management, in Stary, C., Baranchini, F., Hawamdeh, S., (Eds.) Knowledge Management: Innovation, Technology and Cultures, pp. 363-372, World Scientific

[Liu et. al., 2006] Liu, C., Chen, C., Han, J., and Yu, P. S., (2006) GPLAG: Detection of Software Plagiarism by Program Dependence Graph Analysis, Proceedings of 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 872-881, Philadelphia, USA http://www.ews.uiuc.edu/~chaoliu/papers/kdd06liu.pdf

[Mödritscher et. al., 2005] Mödritscher, F., García-Barrios, V. M., Maurer, H., (2005) The Use of a Dynamic Background Library within the Scope of adaptive e-Learning, Proceedings of eLearn 2005, AACE, Vancouver, Canada

[Maurer et al, 2006] Maurer, H., Kappe, F., Zaka, B., (2006) Plagiarism- a Survey. Journal of Universal Computer Science, Vol. 12, No. 8, 2006, pp. 1050-1084.

[Maurer, Zaka, 2007] Maurer, H., Zaka, B., (2007) Plagiarism- A Problem and How to Fight It, Proceedings of Ed-Media 2007, pp. 4451-4458, AACE, Vancouver, Canada, http://www.iicm.tugraz.at/iicm_papers/plagiarism_ED-MEDIA.doc

[Midolo, Scott, 2003] Midolo, J., Scott, S., (2003) Teach Them to Copy and Paste: Approaching Plagiarism in the Digital Age, Resourcing the Curriculum, 14 August 2003 http://www.det.wa.edu.au/education/cmis/eval/curriculum/copyright/islandjourneys/documents/paper.pdf

[Ong, Kulathuramaiyer, 2006] Ong, S. C., Kulathuramaiyer, N., Yeo, A. W., (2006) Automatic Discovery of Concepts from Text, Proceedings of the IEEE/ACM/WIC Conference on Web Intelligence 2006 pp.1046-1049

[Pannepacker, 2007] Pannepacker, S., (2007) Is Plagiarism on the Rise? Why?, http://muweb.millersville.edu/~jccomp/acadintegrity/plagnotfinalnewsp.html,

[Paper Mills, 2006] Paper Mills, (2006) Cheating 101: Internet Paper Mills, Kimbel Library: Presentations, Coastal Carolina University, http://www.coastal.edu/library/presentations/mills2.html

[Plagiarism Statistics, 2007] Plagiarism Statistics, (2007) http://www.plagiarism.org/facts.html

[Sathya Sai Baba, 2001] Sathya Sai Baba, (2001) Proceedings of First Conference of Sri Sathya Sai Schools, Prasanthi Nilayam, India, http://www.srisathyasai.org.za/programs/educare.asp

[Standing, Gorassini, 1986] Standing, L., Gorassini, D., (1986) An Evaluation of the Cloze Procedure as a Test for Plagiarism Teaching of Psychology, Vol. 13, No. 3, 1986, pp 130-132

[Valiathan, 2002] Valiathan, P., (2002) Blended Learning Models, http://www.learningcircuits.org/2002/aug2002/valiathan.html

[WBT, 2006] WBT Master White Paper (2006), http://www.coronet.iicm.edu

[Weber, 2006] Weber, S., Das Google-Copy-Paste-Syndrom, (2006) Wie Netzplagiate Ausbildung und Wissen gefährden, Heise, Hannover

Section 7: Dangers posed by Google, other search engines and developments of Web 2.0

This chapter is contributed by Hermann Maurer.

It is now some 18 months ago that we found out that Google proved to be unwilling to support high power plagiarism detection services. The information we collected since then is indeed quite scary.

It has become apparent that Google has amassed power in an unprecedented way that is endangering our society. We will first summarize the main points here that are amply supported by Appendices 4 and 5 written by members of the authoring team. We also recommend some measures to fight against the big dangers ahead of us. However, if none of the persons and institutions receiving this report is going to act also, I am afraid little will happen, with potentially very dire consequences.

Here is the summary, first:

Google as search engine is dominating (Convincing evidence on this is easily available and presented in Section 1). That on its own is dangerous, but could possibly be accepted as "there is no real way out", although this is not true, either. (We would rather see a number of big search engines run by some official non-profit organisations than a single one run by a private, profit driven company.) However, in conjunction with the fact that Google is operating many other services, and probably silently cooperating with still further players, this is unacceptable.

The reasons are basically:

- Google is massively invading privacy. It knows more than any other organisation about people, companies and organisations than any institution in history before, and is not restricted by national data protection laws.

- Thus, Google has turned into the largest and most powerful detective agency the world has ever known. I do not contend that Google has started to use this potential, but as commercial company it is FORCED to use this potential in the future, if it promises big revenue. If government x or company y is requesting support from Google for information on whatever for a large sum, Google will have to comply or else is violating its responsibilities towards its stockholders.

- Google is influencing economy by the way advertisements and documents are ranked right now in a way that has become unacceptable to the extent that the European Commission has even started to look at anti-trust measures with the beginning of November 2007 (Added after the official deadline of the report.): the more a company pays, the more often will the add be visible. Google answers that result from queries are also already ranked when searches are conducted (we give strong evidence for this in Section 1): Indeed we believe it cannot avoid ranking companies higher in the future who pay for such improved ranking: it is responsible to stockholders to increase the company's value. Google is of course doing this already for ads.

- Since most material that is written today is based on Google and Wikipedia, if those two do not reflect reality, the picture we are getting through "googeling reality" as Stephan Weber calls it, is not reality, but the Google-Wikipedia version of reality. There are strong indications that Google and Wikipedia cooperate: some sample statistics show that random selected entries in Wikipedia are consistently rated higher in Google than in other search engines.

- That biased contributions can be slipped into Wikipedia if enough money is invested is well established.

- Google can use its almost universal knowledge of what is happening in the world to play the stock market without risk: in certain areas Google KNOWS what will happen, and does not have to rely on educated guesses as other players in stock market have to. This is endangering trading on markets: by game theory, trading is based on the fact that nobody has complete information (i.e. will win sometimes, but also loose sometimes). Any entity that never looses rattles the basic foundations of stock exchanges!

- It has to be recognized that Google is not an isolated phenomenon: no society can leave certain basic services (elementary schooling, basic traffic infrastructure, rules on admission of medication, …) to the free market. It has to be recognized that Internet and the WWW also need such regulations, and if international regulations that are strong enough cannot be passed, then as only saving step an anti-Trust suite against Google has to be initiated, splitting the giant in still large companies, each able so survive, but with strict “walls” between them.

- It has to be recognized that Google is very secretive about how it ranks, how it collects data and what other plans it has. It is clear from actions in the past (as will be discussed in this report) that Google could dominate the plagiarism detection and IPR violation detection market, but chooses not to do so. It is clear that it has strong commercial reasons to act as it does.

- Google’s open aim is to “know everything there is to know on Earth”. It cannot be tolerated that a private company has that much power: it can extort, control, and dominate the world at will.

Lest we are misunderstood: we are not against new technologies: But we have to watch and keep in check new technology monopolies that endanger the very fabric of our society. I thus call for immediate action. In a nutshell, and some are this:

- Break the monopoly of Google as search engine by developing specialized search engines (from carpentry to medicine). This MUST be supported by governments, initially: I am herewith challenging the major powers in Europe including the European Commission and the European parliament. After initial funding those search engines can be self supporting due to two reasons: they can perform better in their area due to area specific features (like terminology) and they can generate income by offering plagiarism and IP rights violation detection services. Such search engines should be run by non-profit organisations. Universities that are publicly funded and under constant control might be obvious places, other governmental agencies might be just as suitable.

- It is necessary to initiate anti.trust measures on all levels against ANY company that combines sophisticated search engines with other powerful tools for data mining, exactly as Google is demonstrating with Gmail, Google Files, Google Earth, YouTube, Cheap or free Internet Access (to get to the real identity of users might be intended), etc. Observe that this DOES NOT endanger Google as search engine: revenue through ads in 2006 generated by just the search engine was in the multi billion dollar range, i.e. are a solid and sound foundation to continue the good work Google is doing as search engine. However, it has to be separated from other efforts like the ones I mentioned so as not to overpower the market, the economy and all of us.

Please study Appendices 4 and 5 carefully for more detail!

Section 8: Emerging Data Mining Applications: Advantages and Threats

In this section we want to make it clear that data mining (i.e. search engines and the like) are indeed both helpful and dangerous, as has been already been mentioned in Part 2 (Executive Summary) and expanded a bit on in the preceding Part 7.

Current national data protection laws and data collection laws are totally inadequate in a global world. To rule in what is happening in cases like Google would need a broad range of international laws. However, to reach a consensus on such laws will prove to be even more difficult to come to an agreement on not spreading nuclear arms, on global disarmament or on effective measures for climate control.

The conclusion of this and the short Section 9-11 will be that we do need government intervention. Since this will not be possible by international laws, at least not in the short run, the alternative is clear: anti-trust codes against companies whose power has become too large and diversified. This has to be done in a way that the parts of the companies remaining will still be viable enterprises, but that unlimited data- and information flow and cooperation beyond certain limits will not be possible any more.

(Note: The rest of this section was joint work between N. Kulathuramaiyer and H. Maurer)

Abstract:

Data Mining describes a technology that discovers non-trivial hidden patterns in a large collection of data. Although this technology has a tremendous impact on our lives, the invaluable contributions of this invisible technology often go unnoticed. This paper addresses the various forms of data mining while providing insights into its expanding role in enriching our life. Emerging forms of data mining are able to perform multidimensional mining on a wide variety of heterogeneous data sources, providing solutions to many problems. This paper highlights the advantages and disadvantages arising from the ever-expanding scope of data mining. Data Mining augments human intelligence by equipping us with a wealth of knowledge and by empowering us to perform our daily task better. As the mining scope and capacity increases, users and organisations become more willing to compromise privacy. The huge data stores of the ‘master miners’ allow them to gain deep insights into individual lifestyles and their social and behavioural patterns. The data on business and financial trends together with the ability to deterministically track market changes will allow an unprecedented manipulation of the stock market. Is it then possible to constrain the scope of mining while delivering the promise of better life?

1. Introduction

As we become overwhelmed by an influx of data, Data Mining presents a refreshing means to deal with this onslaught. Data Mining thus holds the key to many unresolved age-old problems. Having access to data thus becomes a powerful capability which can be effectively be harnessed by sophisticated mining software. Data at the hands of credit card companies will allow them to profile customers according to lifestyles, spending patterns and brand loyalty. Political parties on the other hand are able to predict with reasonable accuracy how voters are likely to vote. [Rash, 2006]

According to [Han and Kamber, 2006] data mining is defined as the extraction of interesting (non trivial, implicit, previously unknown and potentially useful) information or patterns from data in large databases. We take a broad view of data mining, where we also include other related machine based discoveries such as deductive query processing and visual data mining. Databases may include both structured data (in relational databases), semi structured data (e.g. metadata in XML documents) as well as unstructured documents such as text documents and multimedia content.

Despite the success stories in areas such as customer relationship modelling, fraud detection, banking, [KDD, 2005], the majority of applications tend to employ generic approaches and lack due integration with workflow systems. As such, Data Mining is currently at a chasm state and has yet to become widely adopted by the large majority [Han and Kamber, 2006].

2. Data Mining Process

Data Mining typically describes an automated acquisition of knowledge from a large collection of data. This traditional view corresponds to the knowledge creation phase in knowledge management. Current developments of data mining have expanded this to also cover support for knowledge organization and consolidation with respect to existing domain knowledge, and data visualization for iterative learning. Data Mining can thus be seen as complementing and supplementing knowledge management in a variety of phases.

In order to describe the processes involved in performing data mining we divide it into 3 phases: domain focusing, model construction (actual mining using machine learning algorithms), and decision making (applying the model to unseen instances).

Data focusing phase involves the application of some form of clustering or may incorporate intensive knowledge engineering for complex applications. At the model construction phase, a model of generalized patterns is constructed to capture the intrinsic patterns stored in the data.

The model generated is then employed for decision-making. Simplistic applications of data mining tend to merely employ the model to predict the likelihood of events and occurrences, based largely on past patterns. Amazon, for example, is able to recommend books according to a user’s profile. Similarly, network operators are able to track fraudulent activities in the usage of phone lines by tracking deviation patterns as compared to standard usage characterization.

3.. Applications of Data Mining

3.1 Web Search As Data Mining

Search engines have turned the Web into a massive data warehouse as well as a playground for automated discovery of hidden treasures. Web Search is thus viewed as an extensive form of multidimensional heterogeneous mining of a largely unstructured database for uncovering an unlimited number of mind-boggling facts. The scale of data available is in the range of peta bytes, and it much greater than the terra bytes of data available at the hands of large global corporations such as Walmart.

Search engines can either simultaneously or incrementally mine these datasets to provide a variety of search results which include phone contacts, street addresses, news feeds, dynamic web content, images, video, audio, speech, books, artifacts. In performing domain focusing, a model allowing to characterize aggregated user search behaviour is used [Coole et al, 1997]. This phase could involve associational subject link analysis, requiring a contextual domain analysis (for mobile users). This Mining phase involves the derivation of aggregated usage profiles based on a multidimensional mining of usage patterns according to clustered characterization. By analyzing search history over a period of time, search engines have access to a great deal of insights into lives of presumably ‘anonymous’ searchers. A search query can indicate the intent of a user to acquire particular information to accomplish a task at hand. Search engines can thus track patterns and drifts in global user intentions, moods, and thoughts.

3.2 Environmental Modelling Application

There are complex problems for which data mining could be used to provide answers by uncovering patterns hidden beneath layers of data. In many cases, domain focusing has in the past has been the biggest challenge. Data mining could be employed for the modeling of environmental conditions in the development of an early warning system to address a wide range of natural disasters such as avalanches, landslides, tsunami and other environment events such as global warming. The main challenge in addressing such a problem is in the lack of understanding of structural patterns characterizing various parameters which may currently not be known.

As highlighted by [Maurer et al], although a large variety of computer-based methods have been used for the prediction of natural disasters, the ideal instrument for forecasting has not been found yet. As highlighted in their paper, there are also situations whereby novel techniques have been employed but only to a narrow domain of limited circumstances.

Integration of multiple databases and the compilation of new sources of data are required in the development of full-scale environmental solutions. As advances in technology allow the construction of massive databases through the availability of new modes of input such as multimedia data and other forms of sensory data, data mining could well provide a solution. In order to shed insights on a complex problem such as this, massive databases that were not previously available need to be constructed e.g. data about after event situations of the past [Maurer et al, 2007]. Such data on past events could be useful in highlighting patterns related to potentially in-danger sites. Data to be employed in this mining will thus comprise of both of weather and terrestrial parameters together with other human induced parameters such as vegetation or deforestation over a period of time. [Maurer et al, 2007]

Domain focusing will be concerned with discovery of causal relationships (e.g. using Bayes networks) as a modeling step. Multiple sources of data which include new sources of data need to be incorporated in the discovery of likely causal relationship patterns. A complex form of data mining is required even at the phase of domain focusing. This will involve an iterative process whereby hypothesis generation could be employed to narrow the scope of the problem to allow for a constrained but meaningful data collection. For complex domains such as this, unconstrained data collection may not always be the best solution. Domain focusing would thus perform problem detection, finding deterministic factors and to hypothesize relationships that will be applied in the model. [Beulens et al, 2006] describe a similarly complex representation for an early warning system for food supply networks.

Subsequently, the model construction phase will employ a variety of learning algorithms, to profile events or entities being modeled. As this stage may negate model relationships, domain focusing will need to be repeated and iteratively performed. The model construction phase will allow the incremental development of a model, based on a complex representation of the causal networks. [Beulens et al, 2006]

Mining methods such as clustering, associational rule mining, neural networks will be used to verify the validity of causal associations. Once a potential causal link is hypothesized, verification can be done via the application of data mining methods. [Beulens et al, 2006], have proposed a combination of approaches which include deviation detection, classification, dependence model and causal model generation.

The Decision Making phase will then apply the validated causal relationship model in exploring life case studies. An environment for an interactive explorative visual domain focusing is crucial, to highlight directions for further research. Data mining could serve as a means of characterization of profiles for both areas prone to disasters or those that are safe.

3.3 Medical Application

We will briefly discuss another form of mining that has a high impact. In the medical domain, data mining can be applied to discover unknown causes to diseases such as ‘sudden death’ syndrome or heart attacks which remain unresolved in the medical domain. The main difficulty in performing such discoveries is also in collecting the data necessary to make rational judgments. Large databases need to be developed to provide the modeling capabilities. These databases will comprise of clinical data on patients found to have the disease, and those who are free of it. Additionally non-traditional data such as including retail sales data to determine the purchase of drugs, and calls to emergency rooms together with auxiliary data such as micro array data in genomic databases and environmental data would also be required. [Li, 2007]

Non traditional data could also incorporate major emotional states of patients by analyzing and clustering the magnetic field of human brains which can be measured non invasively using electrodes to a persons’ heads. [Maurer et al, 2007] Social patterns can also be determined through profile mining as described in the previous section to augment the findings of this system. Findings of functional behavior of humans via the genomic database mining would also serve as a meaningful input.

The development of large databases for medical explorations will also open possibilities for other discoveries such as mining family medical history and survival analysis to predict life spans. [Han and Kamber, 2006]

4. The Advantages of Data Mining

Data mining has crept into our lives in a variety of forms. It has empowered individuals across the world to vastly improve the capacity of decision making in focussed areas. Powerful mining tools are going to become available for a large number of people in the near future.

The benefits of data mining will include preserving domestic security through a number of surveillance systems, providing better health through medical mining applications, protection against many other forms of intriguing dangers, and access to just-in-time technology to address specific needs. Mining will provide companies effective means of managing and utilizing resources. People and organizations will acquire the ability to perform well-informed (and possibly well-researched) decision-making. Data mining also provides answers through sifting through multiple sources of information which were never known to exist, or could not be conceivably acquired to provide enlightening answers. Data Mining could be combined with collaborative tools to further facilitate and enhance decision-making in a variety of ways. Data mining is thus able to explicate personal or organizational knowledge which may be locked in the heads of individuals (tacit knowledge) or in legacy databases, to become available. Many more new benefits will emerge as technology advances.

5. Disadvantages of Data Mining

A great deal of knowledge about users is also being maintained by governments, airlines, medical profiles or shopping consortiums. Mining applications with dramatic privacy infringement implications include search history, real-time outbreak and disease surveillance program, early warning for bio-terrorism [Spice, 2005] and Total Information Awareness program.[Anderson, 2004] For example, search history data represents an extremely personal flow of thought patterns of users that reflects ones quest for knowledge, curiosity, desires, aspirations, as well as social inclinations and tendencies. Such logs can reveal a large amount of psychographic data such as user’s attitudes towards topics, interests, lifestyles, intents and beliefs. A valid concern would be that the slightest leak could be disastrous. The extent of possible discoveries has been clearly illustrated by the incidence where AOL released personal data of 658,000 subscribers [Jones, 2006].

Another common danger is profiling where there is a possibility of drastic implications such as a conviction being made based on the incriminating evidences of mining results. There is also a danger of over-generalization based on factors such as race, ethnicity, or gender. This could result in false positives, where an entirely innocent individual or group is targeted for investigation based on a poor decision making process. For the domestic security application, a reasonable data mining success rate of 80% implies that 20% of all US citizens (or 48 million people) would be considered false positives [Manjoo, 2002].

Data mining will further empower mining kingpins to be able to go beyond the ability to PREDICT what is going to happen in a number of areas of economic importance, but actually have the power to KNOW what will happen, hence can e.g. exploiting the stock market in an unprecedented way. They also have the capacity to make judgments on issues and persons with scary accuracy.

6. What can we do?

A majority of related works are more concerned about privacy, but it is no longer the main issue of concern. [Kovatcheva, 2002] has a proposed a means of protecting the anonymity by the use of anonymity agents and pseudonym agents to avoid users from being identified. Their paper also proposed the use of negotiation and trust agents to assist users in reviewing a request from a service before making a rational decision of allowing the use of personal data.

A similar agent-based approach is described by [Taipale, 2003] via rule-based processing. An "intelligent agent" is used for dispatching a query to distributed databases. The agent will negotiate access and permitted uses for each database. Data items are labeled with meta-data describing how that item must be processed. Thus, a data item is protected as it retains relevant rules by which it describes the way it has to be processed. The main challenge then lies in coming up with guidelines and rules such that site administrators or software agents can use to direct various analyses on data without compromising the identity of an individual user. This approach however is not applicable for areas such as Web search, where a standard framework for conscientious mining is far from sight. Furthermore, the concern with the emergence of extensive mining is no longer solved by addressing privacy issues only. As more and more people are willing to compromise privacy, the questions that we pose are: Who do we trust as the gatekeeper of all our data? Do we then trust all our private data at the hands of a commercial global company?

One approach to overcome the concerns mentioned above is by employing a distributed data mining approach, where separate agencies will maintain and become responsible and accountable for different (independent segments of) data repositories. This proposal further ensures that no central agency will have an exclusive control over the powerful mining technology and all resources. [Kulathuramaiyer, Maurer, 2007] In order to realize this solution, governments have to start playing a more proactive role in maintaining and preserving national or regional data sources. We are also beginning to see partnerships between global search engines and governments in this respect. Such partnerships should however be built upon a balanced distribution of earnings. Such a solution can be complemented by the nurturing of a larger number of autonomous context-specific or task specific miners.

7. Conclusion

As data mining matures and becomes widely deployed in even more encompassing ways, we need to learn to effectively apply it to enrich our lives. At the same time, the dangers associated with this technology needs to be minimized by deliberate efforts on the part of the enforcement agencies, data mining agencies and the users of the system. There should be strict regulations to prevent the abuse or misuse of data. Users should also be made aware of the privacy policies in order to make an informed decision about revealing their personal data. The success of such regulations and guidelines can only be guaranteed if they are backed up by a legal framework

References

[Anderson, 2004], Anderson, S.R., Total Information Awareness and Beyond, Bill of Rights Defense Committee; White paper. The Dangers of Using Data Mining Technology to Prevent Terrorism, July 2004

[Beulens et al, 2006] Beulens, A., Li, Y., Kramer, M., van der Vorst, J., Possibilities for applying data mining for early Warning in Food Supply Networks, CSM'06, 20thWorkshop on Methodologies and Tools for Complex System Modeling and Integrated Policy Assessment, August, 2006 http://www.iiasa.ac.at/~marek/ftppub/Pubs/csm06/beulens_pap.pdf

[Coole et al, 1997] Coole, R. Mobasher, B., Srivastava, J., Grouping Web Page References into Transactions for Mining World Wide Web Browsing Patterns, Proceedings of the 1997 IEEE Knowledge and Data Engineering Exchange Workshop Page: 2, 1997 ISBN:0-8186-8230-2 IEEE Computer Society

[Han and Kamber, 2006] Han, J., and Kamber, M., Data Mining: Concepts and Techniques, 2nd ed., The Morgan Kaufmann Series in Data Management Systems, Jim Gray, Series Editor , Morgan Kaufmann Publishers, March 2006.

[Jenssen, 2002] Jenssen, D., Data mining in networks. Invited talk to the Roundtable on Social and Behavior Sciences and Terrorism. National Research Council, Division of Behavioral and Social Sciences and Education, Committee on Law and Justice. Washington, DC. December 11, 2002

[Jones, 2007] Jones, K .C., Fallout From AOL's Data Leak Is Just Beginning , http://www.informationweek.com/news/showArticle.jhtml?articleID=191900935, accessed 2007

[KDD, 2005] KDDnuggets : Polls: Successful Data Mining Applications, July 2005 http://www.kdnuggets.com/polls/2005/successful_data_mining_applications.htm

[Kovatcheva, 2002] Kovatcheva, E., Tadinen ,H., The technological and social aspects of data mining by means of web server access logs http://www.pafis.shh.fi/~elikov02/SFISWS2/SFIS2.html 18 January 2002

[Kulathuramaiyer, Balke, 2006]Kulathuramaiyer, N., Balke, W.-T., Restricting the View and Connecting the Dots — Dangers of a Web Search Engine Monopoly, J,UCS Vol. 12 , Issue 12, pp.1731 – 1740, 2006

[Kulathuramaiyer N., Maurer, H., 2007], Kulathuramaiyer, N., Maurer, H., "Why is Fighting Plagiarism and IPR Violation Suddenly of Paramount Importance?" Proc. of International Conference on Knowledge Management, Vienna, August 2007.

[Li, 2007] Li, C. S., Survey of Early Warning Systems for Environmental and Public Health Applications, in Wong, ,S., Li,, C. S.,(eds.), Life Science Data Mining, Science, Engineering, and Biology Informatics- Vol. 2, 2007 http://www.worldscibooks.com/compsci/etextbook/6268/6268_chap01.pdf

[Manjoo, 2002] Manjoo, F., Is Big Brother Our Only Hope Against Bin Laden?, Dec. 3, 2002 http://www.salon.com/tech/feature/2002/12/03/tia/index_np.html

[Maurer et al., 2007] Maurer, L., Klingler, C., Pachauri, R. K., and Tochtermann, K,. Data Mining as Tool for Protection against Avalanches and Landslides, Proc. Environmental Informatics Conference, Warsaw, 2007

[Milne, 2000] Milne, G. R., Privacy and ethical issues in database/interactive marketing and public policy: A research framework and overview of the special issue, Journal of Public Policy & Marketing, Spring 2000

[Mobasher, 2005] Mobasher, B., Web Usage Mining and Personalisation, in Singh, M. P. (ed.) Practical Handbook of Internet Computing, Chapman & Hall/ CRC Press, 2005 http://maya.cs.depaul.edu/~mobasher/papers/IC-Handbook-04.pdf

[Rash, 2006] Rash, W., Political Parties Reap Data Mining Benefits ,eWeek.com enterprise News and reviews, November 16, 2006; http://www.eweek.com/article2/0,1895,2060543,00.asp

[Spice, 2003] Spice, B., Privacy in age of data mining topic of workshop at CMU, March 28, 2003 http://www.post-gazette.com/nation/20030328snoopingnat4p4.asp

[Taipale, 2003] Taipale, K.A. "Data Mining and Domestic Security: Connecting the Dots to Make Sense of Data". Colum. Sci. & Tech. L. Rev. 5 (2). SSRN 546782 / OCLC 45263753, December 15, 2003.

Section 9: Feeble European attempts to fight Google while Google's strength is growing

Some European officials have realized the danger emanating from an Internet world dominated by a private and even Non-European company. However, concrete efforts have been few and have more or less failed, or changed direction.

The first project, a joint undertaking between France and Germany called "Quaero" was well funded and was seen as an initiative to curb the power of Google. However, after an interesting start the partners have basically split. Although much has been said against Google and Wikipedia let us still quote the first lines one finds in Wikipedia if one searches for Quaero:
*"**Quaero** (lateinisch ich suche) ist ein französisches Projekt mit deutscher Beteiligung zur Finanzierung der Erforschung von Suchmaschinen.*
Seit Oktober 2004 ist die zum Quaero-Projekt gehörige Internet-Suchmaschine Exalead mit Sitz in Paris (Frankreich) online. Ansonsten befindet sich das Projekt im Stadium wechselnder politischer Verlautbarungen. Eine Suchmaschine unter dem Namen quaero existiert derzeit (Anfang September 2007) nicht. Ein bereits bestehendes öffentliches Angebot einer experimentellen Suchmaschine unter diesem Namen im Internet wurde etwa Ende 2006 wieder auf einen engeren Kreis von Zugriffsberechtigten beschränkt. Der Name Quaero selbst wurde von den Projektbetiligten offensichtlich nicht geschützt; gleichnamige Domains im Internet werden offensichtlich von unbeteiligten Dritten genutzt.
Das geplante Vorhaben Quaero wurde im April 2005 von Jacques Chirac und Gerhard Schröder bekanntgegeben und Anfang 2006 eingeleitet. Am 26. April 2006 kündigte Jacques Chirac dazu ein auf fünf Jahre angelegtes Entwicklungsbudget von 250 Millionen Euro seitens der Agence de l'innovation industrielle und der Industrie an. Ursprünglich sollte sich die Gesamtfördersumme auf gut 400 Millionen Euro belaufen, wovon 240 Millionen Euro von der deutschen Bundesregierung stammen sollten.[1]
Beim ersten deutschen „IT-Gipfel" in Potsdam am 18. Dezember 2006 sagte Staatssekretär Hartmut Schauerte, die Bundesregierung werde sich aus dem Quaero-Konsortium zurückziehen und sich stattdessen auf das rein deutsche Forschungsprogramm unter dem Namen „Theseus" konzentrieren."
Thus, Germany is now investing on its own in a project „Theseus". The German government denies that Theseus is a search engine project, it is an investment in knowledge infrastructure and semantic technologies:

"Das THESEUS-Konsortium begrüßt die Entscheidung der EU-Kommission, die öffentliche Förderung des Forschungsprogramms THESEUS durch das Bundesministerium für Wirtschaft und Technologie (BMWi) zu genehmigen. Das Programm hat eine Laufzeit von fünf Jahren und wird vom BMWi mit ca. 90 Mio. Euro gefördert. Die für Forschung und Entwicklung zur Verfügung stehenden Mittel verteilen sich je zur Hälfte auf Wissenschaft und Wirtschaft. Zusätzliche 90 Mio. Euro werden als Eigenmittel der beteiligten Partner aus Industrie und Forschung aufgebracht, so dass insgesamt rd. 180 Mio. Euro in die zukunftweisenden Forschungsarbeiten fliessen.

Zahlreiche Unternehmen, Forschungseinrichtungen und Universitäten starten in den kommenden Wochen und Monaten zahlreiche und vielfältige Forschungsprojekte zur Entwicklung anwendungsorientierter Basistechnologien und technischer Standards für eine neue internetbasierte Wissensinfrastruktur. Diese Basistechnologien werden von den Industriepartnern im Konsortium in 7 Anwendungsszenarien prototypisch umgesetzt und erprobt. Dabei soll überprüft werden, wie diese neuen Technologien zeitnah in innovative Werkzeuge, marktfähige Dienste und erfolgsversprechende Geschäftsmodelle für das World Wide Web umgesetzt werden können.

Das THESEUS-Konsortium wird durch die empolis GmbH, einer Tochtergesellschaft der arvato AG, koordiniert. Zum Konsortium gehören Siemens, SAP, empolis, Lycos Europe, Deutsche Nationalbibliothek, sowie Deutsche Thomson oHG, intelligent views, m2any, Moresophy, Ontoprise, *Festo, Verband Deutscher Maschinen und Anlagenbau (VDMA) und das Institut für Rundfunktechnik. Dabei arbeiten die industriellen Forschungs- und Entwicklungsabteilungen eng mit den öffentlichen Forschungspartnern zusammen. Dazu gehören international anerkannte Experten des Deutschen Forschungszentrums für Künstliche Intelligenz (DFKI), des Forschungszentrums Informatik (FZI), der Ludwig-Maximilians-Universität (LMU) und Technischen Universität (TU) München, der TU Darmstadt, der Universität Karlsruhe (TH), der TU Dresden und der Universität Erlangen.*

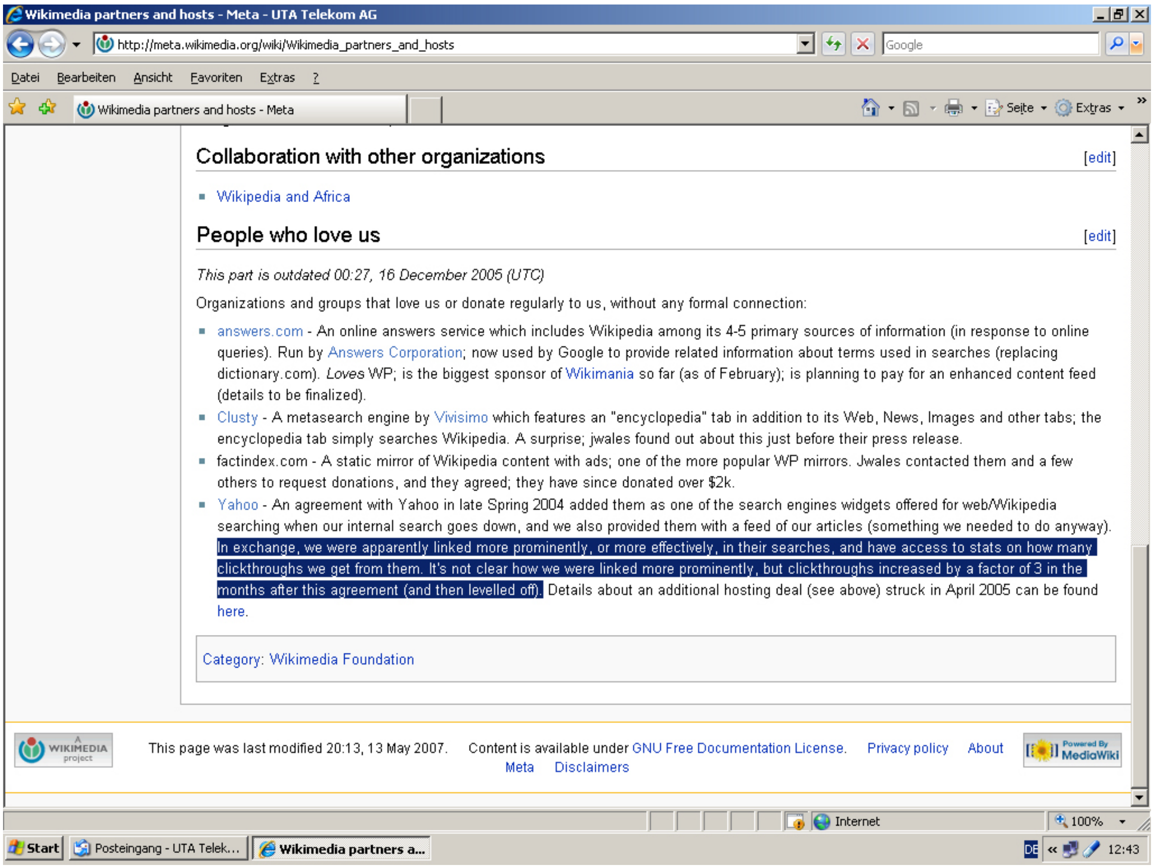
Im Fokus des Forschungsprogramms stehen semantische Technologien, die die inhaltliche Bedeutung der Informationen (Wörter, Bilder, Töne) erkennen und einordnen können. Mit diesen Technologien können Computerprogramme intelligent nachvollziehen, in welchem inhaltlichen Kontext Daten genutzt und verarbeitet werden. Darüber hinaus können Computer durch Anwendung von Regeln und Ordnungsprinzipien aus den Inhalten logische Schlüsse ziehen und selbständig Zusammenhänge zwischen unterschiedlichen Informationen aus mehreren Quellen erkennen und herstellen. Dabei werden künftig Internet-Nutzer mit Hilfe der von THESEUS erarbeiteten Standards und Basistechnologien ("semantischer Werkzeugkasten") selbst Inhalte, Regeln und Ordnungen erstellen und bearbeiten sowie multimediale Inhalte intelligent aufbereiten, sammeln und verknüpfen können. Auf diese Weise wird aus dem heutigen Web 2.0 mit seiner offenen, interaktiven und sozialen Vernetzungsphilosophie durch die Verknüpfung mit semantischen Methoden das Internet der nächsten Generation.

Zu den Basistechnologien, die von den Forschungspartnern entwickelt werden, gehören unter anderem Funktionen zur automatisierten Erzeugung von Metadaten für Audio-, Video-, 3D- und Bilddateien und Mechanismen für die semantische Verarbeitung multimedialer Dokumente und der damit verknüpften Services. Im Fokus der Forschung steht auch die Entwicklung von Werkzeugen für das Management von Ontologien. Darüber hinaus entwickeln die Forschungspartner neue Methoden des maschinellen Lernens und der situationsbewussten Dialogverarbeitung. Gleichzeitig wird auch an innovativen Benutzeroberflächen und Interfaces gearbeitet. Neue Verfahren des Digital Rights Management (DRM) sollen die Urheber- und Vermarktungsrechte am geistigen Eigentum multimedialer Inhalte künftig besser schützen.“

Thus, de facto no large European Initiative against Google is visible. In the meantime Google continues to do well, and is working on still more dangerous (powerful) tools, e.g. combining human and machine intelligence.
(Note: The rest of this section was compiled by S. Weber after discussions with H. Maurer, who did a final bit of editing)

1. Reported or suspected cases of manipulation from inside: When Google announced its new commentary function for Google News in August of 2007 [see N. N. 2007b], for the first time it became clear that not only mathematical algorithms and software, but also human brains in the Google headquarter will edit information processed by Google and decide what will go online and in which form. Thus we were witness of a small revolution: If Google will step-by-step not only allow algorithms to process information (or bring information into a hierarchy), but also the Google editorial staff (as it is in the case of the commentary function of Google News) will do that, the paradigm changes: Google opens the door for possible interventions by will, by human mind. Cases of inner manipulation of the Google search engine results (amongst them the most prominent case of www.google.cn) are reported in [N. N. 2006b]. There were rumours about a secret collaboration of Wikipedia and Google because Wikipedia key term entries usually appear very high on the list of Google search results (if not on top rank). In the empirical part of this study we were able to prove this fact. Of course this is no proof of a hidden collaboration! But it is interesting that for example Wikipedia admits on its web site that they had arranged a collaboration with Yahoo!, and the months after the contract Wikipedia links climbed up the Yahoo! search results list:

Figure 39: Wikimedia admitting that Yahoo! improved search results for them


Wikimedia partners and hosts - Meta - UTA Telekom AG
http://meta.wikimedia.org/wiki/Wikimedia_partners_and_hosts
Google
Datei Bearbeiten Ansicht Favoriten Extras ?
Wikimedia partners and hosts - Meta
Seite
Extras
Internet
100%
Start
Posteingang - UTA Telek...
Wikimedia partners a...
DE
12:43

Collaboration with other organizations [edit]

- Wikipedia and Africa

People who love us [edit]

This part is outdated 00:27, 16 December 2005 (UTC)

Organizations and groups that love us or donate regularly to us, without any formal connection:

- answers.com - An online answers service which includes Wikipedia among its 4-5 primary sources of information (in response to online queries). Run by Answers Corporation; now used by Google to provide related information about terms used in searches (replacing dictionary.com). *Loves* WP; is the biggest sponsor of Wikimania so far (as of February); is planning to pay for an enhanced content feed (details to be finalized).
- Clusty - A metasearch engine by Vivisimo which features an "encyclopedia" tab in addition to its Web, News, Images and other tabs; the encyclopedia tab simply searches Wikipedia. A surprise; jwales found out about this just before their press release.
- factindex.com - A static mirror of Wikipedia content with ads; one of the more popular WP mirrors. Jwales contacted them and a few others to request donations, and they agreed; they have since donated over $2k.
- Yahoo - An agreement with Yahoo in late Spring 2004 added them as one of the search engines widgets offered for web/Wikipedia searching when our internal search goes down, and we also provided them with a feed of our articles (something we needed to do anyway). In exchange, we were apparently linked more prominently, or more effectively, in their searches, and have access to stats on how many clickthroughs we get from them. It's not clear how we were linked more prominently, but clickthroughs increased by a factor of 3 in the months after this agreement (and then levelled off). Details about an additional hosting deal (see above) struck in April 2005 can be found here.

Category: Wikimedia Foundation

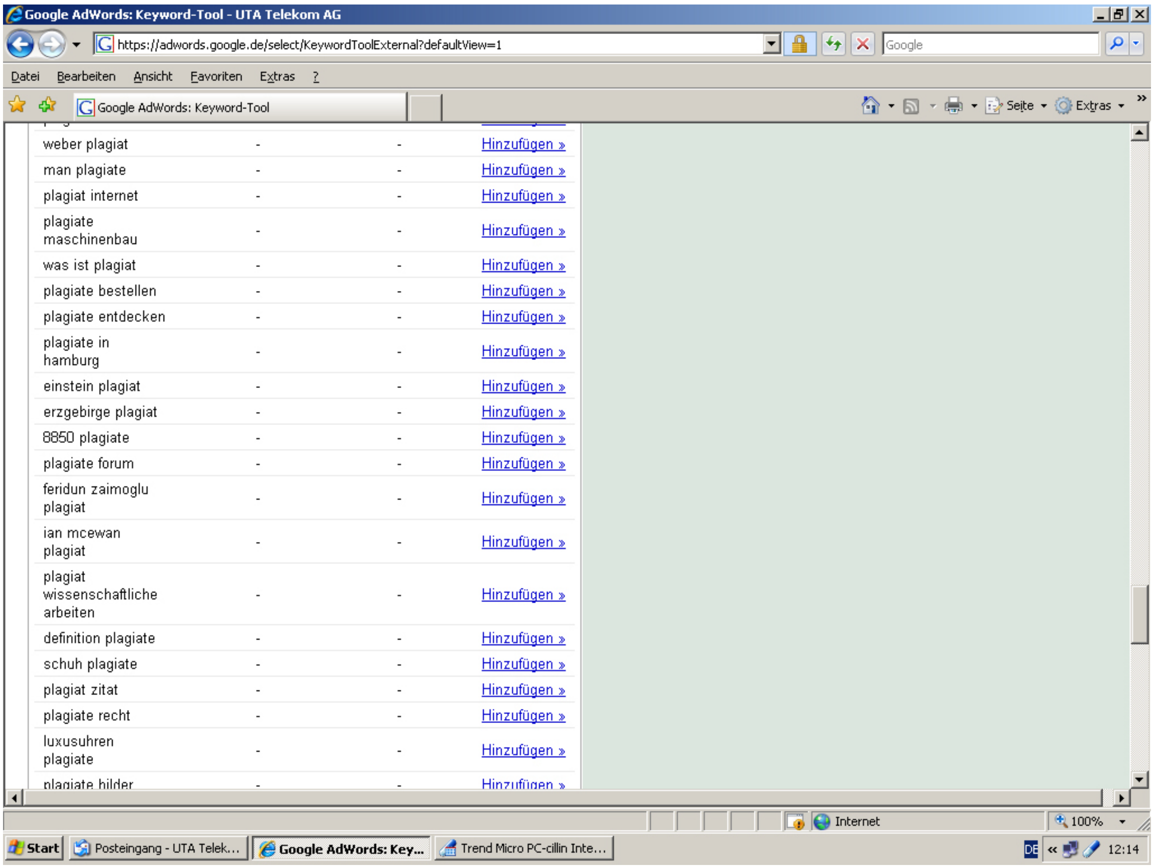
WIKIMEDIA project This page was last modified 20:13, 13 May 2007. Content is available under GNU Free Documentation License. Privacy policy About Meta Disclaimers Powered By MediaWiki

[Personal Screenshot, 15/5/07]

2. Reported or suspected cases of manipulation from outside: As is commonly known, Google research results can be manipulated from outside. "Google Bombs" (influencing rankings with intentionally set hyperlinks), "Google Spamming" (putting websites on top of the ranking) and "Google Bowling" (putting websites out of the ranking) [see Salchner 2007] were easier to realise in the past, but the problem of willingly influencing the results ranking remains. And if you think of the complaints of corporations about cases of suspected click betrayal or non-transparent billing done by Google, you can clearly see that the problem still exists and is often played down by Google representatives.

3. Endangering or privacy with the possibility of data mining: Google already is the world's biggest detective agency. So far there are no reports about Google passing personal data over to governments or private companies. On the contrary, Yahoo! and MSN did it at least once for the American ministry of Justice [see Rötzer 2007]. But Google already is able to collect and combine all these data packages from individuals all over the world for personal user profiles for their own use. A Google user should not be naive to believe that Google even today doesn't publish his or her search results. If you type in a text string, and some other people use the same phrase, your search term might appear on the AdWords keywords list. This is already now a very easy way for everybody to look what mankind is or was googling. If you take the German search term "Plagiat", the results are surprising:

Figure 40: The Google keyword tool offers insights in search terms of users



Google AdWords: Keyword-Tool - UTA Telekom AG
https://adwords.google.de/select/KeywordToolExternal?defaultView=1
Google
Datei
Bearbeiten
Ansicht
Favoriten
Extras
?
Google AdWords: Keyword-Tool
Seite
Extras
Internet
100%
Start
Posteingang - UTA Telek...
Google AdWords: Key...
Trend Micro PC-cillin Inte...
DE
12:14

weber plagiat	-	-	Hinzufügen »
man plagiate	-	-	Hinzufügen »
plagiat internet	-	-	Hinzufügen »
plagiate maschinenbau	-	-	Hinzufügen »
was ist plagiat	-	-	Hinzufügen »
plagiate bestellen	-	-	Hinzufügen »
plagiate entdecken	-	-	Hinzufügen »
plagiate in hamburg	-	-	Hinzufügen »
einstein plagiat	-	-	Hinzufügen »
erzgebirge plagiat	-	-	Hinzufügen »
8850 plagiate	-	-	Hinzufügen »
plagiate forum	-	-	Hinzufügen »
feridun zaimoglu plagiat	-	-	Hinzufügen »
ian mcewan plagiat	-	-	Hinzufügen »
plagiat wissenschaftliche arbeiten	-	-	Hinzufügen »
definition plagiate	-	-	Hinzufügen »
schuh plagiate	-	-	Hinzufügen »
plagiat zitat	-	-	Hinzufügen »
plagiate recht	-	-	Hinzufügen »
luxusuhren plagiate	-	-	Hinzufügen »
plagiate bilder	-	-	Hinzufügen »

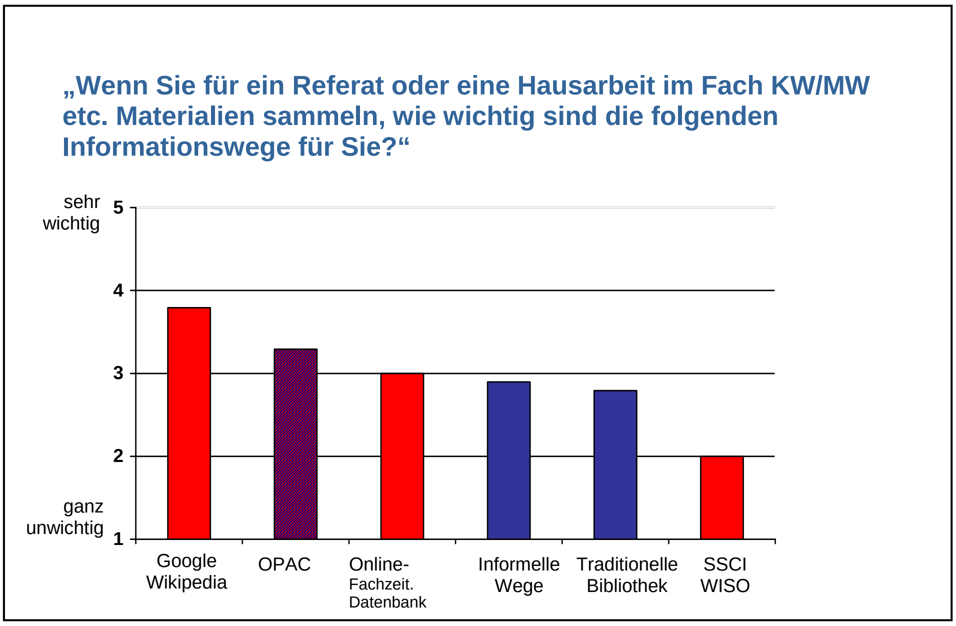
[Personal Screenshot, 11/5/07]

Nearly each day we pass over secret data to Google: maybe about our sexual preferences, maybe about our diseases (the so-called "Dr. Google effect"). With our IP addresses Google is able to make up a personal user profiles and to do cluster analysis of large groups of people. If you also use the Google toolbar or various devices which give Google the possibility to crawl on your desktop, the problem increases. Just think of all the information you pass over to Google: when you search, what you search, how long you scroll through the search results, and on which web site you klick when you leave Google. The dossier on our very own Googling behavior would fill books. If you are able to extract all relevant information by data mining, you know probably more about us than we do [also see the critique in Conti 2006; for a further discussion on Google and privacy concerns see Privacy International 2007 and N. N. 2007a].

4. Turning web searchers into "Google addicts": The problem of Google addiction so far is nearly completely neglected by social sciences and media psychology. On a recent ARTE documentation on Google, the phenomenon of "Google addiction" was covered: People who are Google addicts not only google the terms of their interest nearly the whole day (in an endless loop to see if something has changed), but they also use googling in the same way as chatting, doing phone calls or watching TV: to fill up time in which they do not work or do not want to work. So far there is no empirical study on Google addiction with convincing hard facts. We would strongly recommend to conduct such an investigation.

5. Cognitive and sociocultural consequences: The problem of "googling the reality" was already described in the context of the Google Copy Paste Syndrome which currently threatens science and education and introduces a tendency towards a culture of mediocrity – at least in some disciplines, on some universities and on some schools [for the dangers of the Google Wikipedia monopoly also see Rolfes 2007]. An American survey has revealed that search engine users are naive and tend to be cursorily in their research [see Fallows 2004 and Fallows 2005]. A recent German study revealed that the way to Google and Wikipedia already is on the top of the ranking of all research possibilities – online as well as offline:

Figure 41: The current Google Wikipedia information monopoly of German students

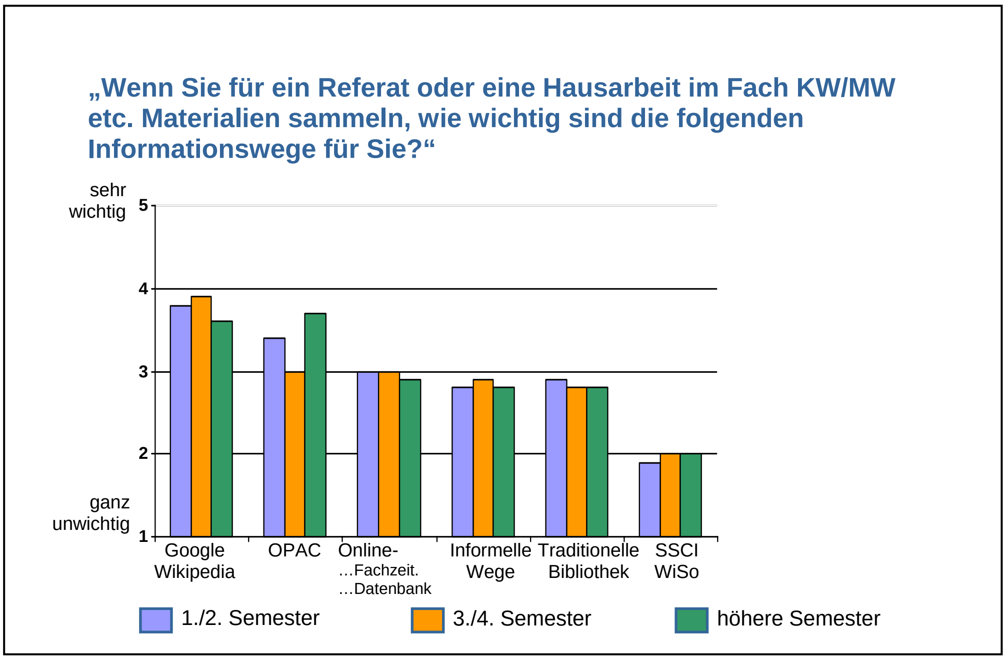


„Wenn Sie für ein Referat oder eine Hausarbeit im Fach KW/MW etc. Materialien sammeln, wie wichtig sind die folgenden Informationswege für Sie?“

[Source: Slide from Emmer & Wolling 2007]

This monopoly remains invariant which means that the students still primarily consult Google and Wikipedia in the higher semesters:

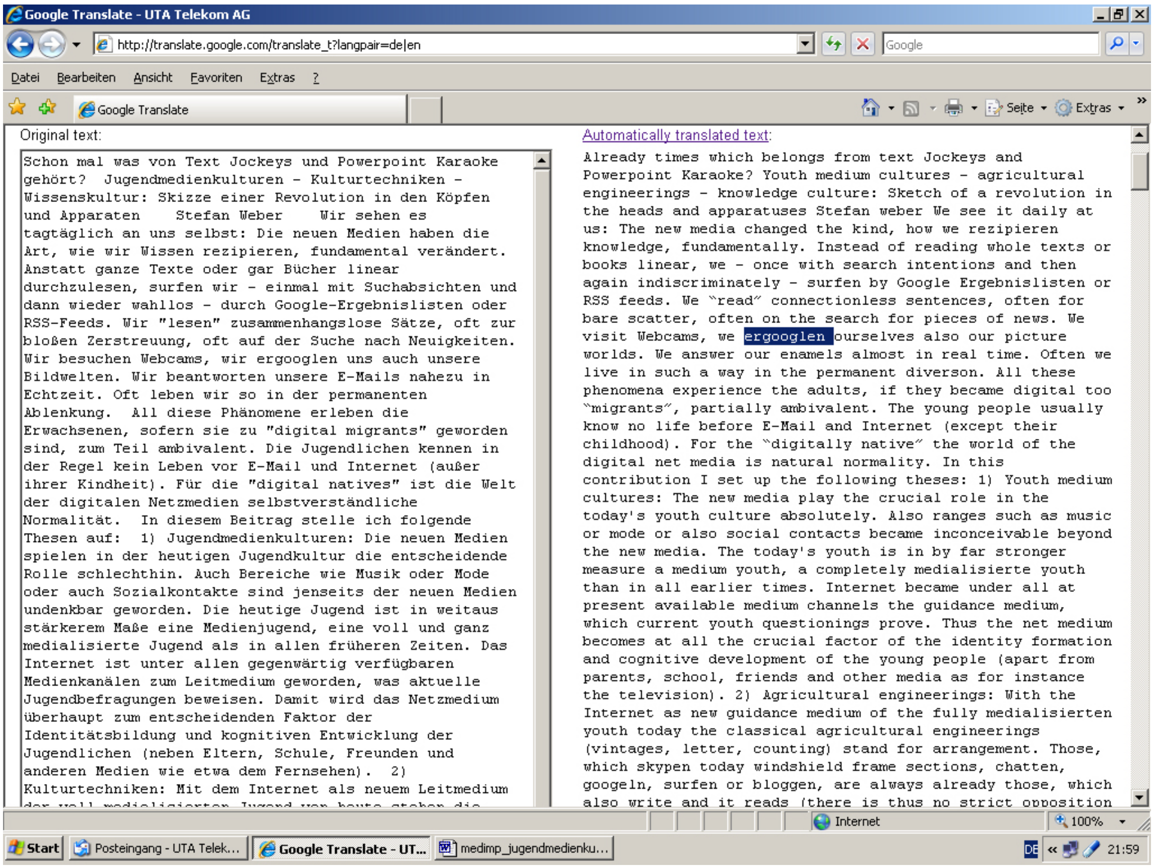
Figure 42: Invariancy of the Google Wikipedia information monopoly of German students



[Source: Slide from Emmer & Wolling 2007]

6. Encouraging translation plagiarism and not preventing plagiarism at all: It is surprising that Google offers so many tools, but no efficient tool to fight against plagiarism. On the contrary Google encourages and simplifies plagiarism of texts from other languages with its new and already quite well-working translation tool http://translate.google.com/translate_t.

Figure 43: Testing the Google translation tool


Google Translate - UTA Telekom AG
http://translate.google.com/translate_t?langpair=de|en
Google
Datei Bearbeiten Ansicht Favoriten Extras ?
Google Translate
Seite Extras
Internet
100%
Start
Posteingang - UTA Telek...
Google Translate - UT...
medimp_jugendmedienku...
DE
21:59

Original text:

Schon mal was von Text Jockeys und Powerpoint Karaoke gehört? Jugendmedienkulturen - Kulturtechniken - Wissenskultur: Skizze einer Revolution in den Köpfen und Apparaten Stefan Weber Wir sehen es tagtäglich an uns selbst: Die neuen Medien haben die Art, wie wir Wissen rezipieren, fundamental verändert. Anstatt ganze Texte oder gar Bücher linear durchzulesen, surfen wir - einmal mit Suchabsichten und dann wieder wahllos - durch Google-Ergebnislisten oder RSS-Feeds. Wir "lesen" zusammenhangslose Sätze, oft zur bloßen Zerstreuung, oft auf der Suche nach Neuigkeiten. Wir besuchen Webcams, wir ergooglen uns auch unsere Bildwelten. Wir beantworten unsere E-Mails nahezu in Echtzeit. Oft leben wir so in der permanenten Ablenkung. All diese Phänomene erleben die Erwachsenen, sofern sie zu "digital migrants" geworden sind, zum Teil ambivalent. Die Jugendlichen kennen in der Regel kein Leben vor E-Mail und Internet (außer ihrer Kindheit). Für die "digital natives" ist die Welt der digitalen Netzmedien selbstverständliche Normalität. In diesem Beitrag stelle ich folgende Thesen auf: 1) Jugendmedienkulturen: Die neuen Medien spielen in der heutigen Jugendkultur die entscheidende Rolle schlechthin. Auch Bereiche wie Musik oder Mode oder auch Sozialkontakte sind jenseits der neuen Medien undenkbar geworden. Die heutige Jugend ist in weitaus stärkerem Maße eine Medienjugend, eine voll und ganz medialisierte Jugend als in allen früheren Zeiten. Das Internet ist unter allen gegenwärtig verfügbaren Medienkanälen zum Leitmedium geworden, was aktuelle Jugendbefragungen beweisen. Damit wird das Netzmedium überhaupt zum entscheidenden Faktor der Identitätsbildung und kognitiven Entwicklung der Jugendlichen (neben Eltern, Schule, Freunden und anderen Medien wie etwa dem Fernsehen). 2) Kulturtechniken: Mit dem Internet als neuem Leitmedium der voll medialisierten Jugend von heute stehen die

Automatically translated text:

Already times which belongs from text Jockeys and Powerpoint Karaoke? Youth medium cultures - agricultural engineerings - knowledge culture: Sketch of a revolution in the heads and apparatuses Stefan weber We see it daily at us: The new media changed the kind, how we rezipieren knowledge, fundamentally. Instead of reading whole texts or books linear, we - once with search intentions and then again indiscriminately - surfen by Google Ergebnislisten or RSS feeds. We "read" connectionless sentences, often for bare scatter, often on the search for pieces of news. We visit Webcams, we ergooglen ourselves also our picture worlds. We answer our enamels almost in real time. Often we live in such a way in the permanent diverson. All these phenomena experience the adults, if they became digital too "migrants", partially ambivalent. The young people usually know no life before E-Mail and Internet (except their childhood). For the "digitally native" the world of the digital net media is natural normality. In this contribution I set up the following theses: 1) Youth medium cultures: The new media play the crucial role in the today's youth culture absolutely. Also ranges such as music or mode or also social contacts became inconceivable beyond the new media. The today's youth is in by far stronger measure a medium youth, a completely medialisierte youth than in all earlier times. Internet became under all at present available medium channels the guidance medium, which current youth questionings prove. Thus the net medium becomes at all the crucial factor of the identity formation and cognitive development of the young people (apart from parents, school, friends and other media as for instance the television). 2) Agricultural engineerings: With the Internet as new guidance medium of the fully medialisierten youth today the classical agricultural engineerings (vintages, letter, counting) stand for arrangement. Those, which skypen today windshield frame sections, chatten, googeln, surfen or bloggen, are always already those, which also write and it reads (there is thus no strict opposition

[Personal Screenshot, 9/6/07]

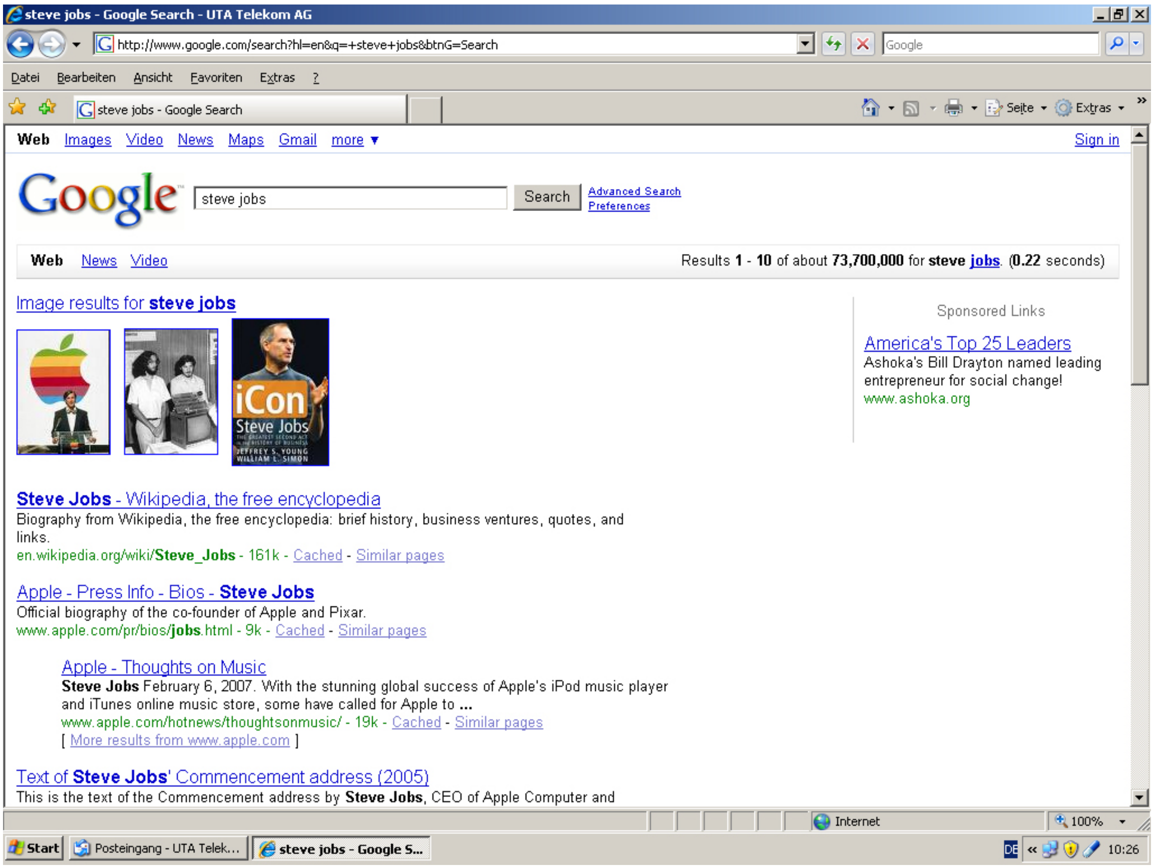
Please note that the marked term "ergooglen" (German for "googling") could not be translated by Google into English language! – For this experiment we just took a text from one of the authors of this study which was also published online. Just imagine how quick and efficient you can produce a scientific paper with doing Google-Copy-Paste and than translating and editing the text! With the new feature of the "Google notebook" you can easily copy text fragments from the Internet directly into a mask for further use – also this feature encourages plagiarism. – It remains one of the central problems that Google does not see the problem of plagiarism – or does not want to see it.

7. Google defining public opinion: A big problem of the reality construction by the Google results list is the fact that often blog entries or commentaries from web discussion groups can be found on top of the ranking. So Google is permanently biasing public opinion. For example if you type in the German word "Plagiatsjäger" ("plagiarism hunter") in German Google, the first entry is not one of the web sites of the most prominent plagiarism hunters in Germany and Austria (Debora Weber-Wulff in Berlin and Stefan Weber in Salzburg), but a blog entry with a rather unmotivated and stupid commentary on Stefan Weber: https://wwwu.edu.uni-klu.ac.at/drainer/wordpress/?p=8 [last test 28/8/07].

Not only Google, but in general the net gives rumours or personal attitudes of contingent people a new voice: by turning the surfer to the commentaries, blog entries and discussion boards of the net. If Google now starts to do their own edited discussion boards in connection with Google News, the whole reality construction of the world's news could change: Biasing the news becomes standard, and manipulations from private persons or private companies will become common. So "web literacy" will turn into a key competence to know how to deal with all these levels of information.

8. Google as the environment, the Internet itself: A few months ago Google announced "Universal Search" which means the integration of the so-far separated systems Google web search (text-based), image search, video search, groups search, and news search to one unified search list. But this development doesn't look that spectacular, it obviously seemed merely to be a marketing trick to gain news coverage.

Figure 44: Google announcing Universal Search – merely a marketing trick?


steve jobs - Google Search - UTA Telekom AG
http://www.google.com/search?hl=en&q=+steve+jobs&btnG=Search
Google
Datei Bearbeiten Ansicht Favoriten Extras ?
steve jobs - Google Search
Seite Extras
Web Images Video News Maps Gmail more ▾
Sign in
Google
steve jobs
Search
Advanced Search
Preferences
Web News Video
iCon Steve Jobs
Jeffrey S. Young William L. Simon
Internet
100%
Start
Posteingang - UTA Telek...
steve jobs - Google S...
DE
10:26

Results **1** - **10** of about **73,700,000** for **steve jobs**. (**0.22** seconds)

Image results for **steve jobs**

Sponsored Links

America's Top 25 Leaders
Ashoka's Bill Drayton named leading entrepreneur for social change!
www.ashoka.org

Steve Jobs - Wikipedia, the free encyclopedia
Biography from Wikipedia, the free encyclopedia: brief history, business ventures, quotes, and links.
en.wikipedia.org/wiki/**Steve_Jobs** - 161k - Cached - Similar pages

Apple - Press Info - Bios - **Steve Jobs**
Official biography of the co-founder of Apple and Pixar.
www.apple.com/pr/bios/**jobs**.html - 9k - Cached - Similar pages

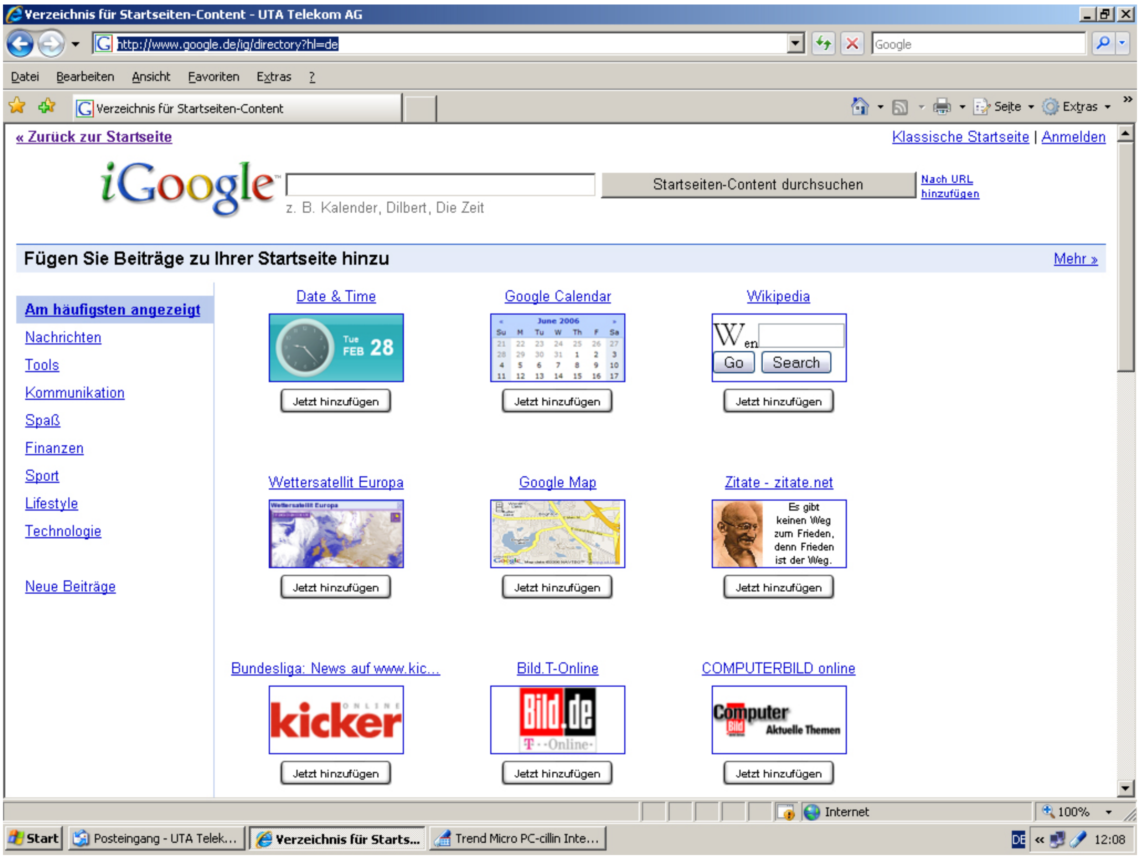
Apple - Thoughts on Music
Steve Jobs February 6, 2007. With the stunning global success of Apple's iPod music player and iTunes online music store, some have called for Apple to **...**
www.apple.com/hotnews/thoughtsonmusic/ - 19k - Cached - Similar pages
[More results from www.apple.com]

Text of **Steve Jobs**' Commencement address (2005)
This is the text of the Commencement address by **Steve Jobs**, CEO of Apple Computer and

[Personal Screenshot, 18/5/07]

With iGoogle, Google started to personalise the main page:

Figure 45: The personalisation of the Google main page with iGoogle

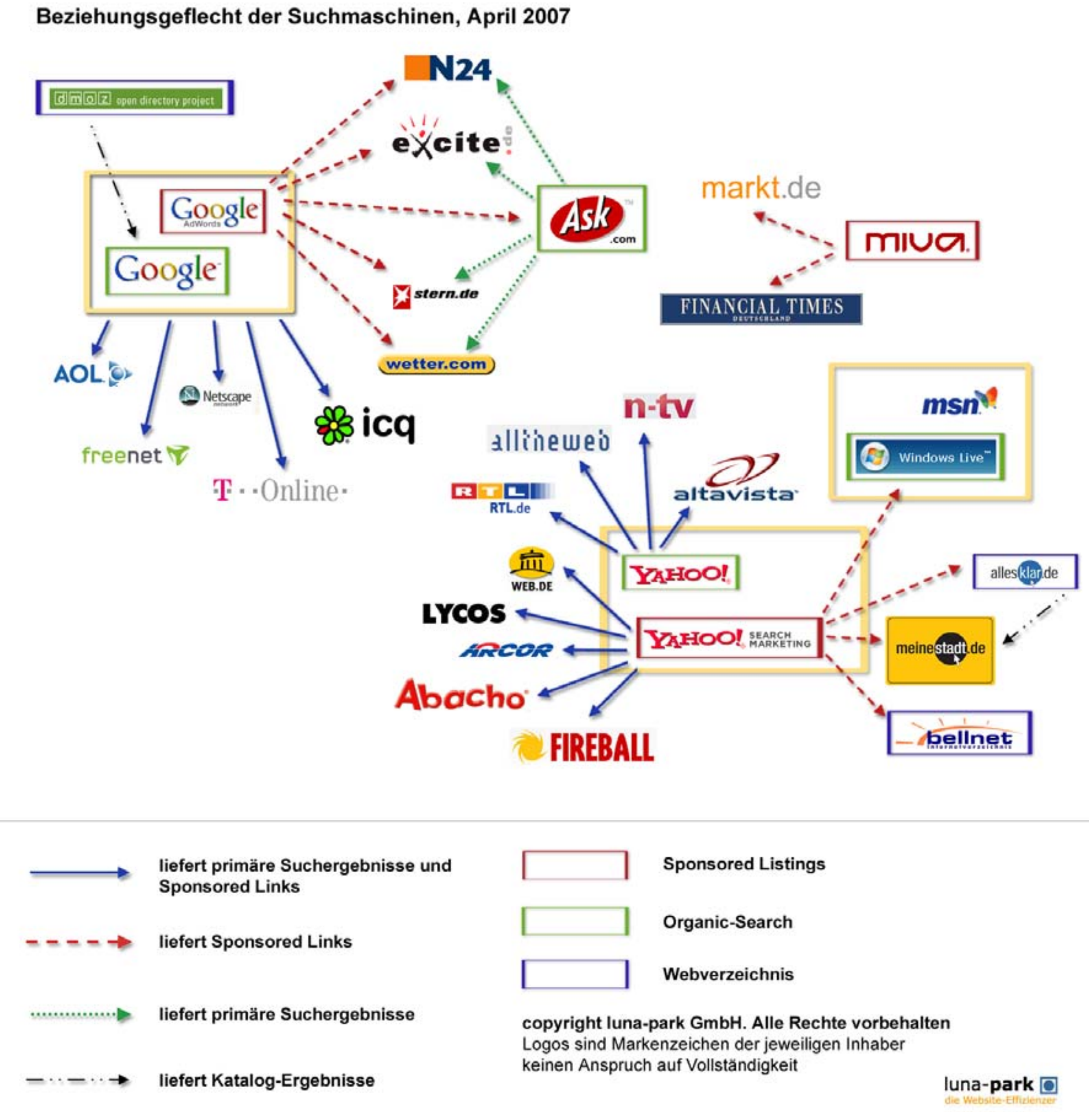

Verzeichnis für Startseiten-Content - UTA Telekom AG
http://www.google.de/ig/directory?hl=de
Google
Datei Bearbeiten Ansicht Favoriten Extras ?
Verzeichnis für Startseiten-Content
Seite Extras
« Zurück zur Startseite
Klassische Startseite | Anmelden
iGoogle
Startseiten-Content durchsuchen
Nach URL hinzufügen
z. B. Kalender, Dilbert, Die Zeit
Fügen Sie Beiträge zu Ihrer Startseite hinzu
Mehr »
Am häufigsten angezeigt
Nachrichten
Tools
Kommunikation
Spaß
Finanzen
Sport
Lifestyle
Technologie
Neue Beiträge
Date & Time
Tue FEB 28
Google Calendar
June 2006
Wikipedia
Go
Search
Wettersatellit Europa
Google Map
Zitate - zitate.net
Es gibt keinen Weg zum Frieden, denn Frieden ist der Weg.
Bundesliga: News auf www.kic...
kicker
Bild.T-Online
Bild.de
T-Online
COMPUTERBILD online
Computer Bild Aktuelle Themen
Jetzt hinzufügen
Internet
100%
Start
Posteingang - UTA Telek...
Verzeichnis für Starts...
Trend Micro PC-cillin Inte...
DE
12:08

[Personal Screenshot, 11/5/07]

With this feature, Google is already moving into the direction of the real new "Internet desktop".

Please also note in this context how many other search engines are already "powered by Google":

Figure 46: More and more on the Internet is "powered by Google"


Beziehungsgeflecht der Suchmaschinen, April 2007
dmoz open directory project
Google AdWords
Google
N24
excite.de
Ask.com
stern.de
wetter.com
AOL
Netscape
freenet
T-Online
icq
markt.de
miva
FINANCIAL TIMES DEUTSCHLAND
msn
Windows Live
n-tv
alltheweb
RTL.de
altavista
YAHOO!
YAHOO! SEARCH MARKETING
allesklar.de
meinestadt.de
bellnet
WEB.DE
LYCOS
ARCOR
Abacho
FIREBALL
liefert primäre Suchergebnisse und Sponsored Links
liefert Sponsored Links
liefert primäre Suchergebnisse
liefert Katalog-Ergebnisse
Sponsored Listings
Organic-Search
Webverzeichnis
copyright luna-park GmbH. Alle Rechte vorbehalten
Logos sind Markenzeichen der jeweiligen Inhaber
keinen Anspruch auf Vollständigkeit
luna-park
die Website-Effizienzer

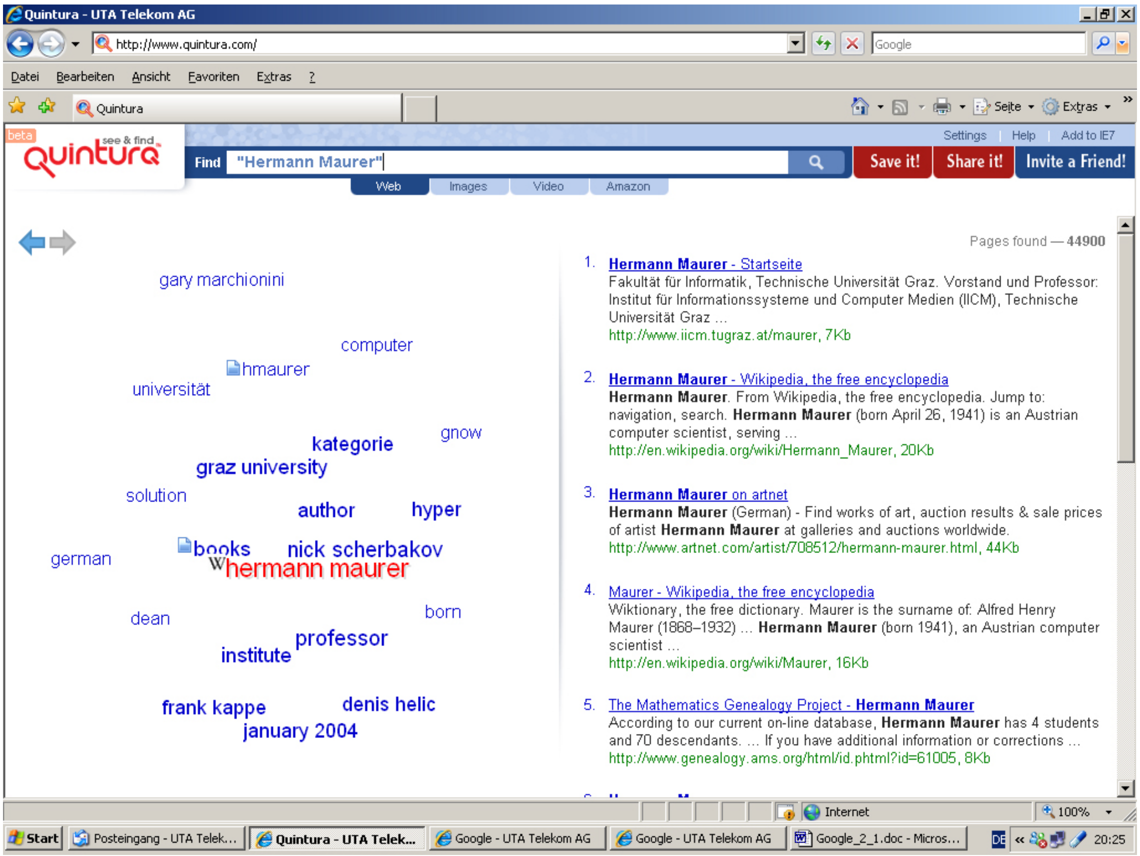
[Source: http://www.luna-park.de/fileadmin/lp/img/ beziehungsgeflecht_suchmaschinen_gross_0407.jpg, visited 28/8/07]

9. Google not only as the main Internet environment, but even more: With "Google Docs and Spreadsheets" (in German "Google Text und Tabellen") and "Picasa" Google for the first time gives a proof that they also want to become a player in the worldwide software market. Maybe the future of software is online at all: People could tend not to download new software updates on their stationary PC, but to use more and more often software just temporarily from the Internet. In this case Google once again would be a pioneer of a beginning development. But the problem is that Google also once again is becoming bigger: Google will not only constitute the new desktop of the Internet, but also the new desktop for every personal computer with various software applications.

If we think all dangers together and put all aspects into one big mosaic, there is a warranted fear that Google will desire world supremacy [see Reischl 2007]. There are several strategies to prevent or at least fight against this: Just think of the French search engine http://www.exalead.com/search which was invented to compete with Google; think of the German semantic search project "Theseus" (http://theseus-programm.de), think of several attempts for personal customized search engines where users can define and limit the scope of search by themselves [see Bager 2007], also think of alternative digitalisation and full text search initiatives of publishers to compete with Google Book Search [see N. N. 2006a].

One future of search engines could also be the representation of found web sites not in a hierarchical order, but in tag clouds as known from folksonomy. The acceptance of this kind of knowledge representation depends on the "graphical mind" of the Internet users. We typed in "Hermann Maurer" with http://www.quintura.com, a search engine which uses the results of other big search engines to visualise the findings in tag clouds. If you click on a tag, the hierarchical list on the right and also the cloud itself changes (it is becoming more concrete in the vertical dimension). – Here are the results:

Figure 47: Tag cloud of search results



Pages found — **44900**

1. **Hermann Maurer** - Startseite
Fakultät für Informatik, Technische Universität Graz. Vorstand und Professor: Institut für Informationssysteme und Computer Medien (IICM), Technische Universität Graz ...
http://www.iicm.tugraz.at/maurer, 7Kb

2. **Hermann Maurer** - Wikipedia, the free encyclopedia
Hermann Maurer. From Wikipedia, the free encyclopedia. Jump to: navigation, search. **Hermann Maurer** (born April 26, 1941) is an Austrian computer scientist, serving ...
http://en.wikipedia.org/wiki/Hermann_Maurer, 20Kb

3. **Hermann Maurer** on artnet
Hermann Maurer (German) - Find works of art, auction results & sale prices of artist **Hermann Maurer** at galleries and auctions worldwide.
http://www.artnet.com/artist/708512/hermann-maurer.html, 44Kb

4. Maurer - Wikipedia, the free encyclopedia
Wiktionary, the free dictionary. Maurer is the surname of: Alfred Henry Maurer (1868–1932) ... **Hermann Maurer** (born 1941), an Austrian computer scientist ...
http://en.wikipedia.org/wiki/Maurer, 16Kb

5. The Mathematics Genealogy Project - **Hermann Maurer**
According to our current on-line database, **Hermann Maurer** has 4 students and 70 descendants. ... If you have additional information or corrections ...
http://www.genealogy.ams.org/html/id.phtml?id=61005, 8Kb

[Personal Screenshot, 28/8/07]

The real way to fight the Google monopoly is to strengthen the research techniques beyond the cursorily Google search – as well as online (data bases, special scientific search engines) as offline (real books in the real library). This is a main task for educational institutions, such as schools and universities in the present and in the near future.

Cited References

[Bager 2007] Jo Bager. "Der persönliche Fahnder. Gezielt suchen mit individuellen Suchmaschinen". c't 1, 2007, pp. 178-183.

[Conti 2006] Gregory Conti. "Googling Considered Harmful". Paper from New Security Paradigms Workshop 06. http://www.rumint.org/gregconti/publications/20061101_NSPW_Googling_Conti_Final.pdf [visited 26/8/07]

[Emmer & Wolling 2007] Martin Emmer and Jens Wolling. "Was wir schon immer (lieber nicht) über die Informationswege und -quellen unserer Studierenden wissen wollten". Powerpoint document, presentation in Bamberg on 17/5/2007.

[Fallows 2004] Deborah Fallows a. o. "Data Memo: The popularity and importance of search engines". http://www.pewinternet.org/pdfs/PIP_Data_Memo_Searchengines.pdf [visited 4/8/07]

[Fallows 2005] Deborah Fallows. "The PEW Internet & American Life Project: Search Engine Users". http://www.pewinternet.org/pdfs/PIP_Searchengine_users.pdf [visited 4/8/07]

[N. N. 2006a] N. N. "Verlage feilen an Antwort auf Google". Financial Times Deutschland, 4/10/06, p. 8.

[N. N. 2006b] N. N. "Der YouTube-Coup. Google: Gut oder böse – wird der Web-Gigant zu mächtig?" FOCUS 42, 16/10/06, pp. 220-232.

[N. N. 2007a] N. N. "Google im Visier der EU-Datenschützer". Die Presse, 26/5/07, p. 7.

[N. N. 2007b] N. N. "Google News to enable comments". http://www.e-consultancy.com/news-blog/363981/google-news-to-enable-comments.html. News report from 9/8/07 [visited 26/8/07]

[Privacy International 2007] Privacy International. "A Race to the Bottom: Privacy Ranking of Internet Service Companies". Report from June 2007. http://www.privacyinternational.org/article.shtml?cmd%5B347%5D=x-347-553961 [visited 26/8/07]

[Reischl 2007] Gerald Reischl. "Google sucht die Weltherrschaft". Kurier, 25/3/07.

[Rolfes 2007] Christian Rolfes. Der Geschichtsunterricht und das "vernetzte Wissen" – Google und Wikipedia als Bedrohung oder Unterstützung? Hausarbeit, Universität Oldenburg [79 pages, copy received from the autor].

[Rötzer 2007] Florian Rötzer. "Google will ein bisschen weniger böse sein". Telepolis, 2007, online only. http://www.heise.de/tp/r4/artikel/24/24854/1.html [visited 26/8/07]

[Salchner 2007] Christa Salchner. "Nicht einmal über Google zu finden. Aspekte des Lebens im Google-Zeitalter." Telepolis, 2007, online only. http://www.heise.de/tp/r4/artikel/24/24720/1.html [visited 26/8/07]

Further References

[Kaumanns & Siegenheim 2007] Ralf Kaumanns and Veit Siegenheim. Die Google-Ökonomie. Wie Google die Wirtschaft verändert. Norderstedt: Books on Demand, 2007.

[Kuhlen 2005] Rainer Kuhlen. "Macht Google autonom? Zur Ambivalenz informationeller Autonomie". http://www.inf-wiss.uni-konstanz.de/People/RK/Publikationen2005/google-nachwort_final100205.pdf [visited 26/8/07]

[Kulathuramaiyer & Maurer 2007] Narayanan Kulathuramaiyer and Hermann Maurer. "Market Forces vs. Public Control of Basic Vital Services". To be published. http://www.iicm.tu-graz.ac.at/market_vs_public.doc [visited 26/8/07]

[Spiegel Special 2007] Spiegel Special "Leben 2.0. Wir sind das Netz. Wie das neue Internet die Gesellschaft verändert". 3/2007 [138 pages].

[Vise & Malseed 2006] David A. Vise and Mark Malseed. Die Google-Story. Hamburg: Murmann, 2006.

Section 10: Minimizing the Impact of Google on Plagiarism and IPR Violation tools.

As has been pointed out the attempts to compete with Google by establishing a European initiative look futile, at best, for the time being. Companies like Yahoo and Microsoft who invest billions of dollars to catch up with Google are making only little progress, so any attempt to directly compete directly with Google on the level of search engines without a new twist seem pointless. We will propose in Section 11 a potential way to reduce at least the importance of Google as search engine. In this chapter we will, however, claim that as far as plagiarism and IPR violation detection Google can be by-passed by a "collaborative" or "syndicated approach" as follows.

(Note: The rest of this Section is mainly based on work of Bilal Zaka with inputs from H. Maurer)

Empowering plagiarism detection with a web services enabled collaborative network

Abstract: This Section explains how collaborative efforts in terms of technology and content, can help improve plagiarism detection and prevention. It presents a web service oriented architecture, which utilizes the collective strength of various search engines, context matching algorithms and indexing contributed by users. The proposed framework is an open source tool, yet it is extremely efficient and effective in identifying plagiarism instances. By creatively using distributed processing capabilities of web services, this tool offers a comprehensive mechanism to identify pirated contents. With an aim to extend current plagiarism detection facilities, the proposed framework not only tries to reduce known deficiencies but also aims to provide plagiarism protection mechanism. The distributed indexing approach adapted in the system provides scalability to examine deep web resources. Network nodes with more focused indexing can help build domain specific information archives, providing means of context aware search for semantic analysis.

1. Introduction

Due to mass digitization and increasing use of digital libraries, scholarly contents are more vulnerable to plagiarism and copyright infringements. Surveys and studies conducted by various researchers [Plagiarism, 07] indicate that the use of contents without proper attribution to the original source is becoming widespread. The policies, consequences and penalties for plagiarism vary from institute to institute and case to case. Different prevention, detection and punishment methods are being practiced globally. Some institutes rely more on grooming and ethical motivation to fight the problem and some place more emphasis on a policing approach. However, the most effective way is a balanced combination of all methods.

Many institutes provide well documented information and rules dealing with academic misconduct and plagiarism during the enrolment phase. The information provision is made possible by means of brochures, web sites and training sessions to improve writing and communication skills. Honor codes are used to add moral bindings. Departments and even teachers on an individual level are undertaking efforts to educate their students, research staff and faculty. Tutorials and guides are available to explain plagiarism, citation rules, and writing standards. The other aspect of plagiarism prevention is detection and penalization. The penalties for plagiarism start from warnings, grade reductions, failing grades and can end up in suspension, expulsion or even revoking of title or degree.

Detection of plagiarism is done using a number of techniques. These techniques include stylometric analysis of writing, manual search of characteristic terms in writing and use of automation tools to compare documents for similarities, within a local corpus or across the internet. It is becoming a common practice to use software and services that automate the processes of plagiarism detection. The majority of these applications are based on the document source comparison method. The detection process in such programs generally starts with submission of the suspected document to the system via a desktop application or web based form. The document is converted into plain text format removing any formatting information, images etc. The text information is broken down into moderately sized segments (referred to as fingerprints). The fingerprints are compared for similarity with index of available documents. Finally a report is presented highlighting the matched sources and copy percentage. The comparison index can be a corpus of local documents processed in similar way, or it may include the public/private internet index.

As mentioned before, the ease with which digitized contents can be accessed accounts for the rise in plagiarism. Without a doubt, the ease of content availability is an attribution of internet usage. Naturally the most popular tools used to detect plagiarism are also built on the idea of efficiently checking for document source availability over the internet. The commercial services claim to use personalized crawlers and up-to-date internet indexes for a comprehensive check. Over the years these programs and services have indeed proven their effectiveness in educational and industrial environments. However, there is still room for considerable improvements. A recent survey on plagiarism [Maurer et al. 06] is a good starting point for a better understanding of various plagiarism detection strategies and strengths/weaknesses of available tools. Experimental results in the referenced survey suggest that in order to have a more precise plagiarism detection tool, the inspection system requires broader and an up-to-date content index, added semantic elements for similarity check, cross language content similarity detection and finally a mechanism to verify the findings. Existing tools following either desktop applications or software as a service approach lack these capabilities. Albert Einstein once said “The secret to creativity is knowing how to hide your sources", and yes, plagiarists today are more creative. Copied contents are often not publicly available or modified in a way which is hard to detect using existing applications and approach. Further experiments to benchmark capabilities of popular plagiarism detection services revealed that intelligent use of good search engines can greatly add value to plagiarism detection applications [Maurer & Zaka, 07].

As an attempt to fulfil the needed requirements in plagiarism detection systems, collaborative service oriented architecture for plagiarism detection is presented. The proposed service oriented collaborative network openly available to educational community aims at extending the existing similarity check methods in the following ways:

It offers a seamless, combined use of multiple search services. This technique provides a broader and more context aware internet search, which proves to be more revealing than any single searching and querying approach.

Collaborative authoring and indexing of document sources at each node enhances the search capabilities with addition of documents not available publicly. This also provides users an option to add intellectual contents for protection against copyright infringements. Participating institutes allow access to deep web, hidden from normal search engines.

The system provides multiple services for search result analysis. More services can be added to the system due to its modular nature. The user has an option to use mere text matching to deduce similarity or can apply writing structure analysis, semantic or cross language analysis.

The system offers the possibility of synonym normalization and translation in collaborative search service and peer node indexes. This adds semantic matching capabilities not possible in conventional internet searches.

This system makes use of off-the-shelf tools (web services) and user contributed contents to extend plagiarism detection. Its pluggable services constitute composite web applications offering flexibility and variety in use.

Having described the basic idea behind the service oriented collaborative plagiarism detection network, the following section describes the conceptual design of the system. Section 3 describes a practical realization of the architecture and compares results of the prototype with other services. Section 4 presents future directions of work. Section 5 examines the potential of the presented approach and concludes the paper.

2. Concepts behind service oriented collaborative architecture

Service Oriented Architecture (SOA) can be described as a heterogeneous environment of applications with self describing and open components which allow inter application communication. SOA offers distributed and collaborative computing infrastructure over the network or internet. A research study for the future of flexible software [Bennet et al. 00] provides a vision of personalized, self adapting and distributed software environment. The software is structured in small simple units which co-operate through rich communication structures. The collaborative units work in a transparent way to provide a single abstract computing environment. The study shows interdisciplinary approach would be critical to developing a future vision of software. A significant proportion of software and associated data does not exist in isolation but in a political, social, economic and legal context. In order to have applications with high level of productivity and quality, it is essential that they don't have rigid boundaries but offer rich interaction and interlinking mechanisms with users as well as other applications.

The service oriented approach has been in use for almost a decade and adds the aforementioned functionalities in software systems. These integration technologies exist in the form of Component Object Model (COM), Distributed Component Object Model (DCOM), Enterprise JavaBeans (EJB) and Common Object Request Broker Architecture (CORBA). However what really boosted the concept recently is the emergence of the next generation of SOA based on "web services". Web services are built using standardized and platform independent protocols based on XML. The service components enable us to build a user-tailored, distributed and collaborative web application environment. A framework built on top of web services will offer the extension and flexibility to plagiarism detection as described in the introductory part.

2.1 Web service model

"A Web service is a software system designed to support interoperable machine-to-machine interaction over a network. It has an interface described in a machine-process able format. Other systems interact with the Web service in a manner prescribed by its description using SOAP3 messages, typically conveyed using HTTP with an XML serialization in conjunction with other Web-related standards" [W3C, 04]. A typical web service can be described using three components

i. Description: (XML based service description, specifically WSDL4)
ii. Publishing and Discovering (Registry, index or peer-to-peer approach of locating services, e.g. UDDI5)
iii. Messaging (XML based message exchange over the network, specifically SOAP or REST6)

The proposed collaborative plagiarism detection framework consists of composite web applications to search the internet and shared document sources. These network distributed applications use a set of web services for searching and sharing documents.

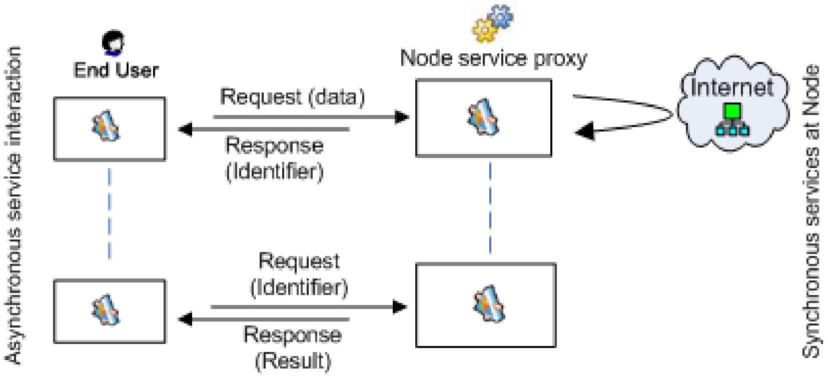
Web service interaction can be either synchronous or asynchronous. Commonly available and popular internet search web service APIs use synchronous request/response communications. This approach works well in limited use environments where the web service can process a request in quickly. However, in plagiarism detection, search normally requires exploring the internet for a large number of queries (moderate size finger prints of a document) or browsing through document signatures from a number of distributed nodes. In this scenario using asynchronous service interaction for the user is the better solution.

The proposed framework consists of a service proxy that enables asynchronous use of synchronous internet search APIs. The time independent interaction model (asynchronous) is implemented using multiple synchronous request/response web services. The first service initiates processing from the end user by sending the information parameters. The service sets an identifier of the submitted job and

[3] Simple Object Access Protocol, http://www.w3.org/TR/soap
[4] Web Services Description Language, http://www.w3.org/TR/wsdl
[5] Universal Description Discovery and Integration, http://www.uddi.org/
[6] REST: http://en.wikipedia.org/wiki/Representational_State_Transfer

responds to the end user with same. The end user can then use the second service and the identifier as a parameter to check if the submitted job is complete, pending or failed. [Hogg et al. 04] The first request in asynchronous communication mode validates and acts as a buffer between the end user and the synchronous internet search service. The submitted job is processed using search and analysis services at the respective network node. The similarity detection results are stored and the job identifier status is updated for later reference of the end user. Figure 48 shows the described service model.

Figure 48: Asynchronous service interaction model in framework



2.2 Mashup of search and analysis web services

One of the major strengths of the system is the next generation search capabilities termed “Search 2.0” by Ezzy [Search2.0, 06]. It is defined as a search "designed to combine the scalability of existing internet search engines with new and improved relevancy models; they bring into the equation user preferences, collaboration, collective intelligence, a rich user experience, and many other specialized capabilities that make information more productive" [Search2.0, 06]. In the concept described here, users are given the option to select a number of system compatible internet & collaborative search services. The search results are processed and passed through further analysis algorithms in order to detect content and context similarities. Combining the strengths and scalability of existing internet search engines broadens the web search scope compared to searching via a single source. Further mashup with collaborative search API built using full text query mechanism on user contributed finger print data and local node resources greatly add to value. The collective search is not conventional meta-search where the user might have to weed through irrelevant matches. The initial result set lists the possible matches by each search service. Analysis services applied to search results produce precise and productive output for the final report.

The system has been tested using a few popular search services. The results of our experiments presented in a later section indicate that using the search services systematically can detect cut paste plagiarism more effectively then any other commercial plagiarism detection service. This is mainly because of recent open access and indexing initiatives by publishers. More and more options are becoming available to do full text search on digital libraries via a particular search engine or a library’s own search mechanism. One significant example of such an initiative is Crossref search pilot. A group of 45 leading journal publishers including ACM, IEEE, Blackwell, Springer, Oxford University press and John Wiley & Sons, are providing full text search options using Google via Crossref gateway [CrossRef, 07]. A plagiarism detection system with up-to-date search capabilities can outperform similar tools of its class. The proposed service oriented approach gives its user an option to integrate any existing search service and any upcoming more powerful search service.

The initial prototype includes an analysis services based on the vector space model [Wikipedia:VSM, 07] approach to measure cosine similarity. The queried text and search engine’s returned matching snippet are converted to word vectors, based upon the vocabulary of both. The angular measure (dot product) of vectors is used as a score to determine similarity between the queried text and any searched result. The combination of the highest similarity scores of the queried text segments represents the percentage of plagiarism in a document. There is a number of other possibilities for similarity analysis within a document or with the search service's detected contents. One such analysis approach tested for the proposed framework involves a structural comparison of suspected documents. This statistical analysis service uses a measure of standard deviation in the document structures (lines, words) to determine a possible match.

Another analysis planned to be part of the framework is stylometric analysis based on Jill Farringdon's CUSUM (cumulative sum) technique [Farringdon, 96]. The CUSUM technique is based on the assumption that every person has some quantifiable writing or speaking habits. The measure of consistency of these habits can be used to determine single or multiple authorships. The numerous factors which determine authorship include checking of sentence length consistencies, checking the use of function words, nouns and other common language practise throughout the document. This technique is used by courts in England, Ireland and Australia to determine authenticity of writings in different cases such as witness statements, suicide notes, ransom notes and copy right disputes. Although this technique may not be considered very effective, especially in the case of multiple authors, it can be beneficial in pointing out any suspicious portion in the text coming from a single author. The suspected parts can then be checked by other more extensive search services. Future research which could be conducted on the system also includes the development of semantic analysis service that uses language ontology. The idea is further described in section 4.

2.3 Collaborative authoring, indexing & searching – Access into the deep web

The ability of collaborative authoring and indexing at participating institute nodes of network is an important feature in extending plagiarism checks. The motive behind collaborative indexing and search approach is the fact that conventional search engines only index the shallow internet contents and do not cover deep web contents. Shallow contents are generally static web pages linked with each other and openly available to search engine spiders. However the deep web consists of unlinked or dynamically generated pages, databases, protected sites, intranets and contents behind firewalls. These contents are invisible to the index of general internet search engines. A study by BrightPlanet in 2001 estimated that the deep web information is 500 times larger than the commonly defined World Wide Web [Bergman, 01]. It is very important to have access to this massive information base for thorough plagiarism checks. Educational institutes usually have a very large collection of un-linked and non-indexed local contents. Institutes and research groups within an institute also have privileged access to, and better knowledge of specific deep web resources. This access and knowledge enables them to gather resources not commonly available. Collaborative networking provides the means of creating a local searchable index of these resources. Any network node run by an institute can setup a search gateway service providing access to its invisible contents and can access to protected digital libraries. A collaborative search API consumes the local search services according to the access policy of each peer node. The collaborative search access produces limited results usable for similarity analysis services. The search results may only contain specific matching portion and associated meta information. A local index can be restricted to a specific domain e.g. an institute specializing in computer science articles. Collaborative searches can be made context aware by selecting domain specific peer indexes of the deep web. This means that in addition to general internet search services; the proposed system also use collaborative search service which harnesses the deep web contents of participating nodes.

The collaborative search channel is also important in terms of reducing the dependency of certain search engines. Researchers have shown concern in recent studies that the search engine monopoly gives them the role of gatekeeper in the information flow. A search engine can determine what is findable and what is kept outside the view of the common user [Kulathuramaiyer & Balke, 06]. The view restriction or any other implication a search engine may apply or is applying can be seen in the form of web search API switching from Google. Shifting from an XML standard and generic SOAP based access to a more restraining AJAX API is not seen as a welcome move by many research and development forums. It is thus imperative to have an alternate and more open channel of searching the intellectual content base.

System users can contribute documents to the associated network node for indexing. User contributed content authoring is done either by conventional indexing and making complete documents openly available. Or by generating moderately sized searchable plain text snippets of submitted document (called fingerprints or signatures in more abstract form). In the case of a search match, only a content snippet and meta information are sent from the index, not the complete document. Any matches found in such snippets point to the original source for further verification. Authoring resources can be tempting for a user or node administrator, because of following reasons

Contributing resources can expose the contents to all network search APIs in a protective manner. This approach helps where users cannot index complete contents in a finished formatting for the public internet.

User contributed authoring acts as a “personal copyright tool” which protects against any possible piracy of articles, personal blogs, assignments, presentations etc. Submitted documents can be indexed with the author’s meta information. Users or node administrators may choose to receive alerts produced by any similarity matches from other sources in the future. This can help authors keep track of legal or illegal use of their contents.

The envisioned framework in its mature form is based on P2P incentive based resource access scheme. Users and nodes with a higher index of shared resources will receive better access to local resources of peer nodes.

2.4 Service publication, discovery and access mechanism

Web services for end users are available as selectable index of compatible searching APIs. No technical details or WSDL is required at the end user level. User can choose any service by simply selecting or providing personal usage credentials e.g. API code or key. Master nodes keep a well descriptive index of available services to share and search. The system administrator of each node can incorporate the available services on a specific node and make them available to the end user. The local document source (collaboratively authored) sharing services at each node uses either an open access policy or implements restrictions on access. Peer nodes may contribute more search services and sample service consuming codes to master service index. Initial implementation uses a plain index approach and open access policy at selected test nodes. Later stages of the project include a more controlled central registry to maintain service descriptions and access mechanisms.

3. Implementation

Based on the abstract architecture which is described in the previous section, a partial implementation is developed as a proof of concept. The prototype serves as a valuable tool to benchmark and test the search engine capabilities, the match detection algorithms and the document source generation. Initial experiments show very promising results closely comparable (better in some cases) to already existing commercial services which detect plagiarism. Prototype named CPDNet7 (Collaborative Plagiarism Detection Network) is available for test purposes, although it is an early stage of development. Users may register for an account with their personal Search API code to test drive the system. CPDNet currently supports Google SOAP search API8, and Microsoft Live Search API9. The server and client for web services are created using PHP SOAP and AJAX technologies. Running nodes can choose any available indexing server to link local contents with collaborative search. Existing CPDNet nodes use Lucene10, an open source Java based indexing and search technology. Result sets are generated in OpenSearch11 standard. The collaborative search client uses a SOAP interface to discover matches from all available service nodes.

[7] Collaborative Plagiarism Detection Network: http://www.cpdnet.org

[8] Google SOAP Search API: http://code.google.com/apis/soapsearch/

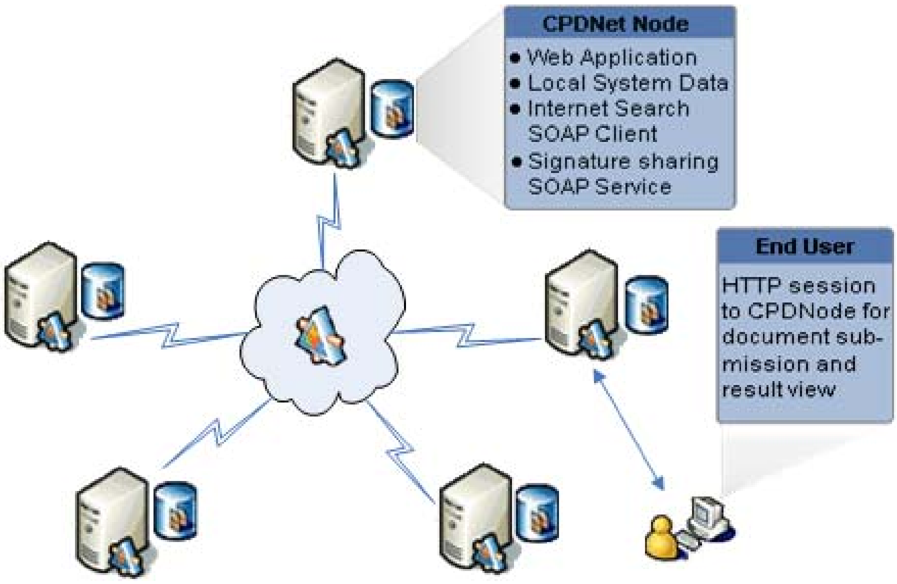
[9] Live Search SOAP API: http://dev.live.com/livesearch/

[10] Lucene Java: http://lucene.apache.org/

[11] OpenSearch: http://opensearch.a9.com/

Figure 49: Collaborative Plagiarism Detection Network, System overview

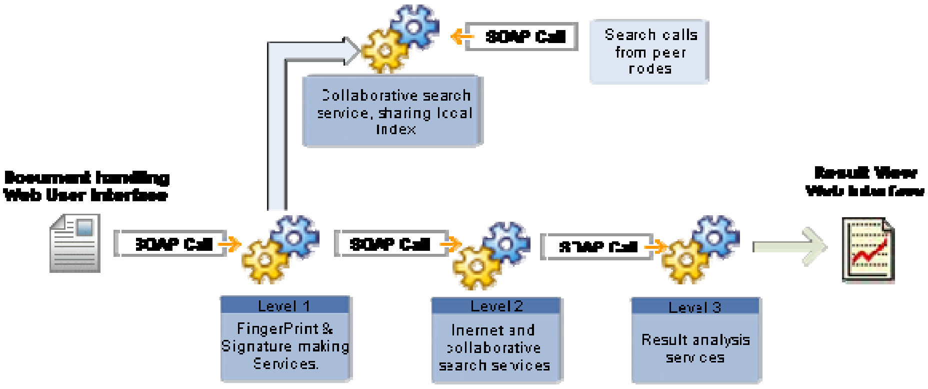
The process of detecting plagiarism includes the following steps


CPDNet Node
• Web Application
• Local System Data
• Internet Search SOAP Client
• Signature sharing SOAP Service
End User
HTTP session to CPDNode for document sub-mission and result view

The submitted document is broken down into moderately sized text chunks also called fingerprints. This document source can also be marked for indexing in the local database, accessible via a collaborative search API.
The plagiarism check is initiated by querying the internet using the fingerprint data. The selected search APIs searche the web. Locally indexing document sources (signature in more abstract form) and that of peer nodes can also be queried if collaborative search service is enabled.
Most relevant matches obtained via the search services are passed to similarity analysis service. The existing active service uses word vector based similarity detection as described earlier.
Fingerprint similarity scores of a document are calculated to determine the plagiarism percentage. The document text linked with the similarity scores, matching contents and source links, is presented to the user within a final report.

The described process steps are visualized in figure 50. Here you can see the various services used in the system.

Figure 50 : Web Service flow in CPDNet


SOAP Call
Search calls from peer nodes
Collaborative search service, sharing local index
Document handling Web User Interface
SOAP Call
SOAP Call
SOAP Call
Result View Web Interface
Level 1
FingerPrint & Signature making Services.
Level 2
Inernet and collaborative search services
Level 3
Result analysis services

The architecture is flexible to accommodate numerous services at each level. The running services in the current prototype can be further explored at the project portal12.
Following request and response SOAP messages show analysis service communication. The request message contains the source (fingerprint) being checked and matching source (snippet) found via search service. The response contains the calculated percentage of similarity.

Request:

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<SOAP-ENV:Envelope SOAP-
ENV:encodingStyle="http://schemas.xmlsoap.org/soap/encoding/"
xmlns:SOAP-ENV="http://schemas.xmlsoap.org/soap/envelope/"
xmlns:xsd="http://www.w3.org/2001/XMLSchema"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xmlns:SOAP-
ENC="http://schemas.xmlsoap.org/soap/encoding/"
xmlns:ns2114="Array">
<SOAP-ENV:Body>
<ns2114:CalculateSimilarity xmlns:ns2114="Array">
<source1 xsi:type="xsd:string">The VoiceXML grammar specification
provides 2 text formats for writing speech recognition grammars XML
or ABNF. XML is Web</source1>
<source2 xsi:type="xsd:string">The VoiceXML 2.0 grammar
specification provides two text formats for writing speech
recognition grammars: XML or ABNF. XML is a Web standard for
representing structured data.</source2>
</ns2114:CalculateSimilarity>
</SOAP-ENV:Body>
</SOAP-ENV:Envelope>
```

Response:

<?xml version="1.0" encoding="ISO-8859-1"?>

```
<SOAP-ENV:Envelope SOAP-
ENV:encodingStyle="http://schemas.xmlsoap.org/soap/encoding/"
xmlns:SOAP-ENV="http://schemas.xmlsoap.org/soap/envelope/"
xmlns:xsd="http://www.w3.org/2001/XMLSchema"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xmlns:SOAP-
ENC="http://schemas.xmlsoap.org/soap/encoding/">
<SOAP-ENV:Body><ns1:CalculateSimilarityResponse xmlns:ns1="Array">
<res xsi:type="xsd:string">86</res>
</ns1:CalculateSimilarityResponse>
</SOAP-ENV:Body>
</SOAP-ENV:Envelope>
```

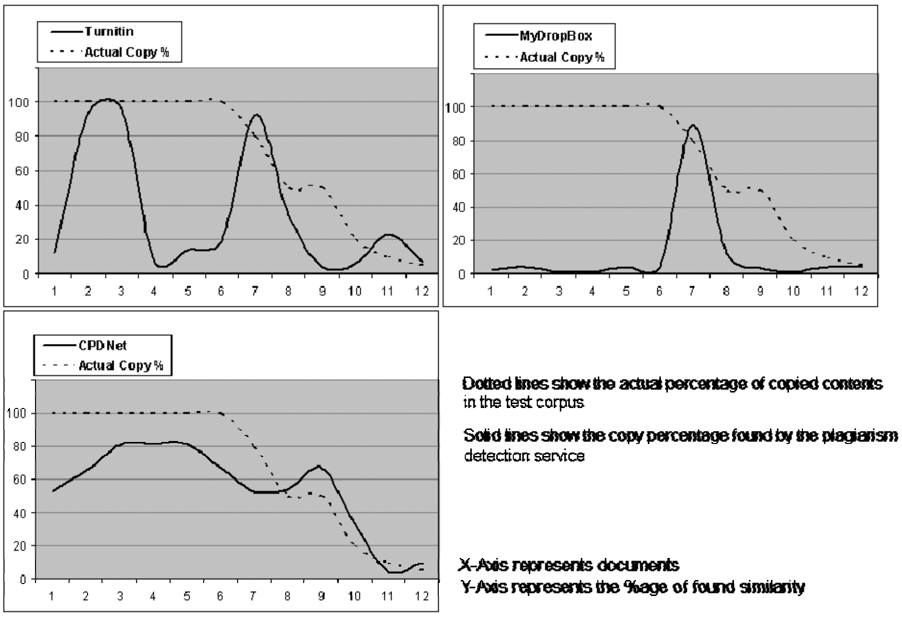
A sample message exchange of the similarity analysis service, Online at http://fiicmpc140.tu-graz.ac.at/cpdnet/webservice/service_wva.php?wsdl

3.1 Results of prototype

In order to benchmark the system, a document corpus is generated with various proportions of plagiarized contents from both deep and shallow internet. Test results from the selected corpus show significantly better similarity detection capabilities of the system then any other service. The graphs shown in figure 51 compare the plagiarism detection capabilities of CPDNet with other two leading plagiarism detection services.

[12] http://www.cpdnet.org/nservices.php?

Figure 51 : Comparison of search and similarity detection capabilities



Dotted lines show the actual percentage of copied contents in the test corpus

Solid lines show the copy percentage found by the plagiarism detection service

X-Axis represents documents
Y-Axis represents the %age of found similarity

Better plagiarism detection by the system developed to date, is due to the enhanced searching capabilities added to the system.

4. *Towards a semantic plagiarism detection service*

To trace paraphrased and cross language plagiarism, algorithms are required to discover similarities on the semantic level. This kind of analysis requires detecting similar word replacement (synonymizing), word deletion, word addition and translation etc. The application of these checks on a large scale with conventional internet indexes and current search APIs seems far fetched and computationally very expensive. However, the proposed framework provides a mean of indexing submitted contents in a normalized form. The normalized contents which are shared at each node can be queried using a collaborative search API of peer nodes. The queried finger print is also normalized in order to determine its conceptual equivalence. The semantic level similarity check can certainly help its users in more then just plagiarism detection. The findings can also be used by knowledge workers to discover relevant contents already available on internet. In the near future, the focus of this project's research will include following:

4.1 Fingerprint normalization into generic signatures

The development and introduction of signature generation (semantic fingerprints) service for the system. The submitted contents will be normalized to a root level, with the use of WordNet13 synonym sets. In order to develop a practical service, benchmarking is required. This will help determine the following:

i. The best suited Part of Speech (POS) tagger that identifies sentence components.

ii. The response of the search system with the normalized verbs and/or nouns in the initial stages as well as inclusion of function words (words with little lexical meanings such as articles, pronouns, conjunctions etc.) at later stages.

[13] WordNet: An Electronic Lexical Database; http://wordnet.princeton.edu/

The collaborative search API in semantic mode will normalize the text being checked and search for matches in local and peer indexes. A fingerprint normalization service based on Wortschatz14 API is currently being tested internally for a later integration with the system.

4.2 Introduction of translation and normalized signature search service

In order to check plagiarism across language barrier, another service at a different abstraction layer is required. This must translate the normalized indexes and queries into standard English. Each node can provide translation into and from a specific language depending on the local resources. This service will compliment the normalization on a more global and conceptual level. Such abstraction may produce undesired and false results at some levels. However it is worth experimenting with the cross language similarity checks, because of the large availability of intellectual contents in non-English languages.

4.3 Noise reduction in plagiarism detection with domain specific searches and citation checks

Similarity detection on an abstract level may introduce unnecessary noise in generated matches. It would be helpful to restrict the semantic level search and analysis to a domain specific index. Subject specific searching capability will be introduced by means of …

i. Setting up specialized indexes of certain domains. The participating institute's local index can be categorized based on various departments or research groups.
ii. Using topic maps to categorize subject domains and grammar to link contextual queries.
iii. Introducing a service before performing search that determines the context of the document being analyzed. One such example is the use of Yahoo Term Extraction service. This service provides its users the context aware relevance technology behind Y!Q [Yahoo:TermExtraction, 07]

Another level of service is required to decrease false positives while detecting plagiarism. Some plagiarism detection services give their users the option of ignoring texts found within quotation. This approach however is not sufficient in determining proper citations. There is a need to automatically compare the referenced articles and remove any plagiarism score coming from these sources. Such automation can be achieved by scanning through the referenced articles and creating an index of referenced resources in a document. The user can then choose to ignore the similarity matches which are available in the reference index. The automated detection of proper attribution or citation in a document will save the examiner both time and effort. Developing such a reference index may require a persistent and universally accessible resource identifier associated with the citation. The increasing web publication trend and the emergence of common linking and resource identification standards like DOI [Warren, 05] are encouraging factors which will lead to further investigations in this area.

5. Conclusions

A collaborative web service oriented architecture substantially extends current plagiarism detection systems. With flexible and extendable services, rich web user interface, standardized XML based inter application communication and collaborative authoring, it brings us a step closer towards Web 2.0 applications. A survey [Maurer et al. 06] pointed out that existing plagiarism detection systems fail in the following areas :

(1) "When systematic attempts are made to combat plagiarism tools by e.g. applying extensive paraphrasing through synonymizing tools, syntactic variations or different expressions for same contents.
(2) When plagiarism is based on documents that are not available electronically or archive is not available to detection tool.
(3) When plagiarism crosses language boundaries."

[14] Wortschatz, Universität Leipzig; http://wortschatz.uni-leipzig.de/

Based on experimental results, it can be safely stated that the platform presented addresses the second issue effectively. This is due to the additional support of internet searching API mashup and the collaborative indexing approach. Moreover, the availability of various analysis services, such as vector space similarity, structural evaluation of suspicious documents and fingerprint normalization in the system is an attempt to handle issues 1 and 3. The technology industry has a rapidly growing interest in web services. Many companies and service providers already have web service components available with their applications. Almost every software and internet organization focuses on web services as a core element in future strategies. This tendency suggests that the proposed web services enabled platform is best suited to carry out multiphase plagiarism detection. It will offer the flexibility to incorporate any new processing, discovery or indexing components that may become available to its users. The user-centred collaborative nature of this system makes it an ideal choice to build specialized indexes which are capable of handling semantic considerations in the similarity detection process.

References

[Bennet et al. 00] Bennett, K., Layzell, P., Budgen, D., Brereton, P., Macaulay, L., and Munro, M. “Service-based software: the future for flexible software” In Proceedings of the Seventh Asia-Pacific Software Engineering Conference (December 05 - 08, 2000). APSEC. IEEE Computer Society, Washington, DC, 214.

[Bergman, 01] Bergman, Michael K., “The deep web: Surfacing hidden value” The Journal of Electronic Publishing. Michigan University Press. July 2001; Online at http://www.press.umich.edu/jep/07-01/bergman.html (Accessed April 07, 2007)

[CrossRef, 07] Crossref Search: Online at http://www.crossref.org/crossrefsearch.html (Accessed April 04, 2007)

[Farringdon, 96] Farringdon Jill M. with contributions by A. Q. Morton, M. G. Farringdon and M. D. Baker; “Analysing for Authorship: A Guide to the Cusum Technique” University of Wales Press, Cardiff, 1996. ISBN 0-7083-1324-8

[Hogg et al. 04] Hogg, K., Chilcott, P., Nolan, M., and Srinivasan, B. 2004. “An evaluation of Web services in the design of a B2B application” In Proceedings of the 27th Australasian Conference on Computer Science - Volume 26 (Dunedin, New Zealand). Estivill-Castro, Ed. ACM International Conference Proceeding Series, vol. 56. Australian Computer Society, Darlinghurst, Australia, 331-340.

[Kulathuramaiyer & Balke, 06] Kulathuramaiyer, N., Balke, Wolf-T., “Restricting the View and Connecting the Dots – Dangers of a Web Search Engine Monopoly” Journal of Universal Computer Science, vol. 12, no. 12 (2006), 1731-1740

[Maurer et al. 06] Maurer, H., Kappe, F., Zaka, B. “Plagiarism- a Survey”. Journal of Universal Computer Science 12, 8, 1050-1084.

[Maurer & Zaka, 07] Maurer, H., Zaka, B. “Plagiarism – a problem and how to fight it” To appear in proceedings of ED-MEDIA 07, Vancouver, Canada, June 25-28

[Plagiarism, 07] Plagiarism.org: Statistics. Online at http://www.plagiarism.org/plagiarism_stats.html (Accessed March 15, 2007)

[Search2.0, 06] Search 2.0 vs. Traditional Search: Written by Ebrahim Ezzy and edited by Richard MacManus. / July 20, 2006 Online at http://www.readwriteweb.com/archives/search_20_vs_tr.php (Accessed March 14 2007)

[Warren, 05] Warren Scott A., “DOIs and Deeplinked E-Reserves: Innovative Links for the Future” Technical Services Quarterly, Vol. 22, number 4, 2005. DOI: 10.1300/Jl24v22n04_01

[Wikipedia:VSM, 07] Vector Space Model, In Wikipedia, The Free Encyclopedia. Online at http://en.wikipedia.org/w/index.php?title=Vector_space_model&oldid=113611338 (Accessed March 15, 2007).

[W3C, 04] W3C: Web Services Architecture, W3C Working Group Note 11 February 2004. Online at http://www.w3.org/TR/ws-arch/ (Accessed March 13, 2007)

[Yahoo:TermExtraction, 07] Yahoo Content Analysis Web Services: Term Extraction, Online at http://developer.yahoo.com/search/content/V1/termExtraction.html (Accessed April 10, 2007)

Section 11: Reducing the influence of Google and other large commercial search engines by building 50- 100 specilized ones in Europe.

(Note: This Section is based on an expose by H. Maurer)

Summary:

Google is in the process of compiling a full dossier on many people and on important economic issues. The first has fatal consequences for privacy and gives Google a handle on being the major power for distributing very individualized ads (if not products), the second has far reaching consequences for economy: Google is starting to be able to predict how markets will change, hence able to use e.g. the stock market for systematic financial gain.

As has been made clear in previous Google has to be stopped. This is not possible with a head-on strategy. Google is already too powerful for this. However, Google's effect can be minimized by introducing very powerful special-purpose search engines, better in their area of application than Google is. While the first few dozen of those will have to be supported by public funds, the following ones will hopefully finance themselves, once their potential becomes apparent. Also, since such specialized search engines are the only way to combat plagiarism and IPR violations, money can be generated by providing services for plagiarism and IPR violation detection. This will support more and more special purpose search engines, distributed all over Europe. Once enough are available, a "hub" tying them together will be able to compete better and better against Google or other commercial search engines at least in parts of the Web: those parts can be overlapping: search engines for some areas of science, others for SMEs , others for the aging, some even geographical oriented ("all that you ever want to find in Berlin").

Extended Summary: Killing two birds with one stone

Point 1: Plagiarism and IPR rights violations are rampant due to copy-and-paste from WWW and DBs

- Endangers research and scientific publications
- Financial loss for companies and organisations
- Detection tools (Turnitin, Mydropbox, etc.) fail for five reasons:
- Not stable against synonyms
- Not designed to detect translations
- Limited access to paid publications
- Material often available not as text but in image form
- The best search engine Google does NOT cooperate
- Additional problem: all plagiarism detection usually limited to textual material (some applicable to computer programs, but all fail for formulae, pictures, ideas, …)

Point 2: Big search engines are too powerful and starting to misuse their power, Google at the moment as the most glaring example

- Google does not allow to use their API or such for professional applications (like plagiarism detection, searches in patent archives, …)
- Google has given up to only use objective ranking (based on link structure) and is now putting on top of this opportunistic considerations. Since we are "googeling our reality" (Stefan Weber) our reality starts to be distorted!

- Google is analyzing billions of queries (and connects this with the use of Gmail, YouTube, Fileservices, Google Earth, etc): Google has a huge profile on persons and issues and knows about many economic developments
- Google is doing nothing illegal, any other company would have to act the same way, due to corporate laws.

Solution: Develop search engines that are very powerful, but each specialized to a restricted domain. Such engines will be welcome by corporations, industry and SME's. They can also support a (distributed) European Center for Plagiarism and IPR Violation Detection (ECPIRD[15]) attached to non-profit organizations (like universities or commercially independent research centers). One of the centers will provide a 'hub' for the others to allow a single input to be searched in deep way on a variety of specialized search engines.

Point 3: How does ECPIRD work and help?

- ECPIRD uses a European Search Engine that can index a subset of WWW sites and DBs relevant to the topic at hand using specialized terminology (e.g. ontologies). ECPIRD is a distributed effort: one center in e.g. Austria might deal with computer science, one in Romania with medicine, one in Germany with car production, another one with carpentry, etc. etc. This is in contrast to Google, and will beat Google in the areas covered. To index all servers that have to do with a specific subject matter – like computers, medicine, material science etc.- an doing this with suitable ontologies that fit the subject is a huge but still doable task. It is also desirable that the result of a search is not a list of items, but a much shorter list of reports generated automatically based on the query results.

- Good partners for this may be a German Fraunhoferinstitutes (the new institute run by Fellner in Darmstadt has quite a bit of experience in searching unorthodox data-sets and in digital libraries), L3S in cooperation with the Germany Library in Hannover, FAST in Olso, the University of Middlesex (Compton) and TU/Hyperwave in Graz: FAST has a list of WWW servers with metadata describing servers and can be extended by good synonym dictionaries, ontologies and semantic net tools. Hyperwave allows intelligent document management combining query results into more or less coherent documents. TU Graz with the center for knowledge management has much experience in aggregating data, and Middlesex has might prove a cornerstone when it comes to English ontologies and synonym dictionaries.

- A typical ECPIRD server may consist of hardware with storage in the two-digit terabyte range, with FAST as search engine and with Hyperwave to manage the documents: this will allow to not just produce a list of very good hits better than with Google, but a more coherent presentation.

- This proposal has been discussed with the CEO of FAST, John Markus Lervik and the CEO of Hyperwave Pieter Duijst, the technical teams involved, and a variety of other potential partners

- In addition to the great value of those specialized search engines they can also be used together with other tools for plagiarism and IPR violation detections as briefly explained in what follows:

- Dealing with text in image form is clearly doable using OCR

- The trick to deal with the other problems is to extract an English signature for each document and store all signatures on servers of ECPIRD using FAST to compare the signature of a given document with the universe of signatures available

[15] Pronounce ECPIRD similar to Expert!

- Note: Publishers will be willing to supply at no or little cost signatures of their papers for two reason: Plagiarism detection never detects, but just mentions suspicion, hence manual check in cases of plagiarism are necessary, potentially increasing hits for publishers; more important: publishers want to have a certificate on their publications 'plagiarism free' (result of discussion at APE 2007 in Berlin in January)

- Signatures for a document in language X are derived as follows. Removal of fill-words (articles, pronouns,…), stemming, reduction to a stream of words each in 'normal form'. Each normal form is replaced by one representatives of synonym class; this list of words is translated into English, and again each word is replaced by one representative in synonym class (for English documents situation is of course a bit simpler)

- Signatures eliminate synonym and language problem

- Remaining difficulties: suspected plagiarism cases have to be checked by domain and language experts

- (Remaining research: application to non-textual material and to plagiarism of concepts and ideas: this is not part of the proposed undertaking, but should be dealt with by the community with research support separately)

- W propose a two stage process is proposed: first stage, domain restricted to one (e.g. computer science), language restricted to two or three: English, German, French

- After success, other domains and languages are added

- Note that the search engines developed for ECPIRD will be the starting point of serious competition to Google. Not all domains can be (initially) supported: this is not part of the proposed undertaking

- ECPIRD must be associated with universities or other non-commercial institutions to avoid commercial exploitation of rank-shifting ("opportunistic ranking") and search analysis

- Aim: The part of ECPIRD used for plagiarism and IPR violation detection will be self sustainable after 2-3 years (institutions send set of papers for analysis to ECPIRD, receive a printed report as result. Assuming some 5.000 institutions at 2.000 € per year gives 10 million € per year as operating budget after implementation phase); however, there will be many other specialized search centers for e.g. non-scientific areas, or for certain segments of the population (like SMEs, schools, children, or the aging population).

- A guessimate of initial cost for systems for a few languages (German and English, possibly also French) and a dozen large domains is: 1000 million € over three years, 15% for FAST development and licenses, 40% for Software Development (distributed) and Hyperwave licenses, 15% for Synonym dictionaries, dictionaries, stemming algorithms (some purchased, some to be developed), 30% for the manual evaluation of suspected cases; if necessary, a proof o concept phase with just two languages and one or at most two areas might be feasible for 10 million €

- Graz University of Technology offers itself as host for running a first server within the ECPIRD framework; L3S in Hanover willing to run another one; the university of Middlesex has also indicated willingness to take on one server; some institution will have the overall coordination and play the role of the hub described earlier (Fraunhofer or L3S?). Note that the actual servers and areas associated with them should be distributed all over Europe to deal with the myriad of topics where better search engines than currently exist are available. In combination those search engines may pose the necessary challenge to Google.

- Note that all ECPIRD servers must be associated with non-profit organization because of the value of information in query analysis. To implement a real rival to Google a substantial effort has to be also invested into ranking algorithms and agglomeration algorithms (4 million extra?).

- It would also make sense to incorporate soon tools for non textual material such as information in tables, material containing formulae (mathematics, chemistry, physics,…) (ideal application for L3S, 3 million extra?), pictures, drawings, diagrams (ideal application for Fellner’s Fraunhofer Computer Graphics Group) and for conceptual search (some experience in Graz and Curtin), and for stylometry (ideal place Faculty of Media, Bauhaus University, Weimar). All those areas are currently neglected by Google and may well provide the handle to beat Google on parts of its turf.

Point 4: How to proceed

Maurer has had talks with director Forster in Luxembourg and has hopefully stirred some interest. He is in contact with the CEOs of FAST and Hyperwave. He has contact Dr. Görderler (Ministry of Economy in Berlin) who is in charge of the Theseus project and is hopefully a good interface to German government.

Dedicated servers and services for the aging may also fit well into a networked initiated under "ICT for Ageing People" coordinated by the University of Jyväskylä, Department of Computer Science and Information Systems (Pertti Saariluoma) and Information Technology Research Institute (Hannakaisa Isomäki & Hanna Parkkola).

Professor Hasebrook (University Luebeck) has indicated that SMEs supported by a consortium of banks in Germany may be interested in specialized servers and welcome the concept.

The Austrian Chamber of Commerce has agreed to support similar initiatives and will announce so in appropriate form.

It should be noted that the proposed approach is less expensive than any other suggested to date, but will proved the strong and necessary competition to commercial search engines that is desperately needed.

This is not a “run of the mill”, a “08/15” project, but an undertaking that may determine if Europe and other parts of the world can withstand the onslaught of commercially rising stars like Google.

Section 12: Market Forces vs. Public Control of Basic Vital Services

(Note: This is joint work between N. Kulathuramaiyer and H. Maurer and will still be edited if it is decided to publish ist)

In this Section we want to argue that there are basic services that can never be left to the free market but require government control, intervention and regulation. Just think of the curriculum for elementary schools, simple medical help for everyone, some road infrastructure, etc.

We want to make it clear to decision makers that the internet and particularly search engines or other data mining activities belong into this domain an thus do need governmental regulation and control. This is why completely market driven phenomena like Google cannot be tolerated.

Abstract:

No matter what economic system you believe in, there are certain aspects which are never or rarely left entirely to the free market: they are considered to be sufficiently important that the public (some government agency) has to have control and get involved. Typical instances are school education up to a certain age, national parks, basic health insurance, road systems, electric power and water supply, parts of the telecommunication infrastructure, etc. It is our belief that certain tools concerning the WWW have to be carefully examined if they do not also belong into this category. Prime candidates seem to be certain aspects of data mining which includes search engines. In this short paper we show that public involvement is important in areas such as the ones mentioned, and that new developments in the WWW should not be overlooked. We know it sounds exaggerated, but we feel that such an oversight might endanger even world economy as we now know it!

1. Introduction

One of the great challenges of sustainable development has been the ability to combine economic prosperity and security provided by the public, by government help or agencies. The reconciliation of the power of markets with the reassuring protection of citizen well-being has been a major challenge throughout history.

Communism places the control of all major services at the hands of the government. Although communism was able to reduce the lowest levels of poverty, it could not advance the general material welfare. [Sarup, 2006]   Apart from primitive native tribes such as American Indians, who communally owned most possessions, all extreme communist economies have thus failed. The failure of such economies has been attributed to the inability to generate monetary incentives. [Sarup, 2006] Although, capitalism has shown better results than any other system, current forms of variations of capitalism have resulted in unfavorable side-effects, such as depressions, excessive inequality, poverty, etc. [Sarup, 2006]

This paper explores the responsibility of governments in combination with the role of market forces in shaping the provision and delivery of public services in an attempt to show that modern Information and Communications Technology (IKT) is probably calling also for public attention to entirely new aspects To be more specific, the reconciliation of the powers of market forces together with concerns of public welfare is a major concern in addressing emerging questions such as ‘Should access to information and knowledge discovery service not be considered a right of every citizen?’ and ‘Should search engines be left entirely at the hands of market forces ?’

In the management of public services, there are a number of models that can be adopted. Publicly controlled services are generally seen as inefficient or leading to redundancies. On the other hand, control based entirely on a free market may not satisfy the needs of marginalized groups. Partnerships between the two parties may also be explored as an option. A mixed model may involve the private sector in providing for the effective needs of the majority, while the government provides subsidies, or concessions to address marginalized groups. Government-owned national corporations may also be formed to ensure that government remains in control while providing cost efficient solutions. A particular model that works well for a particular country or community may fail in other environments. Determining the best model has in the past been a major challenge. Government owned corporations, have over time become privatized for greater efficiency. At the same time there have been circumstances that have lead to privatized corporations becoming repossessed by the government.

This paper presents a review of critical success factors of models that have worked well and tries to highlight reasons for success or failure. We discuss numerous case studies on the management of critical resources such as water, electricity, healthcare, telecommunication services, bandwidth, to explain the implication of control. The insights gained from these case studies will, we hope, shed light on effective strategies for the management of telecommunications and information resources.

As certain services such as water, rice (or staple food) and even electricity and information-access could be considered as the rights of every citizen, extreme care and due diligence is absolutely essential for a fair and equitable provision. Both the private and the public sector have a role to play in carrying out a responsibility for ensuring fair-use and to avert exploitation of vital resources for the benefit of all. The next section discusses the implications of both public control and market-forces control of critical public services.

2. Vital Public Services

2.1 Water

Access to free clean drinking water is considered a right of every citizen as water is need for life. The United Nations Committee on Economic, Cultural and Social Rights declared access to water a human right calling for an obligation to progressively ensure access to clean water, "equitably and without discrimination"[Capdevila, 2007]. Despite this, more than one billion people don't have access to clean water. Dirty water has been reported to kill 2 children every minute [Kaufman and Snitow, 2006]. Furthermore, nearly 2 billion people do not have adequate sewage and sanitation facilities. [Kaufman and Snitow, 2006]

Water is traditionally a resource controlled by the public. However, publicly controlled water has been sometimes the target of complaints as not being able to meet the requirements of consumers cost-effectively. The propagation of water-privatization projects was thus supported by the World Bank and other international institutions as part of policies to transform developing nations into more market-oriented economies.[Thomas,2004] Ambitious World Bank-funded privatization schemes, intended to be a model for how the world's poorest communities [Vidal,2005] has been undertaken in a number of developing countries in Africa, South America and Asia. International corporations have failed to improve the supply for millions of people. [Hall,2005] Despite the support through private water concessions in South America, they performed no better than public sector operators in terms of expanding services to the poor. [Elwood,2003]. Even in Manila and Jakarta, private operators, have performed worse as compared to the majority of cities where water has been publicly managed.[Hall,2005a]

In Bolivia, privatization has led to the water war causing major protests and riots. Bolivia and other countries such as Peru, Ghana, Uruguay have thus learnt to resist World Bank's policies on privatizing the water system.[Hall et al, 205] In Africa, a referendum banning water privatization was passed in October 2004.[Elwood, 2003]

The World Bank has since acknowledged the failure of privatization to deliver investments in extending water services.[Hall,2005a] According to the World Bank, regulators have lacked the authority and competence to control companies' behavior. [Thomas, 2004] Uncertainties such as currency movements and economic crises have resulted in corporations not being able to provide cost-effective water supply.[Thomas, 2004] There are companies involved in privatization that have become bankrupt.

Even in the US, efforts to privatize water has been carried out in almost every city. This privatization of water has led to brown-water days in e.g. Atlanta, US [Elwood, 2003]. Such occurrences were previously known to happen only in less developed countries such as Argentina and South Africa. In Florida, water and sewage services was returned back to the government after the discovery of negligence in maintenance.[Elwood, 2003]

The main problem with projects of commercialization of water has been due to the unbundling of public accountability from service provision and delivery. This separation can lead to great dangers as in the following case. In Canada, the government of Ontario, had reduced water quality inspection staff and shut-down government testing labs in favour of private labs [Elwood, 2003]. The result was disastrous: 7 people died and hundreds became ill due to contaminated drinking water. Another tragic situation happened in Houston, when a fire destroyed an important cotton seed mill in 1886 “while firemen stood by helplessly because the hydrants were dry” [as quoted in Hausman, 1986].

Despite the benefits of market-forces in taking over public services to ensure the reduction of financial loads on governments, the impact on the people has been severe and even led to fatalities. Public control of vital services is better able to bundle in accountability and address emergency measures under challenging situations. Leaving the handling of such resources entirely at the hands of market forces could thus lead to great dangers. When water is treated as a commodity, rather than a rightfully endowed resource of all citizens, there is a potential for exploitation and abuse.

An alternative to privatization is to forge a partnership between the government and private sector. The Penang Water Authority (PBA) in Penang, Malaysia highlights a public-public partnership in water management. PBA is legally privatized with the state owning 75% stake in the company. Penang has shown a remarkable achievement of 99% universal access to drinking water at the lowest prices (compared to 65 other cities [Santiago,2005]) with a 98% revenue efficiency. [Santiago,2005] Penang has also managed to provide an interest-free loan of RM1,000 to poor communities for the purposes of connection to water services. In order to achieve self-reliance, profits of the water utility were reinvested and new infrastructure investments are acquired and maintained. [Santiago,2005] The critical success factor has been reported to be the close working relationship between PBA and the government and its strong commitment to provide quality public service.

The onus to provide basic crucial services should be seen as the responsibility of all parties including the community and private sector. As an example, a Non-Governmental Organisation(NGO) single-handedly addressed the severe water shortage in large areas of India [Manu Rao,2004]. The Sri Sathya Central Trust undertook the development of facilities to provide water to the entire city of Chennai (fourth largest city in India) affected by a succession of water crises for over 50 years. This project has then been handed over to the State Government to maintain. This demonstrates for the communitiy and private sector to proactively engage and collaborate in addressing basic needs, for the common well-being.

A great deal of responsibility is needed for the proper utilization of scarce natural resources such as water. Value-based water education is being emphasized in Africa and other developing countries to avoid abuses and exploitation of scarce resources. Education serves as important measure in bringing about long-term benefits with the cooperation of all parties.

2.2 Electricity

Safe, reliable and affordable electricity is a fundamental building block for all modern societies. It is one of the key components of sustainable development, providing the backbone for society’s social and economic well-being. In developed countries, shortages and blackouts can be disastrous. However, in less-developed countries shortages and blackouts may be considered as the norm.

Electricity liberalization refers to the control of electrical markets by market forces. As electricity supply tended to be a monopoly, a complex systems of regulations is enforced when partial privatization occurs. [Wikipedia, Electricity Liberalization]

Still, privatization of electricity (liberalization) has also lead to unreliable poor service which includes blackouts in a number of places. The difficulty in ensuring cost-effective solutions has largely resulted in compromising quality of service.

The study [Thomas,2004] indicates that privatization has been running into serious problems all over the world. Electricity has been designated as a ‘service of general economic interest’, with stringent requirements specified on price, quality and accessibility making it difficult for a free market to guarantee deliver. Apart from that this business can be risky in that the wrong choice of technology or the failure to complete the project on time can be disastrous. [Thomas,2004] These characteristics of the electricity sector makes it difficult, if not impossible, to impose market dynamics. [Thomas,2004] Public controlled service is still the best option for this sector.

Liberalization of electricity tends to substantially benefit large consumers (mainly industrial users), but the benefits for domestic consumers compared with a public monopoly or a regulated private monopoly are questionable. [Wikipedia, Electricity Liberalisation] Doubts has been raised as to whether the electricity generation system can ensure long-term security and timeliness of supply with the provision of sufficient incentives.

The unbundling of key service provision that results from privatization compounded by weak governance has made regulatory task complex.[Thomas,2004] According to the World Bank, ‘Regulatory weaknesses as the cause of most failed attempts at infrastructure reform and privatization in developing countries.’ The regulatory processes would need to encourage competition, be open and transparent, and be designed prior to privatization. [Thomas,2004]. This again calls for the joint responsibility of the public and private sectors, working closely to ensure citizen welfare.

2.3 Natural Resources

National parks are examples of natural resources that need to be protected against human development and pollution. In most countries they are owned by the government and preserved as protected areas. Well-managed conservation reserves are important for the long-term benefit of mankind. Apart from the government who take care of conservation, the community also has a part to play in maintaining the welfare of these parks.

The National Park Service, [Wikipedia, National Park Service] in the US employs various forms of partnerships, or concessions, with private businesses to bring recreation, resorts, and other amenities to their parks. The adaptive reuse, as proposed by private agencies, has raised controversy from public on the preservation of natural sites. In Australia, non-profit community organizations such as National Parks Association of NSW takes on the role of protecting and conserving the diversity of species, natural habitats, features and landscapes.[National Parks Association, NSW]

Apart from protection of ecological resources, there are situations where governments may also choose to own critical resources such as oil operating on their soil as in most OPEC countries. It has to be noted that the government, has to make a judgment as to the choice of resources that needs to be under public control for the nation’s well-being.

2.4 Healthcare

Health care can be entirely be government funded (as in United Kingdom), sustained by private insurance systems (as in United States), or somewhere in between. In the case of a welfare state, healthcare is linked to the level of taxation. No matter which approach is taken, a balance is needed to ensure the continuation of quality service regardless of commercial, environmental, or other external pressures.

In the US, it was noted that the government's privatized emergency relief agencies were not able to coordinate well with the military (which would often turn back relief trucks) resulting in the poor response during the Katrina hurricane incident. [Wikipedia, Privatization] The unbundling of responsibility had once again severe consequences, although private companies cannot be entirely held accountable for negligence in addressing human rights in times of unexpected natural disasters.

In India, due to the deficiencies in the public sector health systems, the poor in India are forced to seek services from the private sector, under immense economic duress.[Venkat Raman, and Bjorkman, 2007] Although the government is providing subsidies to address this concern, it does not effectively solve the problem. The poorest 20% of population benefit only 10% of the public (state) subsidy on health care, while richest 20% benefit from 34% of the subsidies. [Venkat Raman, and Bjorkman, 2007]

Furthermore, in both public and private controlled medical services, a challenging problem is the quality of medical service and personnel given to remote to rural areas. This is especially a problem in developing countries. Health sector reform strategies that have been proposed to address these problems include alternative financing (e.g. health insurance, community financing), de-centralized institutional management public sector reforms and collaboration with the private sector. [Venkat Raman, and Bjorkman, 2007] A well-coordinated partnership between the various parties, with clearly defined roles will be essential in addressing the needs of all segments of the population.

In order to realize partnerships the core elements that need to be addressed include the ensuring of mutual benefits, autonomy in carrying out assigned roles, shared decision-making and accountability and fair returns.[Venkat Raman, and Bjorkman, 2007]

2.5 Education

Another case study that is interesting to consider is education. It is largely subsidized by many governments (at least at the elementary school level), to avoid (high) fees. Institutions depend on government subsidies to keep fees low. Governments involvement in the educational sector is essential to ensure a more equitable distribution of goods and services as well as to ensure that the society will have the expertise it needs to remain competitive.[IAESBE,2004] Despite the public control, academic institutions should be endowed with a free hand in building an academic culture of excellence, in shaping future generations. Over- regulation and political intervention affect the quality of education in a number of countries.

In the education domain, the role of private agencies to participate in alleviating the financial burdens of the nation is vital. As education is the basis for addressing the man-power needs of both the public and private sectors, a strong commitment from both party is required. A strategic partnership could thus play a pivotal role in shaping the education ecology of a country.

3. Control Implications on ICT Resources

The previous section has highlighted the implications of control of a number of public services both at the hands of the public and market forces. A number of lessons have been highlighted which also apply to the area of telecommunications. This section will explore the control over telecommunications resources such as telephone, internet and bandwidth.

3.1 Telephone

Telecommunications play an important role in the development of societies. The industry's revenue is estimated at $1.2 trillion or just under 3% of the gross world product. [Wikipedia, Telecommunications] Despite the similarities in the pricing problems faced by telephone and water privatization, the provision of local telephone service has been privatized to a large extent.[Croker and Masten, 2000]

The main problem with the mainly corporate controlled telecommunications services has been the increasing digital divide. Access to telecommunication systems is not being equally shared amongst the world's population. In most sub-Saharan countries not even 1% of the population have landline-connected telephones. That compares with more than 10 lines per 100 people in Latin America and more than 64 per 100 in the United States. [Mbarika and Mbarika, 2006]

Even today, the telephone penetration in the US, is currently about 95%. Thus, there are people in remote parts who do not have access to basic facilities.[Wikipedia, Universal Service] In developing countries, the divide can be much more worse. For example, the country of Senegal has about 140 000 lines, with 65% of those lines situated in the capital city, Dakar.[UK E-Society]. In contrast, countries like Austria and Germany, where telecommunications remained a monopoly till the mid 90s telephone coverage has been close 100%: even remote farm houses are connected to the phone system, even if this was economically not sensible: the cost was spread to other users. Thus, in the mid 90s countries like Germany and Austria were able to achieve a very high percentage of telephone coverage. Phone calls were however, more expensive than in countries with privatized systems. This situation clearly highlights the possibility of addressing digital divide by an equitable distribution of the cost.

This however cannot be achieved under the control of market forces as the market forces alone cannot address the cost of providing access to the remote minorities. Universal Service describes a 'commitment' enforced upon service providers to citizens in areas whereby basic telephone service has been difficult to access or unavailable.[Wikipedia, Universal Service] The responsibility of regulating universal service is handled by the government or government-owned agency. Universal Service also relates to telecommunication access to schools, libraries and rural health care providers, while promoting open access and competition among telecommunication companies.

A provider engaged in universal service is given subsidies from a fund to economically provide necessary services. Most countries maintain complex legislature and regulation structure to guarantee the service with an adequate subsidy mechanisms to implement universal service.[Wikipedia, Universal Service]

3.2 Mobile Phone

Mobile phones have had a significant impact on telephone networks. Mobile phone subscriptions now outnumber fixed-line subscriptions in many markets. [Wikipedia, Telecommunications,] Sub-Saharan Africa is now the world's fastest-growing wireless market. This tremendous growth in wireless market has arisen because its national telecommunications monopolies are poorly managed and corrupt, and they can't afford to lay new lines or maintain old ones. [Mbarika and Mbarika, 2006] Telkom Kenya Ltd., for example, which has only 313 000 landline subscribers in a country of 30 million people, will not invest in new lines. Instead, it is partnering with foreign companies that are putting in new wireless infrastructure.[Mbarika and Mbarika, 2006]

The African cell phone revolution thrives on a combination of a novel business model and clever calling tactics that have made mobile phone usage affordable for a large segment of the population. [Mbarika and Mbarika, 2006] As a result, growing numbers of Africans are gaining access to phones. The question raised by [Mbarika and Mbarika, 2006] with regards to such development is whether the region can sustain the wireless sector's phenomenal growth rates in bringing about sustainable development and wealth. Despite the short term benefits of infusing the desperately needed foreign investment, this could well be a form of neocolonialism dressed up as a market liberalization. [Mbarika and Mbarika, 2006]

3.3 Internet and Bandwidth

The Internet has become a powerful community resource in bringing about social change through community knowledge building, community well-being and development. The internet is currently the fastest growing network with an estimated penetration rate of 16.9% [Internet World Stats].

Although the Internet is not centrally controlled by any agency, an American, non-profit corporation called the Internet Corporation for Assigned Names and Numbers (Icann), holds the right to regulate domain names and allocate addresses. The European Union and other countries have raised concerns to water down the control of the US on the Internet, as it is an international resource. [Smith,2005]
The Internet has resulted in a massive diversification of media produced by millions of web content providers. That evolution, described as citizen journalism makes it possible for practically everybody to be a media creator, owner and actor, instead of merely being a passive user. The Web2.0 has become a read-write channel by engaging the participation of millions of Web users.
The control over bandwidth is also a widely contentious issue. The network provider companies, the ISPs are now becoming interested in controlling bandwidth utilization based on usage. There is an imbalance in the current use of bandwidth. More than half of all Internet bandwidth is currently being used for peer-to-peer traffic such as BitTorrent traffic, [Cringely, 2007] which is mainly video. They only accounts for less than 5% of all Internet users. Broadband ISPs have a problem with these super users and are exploring ways to isolate and address them appropriately.
As more and more people will be downloading movies and television shows over the net, bandwidth utilisation is looking to explode from the current broadband usage of 1-3 gigabytes per month, to 1-3 gigabytes per DAY. [Cringely, 2007] This implies a 30-times increase that will place a huge backbone burden on ISPs. Network Neutrality [Wikipedia Net Neutrality] is a slogan used by ISPs to charge content providers based on their control of network pipes. Internet content providers such as Google are however strongly opposed to Net Neutrality as this will mean that they will have to pay more for greater bandwidth usage.

3.4 Web Search

In today's world we find that people are relying more and more on search engines for almost everything. Search engine companies are becoming more and more in control of larger information sources. [Kulathuramaiyer and Balke, 2006] have shown that the monopoly of Web search engine can result in a severe disadvantages leading to economic and social imbalance. Private companies are mainly concerned with maximization of profits. A private company may tend more to serve the needs of those who are most willing (and able) to pay, as opposed to the needs of the majority. To makes things worse, there is however, no regulation that controls the mining ability of these companies. [Kulathuramaiyer and Balke, 2006]

The uncontrolled agglomeration of services of global search engines is of serious concern. As an alternative, [Kulathuramaiyer and Maurer,2007] have proposed that distributed services being managed by multiple sites as a potential solution to the monopoly by global players. Highly specialized domain or category specific mining has been proposed.

3.5 Discussion

The paper has highlighted the issues surrounding the control of various critical resources. As seen with the management of water and electricity, there are resources that are best managed by the public sector. The public sector is able to bundle in the responsibility of effecting public welfare, and take into consideration the rights of citizen. A partnership between the public and private sectors are also presented as a viable solution for managing a number of resources. The balance between public and private determines the effectiveness and efficiency of the distribution of resources, benefits to citizens, and the long-term viability of a solution. The elements of successful partnerships as discussed for the health scenario needs to be taken into consideration in designing an effective partnership. A careful strategy in assigning and monitoring roles and responsibilities would be required for a fair and equitable service provision.

For resources, such as mobile phones with a strong economic model, the control by market forces has been seen to perform well. Despite this achievement, a partnership with the public sector is required to address the needs of the marginalized population. Well-coordinated effort and strong regulation is critical for the management of such resources.

Efforts in liberalization of resources such as water and electricity by the World Bank and other international institution have largely failed. This is seen as the consequence of globalization in connection with privatization of public services[Hall, 2001]. The limiting of social and economic solidarity in the financing of services, has resulted in the undermining of the basic principle of public services. [Hall, 2001] The provision of commercialized service dictated by profits, cannot be expected to meet the needs of the population.

As described in [IAESBE, 2004], government control may provide the best, or even the only, means by which particular goods or services may be produced. When the public well-being component is significant— as it is with the military, fire fighting, mass transit, water, electricity—the government can often provide the best solution to market failures [IAESBE, 2004]. Market failure refers to the inability of the private control to ensure that the well-being of all citizen is met adequately.

4. Social and Economic Implications of Control of the WWW

4.1 Control over Powerful Mining and the Bundling of Services

With regards to all telecommunications and other emerging services, there is a need to scrutinize the social and economic issues that arise from an entirely private control. [Kulathuramaiyer and Balke, 2006] have emphasized the need for international legislative intervention to curtail the expansive mining capabilities of agencies with powerful mining capability such as search engine companies. Based on billions of daily queries, these global kingpins are able to predict with astonishing accuracy what is going to happen in a number of areas of economic importance. For example, the surge of interest (or a sudden drastic drop in interest) in a particular stock can be monitored closed and tracked easily. As illustrated by [Trancer, 2007], incidents such as social and market trends can be predicted well before they happen. It would thus be possible to foresee events such as unemployment and stock market crash. To dramatize the potential implications further, we consider a situation where such a mining powerhouse artificially causes a hype leading to the surge or severe drop in particular stocks. What worries us most is that there is currently no way to avoid such a hypothetical situation from actually happening.

With their broad information base together with the means to shift public awareness and the continued agglomeration of additional services, such a company could analyze user behavior on a very personal level. It would thus be possible for individual persons to be targeted for scrutiny and manipulation with high accuracy resulting in severe privacy concerns.

4.2 Digital Divide and Knowledge Gap

Digital divide is widely addressed as the separation between those who have and those who do not with regards to information services. The digital divide represents an important social problem accompanying the diffusion of technologies such as the internet. The access-divide, which has been widely addressed is expected to eventually lead to a learning-divide, a content-divide, and other types of divides. [Rogers and Shukla, 2003] The internet and other Web technologies will continue to advantage particular individuals, and disadvantage many others.

The study on e-societies [UK E-society], has compared the distribution of e-unengaged population with e-experts across UK. This study clearly highlights that there are layers of digital and information divides that has to be considered in the design of information service management (beyond access). As there is a big gap between access and effective utilisation, there is thus an even wider gap between the e-experts as compared to e-unengaged. (or e-isolated). These gaps are likely to further expand with the advent of the Web2.0, the participatory Web and other emerging technology.

A study conducted in South Korea [Hwang, 2004] highlighted a wide access gap between urban and rural areas. The study also showed that there was a substantial engagement gap, between urban and rural users with access. [Nielsen, 2006a] describes a usability divide highlighting the differences between power users and naïve users. According to him, there are naïve users who are not able to distinguish between ranked search results and sponsored links.

Participation inequality [Nielson, 2006b] highlights the social problems arising from the empowerment divide that has held constant throughout the years of Internet growth. In social networks and community systems (Web-based), about 90% of users don't contribute, 9% contribute sporadically, and a tiny minority of 1% accounts for most contributions. (described as the 90-9-1 rule) With the advent of Web2.0 the problem has become even more skewed with Blog participation having a 95-5-0.1 ratio and Wikipedia having a 99.8-0.2-0.003 participation ratio.

The knowledge divide is a phrase used to indicate that a section of the world's people with an abundance of this knowledge, while the other section is severely lacking. The gap between those with the knowledge and those without keeps increasing. [Addy, 2006] Scientific knowledge, indigenous knowledge is more and more being controlled by a few global players. Individual global companies' R&D capacity is going to be many times more that that of not only developing nations but also developed nations.

With the control and power of large power-houses such as search engine companies, it is no longer, a situation of North countries having an upper hand "selling" knowledge to South countries. It is a situation where a few global companies are having the control over crucial information and knowledge resources selling knowledge to the entire world. This results in a drastic compromise to sustainable development [Addy, 2006], as these players are only accountable to their shareholders. The emerging global but not integrated order, links us together, while maintaining a deep divide between a selected few and the rest of humanity. This will go beyond the 'fractured global order' [Sagasti, 1989] of globalization, to become a 'global feudal order' of the Web Civilization.

5. Conclusion

In managing the control of vital resources, responsibility and accountability has to be bundled into the delivery process. The best strategy for control, has to consider long term benefits to all segments of the population and it has to be sustainable. The social issues mentioned in the previous section have also to be addressed effectively, in coming up with this strategy.

Access to resources such as water, energy, health care, natural heritages should be maintained as a right of every citizen. Similarly access to clean, unbiased, non-commercially slanted information (and knowledge) should also be made a right of every citizen. This paper highlights the importance of state-intervention and responsibility in protecting the gateway to knowledge of the population and to assure that data mining does not provide insights into persons, transactions and economic developments which can be exploited dramatically for gain by a company, yet considerably hurt society. Sound policies and government regulation is required in realizing this.

Governments have found it necessary to intervene in freely functioning markets in order to protect disadvantaged citizens from those with superior information as done for the regulated medical licensing, controlling of financial institutions and avoiding unfair monopoly through antitrust laws. Similarly, governments will need to act swiftly to effect a more equitable distribution of information and knowledge goods and services to ensure that the government, society and citizens will have the expertise needed to remain competitive.

References

Addy, M. E.: The Knowledge Divide as a Challenge to Sustainable Development, 2006, http://www.sciforum.hu/file/Addy.doc

Balanyá,B., Brennan,B., Hoedeman,O., Kishimoto, S. and Terhorst , P., (eds): Reclaiming Public Water: Achievements, Struggles and Visions from Around the World, Transnational Institute and Corporate Europe Observatory, January 2005.

Capdevila, G.: RIGHTS: UN Consecrates Water As Public Good, Human Right, Inter Press Service News Agency, March 31, 2007, http://www.ipsnews.net/interna.asp?idnews=14204

Crocker, K.J., Masten, S.E.: Prospects for Private Water Provision in Developing Countries: Lessons from 19th Century America, to be published in Thirsting for Efficiency: The Reform of Urban Water Systems, June 2000 http://www.isnie.org/ISNIE00/Papers/Crocker-Masten.pdf

Cringley, R.X.: When Being a Verb is Not Enough: Google wants to be YOUR Internet. January 19, 2007, http://www.pbs.org/cringely/pulpit/2007/pulpit_20070119_001510.html

Dickson, D. : Scientific output: the real 'knowledge divide'; 19 July 2004 http://www.scidev.net/Editorials/index.cfm?fuseaction=readEditorials&itemid=122&language=1

Hall, D.: Introduction in Balanyá,B., Brennan,B., Hoedeman,O., Kishimoto, S. and Terhorst , P. (eds): Reclaiming Public Water: Achievements, Struggles and Visions from Around the World, Transnational Institute and Corporate Europe Observatory, January 2005.

Hall, D.: Globalization, privatization and healthcare: a preliminary report , PSIRU, January 2001 Public Services International; www.psiru.org/reports/2001-02-H-Over.doc

Hall, D., Lobina, E., De La Motte, R.: Public resistance to privatization in water and energy June 2005, PSIRU Report, Development in Practice, Volume 15, Numbers 3 & 4, June 2005 http://www.psiru.org/reports/2005-06-W-E-resist.pdf

Hausman, W. J., Kemm, D. M., Neufeld, I. J.: 1986, the Relative Economic Efficiency of Private vs. Municipal Waterworks in the 1890s, Business and Economic History, Vol 15, p.p. 13-28. http://www.thebhc.org/publications/BEHprint/v015/p0013-p0028.pdf

Hwang ,J. S.: Digital Divide in Internet Use within the Urban Hierarchy: The Case of South Korea , Urban Geography, Volume 25, Number 4, May-June 2004, pp. 372-389(18), Bellwether Publishing

[IASBE,2004] Institute for Applied Economics and the Study of Business Enterprise, Balancing Public and Private Control: Germany and the United States in the Post-War Era, International colloquium on “Balancing Public and Private Control: The United States and Germany in the Post-World War II Era.” http://www.jhu.edu/~iaesbe/PUB-PRIV%20introduction%20DRAFT.pdf

Kaufman, D. Snitow, A.: Market forces seek to control the essence of life – water, San Fransisco Chronicle Open Forum, March 20, 2006 http://www.sfgate.com/cgi-bin/article.cgi?f=/c/a/2006/03/20/EDGU9GJD0K1.DTL&hw=Market+forces+water&sn=001&sc=1000

Kulathuramaiyer N., Balke, W.T.: Restricting the View and Connecting the Dots - Dangers of a Web Search Engine Monopoly; Journal of Universal Computer Science, Vol. 12, No. 12, pp. 1731-1740, 2006 http://www.jucs.org/jucs_12_12/restricting_the_view_and

Kulathuramaiyer, N., Maurer, H.: Why is Plagiarism Detection and IPR Violation Suddenly of Paramount Importance; submitted to International Conference in Knowledge Management , Vienna 2007

Manu Rao, B. S.: Super-speciality hospital touches 2.5 lakh cases; Times of India, 29 Apr, 2004 http://timesofindia.indiatimes.com/cms.dll/articleshow?msid=646815

Mbarika, V.W.A. and Mbarika, I.: Africa Calling Burgeoning wireless networks connect Africans to the world and each other, IEEE Spectrum, May 2006.
http://www.spectrum.ieee.org/may06/3426

Nielsen, J.: Digital Divide: The Three Stages; Jakob Nielsen's Alertbox, November 20, 2006: http://www.useit.com/alertbox/digital-divide.html

Nielsen, J.: Participation Inequality: Encouraging More Users to Contribute, Jakob Nielsen's Alertbox, October 9, 2006 http://www.useit.com/alertbox/participation_inequality.html

NPA, NSW, National Parks Association of New South Wales, http://www.npansw.org.au/web/about/about.htm

Rogers, E.V. and Shukla P.: In: the Role of Telecenters in Development Communication and the Digital Divide, 2003, http://wsispapers.choike.org/role_telecenters-development.pdf

Santiago, R.: Public-Public Partnership: An Alternative Strategy in Water Management in Malaysia; In: Balanyá,B., Brennan,B., Hoedeman,O., Kishimoto, S. and Terhorst , P., (eds), Reclaiming Public Water: Achievements, Struggles and Visions from Around the World, Transnational Institute and Corporate Europe Observatory, January 2005.

Sagasti, A.: The Knowledge Explosion and The Knowledge Divide, Human Development Report, United Nations Development Programme, 2001hdr.undp.org/docs/publications/background_papers/sagasti.doc

Sarup, K.: Democracy vs. Communism, American Chronicle, November 5, 2006 http://www.americanchronicle.com/articles/viewArticle.asp?articleID=16031

Smith, S.: Debate on Internet ownership continues, December 2, 2005 http://edition.cnn.com/2005/TECH/12/02/spark.internet.ownership/index.html

Trancer, B.: July Unemployment Numbers (U.S.) - Calling All Economists, http://weblogs.hitwise.com/billtancer/2006/08/july_unemployment_numbers_us_c.html Accessed 17 January 2007

Thomas, S.: Electricity liberalisation: The beginning of the end, PSIRU, University of Greenwich, http://www.psiru.org/reports/2004-09-E-WEC.doc

UK Geography of the E-Society: A National Classification , http://www.tni.org/books/publicwater.pdf

UK Geography of the E-Society: A National Classification; http://www.spatial-literacy.org/inc/resources/e_soc.pdf

Vidal, J.: Flagship water privatization fails in Tanzania, The Guardian, May 25, 2005 http://society.guardian.co.uk/aid/story/0,14178,1491729,00.html

Venkat Raman, A., Björkman, J.W.: Public-Private Partnership in the provision of Health Care Services to the Poor in India, Eighth Annual Global Development Conference: Shaping a New Global Reality: The Rise of Asia and its Implications; Beijing January 2007

Elwood, W. (Eds): New Internationalist 355, April 2003 Privatization / The FACTS, http://www.newint.org/issue355/facts.htm

Weber, S.: Das Google-Copy-Paste-Syndrom- Wie Netzplagiate Ausbildung und Wissen gefährden; Heise, Hannover, 2006

Wikipedia, Privatization, http://en.wikipedia.org/wiki/Privatization

Wikipedia, Public Ownership, http://en.wikipedia.org/wiki/Public_ownership

Wikipedia, Nationalization, http://en.wikipedia.org/wiki/Nationalization

Wikipedia, Municipalization, http://en.wikipedia.org/wiki/Municipalization

Wikipedia, Public-Private Partnership, http://en.wikipedia.org/wiki/Public-private_partnership

Wikipedia, Network Neutrality, http://en.wikipedia.org/wiki/Network_neutrality

Wikipedia, Free Market, http://en.wikipedia.org/wiki/Free_market

Wikipedia, Social Democracy, http://en.wikipedia.org/wiki/Social_democracy

Wikipedia, State-owned enterprise, http://en.wikipedia.org/wiki/State-owned_enterprise

Wikipedia, Electricity Liberalization, http://en.wikipedia.org/wiki/Electricity_liberalization

Wikipedia, Universal Service, http://en.wikipedia.org/wiki/Universal_service

Wikipedia, Government Monopoly, http://en.wikipedia.org/wiki/Government_monopoly

Wikipedia, National Park Service, http://en.wikipedia.org/wiki/National_Park_Service

Wikipedia, Telecommunications, http://en.wikipedia.org/wiki/Telecomunications

Wikipedia, Concentration of Media Ownership, http://en.wikipedia.org/wiki/Concentration_of_media_ownership

World Internet Users and Population Stats, internetworldstats.com, 2006 http://www.internetworldstats.com/stats.htm

Section 13: The three-dimensional Web and „Second Life“

(This is a contribution provided by Frank Kappe)

Virtual Words, also known as Massively Multiplayer Online Role-Playing Games (MMORPGs) or Virtual Interactive Communities have existed for quite some time. However, until recently they were seen as games rather than media technology.

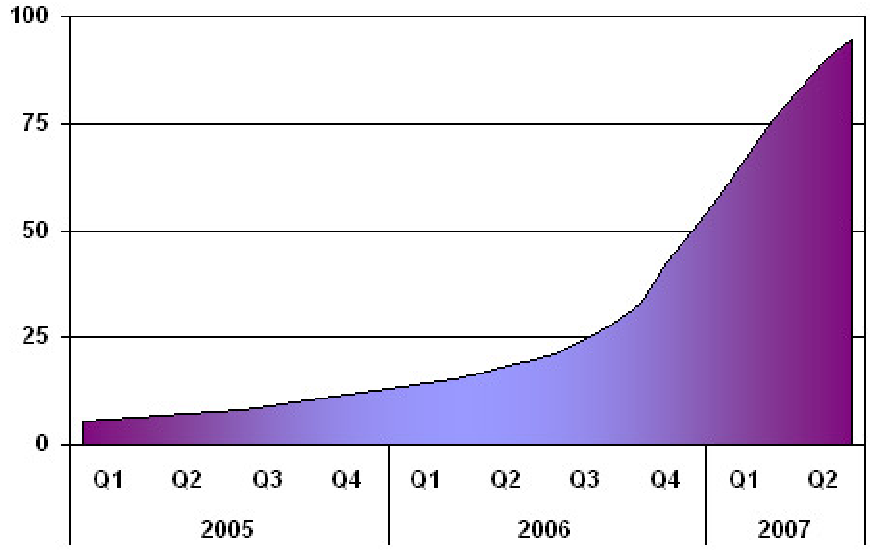
With the hype that surrounded “Second Life” (SL) in the beginning of 2007, this virtual world in particular has grabbed the attention of a broader audience. And in fact SL – developed and operated by the US company “Linden Labs” (LL) –differs from other virtual worlds in significant ways:

- In SL users can create their own world (“user-generated content”). This allows reconstruction of real buildings and in fact whole cities in SL, to stage events and to place advertising. This in turn makes SL a real medium. About 95% of SL has not been created by LL, but by ordinary users. Of course in order to build one needs land to build upon, which has to be bought (directly or indirectly) from LL, and monthly fees have to be paid (this is in fact the main source of income for LL). Only paying customers (“premium users”) can own land, but for users who do not wish to build, usage is free.
- Objects in SL can be programmed (using a proprietary scripting language) and can interact with other objects (e.g. avatars). Of particular interest is the ability of such scripts attached to virtual objects to communicate with applications outside of SL using HTTP, XML-RPC and E-Mail protocols. This functionality allows SL to interact with the real world (e.g. by clicking on a virtual object one can order real-world goods). It is also possible to stream (live) video and audio into SL, which is frequently used for so-called “mixed reality” events (where real-world events can be followed in SL and vice versa).
- The intellectual property rights (IPRs) of objects, scripts, and animations created by users remain with them. This allows creative users to earn virtual money (so-called Linden$) by selling their works to other users.
- The “LindeX” (Linden Exchange) allows conversion of Linden$ to US$ and vice versa (another source of income for LL), with the exchange rate determined by supply and demand, at least in theory. In practice, LL has managed to keep the exchange rate almost constant at 270 Linden$/US$ for the last year. For a few hundred individuals, SL is their main source of income.

The unique combination of these features makes SL interesting as a medium, which is why traditional media companies like Reuters and Bild.T-Online.de have build virtual headquarters in SL and report on SL events. Other companies like IBM claim to use SL intensively to conduct internal meetings and to meet partners.

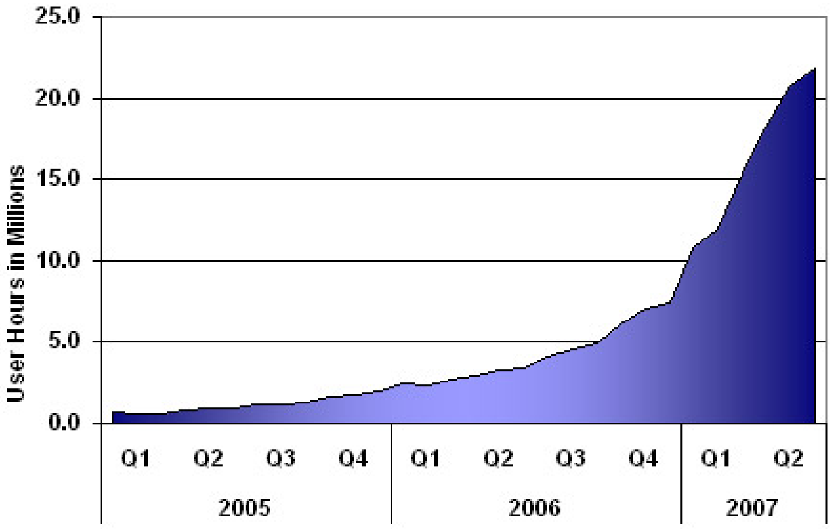
At the end of 2006 and the beginning of 2007, SL has enjoyed significant media coverage, which resulted in dramatically growing user figures (the hype has somewhat faded in the meantime, as usual).

Figure 52: Paying SL users (premium users)



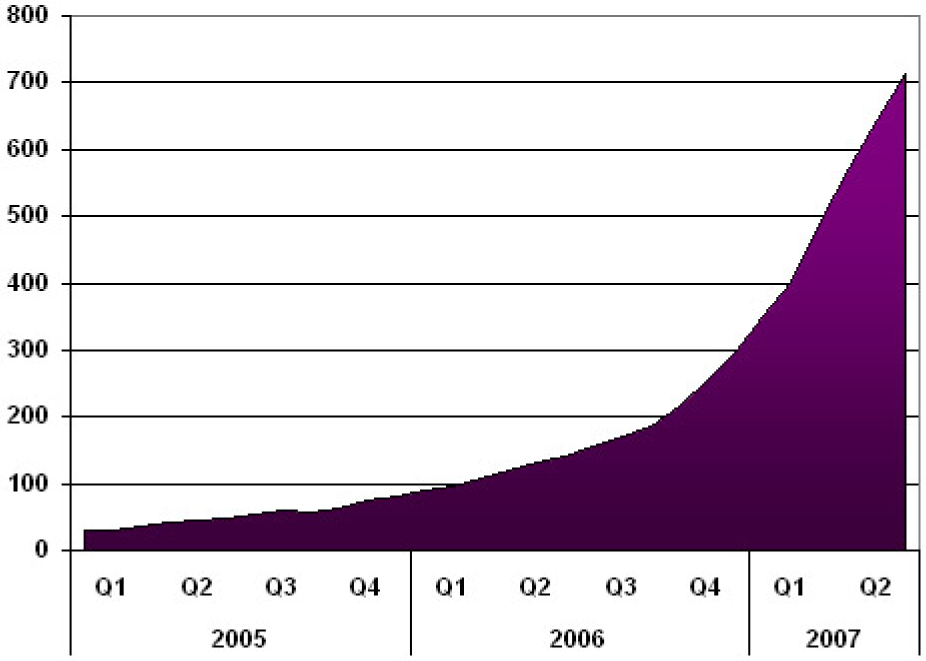
It is a bit difficult to come to realistic usage numbers. The often-quoted number of about 9 million "residents" is misleading, because everybody who ever signed up for the service is counted, regardless of whether the person still uses the service or not (maybe "tourists" would be a better word for this). In order to judge the overall trend we can use the number of paying "premium users" (even though the absolute number is of course much smaller), as well as the total hours users spend in SL.

Figure 53: Total number of hours users spend with SL per month



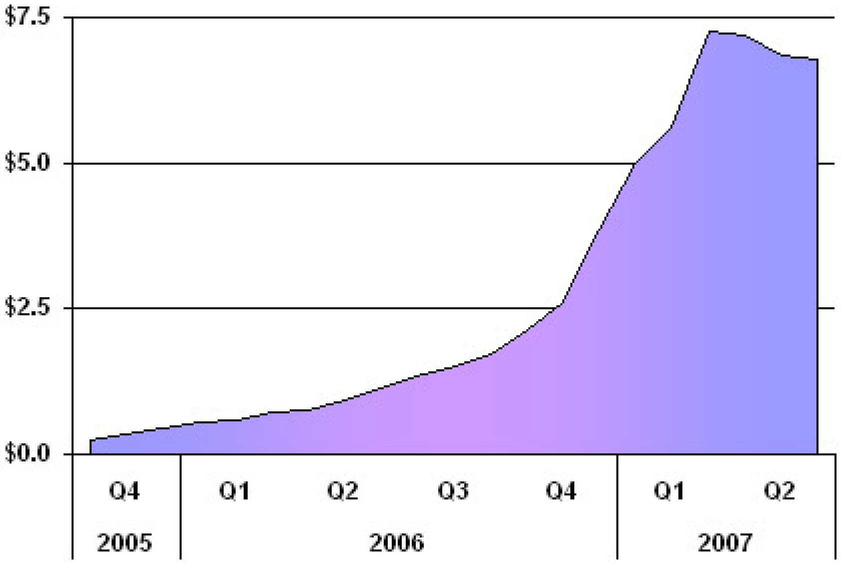
As can be seen from the graphs, SL has enjoyed significant growth over the last two quarters (however, more recent numbers show a drop in premium numbers after LL closed virtual casinos). Also the world itself grows: currently, about 700 km2 of land have been sold.

Figure 54: Second Life Land (km2)



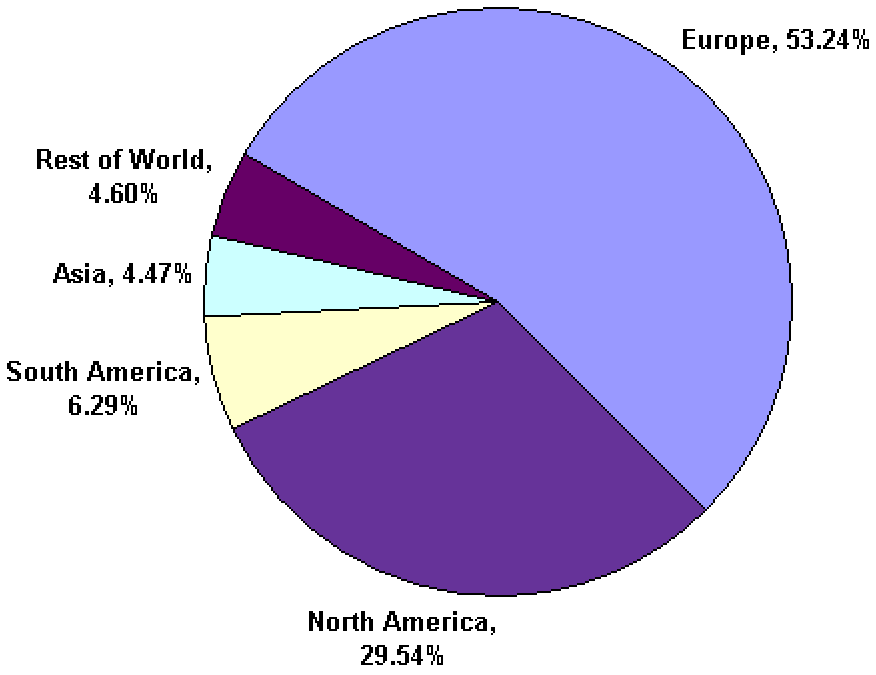
Turnover at the LindeX, however, is flat at about 7 million UD$ / day. The reason for this may be the opening of alternative exchanges for trading US$ and other currencies against the Linden$, and a recent crack-down on money-laundering and a ban of virtual casinos by LL.

Figure 55: LindeX turnover (in millions of US$ per day)



It is worth noting that SL is not a US phenomenon. On the contrary, US users account for only 26% of the total users. Most users are now European, with Germany on rank two with about 10%. In general, SL seems to be very popular in Germany, which many German cities already having a virtual copy in SL.

Figure 56: SL users be region (as of March 2007)



With all the hype, there is also plenty of criticism surrounding SL and LL:

- SL is not a game. There is no clear goal, no victories can be made. Many users get bored quickly and give up.
- A fast Internet connection is necessary. Because everything is user-generated (e.g. textures), nothing can be installed locally but has to be fetched over the Internet when needed.
- Also, the graphics are less sophisticated (for the same reason) compared to other 3D games. However, LL delivered a lot of improvements recently.
- LL has been growing extremely fast and has become a victim to its own success: the software contains too many bugs, runs sometimes rather slow, and customer support is bad.
- LL's terms of service allow LL to freeze the virtual assets of users at any time without compensation (as a number of owners of before-legal-now-illegal virtual casinos have found out recently).
- LL operates under California Law, which limits the freedom of its users, even when e.g. Austrians want to perform virtual business among each other.

Recently, LL has reacted to some of the issues raised above, and separated the software (now called "second life grid") from the service ("Second Life"), and changed their business model accordingly. The software will eventually become completely open-source (currently only the viewer is open-source), allowing anyone to run an SL server (very much like a Web server today). Second Life (the service) will – in the word of LL – be "the virtual world where end users can create an infinite variety of content and participate in a robust virtual economy."

It should be noted that there are also other similar virtual worlds around (but none as user-participatory as SL), as well as initiatives aiming at developing an open-source alternative for LL's SL (the OpenSim software with the DeepGrid service). Most likely, however, these will merge with the second life grid once it becomes open source. Also, IBM has announced to invest 100 million US$ in 35 ideas, Virtual Worlds being one on them.

The Gartner group estimates that in the year 2011 about 80% of the internet users will own one or more accounts in virtual worlds. Most likely these will not be the "Second Life" as we know it today, but it worth gaining experience in virtual world business models today to be prepared then.

It is not known how SL and similar undertakings are going to influence the use of Wikipedia, of search engines and ordinary data mining, and how much SL will also be used to discover person and company profiles. It will certainly be important to keep a close eye on those issues!

References

„Second Life":

http://www.secondlife.com

http://secondlifegrid.net/

http://secondlife.reuters.com/

http://www.theavastar.com

http://www.secondlifeinsider.com

http://www.lifeforyou.tv

„Metaverse 1.0":

http://www.lostinthemagicforest.com/blog/?p=43

„OpenSIM":

http://opensimulator.org/wiki/Main_Page

Section 14: Service oriented information supply model for knowledge workers

(Note: this was joint work between B. Zaka and H. Maurer)

In this section we are trying to show how the techniques that we have gained through plagiarism detection can also be used in the context of knowledge workers, i.e. helping the worker during the work, rather than checking the work afterwards.

Abstract: This paper suggests a powerful yet so far not used way to assist knowledge workers: while they are working on a problem, a system in the background is continuously checking to determine if similar or helpful material has not been published before, elsewhere. The technique described aims to reduce effort and time required to search relevant data on the World Wide Web by moving from a "pull" paradigm, where the user has to become active, to a "push" paradigm, where the user is notified if something relevant is found. We claim that the approach facilitates work by providing context aware passive web search, result analysis, extraction and organization of information according to the tasks at hand. Instead of conventional information retrieval enhancements we suggest a model where relevant information automatically moves into the attention field of the knowledge worker.

1. Introduction

Information search- and retrieval- processes play a vital role in the productivity of a knowledge worker. Every knowledge worker has to do extensive searches at some point in time to find information that may help, or show that certain aspects have already been covered before. Search engines provide the basic means of interaction with the massive knowledge base available on the World Wide Web.

Conventional search technology uses a pull model: i.e. search engines require an input from the user in form of a query consisting of keywords. This active search paradigm has a number of downsides: knowledge workers normally are not trained for really comprehensive searching. They may not know all the tricks required to locate the right sources of information. They may not know how to formulate a query describing all that they want to find.

The formulation of search queries is often difficult due to special terminology, or just the difference of terminology used by authors in various sources. Another constraining factor of typical search engines is the fact that they only cover shallow web contents, ignoring the almost 500 times larger, "deep" or invisible web information base [Bergman, 01]. There are special search interfaces for domain specific searches but not all are commonly known to general users. Thus, any organization or individual pays a high price due to ineffective information discovery. According to an IDC white paper by Susan Feldman, knowledge workers spend 15% – 35% of their time searching for information. The success rate of searches is estimated at 50% or less. The study further states that knowledge workers spend more time recreating information that already exist, simply because it was not found when needed. These factors contribute to a waste of time, costing individuals and organizations a substantial amount of money [Feldman, 06].

2. Searching the web for knowledge acquisition

The primary step in knowledge work is the acquisition of information. Access to first hand information comes by means of reading literature, by meeting subject specialists, by learning technologies and finally but most importantly making use of the immense information space of the internet.

Most knowledge workers consider online technologies to be the most efficient way to access information. Web search engines no doubt are the gateway to this information base. A research study

[Hölscher & Strube, 00] describing behaviour of web search by both experts and newcomers shows that when presented with information seeking tasks, the majority resorts to search engines instead of browsing direct sources such as collections of journals made available by some publishing company. Further experiments show that searchers quite frequently switch back and fourth between browsing and querying. The switching involves reformulation, reformatting of queries and change of search engines, based on previous result sets. The study also states that web search experts make more use of advanced search features such as specialized search engines, Boolean operators, search modifiers, phrase search, and proximity search options. Domain experts tend to show more creativity as far as terminology is concerned and form better and longer queries. The overall information seeking process as described by the experiments in the study shows the difficulty and the reduction of productivity in the majority of cases. A knowledge seeker is required to have good searching capabilities as well as good domain knowledge with rich vocabulary to successfully accomplish the goal. However, this is not always the case: there was and still is a big need for the enhancement of search environments for knowledge seeker.

There are attempts to facilitate searches in the form of

i. Meta search services covering multiple search databases e.g. Jux2, Dogpile, Clusty16 etc.
ii. Web based and desktop tools based on search APIs to facilitate advance searching e.g. Ultraseek, WebFerret17 and many web search API mashups available at ProgrammableWeb portal18
iii. Desktop tools to explore local resources e.g. Google desktop, Yahoo desktop, Copernic19 etc.
iv. Specialized search engines, typically content specific like Google Scholar, Live Academic20, CiteSeer, Scirus, IEEExplore21 etc. or media specific like image search, video search etc.
v. Semantic search e.g. Swoogle, Hakia22 etc.

Meta search approaches provide access to multiple search sources but the process of advanced query generation is more complex due to different query formulation options of the search engines. Some times variations in results by meta search requires greater examining time and effort. Tools to facilitate searches based on APIs provide another meta search option with some query optimization facilities (spelling suggestions, synonyms support, easy use of advance search options, etc.). However, they all suffer from another drawback: the limit of allowed queries in terms of quantity and quality. A quantitative analysis [McCown & Nelson, 07] of results produced using conventional WUI (Web User Interface) and APIs, shows significant discrepancies. Results produced by search engine's own interface and APIs are rarely identical. This seems to suggest that API access probably is restricted to a smaller index. Specialized search engines provide platforms to look for information in a broader context. Locating relevant search sources and seeking information in limited indexes individually is again a time consuming task. More recent attempts to add semantic element to search suffers from the limited scope of available ontologies. The semantic web vision based on the Resource Description Framework (RDF) and Web Ontology Language (OWL) is yet to gain popularity among common content providers and developers. Information systems are still a long way from reasonable data annotation in standardized formats.

Finally, all above automation attempts fall under the same query based active search model (referred to as pull model in introductory part) and only provide surface web search. All regular searchers and particularly knowledge workers feel a strong need to go beyond these restrictions.

[16] http://www.jux2.com/, http://www.dogpile.com/, http://clusty.com/
[17] http://www.ultraseek.com/, http://www.ferretsoft.com/
[18] Mashup & web 2.0 API portal: http://www.programmableweb.com
[19] Copernic search engine for desktop: http://www.copernic.com/
[20] http://scholar.google.com, http://academic.live.com
[21] http://citeseer.ist.psu.edu/, http://www.scirus.com/, http://ieeexplore.ieee.org
[22] http://swoogle.umbc.edu/, http://www.hakia.com/

3. From information retrieval to information supply

In an interview, Yahoo's Vice-President for research and technology describes the next generation search to be a "search without a box". This formulation indicates a move from information retrieval towards information supply, where information comes from multiple sources, in a given context, and without actively searching [Broder, 06]. For an effective information supply, understanding the context is very important. An ideal approach for context aware search would be the use of semantic inferences. However as we have mentioned earlier even after almost eight years --- this is how long the concept of semantic web has already been around---, there is no sign of mass implementation. Billions of web pages and documents still contain no or very few annotations and pieces of meta information. Thus, a mechanism is required to effectively finding the context of information and bridge the gap between conventional and semantic web. Context can be defined as effectively interpreting the meaning of a language unit at issue. An analysis of a large web query logs [Beitzel et al. 04] shows that average query length is 1.7 terms for popular queries and 2.2 terms averaged over all queries. This seems to indicate that input information is not enough to determine the context of a search.

We propose an information supply model for knowledge workers based on similarity detection. The proposed information supply method is utilized at the level where information seekers have an initial draft of their work available in written form. This input document is used to define the information need. It could be either the abstract, the initial draft of the task at hand or some document similar to the area of work currently carried out. The information supply engine performs the following processes to facilitate the search process:

3.1 Term extraction and normalization of variation

Terms can be seen as elements in a language to describe particular thoughts. They are the designators of concepts in any document. That is why automated processing of term recognition and extraction has been a critical aspect of natural language processing research. Term extraction approaches can be mainly categorized as statistical and linguistic. Statistical techniques identify terms and important phrases using factors such as consistency and structural location. Other approach makes use of linguistic morphology and syntactical analysis to identify terms of high importance. Substantial research in both techniques has transformed cutting edge research into usable applications and products. We can find examples (Yahoo Term Extraction23, Topicalizer and TermExtractor24) that successfully apply these methods to identify important terms highlighting concepts behind just textual information. Linguistic resources are used to find lexical variations in terms. Lexical variations (Synsets in WordNet25) are also used for canonical representation of information need. This normalized representation will be used for search result analysis at later stages. Terms can be extracted from knowledge work space on word count basis (fixed text chunks), paragraphs or complete documents. The extracted term batches will form queries to be processed by search services.

3.2 Determine the subject domain with the help of classification systems

In order to enhance the quality of search, additional meta data association can be very useful. There are several standardized classification initiatives in the form of taxonomies and topic maps. A subject domain can not only help to determine additional meta information to enrich search queries, but can also help in the selection of appropriate search services and results. Domain specific selections by adding this semantic element to information supply to search sources and results will reduce the "noise" (i.e. the undesirable information) generated.

3.3 Query expansion

Another approach is to use lexical enhancements. Cognitive synonyms and domain significant co-occurrences found with the help of lexical resources are used to expand queries. The idea behind lexical enhancements is to identify related information even if user defined terms and information

[23] http://developer.yahoo.com/search/content/V1/termExtraction.html

[24] http://www.topicalizer.com/, http://lcl2.di.uniroma1.it/termextractor/

[25] WordNet: Lexical database of English, http://wordnet.princeton.edu/

terms do not match. This expansion provides an improvement in classical search where matching is based on simple content match, vector space, link analysis ranking etc.A still further step is to move to advanced query formation. Cognitive terms and phrases are organized to form complex Boolean search queries. A query routing agent determines the use of AND, OR, phrase, include, exclude, and proximity operators for a specific search interface.

3.4 Distributed search services

Our proposed information supply engine maintains an up-to-date index of general search APIs and deep web resources. Knowledge workers may configure the information supply based on general web services or include domain specific deep web search. Almost every search engine provides web service interfaces and access to its index through the XML standard. Search services combine results of common web search engines, along with deep web search engines. The processed term batches are sent as queries to search interfaces available in distributed framework of system.

A recent book by Milosevic highlights the importance of distributed information retrieval systems. The suggested framework makes use of intelligent agents to establish coordination and apply filtering strategies [Milosevic 07]. In our information supply architecture we propose the use of distributed indexing and sharing to address the restriction of view issue of web search companies and access to deep web. On the basis of success of peer to peer file sharing applications a similar indexing and searching network is envisioned. The distributed nodes working at institutional or personal level provide open search access to deep web resources. Such distributed indexing approach can have numerous other applications; one example is illustrated in Collaborative Plagiarism Detection Network architecture [Zaka, 07][CPDNet, 07]. The Web presents a huge and continuously growing information base, but "search engines that crawl the surface of the web are picking up only a small fraction of the great content that is out there. Moreover, some of the richest and most interesting content can not even be crawled and indexed by one search engine or navigated by one relevancy algorithm alone" [OpenSearch, 07]. Open and distributed search initiatives provide common means of search result syndication from hundreds of shallow and deep web search engines.

3.5 Result analysis with the help of similarity detection

In conventional active search models the iterative process of obtaining and quickly examining the results provided by search engine consumes a lot of time. Information supply for knowledge worker provides an automated analysis of result using similarity detection techniques. The resulting URIs are fetched by an analysis module. The fetching module uses PHP's cURL library to extract textual contents and removes the formatting information (HTML tags, scripts etc.). Term extraction and lexical data services are used again to obtain the gist of contents. Similarity is calculated between information need and processed result contents with the help of vector space mathematics. The similarity detection service is fed with term batch of knowledge workspace and the terms of search result URIs. The terms/words are mapped as vectors against a compound term vocabulary. The dot product of two vectors (cosine similarity) determines the relevance. The semantic space for result analysis is built with the help of two additional services, namely POS (Part of Speech) tagger and Synonym [CPDNet 07]. POS tagger returns the sense of each word (i.e. Verb, Noun etc.) and Synonym service based on WordNet 3.0 returns the related synonyms and Synset IDs. The normalisation or canonical lexical representation of terms introduces a semantic relevance element in similarity detection process. A paper on power of normalized word vectors [Williams 06] presents the described concept in detail. Such analysis capability can replace typical result browsing and filtering in information seeking process.

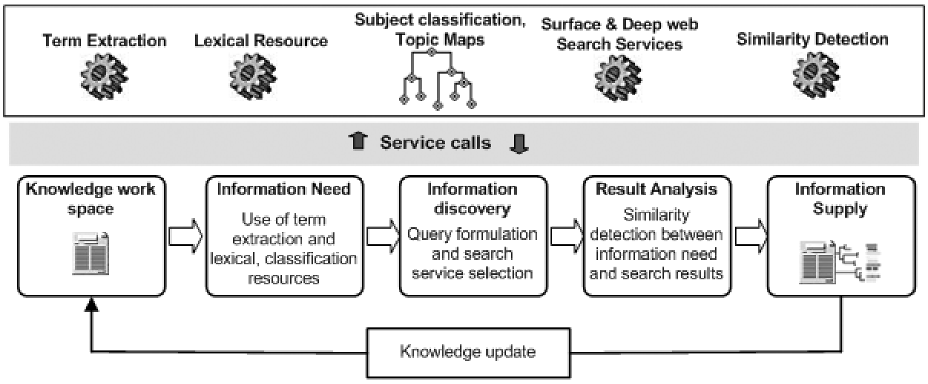
3.6 Result mapping to knowledge space

Key terms and phrases of the knowledge work space (input document) are linked with matching information sources. Result mapping with a rich web user interface provides a clear picture of relevant documents. The links include subject domain, key terms, persistent phrases, and summary of matching source. This function provides far better first hand knowledge about information source then the simple hit-highlighting as is done by a normal search engine. Users will have a higher degree of knowledge whether to further investigate matches suggested or not.

4. *Service oriented model*

The model shown in Figure 57 is a step towards the realization of a comprehensive information supply system. In order to provide the functions described in the previous section, we emphasize the use of off-the shelf tools in form of web services. Combination of external and internal services provides the depth in search and automation in information supply and analysis. The first level in the information supply environment is to determine the information need with the help of term extraction. The knowledge work space is submitted to term extraction services. A term extraction process in information supply model consists of converting formatted contents into plain text. The plain text is passed to a Part of Speech (POS) tagger in order to determine word sense and to eliminate function words. Key terms (verbs and nouns with higher occurrence frequencies) are further enriched using linguistic resources: WordNet and Wortschatz26 lexical databases are used to get synonym terms and synonym group IDs with similar sense. This data is used for query expansion and canonical lexical representation. The initial level also attempts to relate information requirement to a subject domain. Initially the subject domain is determined with the help of predefined standard taxonomies. One example is the use of ACM Computing Classification System27: the keyword index of each category can be checked for similarity with extracted terms. A web service developed to calculate angular measure among two word/term vectors mapped in compound term vocabulary is used for similarity check. [CPDNet 07]

Figure 57: Information supply model for knowledge workers


Term Extraction
Lexical Resource
Subject classification, Topic Maps
Surface & Deep web Search Services
Similarity Detection
Service calls
Knowledge work space
Information Need
Use of term extraction and lexical, classification resources
Information discovery
Query formulation and search service selection
Result Analysis
Similarity detection between information need and search results
Information Supply
Knowledge update

After formation of the consolidated information need, the system selects available search services. The mashup of search services provides a broader coverage of distributed indexes of the shallow and the deep web. A distributed search test bed [CPDNet 07] is used in the proposed information supply model. The distributed search framework acts as a proxy for search services. It not only includes popular search APIs like Google and Microsoft Live but also support OpenSearch and peer search. The search results are filtered at the next level with the use of already mentioned similarity detection techniques. Cosine similarity measure is determined from term vector of knowledge workspace and term vector of returned results from search services. The term vectors can be additionally presented in normalized form (canonical lexical representation) in order to develop semantic relevance. Filtered results with high angular similarity measure are linked with the knowledge workspace. Knowledge workers can see first hand similar data available on the internet. With an update of the knowledge space by incorporating new relevant documents found users can initiate the same process for an updated information supply.

[26] Wortschatz: http://wortschatz.uni-leipzig.de/
[27] ACM Classification: http://www.acm.org/class/

5. *Conclusion and further work*

In this paper we are suggesting the systematic use of web services for an information supply model. The model presented is a step forward from classical IR to proactive search systems. We introduce the use of lexical services and similarity detection to (i) find the context of a search and to map the information need to a specific subject domain, and (ii) provide an automated result scanning service, similar to human judgment. Both elements play an essential role in efficient information supply for knowledge workers. The research indicates that additional meta information found via context of search and lexical resources can prove to be very useful in the creation of automatic search queries. The use of mathematical techniques to determine information relevancy can also eliminate a time consuming and iterative process of manually scanning search results. Individual components/services of model available at CPDNet platform [CPDNet 07] and their performance to accurately link relevant resources provide promising avenue for complete implementation of presented information supply system.

Our experiments of search service mashup indicated very good possibilities of search access across multiple jurisdictions. The research on information retrieval in the internet and role of search engines also pointed out issues of restrictions in general web search APIs. The success of the next generation web or web 2.0 depends not only on the collaborative efforts from users but also on open and honest syndication and standard API provision by enterprises. The discontinuation of application consumable search service by Google, no availability of search syndication by specialized search engines like Google Scholar and Live Academia are examples of undesirable restrictions against which the community should protest before it is too late. There is a strong requirement for a scalable and open search and indexing platform. In order to develop such a platform, use of peer to peer search with user or institute level indexing is worth serious consideration.

References

[Beitzel et al. 04] S. M. Beitzel, E. C. Jensen, A. Chowdhury, D. Grossman "Hourly Analysis of a Very Large Topically Categorized Web Query Log", In Proc. of 2004 ACM Conf. on Research and Development in Information Retrieval (SIGIR-2004), Sheffield, UK, July 2004.

[Bergman, 01] M. K. Bergman. The deep web: Surfacing hidden value, http://www.press.umich.edu/jep/07-01/bergman.html (Accessed April 07, 2007).

[Broder, 06] A. Broder. “Search without a Box” (Interview). Yahoo! Search Blog. March 09, 2006; http://www.ysearchblog.com/archives/000262.html (Accessed April 2007)

[CPDNet, 07] Collaborative Plagiarism Detection Network (CPDNet) Web Services: Online at: http://www.cpdnet.org/nservices.php? (Accessed April 26, 2007)

[Feldman, 06] S. Feldman. “The Hidden Costs of Information Work” (IDC #201334, April 2006) http://www.idc.com/getdoc.jsp?containerId=201334 (Accessed April 2007).

[Hölscher & Strube, 00] C. Hölscher, G. Strube. "Web search behavior of internet experts and newbies" Computer Networks, Vol. 33, Issues 1-6, June 2000, Pages 337-346.

[Maurer & Tochtermann, 02] H. Maurer, K. Tochtermann. “On a New Powerful Model for Knowledge Management and its Applications” J.UCS vol.8., No.1, 85-96.

[McCown & Nelson, 07] F. McCown, M. L. Nelson. “Agreeing to Disagree: Search Engines and their Public Interfaces” Proc. of Joint Conference on Digital Libraries (JCDL) 2007.[Milosevic 07] D. Milosevic “Beyond Centralised Search Engines” An Agent-Based Filtering Framework,VDM Verlag Dr. Müller 2007 ISBN: 978-3-8364-1222-3

[OpenSearch, 07] Open Search: Documentation, Online at http://www.opensearch.org/ (Accessed April 26, 2007)

[Williams 06] R. Williams "The Power of Normalised Word Vectors for Automatically Grading Essays" The Journal of Issues in Informing Science and Information Technology Volume 3, 2006 pp. 721-730

[Zaka, 07] B. Zaka. “Empowering plagiarism detection with web services enabled collaborative network” submitted for publication

Section 15: Establishing an Austrian Portal

(This chapter is based on work mainly done by H. Maurer and P. Diem from a conceptual point of view and Nick and Alexeii Sherbakov from a technical point of view)

Following the ideas that are proposed on a European level in Section 11, i.e. building dedicated search engines that would be better in their area of speciality we have tried to build a server with the aim to eventually have THE portal to all important information on Austria, and experiences of Austrians, inside or outside their home-country.

The server has been operational since Sept. 24, 2007. We first present some of the main features of the server and then add, for completeness sake, the the press-release sent out Sept.24, 2007 day and the mission statement of the server (both of course in German).

The server consists of two parts.

The first is an archive of contributions on Austria that do not exist in this form elsewhere, including the collection of some 20.000 entries collected in a server www.aeiou.at, most of those both in English and German. This archive will be increased by a staff of volunteer editors, most of them still to be recruited: those editors will be specialists in their field, will be known by name and with their CV available on the server for information.

The second part is a collection of contributions from the community. The community consists of everyone interested in reading or contributing to the server. Reading can be done without registration, for contributions members of the community have to register, but can choose an arbitray user name, need only to reveal a working email to the system (needed for sending them a password and providing the possibility to block SPAM or unwanted contributions) but can specify that their E-Mail address is not divulged, i.e. that they can remain “anonymous” within the community.)

Registered users and editors can comment and grade any contribution (and the comments can lead to potentially long strings of arguments, even encouraging authors to change or withdraw their contribution), but only authors themselves can change their contribution, in marked difference to Wikipedia. Further, contributions of editors will be “frozen” after three months: this means they are authored by known person, and stable in time, i.e. can be quoted like a scientific paper. Further, contribution of the community that receive much applause from the community (in terms of comments and grades) can be encouraged to permit that their contributions are moved to the fist part of the server: an editor has to vouch for the quality. Such contribution appear under the name the user chooses with the addition: “Approved by editor xyz”.

The aim of the above separation is to (a) tie in the community as much as possible and (b) to make sure that a good part of the data-base is of reliable high quality.

The system, austria-forum.org, also provides novel search facilities, user selected profiles (to support beginners as well as experts), a host of communication facilities, an other technical innovations that will be part of more technical contributions, elsewhere. It is hoped that anyone who wants to find information on “Austriacas” (items related to Austria) will first turn to this entry point in WWW, rather than to other sources or search engines. Or, turning it around, we hope that looking for an item on Austria in any search-engine, the entry in the Austria-Forum will be ranked fairly high.

Press release
Graz, September 24, 2007

http://austria-forum.org
Neue Wissensplattform mit digitalem Österreich-Lexikon gestartet

Auf dem Server des Instituts für Informationssysteme und Computermedien der TU Graz wurde heute der offizielle Startschuss für die neue österreichische Wissensplattform http://austria-forum.org gegeben.

Die auf die Möglichkeiten des Web 2.0 abgestimmte innovative Internet-Plattform umfasst folgende Teile:

1. *Das immer wieder zur Seite gedrängte, aber weiterhin informationstaugliche Österreich-Lexikon www.aeiou.at in verbesserter Form.*
2. *Eine Wissensplattform für Forschung und Lehre, bestehend aus gezeichneten Beiträgen mit starkem Österreichbezug und einem Schwerpunkt „politische Bildung".*
3. *Eine wissensorientierte Internet-Community - vorwiegend für den Gebrauch durch Wissenschaftler, Lehrer, Schüler und Studenten.*
4. *Zugang zu wichtigen Informationen über Österreich.*

Herausgegeben wird das Austria-Forum von dem weit über die Grenzen Österreichs hinaus bekannten Informatiker Univ.-Prof. Dr. Hermann Maurer, zurzeit Dekan an der TU Graz. Mitherausgeber sind die beiden Publizisten Dr. Trautl Brandstaller und Dr. Peter Diem. Der Qualitätssicherung dient ein hochkarätig besetzter Beirat, dem bereits jetzt der Rektor einer österreichischen Universität und mehrere Professoren aus dem Ausland, darunter Vertreter der Universität Stanford, angehören.

Prof. Maurer erläutert den besonderen Charakter der neuen Internetplattform:

„Durch den starken Österreichbezug und die gesicherte Qualität des Inhalts entsteht hier ein besonders für den Schulunterricht und das Universitätstudium geeignetes Werkzeug der Allgemeinbildung mit einem verlässlichen digitalen Nachschlagwerk. Da die redaktionellen Inhalte als Fachbeiträge gezeichnet sind und nach drei Monaten, die der Community für Diskussionen und Anregungen zur Verfügung stehen, gesperrt werden, sind diese Beiträge des Austria-Forums im Gegensatz zur „Wikipedia" auch wissenschaftlich zitierbar. Es wird also hier erstmals das „Wissen der Experten" mit dem „Wissen der Vielen" sinnvoll kombiniert. Damit wird ein Ziel des kalifornischen Internet-Experten Andrew Keen („Wir wollen keine Diktatur der Experten. Aber auch keine der Amateure.") erstmals verwirklicht.

Die Möglichkeit, die Textbeiträge und das reiche Bildangebot zu kommentieren und zu bewerten sowie der ausführliche Community-Teil des Forums tragen der neuesten Entwicklung des Internets hin zur Interaktivität (Web 2.0) Rechnung. Durch die Einrichtung geschlossener User-Gruppen können etwa Schulklassen ihre Projektarbeit im Austria-Forum online durchführen und dokumentieren. Auch andere Neuerungen verdienen Beachtung: etwa die Tatsache, dass alle Benutzer jederzeit feststellen können, welche Beiträge sie noch nicht gelesen haben."

Auch in der gegenwärtigen Aufbauphase enthält das Austria-Forum über 20.000 Beiträge bzw. wertvolle Links. Dies soll sich bis Ende 2008 auf über 50.000 erhöhen. Bis Jahresende besteht für die Internet-Öffentlichkeit die Gelegenheit, mit der Version 1.0 des Austria-Forums zu experimentieren, wofür nur die eine einfache Registrierung auf der Basis einer gültigen E-Mail-Adresse erforderlich ist.

Die Website ist und bleibt allgemein und kostenlos zugänglich über die URL http://austria-forum.org; die Adresse für Erfahrungsberichte und Anregungen der Besucher der Site lautet office@austria-

forum.org bzw. können Anregungen und Diskussionen zu Vorschlägen auch direkt im System geführt werden. Erfahrungen und Benutzerwünsche werden in weitere Versionen 2008 Schritt für Schritt eingebracht werden.

„Das Austria-Forum wird im Laufe der Zeit DIE zentrale Anlaufstelle für Informationen über Österreich werden", ist das hochgesteckte Ziel dieses Unterfangens.

The Austria-Forum
Mission Statement

Das Austria-Forum ist eine von unabhängigen Wissenschaftern und Publizisten gestaltete und der Allgemeinheit zur Verfügung gestellte Wissens- und Diskussionsplattform mit Schwerpunkt Österreich. Inhaltlich orientiert sich das Austria-Forum am Gedanken des Universal-Lexikons und der Mitgestaltung durch volksbildnerisch engagierte Benutzer. Staatsbürgerliches Wissen und allgemein bildende Inhalte stehen daher im Vordergrund. Die Herausgeber und Editoren sind politisch unabhängig und bekennen sich zur parlamentarischen Demokratie, zum Rechtsstaat und zu den Menschenrechten.

Struktur des Austria-Forums

Das Austria-Forum besteht im Wesentlichen aus dem Österreich-Angebot (weiter auch unter http://www.aeiou.at abrufbar), aus einem „wissenschaftlichen" und einem "Community-Bereich". Die Beiträge des wissenschaftlichen Bereichs werden vom Herausgeber-Team und den Editoren unter Mitwirkung eines Wissenschaftlichen Beirats verfasst, wobei die Beiträge auch nur aus Links zu anderen Websites bestehen können. Im Community-Bereich können registrierte Benutzer unter selbst gewählten Benutzernamen ohne Bekanntgabe ihrer eigentlichen Identität Beiträge verfassen. Sie können auch Kommentare und Bewertungen zu allen Beiträgen abgeben und sich an Diskussionen beteiligen.

Community-Beiträge können, wenn sie von den Herausgebern, den Editoren, den wissenschaftlichen Beiräten - aber auch von einer Vielzahl von Benutzern - als besonders interessant eingestuft werden, in den wissenschaftlichen Teil kopiert werden. Beiträge können auf Grund der Kommentare, Bewertungen und Diskussionen von den Autoren oder Editoren abgeändert werden.

Das Austria-Forum bietet diverse Kommunikationsmöglichkeiten zwischen den Benutzern. Insbesondere können Benutzer die meisten der angebotenen Funktionen auch in "geschlossenen Benutzergruppen" verwenden. Eine solche User-Gruppe wird dadurch eingerichtet, dass ein angemeldeter Benutzer andere Benutzer namentlich einlädt. Damit können die so zusammengefassten Personen (Clubs, Freundeskreise, Schulklassen ...) Informationen für sich zusammen tragen bzw. Diskussionen führen, ohne dass dies für andere sichtbar ist.

Im Hinblick auf die Erfordernisse des Web 2.0 enthält das Austria-Forum eine Reihe wichtiger Innovationen, die das Auffinden von gewünschten Informationen stark erleichtern. Neben einer Suchfunktion mit neuartigen Möglichkeiten kann man auch über hierarchische und lineare Themenlisten nach Informationen suchen.

Teilnahmemöglichkeit

Die Inhalte des Austria-Forums stehen allen Internetteilnehmern zur Verfügung. Sie können unter bestimmten Bedingungen kopiert und beliebig verwendet werden.

Die Abfassung von Beiträgen, die Abgabe von Kommentaren, das Einfügen von Links und die Bewertung von Beiträgen ist nur für registrierte Nutzer möglich. Diese können dabei beliebige Benutzernamen verwenden. Die Registrierung dient ausschließlich zur Vermeidung von SPAM und zur allfälligen Verfolgung von Beiträgen, die urheberrechtlich oder strafrechtlich bedenklich sind.

Wer sich dazu berufen fühlt, als Editor Beiträge auch im wissenschaftlichen Bereich zu verfassen, muss sich namentlich ausweisen, seinem Ansuchen einen kurzen Lebenslauf beifügen und den besonderen Nutzungsbedingungen ausdrücklich zustimmen. Das Herausgeber-Team kann ein solches Ansuchen ohne Begründung abweisen.

Von den Editoren wird erwartet, dass sie regelmäßig Beiträge liefern. Gedacht ist an eine Mitarbeit im Ausmaß von mindestens einer Stunde pro Woche. Spezialisten, die nur in Ausnahmefällen über die Qualität von Beiträgen befragt werden wollen, können sich analog als Mitglieder des wissenschaftlichen Beirates beim Herausgeber-Team bewerben.

Verlinkungen

Die Möglichkeit, weiterführende Internetangebote durch klickbare Verweise (Hyperlinks) zugänglich zu machen, gehört zu den großen Vorzügen des Internets. Auch im Austria-Forum wird davon in vollem Umfang Gebrauch gemacht.

Die Herausgeber des Austria-Forums betonen jedoch ausdrücklich, dass sie keinerlei Einfluss auf die Gestaltung und die Inhalte der verlinkten externen Seiten haben. Deshalb lehnen sie jede Verantwortung für die Einhaltung der gesetzlichen Vorschriften und Internet-Regeln seitens der Betreiber verlinkter externer Seiten inklusive aller Unterseiten ausdrücklich ab.

Werbung

Um die Selbstkosten zumindest teilweise abdecken zu können, behalten sich die Betreiber vor, Werbeeinschaltungen zu akzeptieren. Dem Austria-Forum angebotene Werbung muss jedoch sachlich sein und sich in unaufdringlicher Weise dem Seitendesign angliedern lassen.

Section 16: References for Sections 1-5

[Anderson 2007] Chris Anderson. The Long Tail. Der lange Schwanz. Nischenprodukte statt Massenmarkt. Das Geschäft der Zukunft. München: Hanser, 2007.

[Bager 2007] Jo Bager. "Der persönliche Fahnder. Gezielt suchen mit individuellen Suchmaschinen". c't 1, 2007, pp. 178-183.

[Batelle 2005] John Batlle. The Search. Peguin, 2005

[Bubnoff 2005] Andreas von Bubnoff. "Science in the Web Age: The Real Death of Print". Nature, vol. 438, December 2005, pp. 550-552. http://www.nature.com/nature/journal/v438/n7068/full/438550a.html [visited 17/5/07]

[Coy 2002] Wolfgang Coy. "Computer Augmented Research and Scientific Misconduct". Proceedings of the IFIP 17th World Computer Congress, pp. 131-146. http://waste.informatik.hu-berlin.de/Tagungen/QUALIS/Coy.pdf [visited 17/5/07]

[Dobusch & Forsterleitner 2007] Leonhard Dobusch and Christian Forsterleitner (eds.). Freie Netze. Freies Wissen. Vienna: Echo media, 2007.

[Fedler 2006] Fred Fedler. "Plagiarism Persists in News Despite Changing Attitudes". Newspaper Research Journal, vol. 27, no. 2, Spring 2006, pp. 24-37.

[Giesecke 2007] Michael Giesecke. Die Entdeckung der kommunikativen Welt. Studien zur kulturvergleichenden Mediengeschichte. Frankfurt: Suhrkamp, 2007.

[Gleich 2002] Michael Gleich. Web of Life. Die Kunst, vernetzt zu leben. Hamburg: Hoffmann und Campe, 2002.

[Haber 2005] Peter Haber. "'Google-Syndrom'. Phantasmagorien des historischen Allwissens im World Wide Web". Geschichte und Informatik/Histoire et Informatique. Vom Nutzen und Nachteil des Internet für die historische Erkenntnis, vol. 15, pp. 73-89. http://www.hist.net/datenarchiv/haber/texte/105742.pdf [visited 17/5/07]

[IG Kultur 2007] IG Kultur (ed.). "Plagiarismus und Ideenklau". Kulturrisse 1, march 2007. Vienna.

[Jeanneney 2006] Jean-Noël Jeanneney. Googles Herausforderung. Für eine europäische Bibliothek. Berlin: Wagenbach, 2006.

[Johnson 1999] Steven Johnson. Interface Culture. Wie neue Technologien Kreativität und Kommunikation verändern. Stuttgart: Klett-Cotta, 1999.

[Johnson 2005] Steven Johnson. Everything Bad is Good for You: How Popular Culture is Making Us Smarter. London: Allen Lane, 2005.

[Keel & Bernet 2005] Guido Keel and Marcel Bernet. IAM/Bernet-Studie "Journalisten im Internet 2005". Eine Befragung von Deutschschweizer Medienschaffenden zum beruflichen Umgang mit dem Internet. http://www.iam.zhwin.ch/download/Studie_2005.pdf [visited 17/5/07]

[Keen 2007] Andrew Keen. The Cult of the Amateur: How today's Internet Is Killing Our Culture. Doubleday, 2007).

[Kroker & Weinstein 1994] Arthur Kroker and Michael A. Weinstein. Data Trash. The Theory of the Virtual Class. Montreal: New World Perspectives.

[Kulathuramaiyer & Balke 2006] Narayanan Kulathuramaiyer and Wolf-Tilo Balke. "Restricting the View and Connecting the Dots – Dangers of a Web Search Engine Monopoly". Journal of Universal Computer Science, vol. 12, no. 12 (2006), pp. 1731-1740. http://www.l3s.de/~balke/paper/jucs06.pdf [visited 17/5/07]

[Kulathuramaiyer & Maurer 2007] Narayanan Kulathuramaiyer and Hermann Maurer. "Why is Fighting Plagiarism and IPR Violation Suddenly of Paramount Importance?" Learned Publishing Journal vol.20, No.4. (October 2007), pp.252-258.

[LaFollette 1992] Marcel C. LaFollette. Stealing Into Print. Fraud, Plagiarism, and Misconduct in Scientific Publishing. Berkeley, Los Angeles, and London: University of California Press, 1992.

[Lehmann & Schetsche 2005] Kai Lehmann and Michael Schetsche (eds.). Die Google-Gesellschaft. Vom digitalen Wandel des Wissens. Bielefeld: Transcript, 2005.

[Leißner 2006] Silvia Leißner. Wikipedia: informativ oder qualitativ bedenklich? Eine theoretische und inhaltsanalytische Untersuchung zur Informationsqualität der Einträge in der freien Internet-Enzyklopädie Wikipedia. Master thesis, TU Ilmenau.

[Lindner 2007] Roland Lindner. "Der Besserwisser. Sein Online-Lexikon Wikipedia kennt jeder. Jetzt will er mit einem Suchprogramm Google angreifen – und endlich Geld verdienen." Frankfurter Allgemeine Sonntagszeitung, 21 January 2007, p. 46.

[Lorenz 2006] Maren Lorenz. Wikipedia. Zum Verhältnis von Struktur und Wirkungsmacht eines heimlichen Leitmediums. WerkstattGeschichte, no. 43 (2006), pp. 84-95. http://www.phil-gesch.uni-hamburg.de/hist/hsperson/lorenz13.pdf [visited 17/5/07]

[Machill & Beiler 2007] Marcel Machill and Markus Beiler (eds.). Die Macht der Suchmaschinen/The Power of Search Engines. Köln: Herbert von Halem, 2007.

[Maurer & Kappe & Zaka 2006] Hermann Maurer, Frank Kappe, and Bilal Zaka. "Plagiarism – A Survey". Journal of Universal Computer Science, vol. 12, no. 8 (2006), pp. 1050-1084. http://www.jucs.org/jucs_12_8/plagiarism_a_survey/jucs_12_08_1050_1084_maurer.pdf [visited 17/5/07]

[Maurer & Zaka 2007] Hermann Maurer and Bilal Zaka. "Plagiarism – a problem and how to fight it". Submitted Paper. http://www.iicm.tu-graz.ac.at/iicm_papers/plagiarism_ED-MEDIA.doc [visited 17/5/07]

[Maurer 2007a] Hermann Maurer. "Google: a serious danger, a way to combat it, and solving another big problem as by-product." Unpublished paper, 5 pages.

[Maurer 2007b] Hermann Maurer. "Google – Freund oder Feind?" Unpublished paper, 7 pages.

[Maurer 2007c] Hermann Maurer. "Dossier on Data Mining and Search Engines: A Serious Danger to Society and Economy that Needs Decisive Action." Unpublished paper, 5 pages.

[Meyer zu Eissen & Stein 2006] Sven Meyer zu Eissen and Benno Stein. "Intrinsic Plagiarism Detection". Lalmas et al. (eds.). Advances in Information Retrieval. Proceedings of the 28th European Conference on IR Research, ECIR 2006, London, pp. 565-568. http://www.uni-weimar.de/medien/webis/publications/downloads/stein_2006d.pdf [visited 17/5/07]

[Möller 2005] Erik Möller. Die heimliche Medienrevolution. Wie Weblogs, Wikis und freie Software die Welt verändern. Telepolis. Hannover: Heise, 2005.

[N. N. 2006a] N. N. "7 things you should know about... Google Jockeying". http://www.educause.edu/ir/library/pdf/ELI7014.pdf [visited 17/5/07]

[N. N. 2006b] N. N. "Wikipedia Critic Finds Copied Passages", AP report, 3 November 2006.

[N. N. 2007a] N. N. "A Stand Against Wikipedia", Inside Higher Ed News, http://insidehighered.com/news/2007/01/26/wiki [visited 17/5/07]

[N. N. 2007b] N. N. "'Google ist ein großer Geheimniskrämer'". Stuttgarter Zeitung, 3 February 2007. http://stuttgarterzeitung.de/stz/page/detail.php/1351016 [visited 17/5/07]

[N. N. 2007c] N. N. "Wikia: Größter Medienkonzern der Welt entsteht im Internet", ddp report, 17 May 2007.

[Pfeifer 2007] David Pfeifer. Klick. Wie moderne Medien uns klüger machen. Frankfurt and New York: Campus, 2007.

[Rötzer 2007] Florian Rötzer. "Google will ein bisschen weniger böse sein". Telepolis, 2007, online only. http://www.heise.de/tp/r4/artikel/24/24854/1.html [visited 17/5/07]

[Salchner 2007] Christa Salchner. "Nicht einmal über Google zu finden. Aspekte des Lebens im Google-Zeitalter." Telepolis, 2007, online only. http://www.heise.de/tp/r4/artikel/24/24720/1.html [visited 17/5/07]

[Schetsche 2005] Michael Schetsche. "Die ergoogelte Wirklichkeit. Verschwörungstheorien und das Internet." Telepolis, 2005, online only. http://www.heise.de/tp/r4/artikel/19/19964/1.html [visited 17/5/07]

[Stein & Meyer zu Eissen 2006] Benno Stein and Sven Meyer zu Eissen. "Near Similarity Research and Plagiarism Analysis". Spiliopoulou et. al (eds.). From Data and Information Analysis to Knowledge Engineering. Selected Papers from the 29th Annual Conference of the German Classification Society, 2006, Magdeburg, pp. 430-437. http://www.uni-weimar.de/medien/webis/publications/downloads/stein_2006a.pdf [visited 17/5/07]

[Surowiecki 2004] James Surowiecki. Doubleday, 2004.

[Tapscott & Williams 2006] Don Tapscott and Anthony D. Williams. Wikinomics. How Mass Collaboration Changes Everything. New York: Portfolio, 20006.

[Tufte 2006] Edward R. Tufte. The Cognitive Style of PowerPoint: Pitching Out Corrupts Within. Second edition. Cheshire: Graphics Press, 2006.

[Vise 2005] David A. Vise. The Google Story. Pan books, 2005.

[Weber 2005a] Stefan Weber. "Mit Shake and Paste ans Ziel. Krise der Kulturwissenschaften angesichts des grassierenden Plagiarismus". Telepolis, 2005, online only. http://www.heise.de/tp/r4/artikel/19/19921/1.html [visited 17/5/07]

[Weber 2005b] Stefan Weber. "Kommen nach den 'science wars' die 'reference wars'? Wandel der Wissenskultur durch Netzplagiate und das Google-Wikipedia-Monopol". Telepolis, 2005, online only. http://www.heise.de/tp/r4/artikel/20/20982/1.html [visited 17/5/07]

[Weber 2006a] Stefan Weber. So arbeiten Österreichs Journalisten für Zeitungen und Zeitschriften. Salzburg: Kuratorium für Journalistenausbildung, 2006.

[Weber 2006b] Stefan Weber. "Textueller Missbrauch. Plagiarismus, Redundanz, Bläh-Rhetorik: Zur Krise der Kulturwissenschaften durch den Einzug des Copy/Paste-Paradigmas – Teil 1". Telepolis, 2006, online only. http://www.heise.de/tp/r4/artikel/24/24006/1.html [visited 17/5/07]

[Weber 2006c] Stefan Weber. "Die abschreibende Zunft. Neue Fälle von dreistem Textklau stellen die wissenschaftliche Selbstkontrolle in Frage – Report eines akademischen Whistleblowers und 'Plagiatsjägers' – Teil 2". Telepolis, 2006, online only. http://www.heise.de/tp/r4/artikel/24/24110/1.html [visited 17/5/07]

[Weber 2006d] Stefan Weber. "Wissenschaft als Web-Sampling. Wie an Universitäten in Windeseile eine Textkultur ohne Hirn entstanden ist – Teil 3". Telepolis, 2006, online only. http://www.heise.de/tp/r4/artikel/24/24221/1.html [visited 17/5/07]

[Weber 2007a] Stefan Weber. Das Google-Copy-Paste-Syndrom. Wie Netzplagiate Ausbildung und Wissen gefährden. Telepolis. Hannover: Heise, 2007.

[Weber 2007b] Stefan Weber. "Contentklau in Blogs und anderswo. Was hat das Web 2.0 mit dem Mittelalter zu tun? Teil IV der Serie zum Google-Copy-Paste-Syndrom". Telepolis, 2007, online only. http://www.heise.de/tp/r4/artikel/24/24756/1.html [visited 17/5/07]

[Weber 2007c] Stefan Weber. "Reuse, Remix, Mashup – also: Plagiieren erlaubt! Der Hype um freie Lizenzen könnte für die Textkultur fatale Folgen haben – Teil V zum Google-Copy-Paste-Syndrom". Telepolis, 2007, online only. http://www.heise.de/tp/r4/artikel/24/24771/1.html [visited 17/5/07]

[Weber 2007d] Stefan Weber. "'Die besten Kopisten konnten nicht lesen'. Der Philosoph Konrad Paul Liessmann über Wikis, die Zukunft der Schrift und das Ende der Bildungsidee". Telepolis, 2007, online only. http://www.heise.de/tp/r4/artikel/24/24927/1.html [visited 17/5/07]

[Weber 2007e] Stefan Weber. "Vom Wissensfortschritt mit stummem h. Die wundersame Wanderung einer Biographie im Netz." Telepolis, 2007, online only. http://www.heise.de/tp/r4/artikel/25/25137/1.html [visited 17/5/07]

[Weber 2007f] Stefan Weber. "Schon mal was von Text Jockeys und Powerpoint Karaoke gehört? Jugendmedienkulturen – Kulturtechniken – Wissenskultur: Skizze einer Revolution in den Köpfen und Apparaten". To appear in Medienimpulse – Zeitschrift für Medienpädagogik, June 2007.

[Weber-Wulff 2004] Debora Weber-Wulff. "Fremde Federn Finden". Online Course. http://plagiat.fhtw-berlin.de/ff [visited 17/5/07]

[Wegner 2005] Jochen Wegner. "Die große Googleisierung". Cover # 5, Spring 2005, pp. 78-81.

[Winkler 1997] Hartmut Winkler. "Suchmaschinen. Metamedien im Internet". In: Barbara Becker and Michael Paetau (eds.). Virtualisierung des Sozialen. Die Informationsgesellschaft zwischen Fragmentierung und Globalisierung. Frankfurt and New York: Campus, 1997, pp. 185-202.

[Witten & Gori & Numerico 2007] Ian H. Witten, Marco Gori, and Teresa Numerico. Web Dragons. Inside the Myths of Search Engine Technology. San Francisco: Morgan Kaufmann, 2007.

Web sites (critical blogs, watchblogs, etc., all visited 17/5/07)

http://www.jmboard.com/gw

http://www.google-watch.org

http://www.wikipedia-watch.org

http://andrewkeen.typepad.com

http://copy-shake-paste.blogspot.com

http://weblog.histnet.ch

http://www.regrettheerror.com

Appendix 1: A Comparison of Plagiarism Detection Tools

Hermann Maurer, Bilal Zaka and Frank Kappe

Institute for Information Systems and Computer Media
Graz University of Technology
Austria
{hmaurer, bzaka, fkappe}@iicm.edu

Abstract: The size and availability of online contents is growing, this increases the possibilities and ease of theft of online textual contents. Same online information retrieval techniques are being used to create tools that detects internet assisted plagiarism. In this paper we implement a test bed to empirically compare major plagiarism detection services. Our experiments show response of various services to carefully selected research articles from various sources. The whole exercise illustrates importance of using such tools to facilitate teachers swiftly deal with blatant copy paste problem. The presented results are meant to be a particle guide for educators and publishers interested in using high tech plagiarism detection tools.

1. *Introduction*

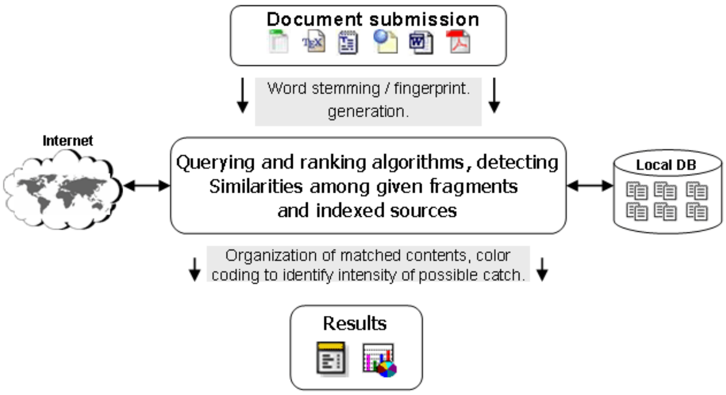
Digital age reformed information access tremendously; educational and intellectual sectors are revolutionized with Internet and World Wide Web. Almost every written resource recently produced is now available in digitized form. Looking at books and library digitization projects initiated by Google, Microsoft and Yahoo, one can easily assume that in coming decade all the contents will be indexed and available electronically. Most probably the best tool to acquire information by any researcher these days is internet and it is very difficult to imagine research without it. This ease and power of access also increased the possibilities and temptation to copy or reuse the contents, which if done without proper acknowledgement to original creator constitutes plagiarism. Defining plagiarism covering all its aspects may require effort and space beyond the scope of this paper; however our recent publication on plagiarism (Maurer et al. 2006) can be a useful resource to develop an understanding about the problem. The gravity of problem is signified by various studies and surveys, a survey by “The Center of Academic Integrity” (McCabe, 2005) indicate that approximately 70% of students admit to some cheating while 40% admits to internet assisted plagiarism. The trends illustrated in mentioned survey and others, show that this percentage is increased in more recent times. A statistical report by US department of education (DeBell & Chapman, 2006) presents that 80% of students starts using computer and 56% internet, right from the primary level of schooling; this percentage is 97% computer and 79% internet for high school students. Report further shows that approximately 50% of students use internet to complete school assignments. These results reflect statistics collected in 2003, one can safely assume a lot larger percentage (may be 100%) in graduate or post graduate level at present.

These figures rightfully present an alarming situation and academia is responding with various reactive and proactive measures. Proactive measures include introduction of honor codes, educating the students about intellectual honesty and integrity, writing guidelines and courses. Reactive measures are comprised of introduction of detection mechanism, enforcing penalties. A balance of both types of preventive measures while dealing with plagiarism is considered to be the most effective one. In this article we focus our attention to internet assisted plagiarism detection tools which are commonly in use by academia all across the glob to fight cut paste plagiarism. While our experiments revealed the fact that they might not be an ultimate and definite answer to deal with plagiarism, there is no doubt in their effectiveness to facilitate a quick validity/originality scan of larger number of documents which otherwise would be very time consuming task. We also observe that these scans and results some times bring false positives and some times completely pass through a clear plagiarized writing. While there is room for improvement in existing plagiarism detection tools there is defiantly no doubt in deterrence potential they posses.

2. *Detection Tools*

A number of free and fee based services are available to detect plagiarism. These tools are available as online services or downloadable applications. Several detection techniques are being applied in different software including characteristic phrase search, statistical comparison of documents and writing style analysis; but most common and widely used technique in major services is the document source comparison. A graphical representation of this method is shown in figure 1, taken from our recent survey (Maurer et al. 2006).

Figure 1: Plagiarism detection with document source comparison


Document submission
Word stemming / fingerprint. generation.
Internet
Querying and ranking algorithms, detecting Similarities among given fragments and indexed sources
Local DB
Organization of matched contents, color coding to identify intensity of possible catch.
Results

Services based on this approach take documents from user, process them to create and store document sources comprising of predefined pieces of text chunks (fingerprints) in local index server. These fragments of text are then compared with already available sources in local index server and internet for any similarity. Results comprised of detected matches are then presented to user for further investigation.

For our analysis we selected Turnitin[28] and its close rival Mydropbox[29] service based on common usage among educational institutes. Turnitin is a web based service by iParadigms LLC, its distribution in UK via Joint Information Systems Committee (JISC 2006) and early operation since 1996 makes it the most widely used service in universities and schools around the world. Turnitin claims to have proprietary algorithms to search extensively indexed 4.5 billion internet resources, 10 million already submitted documents, and ProQuest™ database. SafeAssignment web service from Mydropbox is another popular service and closest rival to Turnitin. This system state to have searching capabilities over 8 billion internet documents, scholastic article archives such as ProQuest™, FindArticles™, and 300,000 articles offered by paper mills.

Since the plagiarist usually use internet to rip contents for the writings in question and according to recent surveys by comScore (comScore, 2006) and Nielsen NetRatings (Nielsen, 2006) Google dominates as mostly used tool to query internet, we developed a system of our own to be added to plagiarism detection tool analysis. This system is based on Google search API and follows the generic process of any document source analysis program. Documents to be checked are submitted using a web portal, where document is converted to a plain text with contents fragments of a moderate size. This word stemming is kept to a moderate size keeping in mind the query limit of Google API. When a document is set to be checked for plagiarism system queries the search engine for each fragment and returned results are analyzed for best match. The matching algorithm of the system uses vector space

[28] http://www.turnitin.com
[29] http://www.mydropbox.com

model to determine similarity between returned results from search engine and queried fragment of document source. The measure of similarity is calculated by converting corresponding result snippet and queried fragment into word vectors and computing cosine similarity of the two vectors (Widdows, 2003). This measure constitutes a percentage of detected similarity; results from all the fragments of a document source are combined to form a total percentage of plagiarism in a document with links to online sources. The results generated by all three services are composed of reformatted contents of submitted document with links to found matches. This linking is colored to show intensity of match, the report also present URLs that contains highest similarities. Generally the evaluator is then required to spend some amount of time to determine the validity of scores. This is done by looking at the document source carefully for quotations and reference markings and contents of linked URLs.

3. Test data

To see the response of these services we selected collection of documents from various sources, these sources include class assignments submitted by students of informatics at Technical University Graz ESA 2006, research articles published at various renowned scholastic databases like ACM[30], IEEE Xplore[31], SpringerLink[32], J.UCS[33].

Table 1: Data collections for plagiarism detection service analysis

Collection	Type	Source	Expected Results
A	Unpublished	Student Assignments	Results are required to present information about plagiarism in documents and original sources
B	Published	ACM	System should detect the published source (100% or very high degree of plagiarism, since the work already exists) or any further sources where the contents are available, used or taken from.
C	Published	IEEE Xplore	System should detect the published source (100% or very high degree of plagiarism, since the work already exists) or any further sources where the contents are available, used or taken from.
D	Published	JUCS	System should detect the published source (100% or very high degree of plagiarism, since the work already exists) or any further sources where the contents are available, used or taken from.
E	Published	SpringerLink	System should detect the published source (100% or very high degree of plagiarism, since the work already exists) or any further sources where the contents are available, used or taken from.

The documents were submitted to three services, namely Turnit, Mydropbox and Google API based.

4. Results

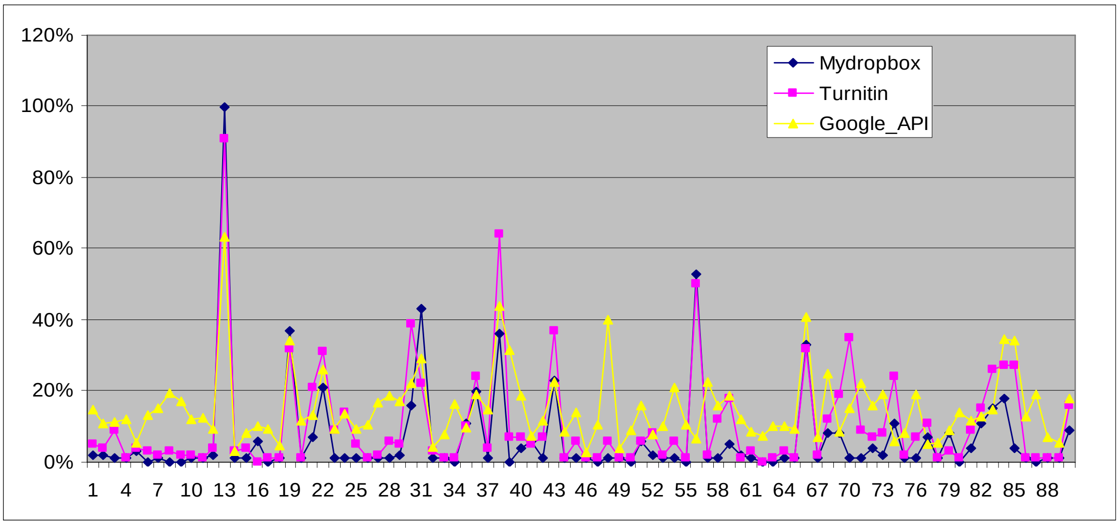
All three systems take some time to process the submitted documents, the turn around time of Turnitin and Mydropbox varies from few minutes to several hours depending on the quantity of documents submitted to system in a single session or within certain period of time or the processing load on the

[30] http://portal.acm.org
[31] http://ieeexplore.ieee.org
[32] http://www.springerlink.com
[33] http://www.jucs.org

main detection engine. In case of self developed Google API based service results are produced real time, only restriction is number of search queries to Google per day for an individual API code. This limits us to process only few papers on daily basis. An individual user with limited number of papers to check may not face any hurdle with this system when the service is used with personal API code.

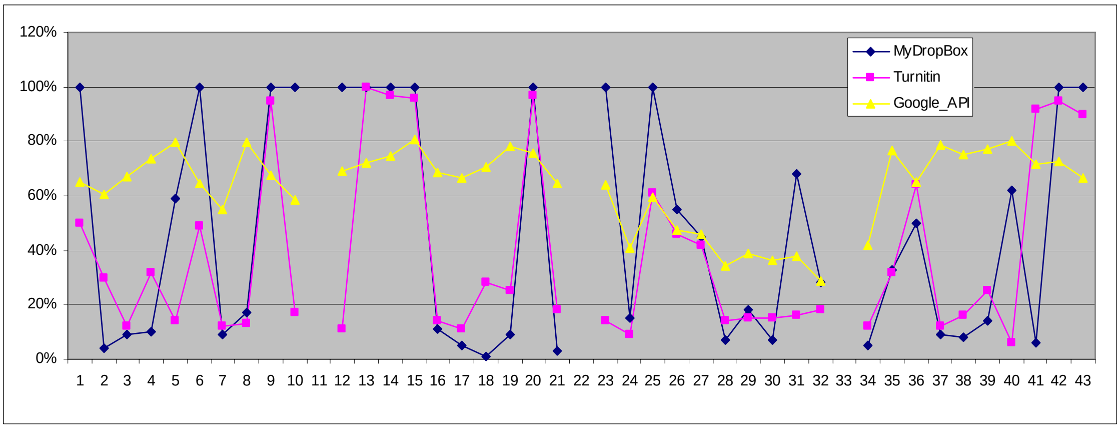
In general all three systems responded with almost similar results for collection A (Figure 2), the generated reports show identical original internet resources for copy incidents found in papers. The composition and weight distribution of plagiarized contents from different original sources also shows identical tendencies. The overall percentage calculated varied in different services because of the difference in weight assignments and calculation factors.

Figure 2: Plagiarism percentage graph for collection A



The response of systems for collection B, C, D and E is very important particularly for higher education institutes. The systems having access to school level essays and paper mill databases might have very less influence on detection performance when it is dealing with higher education or research environment. In later cases most of the article references or copied contents are taken from collections of higher intellect value articles, books journals available in renowned digital libraries. Thus it is very important to see how these systems respond to plagiarism from these sources. In order to test that we randomly selected papers from registered user areas of few famous scholastic archives. Ideally the systems must show a complete plagiarism incident and link to original source. The results (figure 3) indicate that commercial services do not have complete or any access to such archives. The 100% detection in some cases is due to the fact that the article is publicly available on internet at some location other then these protected stores. Although Google API based service do not show 100% but this is because scoring is done based on strict detection patterns. However the search engine based approach overall produce better and higher detection scores, with accurate original source indication. Careful look at reports and matching sources show that a good search engine powered system can present equally helpful results.

Figure 3: Plagiarism percentage graph for collection B, C, D and E



Further investigation in large difference cases revealed the fact that in some cases Google based search returned matches with more updated index of internet resources, and in some cases the commercial services detected matches with information either in their proprietary database index or archived internet index.

5. *Conclusion*

Debates continue at various forums among teachers, students and researchers about positive and negative aspects of these services. We see more and more institutes and individual teachers acquiring these tools and services and on other hand we hear news about negative impacts and discontinuation of use at various places. A recent example of such incident is announcement of terminating use of plagiarism detection service at Kansas University because of cost and legal issues (LJW20, 2006), few days later after an outcry from faculty the decision to cancel use of service is changed (LJW04, 2006). At some places due to rising disapproval from students such services are being turned off basing the fact that such services may create legal privacy and copyright issues and from students and researchers point of view creation of "an antagonistic atmosphere and a culture of guilt, fear and suspicion on campus" (G&M, 2006). Users not satisfied with the tools comment that plagiarists are becoming aware of the fact that even the high priced services are not fool proof and there are ways to work around such detection as demonstrated in "Plagiarism - A Survey" (Maurer et al. 2006).
Academia in general refers to

i. Learning and guiding qualities (when students are made aware of generated reports for correction in work)
ii. Assisting teachers in quickly scanning the internet resources which otherwise seems an impossible task
iii. Deterrent effects

as positive points regarding use of such tools and following as negative factors

i. The "gotcha" and policing approach in educational and research environment.
ii. Copyright and privacy issue by uploading writings on a remote server.
iii. No magic bullet, final results require further manual interpretations and these tools can be fooled as well.
iv. Price.

Results of our experiments show that with the help of any good search engine one can reasonably make an assumption about copied portions in writing but doing a thorough search requires lots of efforts. Thus an automation in search and detect process is inevitable. This automation and ease for checking large volumes of documents is the reason why such services and tools are finding their ways in academic institutes. Although these tools show certain deficiencies like being susceptible to paraphrasing, cross language translations, manual interpretation of found results; whole exercise shows importance of using such tools to create an effective deterrence if not complete defense against plagiarism. So even if no magic bullet a teacher might like to have this detection capability to establish an environment of academic integrity. Even with a low percentage of detection or use, promoting use of such tool or service may act as a proactive measure against plagiarism in an institute, causing improvement in writing standards and citation skills.

References

[comScore, 2006] comScore press release, July 2006, U.S. Search Engine Rankings http://www.comscore.com/press/release.asp?press=984

[DeBell & Chapman, 2006] DeBell, M., and Chapman, C. (2006). Computer and Internet Use by Students in 2003 (NCES 2006–065). U.S. Department of Education. Washington, DC: National Center for Education Statistics.

[G&M, 2006] The Glob and Mail Technology, source: Canadian Press, Turnitin.com turned off at Halifax University http://www.theglobeandmail.com/servlet/story/RTGAM.20060308.gtturnmar8/BNStory/Technology/home

[JISC 2006] Joint Information Systems Committee (JISC) plagiarism Advisory Program website, http://www.jiscpas.ac.uk/ , visited: 22 July 2006

[LJW20, 2006] Lawrence Journal-World, Anti-plagiarism tool pulled from professors' arsenal, September 20, 2006 http://www2.ljworld.com/news/2006/sep/20/antiplagiarism_tool_pulled_professors_arsenal/

[LJW04, 2006] Lawrence Journal-World, KU renews anti-plagiarism software subscription, October 4, 2006 http://mobile.ljworld.com/news/2006/oct/04/ku_renews_antiplagiarism_software_subscription

[Maurer et al. 2006] Mauere H, Kappe F., Zaka B. (2006). Plagiarism-A Survey. Journal of Universal Computer Science, Vol. 12, No. 8, pp. 1050-1084

[McCabe, 2005] The Center for Academic Integrity's Assessment Project Research survey by Don McCabe June 2005, http://www.academicintegrity.org/cai_research.asp

[Nielsen, 2006] Nielsen Netratings Search Engine ratings, July 2006. http://searchenginewatch.com/showPage.html?page=2156451

[Widdows, 2003] Dominic Widdows, Geometry and Meaning : Chapter 5, Word Vectors and Search Engines, 2003 CSLI Publications.

Appendix 2: Plagiarism – a problem and how to fight it

H. Maurer
Graz University of Technology, Graz, Austria
hmaurer@iicm.edu

B. Zaka
Graz University of Technology, Graz, Austria
bzaka@iicm.edu

Abstract: The continued growth of information on the WWW, in databases and digital libraries is making plagiarism by copying, possibly followed by some modification, more and more tempting. Educational institutions are starting to fight this by a bundle of measures: (a) by making students more aware of plagiarism, (b) by enforcing a code of ethics whose violation can lead to drastic measures including the expulsion from universities and (c) by using software that detects suspected plagiarism in the majority of cases. In this paper we show that plagiarism detection applies to much more than just student work: it is relevant in many other situations, including rather unexpected ones. We the briefly describe the two main approaches to plagiarism detection. We cover the main approach, the one based on 'fingerprints', in some detail and compare the two leading commercial packages with a tool developed by us based on the Google API. We also argue that all plagiarism detection tools at this point in time suffer from three major shortcomings whose elimination is possible on principle, but will require a major effort by a number of big players.

1. Introduction

Plagiarism is the use of material, be it verbatim, be it with small changes, or be it the use of a novel idea as such without proper and full disclosure of the source. Based on the fact that many instances that are properly considered plagiarism occur unintentionally by e.g. sloppy referencing, many universities offer now courses to alert student and staff to plagiarism. A short tutorial on this is e.g. [Indiana-Tutorial 2006], a solid survey of plagiarism is the paper [Maurer, Kappe, Zaka 2006].

Note that plagiarism is overtly supported by hundreds (!) of "paper mills", see e.g. [Wikipedia 2006] or [Paper Mills 2006] that offer both "off the rack" reports or will prepare any kind of document on demand. Plagiarism in academic institutions is not a trivial matter any more: A survey conducted by the Center of Academic Integrity's Assessment project reveals that 40% of students admitted to occasionally engaging in plagiarism in 2005, compared to just 10% in 1999 [Integrity 2006]. However, plagiarism is by no means restricted to students in academia: it crops up in many other connections as we will discuss in the next section.

2. Why plagiarism detection is important

In academia plagiarism detection is most often used to find students that are cheating. It is curious to note that as better and better plagiarism detection software is used, and the use is known to students, students may stop plagiarizing since they know they will be found out; or else, they will try to modify their work to an extent that the plagiarism detection software their university or school is using fails to classify their product as plagiarized. Note that there are already anti-anti plagiarism detection tools available that help students who want to cheat: students can submit a paper and get a changed version in return (typically many words replaced by synonyms), the changed version fooling most plagiarism detection tools.

However, plagiarism is not restricted to students. Staff may publish papers partially plagiarized in their attempt to become famous or at least beat the "publish or perish" rule. There are cases known where tenured staff has been dismissed because some important contributions of the staff member has been found to be plagiarized. Sadly, some scientists will go to any length (falsifying data obtained from experiments, plagiarizing, or claiming an achievement they have no right to claim) to preserve or promote their status. If this is any consolation, such attempts are not a new phenomenon that has started because of the internet, but those attempts just continue the large number of proven or suspected cases of cheating by scientists and discoverers. To mention one typical and mysterious case, it will be now soon 100 years that the "North Pole" has been discovered. However, it is still not clear whether Frederick Cook reached the pole one year before Robert Peary did, as was initially assumed, or whether Cook was a cheat: after all, his companion in the first winter-ascent of Mt. McKinley did claim after Peary's return that he and Cook never made it to the top of Mt. McKinley. Unfortunately we do not know whether this sudden change of opinion is due to feelings of guilt or a bribe by Peary. The first author of this paper has studied this case carefully and is indeed convinced that Peary was cheating, not Cook, see [Henderson 2005].

It is interesting to note that sometimes persons accused of plagiarism by actually showing to them that they have copied large pieces of text more or less verbatim sometimes refuse to admit cheating. A tool called Cloze helps in such cases: it erases every fifth word in the document at issue, and the person under suspicion has to fill in the missing words. It has been proven through hundreds of experiments that a person that has written the document will fill in words more than 80% correctly, while persons who have not written the text will not manage more than some 50% correct fill-ins at most!

No plagiarism detection tool actually proves that a document has been copied from some other source(s), but is only giving a hint that some paper contains textual segments also available in other papers. The first author of this paper submitted one of his papers that had been published in a reputable journal to a plagiarism detection tool. This tool reported 71% plagiarism! The explanation was that parts of the paper had been copied by two universities using OCR software on their servers! This shows two things: first, the tools for plagiarism detection can be used also to find out whether persons have copied illegally from ones own documents and second, it can help to reveal copyright violations as it did in this case: the journal had given no permission to copy the paper!

This raises indeed an important issue: plagiarism detection tools may be used for a somewhat different purpose than intended like the discovery of copyright violation. In examining studies conducted for a government organisation for a considerable amount of money each we found that two of the studies were verbatim copies (with just title, authors and abstract changed) of studies that had been conducted elsewhere. When we reported this to the organisation involved the organisation was NOT worried about the plagiarism aspect ("we got what we wanted, we do not care how this was compiled") but was concerned when we pointed out that they might be sued for copyright violation!

It is for similar reasons why some journals or conferences are now running a check on papers submitted in a routine fashion: it is not so much that they are worried about plagiarism as such, but (i) about too much self-plagiarism (who wants to publish a paper in a good journal that has appeared with minor modifications already elsewhere?) and (ii) about copyright violation. Observe in passing that copyright statements that are usually required for submissions of papers to prestigious journals ask that the submitter is entitled to submit the paper (has copyright clearance), but they usually do not ask that the paper was actually authored by the person submitting it. This subtle difference means that someone who wants to publish a good paper may actually turn to a paper mill and order one including transfer of copyrights!

Checking for plagiarism becomes particularly complex when the product published is part of some serious teamwork. It is common in some areas (like in medicine) that the list of authors of papers is endlessly long, since all persons that have marginally contributed are quoted. This is handled in different ways depending on the discipline: in computer science it is quite common that when a team of three or more work on a project, one of the researcher, or a subgroup makes use of ideas and formulations developed by the team without more than a general acknowledgement. This is done since it is often impossible to ascertain which member of the team really came up with a specific idea or formulation first.

Overall, when plagiarism detection software reports that 15% or more of some paper has been found in one or a number of sources it is necessary to manually check whether this kind of usage of material from other sources does indeed constitute plagiarism (or copyright violation) or not. No summary report of whatever tool employed can be used as proof of plagiarism without careful case by case check!

Keeping this in mind we now turn to how plagiarism detection works. In the light of what we have explained "plagiarism warning tools" might be a more exact term for what is now always called "plagiarism detection tools".

3. How plagiarism detection software works

Let us start out by discussion two interesting approaches that are outside the mainstream.

A number of researchers believe in so-called stylometry or intrinsic plagiarism detection. The idea is to check a document for changes in style (that might indicate that parts are copied from other sources) or to compare the style of the document at hand with the style used by the same author(s) in other papers. This requires of course a reliable quantification of linguistic features to determine inconsistencies in a document. Following [Eissen & Stein 2006], "Most stylometric features fall in one of the following five categories: (i) text statistics, which operate at the character level, (ii) syntactic features, which measure writing style at the sentence-level, (iii) part-of-speech features to quantify the use of word classes, (iv) closed-class word sets to count special words, and (v) structural features, which reflect text organization." We believe that further advances in linguistic analysis may well result in a tool whose effectiveness is comparable to the major approach use today, document comparison, an approach described a bit later.

The second little used, yet powerful approach, is to manually pick a segment of the document at issue (typically, 1-2 lines) that seems to be typical for the paper, and just submit it to a search engine. It has been described in detail in [Maurer, Kappe, Zaka 2006] that picking a segment like "Let us call them eAssistants. They will be not much bigger than a credit card, with a fast processor, gigabytes of internal memory, a combination of mobile-phone, computer, camera" from a paper is likely to work well, since eAssistant is not such a common term. Indeed used as input into a Google query (just try it!) finds the paper from where this was copied immediately! Thus, using a few "characteristic pieces" of text with a common search engine is not a bad way to detect plagiarism if one is dealing with just a small number of documents. If a large number of documents have to be checked, the most common method is to compare, using appropriate tools, the document at issue with millions or even billions (!) of other documents. Doing this as will be described in the next paragraph gives an estimate of how much of a document appears in a very similar form in others. If the total percentage is low (typically, below 15%) the idea of plagiarism can usually be dismissed: the list of references in a paper alone will often be the reason why such systems will report 1% 'plagiarism', since the references are also found in other papers. If the total percentage is higher than 15- 20% a careful check has to be made to determine if a case of plagiarism has been detected, or if just self-plagiarism, or using a piece of text created by a team, etc. has been found. All plagiarism detection tools that work with the paradigm of comparing a source document with a large number of documents on the Web or in databases employ the same basic approach:

The source document is split into small segments, called fingerprints. For each such fingerprint a search is performed using some search engine in a very large body of documents. Each fingerprint will usually return a very small number of (if any) documents where something very similar to the fingerprint is found. This "very similar" is determined using the usual "distance approach in high dimensional vector space based on words used" (see [Maurer, Kappe, Zaka 2006] for details). Documents found that fit well (in some cases only the best fit if there is a good one) is retained for each fingerprint. Then the information of all fingerprints is collated, resulting in a list of documents with a certain individual (and if added up total) estimated percentage of similarity. The user is provided with a simple graphical interface which will show which parts of the source document occur elsewhere.

Most of the current free or commercially available plagiarism detection tools work on this basic principle. They differ in the choice of fingerprints (some allowing the user to try various sizes of fingerprints), the choice of search engine, the exact way how similarities found are used, how the results are combined into a total estimate, and finally what databases are accessed.

Probably the leading contenders in the market right now are Turnitin® and Mydropbox®. Turnitin® [Turnitin® 2006] claims to use 4.5 billion pages from the internet, 10 million documents that have been tested previously and the ProQuest® database. In passing let it be mentioned that the fact that documents previously examined are added to the Turnitin® internal collection has proved contentious to the extent that Turnitin® now allows users to prevent their document(s) to be added to the internal Turnitin® database. Mydropbox® [Mydropbox® 2006] claims to search 8 billion internet documents, ProQuest®, FindArticles® and LookSmart® databases and some 300.000 documents generated by paper mills. The document space searched is being enlarged in both cases all the time. It is worthwhile to note that both Turnitin® and Mydropbox® use their own proprietary search engine. We will see in Section 4 why!

4. Comparison of plagiarism detection software

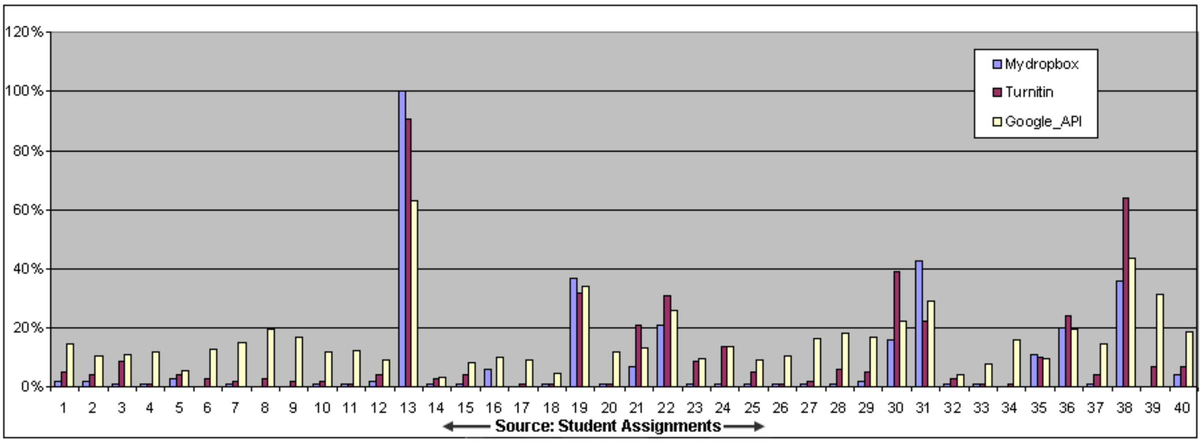
We decided to test various plagiarism detection tools, particularly the leaders in the field, against each other. However, to obtain a better understanding we also developed our own tool which we will call “BPT” (for Benchmark Plagiarism Tool) in what follows. It is built using exactly the same paradigm as the other two tools, but uses the Google API for searching instead of other search engines.

We have tested Turnitin®, Mydropbox® and BPT and other tools with various sets of documents. However, to keep things simple we will just report on the findings for Turnitin®, Mydropbox® and BPT using two very dissimilar sets of documents.

The first set of documents consisted of 90 term papers in the last undergraduate year at our university. The results for the first 40 of those paper is shown in Figure below, the result for the other 50 papers is very similar.

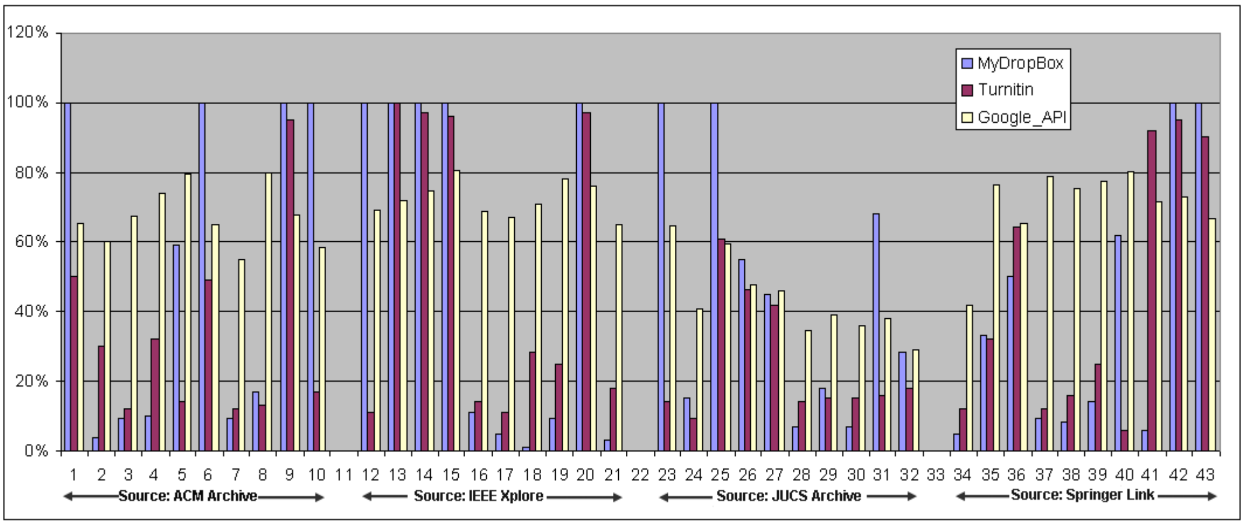
The total percentages of overlap of each student essay with documents on the Web are shown in the figure, the bars showing the result of Mydropbox®, Turnitin® and BPT, respectively. Note that papers 13, 19, 21, 22, 30, 31, 36 and 38 show around 20% or more for each of the tools. It seems surprising that our home-baked solution BPT is doing so well, is actually also identifying 8, 9, 27, 28, 29, 39 and 40 close to or above the 20% threshold!

Figure 4: Comparison with student papers



We will explain this surprising result after discussing Figure 5. It shows the analogous comparison for 40 documents, this time taken from documents in journals that are not available free of charge, and in none of the databases searched by Turnitin® and Mydropbox®.

Figure 5: Comparison of papers not accessible on the Web without charge



Since those documents are actually available verbatim on the web, all tools should show 100% plagiarism! However, as can be seen from the diagram both Turnitin® and Mydropbox® do not recognize more than 15 of the papers as plagiarized. However, BPT shows all documents high above the threshold! Thus, BPT is the most successful tool. As designers of BPT we might be happy with the result. Yet we are not. We are not doing anything better than Turnitin® or Mydropbox®, but we are using Google rather than a home-grown search engine. And Google is evidently indexing many more Web sites than the other search tools are using, including small sites where authors who have published their paper in a journal keep their own copy on their own server: free, and hence detectable by Google.

This leads to the obvious question: why is not everyone using Google. The simple but unpleasant answer to this question is: Google does not allow this (!). Not well-known to the general public, Google only allows 1000 queries per day per user. Fine for the typical user, by far insufficient for a powerful plagiarism tool that sends some 200 queries to the search engine for a 10 page paper. For this reason BPT is NOT a viable service: we can only process some 10 papers per day using two different persons. BPT or similar efforts would only be viable if Google were willing to offer a commercial license. However, we have been unable to get such a license from Google, and we are apparently not the first ones to try in vain. This shows a dark side of Google: Google is keeping its strength in searching as a monopoly, thus allowing Google at some point if it so wishes to offer a plagiarism detection service that would threaten all other services immediately.

It seems to us that Google is ill advised to play this game. It should make profit by charging for the millions of searches required for plagiarism detection but not by threatening to ruin the existence of such services whenever it wants. Overall, the dominance of Google in the field of searching and in others could become a major threat to various IT developments.
This is very much elaborated in the papers [Kulathuramaiyer & Balke 2006] and [Kulathuramaiyer & Maurer 2007a].

The abstracts of the two papers, abbreviated a bit explain the gist of the situation. In the first paper we find: “Everyone realizes how powerful the few big Web search engine companies have become, both in terms of financial resources due to soaring stock quotes and in terms of the still hidden value of the wealth of information available to them. Following the common belief that "information is power" the implications of what the data collection of a de-facto monopolist in the field like Google could be used for should be obvious. However, user studies show that the real implications of what a company like Google can do, is already doing, and might do in a not too distant future, are not explicitly clear to most people.

Based on billions of daily queries and an estimated share of about 49% of the total Web queries, allows predicting with astonishing accuracy what is going to happen in a number of areas of economic importance. Hence, based on a broad information base and having the means to shift public awareness such a company could for instance predict and influence the success of products in the market place beyond conventional advertising or play the stock market in an unprecedented way far beyond mere time series analysis. But not only the mining of information is an interesting feature; with additional services such as Google Mail and on-line communities, user behaviour can be analyzed on a very personal level. Thus, individual persons can be targeted for scrutiny and manipulation with high accuracy resulting in severe privacy concerns.

All this is compounded by two facts: First, Google's initial strategy of ranking documents in a fair and objective way (depending on IR techniques and link structures) has been replaced by deliberatively supporting or down-grading sites as economic or political issues are demanding. Second, Google's acquisition of technologies and communities together with its massive digitization projects such as enable it to combine information on issues and persons in a still more dramatic way. Note that search engines companies are not breaking any laws, but are just acting on the powers they have to increase shareholder value. The reason for this is that there are currently no laws to constrain data mining in any way. We contend that suitable internationally accepted laws are necessary. In their absence, mechanisms are necessary to explicitly ensure “web content neutrality”. We need to raise awareness to the threat that a Web search engine monopoly poses and as a community start to discuss the implications and possible remedies to the complex problem.”

The abstract of the second paper quoted has arguments pointing in a similar yet more general direction: “Plagiarism and Intellectual Property Rights (IPR) Violations have become a serious concern for many institutions and organisations. The revolutionary development of the Web presents numerous opportunities for such infringements to become even more widespread. This situation poses a variety of threats both on the individual level by introducing a ‘culture of mediocrity’ and at a global level in producing a disproportionate power-shift. We believe it is necessary to address those concerns both through institutional efforts and by applying viable technological solutions.”

5. *Shortcomings of plagiarism detection software*

Summarizing what we have seen in Chapter 5, commercially available services for plagiarism detection are doing quite well, but in the light of the “Google experience” described they have to do all they can to continue to extend their databases.

Even if this is done, there are three major obstacles to plagiarism detection that remain: First, all tools are not stable against synonyms: if someone copies a paper and systematically changes words by using synonyms or such, all plagiarism tools we are aware of will fail. Second, all tools are not stable against translation: if someone translates a copy of an e.g. Italian paper into English, no tool will ever detect the translated version as plagiarized. Finally, since many databases and online digital libraries cannot be used free of charge papers resting only in such repositories (or available only in printed form) can often not be used for plagiarism detection.

To solve those three problems together would require reducing each document in a data base and of the document to be examined to what we want to call a “normalized English version”. We refer to [Maurer, Kappe, Zaka 2006] once more for a bit more detail but the gist of the idea is this: a document in an arbitrary natural language (Italian, German,…) is reduced by removing all common words, normalizing all words to their main form (infinitive, first case singular,…). The resulting more or less meaningless string of words is translated word by word into English, using one designated representative in each synonym class.

We have explained in the paper [Maurer, Kappe, Zaka 2006] that "Das azurblaue Cabriolet fiel mit lautem Klatschen in den Bach" (German) and "The deep-blue limousine dropped with a big splash into the river" (English) can be mapped, using the technique described, onto the same sequence of words (the same “signature” as it is called): "blue car fall big noise water".

Since such signatures can only be used for plagiarism detection, and since publishers are also concerned about plagiarized or self-plagiarized material a number of publishers that we have contacted so far would be willing to make available their documents in this form for no or limited cost. If this were the case for all publishers it is clear that a superb plagiarism detection tool could be produced.

6. Outlook

As we have shown, plagiarism detection as it is possible today with commercial tools does by no means guarantee plagiarism detection. Many cases will go undetected because of synonyms, papers that have been translated, and material that is not available to plagiarism detection tools. We have explained that some of this can potentially remedied, if attempts on a large scale are made. Even so, serious problem areas remain. We have only dealt in this paper with plagiarism detection in textual material. The tools mentioned break down if tables, graphs, formulae, program segments etc. are also to be dealt with. First attempts to fight plagiarism in programs can be found in the literature, and some tools that are currently under development for optical recognition of mathematical and chemical formulae may eventually help to extend current techniques. It is clear, however, that a thorny road lies ahead of us if we want to have plagiarism detection tools that work universally!

In summary, it should be clear that big search engines are dangerous in a variety of ways: they support plagiarism and IPR violations; they allow to construct profiles of users with high accuracy; they allow the prediction of economic developments; and, last but not least, may effect how people read (“without deep understanding”) and become incapable of writing coherent essays (since they are used to just gluing together pieces they find with search engines or in the Wikipedia). These points are discussed in the excellent book [Weber 2006]. Readers interested to learn more about search engines are strongly encouraged to study the book [Witten, Gori, Numerico 2007].

However, despite the fact that we have indicated some of the weaknesses and dangers of big search engines we must not forget three important aspects: One, the Web without powerful search engines would not have the value it now has: what are needed are international privacy laws concerning search engines. Two, such privacy laws and other restrictions will take a long time to become accepted, although they are needed desperately not just for search engines but for data mining in general, as is explained in [Kulathuramaiyer & Maurer 2007b]; in the absence of such rules or laws the least we should aim for is to avoid large monopolies and maybe more than a gentle interference by public and non-profit organisations. Three, plagiarism and IPR violation can be fought: even if current day systems do have weaknesses, much work and research is going to make future systems more and more perfect, at the same time creating the necessary awareness that plagiarism is something that ahs to be taken as serious as theft: it is theft of intellectual material.

References

[Eissen & Stein 2006] Eissen, S., Stein, B. (2006). Intrinsic Plagiarism Detection. In: *Proceedings of the 28th European Conference on Information Retrieval;* Lecture Notes in Computer Science vol. 3936, Springer Pub. Co., 565-569.

[Henderson 2005] Henderson, P. (2005). The True North; Norton and Company, New York and London

[Indiana-Tutorial 2006] http://education.indiana.edu/~frick/plagiarism/item1.html (visited November 2006)

[Integrity 2006] http://www.academicintegrity.org/cair_research.asp (visited July 2006)

[Kulathuramaiyer & Balke 2006] Kulathuramaiyer, N., Balke, W.-T. (2006). Restricting the View and Connecting the Dots . Dangers of a Web Search Engine Monopoly; *Journal of Universal Computer Science* 12, 12, 1731 – 1740. See also http://www.jucs.org/jucs_12_12/restricting_the_view_and...

[Kulathuramaiyer & Maurer 2007a] Kulathuramaiyer, N., Maurer, H. (2007). Why is Fighting Plagiarism and IPR Violation Suddenly of Paramount Importance?; *Learned Publishing* 20, no. 4, (Oct. 2007), 252-258, see also: http://www.iicm.tugraz.at/iicm_papers/why_is_fighting_plagiarism_of_importance.doc

[Kulathuramaiyer & Maurer 2007b] Kulathuramaiyer, N., Maurer, H. (2007). Data Mining is becoming Extremely Powerful, but Dangerous; submitted for publication

[Maurer, Kappe, Zaka 2006] Maurer, H., Kappe, F., Zaka, B. (2006). Plagiarism- a Survey. *Journal of Universal Computer Science* 12, 8, 1050-1084. See also http://www.jucs.org/jucs_12_8/plagiarism_a_survey

[Mydopbox® 2006] http://www.mydropbox.com

[Paper Mills 2006] http://www.coastal.edu.library/presentation/mills2.htmlIndiana (visited November 2006)

[Turnitin® 2006] http://www.turnitin.com

[Weber 2006] Weber, S. (2006) . Das Google Copy-Paste-Syndrom; Heise, Hannover, Germany.

[Wikipedia 2006] http://en.wikipedia.org (visited October 2006 with ‘paper mills’)

[Witten, Gori, Numerico 2007] Witten, I., Gori, M., Numerico, T. (2007). Web Dragons- Inside the Myths of Search Engine Technology; Morgan Kaufmann, San Francisco

Appendix 3: Why is Fighting Plagiarism and IPR Violation Suddenly of Paramount Importance?[34]

Narayanan Kulathuramaiyer
Universiti Malaysia Sarawak,

Hermann Maurer
Institut für Informationssysteme und Computer Medien (IICM)
Graz University of Technology

ABSTRACT: Plagiarism and Intellectual Property Rights (IPR) Violations have become a serious concern for many institutions and organisations. The revolutionary development of the Web presents numerous opportunities for such infringements to become even more widespread. This situation poses a variety of threats both on the individual level by introducing a 'culture of mediocrity' and at a global level in producing a disproportionate power-shift. This paper delves on these issues and proposes means of addressing the concerns both through institutional efforts and by applying viable technological solutions.

1. Introduction

Plagiarism as the use of material (text, pictures, movies, etc.) without the exact specification of source; be it in unchanged form or in some kind of derivative. IPR violation however involves the usage or exploitation of works, transgressing the boundaries of its stipulated protection. IPR violations The Web is currently expanding at such a rapid pace, that it becomes a challenge to establish the novelty of contents and artefacts. Web contents are being created, exchanged and transferred at lighting speeds to make it even tougher to detect the degree of originality of contents. Plagiarism and Intellectual Property Rights (IPR) violations are thus concerns that plague many institutions and organisations. For example, educational institutions need to validate the calibre of their students by assessing their academic or literary contributions. Organisations on the other hand need to ascertain the novelty of the articulation and explication of knowledge expressions. Incentive schemes are then formulated to recognise the ability to publish or produce original works.

Plagiarism relates to the theft or borrowing of published work without the proper attribution or acknowledgement of source. We define Plagiarism as the use of material (text, pictures, movies, etc.) without the exact specification of source; be it in unchanged form or in some kind of derivative. IPR violation however involves the usage or exploitation of works, transgressing the boundaries of its stipulated protection. IPR violations constitute the illegal use of material or derivations thereof; be it referenced or not referenced. We view both as violations of acceptable code of conduct and professional ethics. In this paper we treat them together as they are closely related and often lead to on another. As plagiarism is more prevalent, it has therefore been thoroughly deliberated in the academic domain. IPR violation, on the other hand, is treated much more seriously as it has a direct impact on revenues of organisations. This paper highlights adverse implications, calling for an immediate response to overcome imminent dangers and threats.

Although Plagiarism and IPR violations are not a new phenomenon, the new media, particularly the internet is effectively taking it to far greater heights. Beyond print media, infringements can now occur

[34] A first version of some of the ideas in this paper was presented at APE 2007 (Academic Publishing Europe) 2007 by the second author.

in all types of digitised forms, including volatile forms such as SMS, Chat and Mail. This makes the comprehensive addressing of plagiarism and IPR violation much more challenging.

This paper starts off by discussing the forms of plagiarism in the light of current developments. The web has brought about an environment for 'rapid generation of publications' mainly through the instantaneous access to myriad sources of information. We then discuss the implications of such a phenomenon on the quality of 'creative thinking and writing' and its ensuing effect on the quality of life. The control over Web Search and its unlimited mining capabilities puts at the hands of few, the power to represent and characterise reality to influence the lives of millions [Kulathuramaiyer & Balke 2007]. The immense impact of the Web and a resulting culture poses many more dangers than foreseeable [Kulathuramaiyer & Balke 2007]. We then present a discussion on current tools for fighting plagiarism and IPR violations. The subsequent section proposes solutions to effectively address the issue, which includes the establishment of a European Centre for plagiarism and IPR violation, and through the application of viable technologies.

2. Plagiarism and IPR Violation

Plagiarism and IPR violations applies to a variety of forms such as term papers, thesis, and research papers in a university, essays and other written projects in a school as well as in all kinds of publications which includes project papers, news articles, web contents such as blogs, postings in wikis, etc.

Plagiarism is a major concern, particularly in an academic environment, where it could affect both the credibility of institutions as well as its ability to ensure quality of its graduates. Plagiarism has been constantly on the rise [Smithers 2005], [Curtis 2004], largely attributed to the Internet and web. Many students tend to take plagiarism lightly and consider a varying degree of copying to be acceptable [Smith 2006], [Fredericks 2002]. An extreme case has been presented [Jacobs 2004], [Guardian 2004] whereby a student was caught for plagiarising just before he was due to receive his degree. The student who admitted to downloading Internet essays, justified himself by saying, 'If they had pulled me up with my first essay at the beginning and warned me of the problems and consequences, it would be fair enough. But all my essays were handed back with good marks and no one spotted it' [Jacobs 2004]. He then went on to sue his university for not catching him earlier. This clearly highlights the lack of responsibility on the part of students who tend to resort to the easiest means of getting work done, without due considerations of the legibility of their actions. To make the situation worst, there exists a large number of paper mills[8], explicitly supporting students in preparation of term papers. Although there are notices on some of these sites for an ethical use of their services, their services make it far too easy for students to resist. Apart from these paper mills, there are other sources of information that can be used by students; this includes search engines, web directories, Wikipedia, book reviews on online bookstores, scholarly publications, and the list goes on. As highlighted by the Fox Times [Jacob 2005], 'parents are willing to spend $75,000 in high school tuition and $120,000 for a private college, and then pay even more to ensure that their children did not learn anything'. This article is an account in the life of a professional paper writer, who has helped totally uninterested, paying students to gain high marks by producing their term papers.

The biggest danger lies in scholarly publications of our future generations which is happening in schools. Extensive support is now available for high-speed production of published worked. As there is no guilt consiousness and sometimes students are not aware of the ethical codes of conducts.

Plagiarism infringements are however not restricted to students; it can also implicate professors [Dawkins 2005], a college vice president [Associated Press] or even a prime minister [NPR News]. Journals or conferences need to take plagiarism seriously, as papers submitted by authors, could well be largely self-plagiarised (plagiarised from their own past works) [Jacobs 2005]. A 'high degree of self-plagiarism' implies that a major part of the paper has been published before and this could lead a journal to 'unwittingly committing a copyright violation'.

Though journals are concerned about plagiarism, their reason for this is mainly to protect themselves against any legal implications. The survey conducted by Enders and Hoover [Dawkins 2005], has revealed that more than 81 percent of journals in Economics did not have a formal policy regarding plagiarism. A typical journal's copyright protection form checks mainly for authors' ownership of rights to publish a paper, rather than to seek a claim of authorship. This allows a person other than the author to be able to publish a paper, by acquiring rights to do so. Similarly in an academic promotion exercise, an academic is hardly ever required to make a claim of actual authorship of all listed published works. By virtue of owning the copyrights to their published works, academics are assumed to be the actual authors.

Even government and commercial organisations are mainly concerned about IPR violations and unnecessary expenditure, rather than plagiarism. A government organisation was noted to be more perturbed about the possibility of being sued for copyright violation, than in accepting the results of plagiarised works [Maurer & Zaka 2007]. Similarly, the US federal government only takes action on plagiarism that arises from projects funded by itself [Dawkins 2005]. In other words their policies mainly protect their own intellectual property and not plagiarism at large.

The seriousness of overlooking plagiarism has been largely disregarded. E.g. the lack of deliberative actions leads to an increase in dishonesty in research. The case of Frederick Cook vs. Robert Peary illustrates the difficulty in resolving disputes, against possible fraudulent research claims [Wikipedia – Peary]. Plagiarism detection will thus require the support of transaction-based research records management ability to substantiate claims.

The next section highlights the severe implications of plagiarism and its impending influence on future civilisations. A deeper consideration of these implications is required, as plagiarism is becoming a widely accepted culture.

3. Google Copy-Paste Syndrome

The 'Google Copy Paste Syndrome' (GCPS), describes a common activity whereby scientists and journalists alike perform fast, easy and usually "not diligently researched" copying, through the abduction of passages in text [Weber 2006]. Acquiring insights is performed by 'conveniently searching' as opposed to a rigorous process of learning through scientific discovery. Information on the Web is often used without even considering the validity of source.

The GCPS has resulted in a proliferation of plagiarism through the widespread practice of fast and easy, instant-publications. Web mining impedes the inquiry-driven scientific process, as the answer seems to conveniently pop up, with a much lesser effort. The blind trust of naïve users upon the results of Web searches [Kulathuramaiyer & Balke 2007] further dilutes the quality of scientific discovery. This syndrome has thus endangered creative writing and thinking by de-emphasizing the need for 'deliberate and insightful' reasoning [Wikipedia – Peary]. The culture expounded by this emerging phenomenon encourages mediocrity in produced creative works, as a result of the lack of due deliberation and insightful internalisation. As the global brain takes shape by providing answers to all queries, a 'text culture without brains' [Wikipedia – Peary] emerges.

The reality presented by the Web is thus taken to be a substitute for the hours that would otherwise be spent in inquiry and rationalisation. Weber [Weber 2006] aptly states that, 'we are in the process of creating reality by **googeling**'. This statement emphasizes the utmost reliance on content warehouses such as Google and Wikipedia that many of us, particularly the younger generation subject ourselves to.

Web search can in no way construct a valid representation of reality. Search engines tend to intentionally or unintentionally restrict the view of users [1]. Apart from that search results are also not organised to provide an absolutely authentic recording of historical events [Witten, Gori, Numerico 2007]. The ranking algorithm of large search engines tend to be biased towards popular sites.

The new media has thus worsened the quality of life in many ways as Email and emerging communicational means tend to be far too distracting, and thus people become less focussed [Weber2006]. A higher degree of focus is required in promoting a culture of discovery and serendipity.

4. Plagiarism and IPR Violation Detection

Plagiarism detection poses many problems in itself, as plagiarism does not always constitute a blatant copying of paragraphs. There are situations where plagiarism may involve the copying of smaller chunks of content, which could further be transformed adequately to make it extremely difficult to detect. It is also possible that copied text can be translated into another language. One also has to be aware that plagiarized documents may also not always be available in digital form. There are also situations where a document is available in a digital form but it is not accessible by the detection system, e.g. document is locked up in the deep or invisible web [Wikipedia – Deep Web] or a part of the Web that can only be accessed for a fee.

There are numerous images that contain valuable information that need to be protected. A potential difficulty lies in the detection of text that is stored within a collection of images. Google for instance resorts to manual tagging of images, as a means to cluster relevant images rather than to rely on image processing techniques. Innovative approaches are required to expand detection to non-textual resources.

In reality there are always cases of innocent people being accused of plagiarism. There are also situations when plagiarism arises based on genuine negligence in recording and attributing a reference. An individual may unintentionally, miss a reference. For example in the case of the Star Tribune Plagiarism Probe [Foxnews] the author was found not guilty of plagiarising the Hertzberg piece. In this case the author's failure to distinguish between direct quotes and paraphrased ideas in a recorded transcript, had led to non-attribution. Plagiarism detection software has thus to be used with great care.

There are other considerations for plagiarism detection system. For instance, publications involving a team of researchers [Maurer & Zaka 2007], makes it extremely difficult to ascertain plagiarism. In some area such as medicine, a long list of authors may be common. In other areas such as Computer Science, authorship may be attributed to only the main contributors of particular ideas in a paper, rather than to list the names of all research team members. As a journal publication comprises of a collection of ideas, it will then be difficult to ascertain the ownership of individual ideas as contributed by each research team members.

Plagiarism detection can in no way be considered a proof beyond doubt. It is merely employed to indicate that plagiarism may have occurred! As such, if suspicion arises based on the findings of a plagiarism detection system, a manual check is always necessary to verify this. An entirely automated plagiarism detection will result in false positives, which could be disastrous.

When a plagiarism case is raised, the impact on the image of an individual can be devastating even if the complaint is not justified [Leach 2005]. There are further situations whereby authors whose works have been plagiarised choose not to take action, as it would be difficult to establish a proof [Dawkins 2005]. Plagiarism detection software would then be required to also help innocent authors in substantiating an ownership claim to their published works. Plagiarism detection tools (and IPR violation) need to incorporate these considerations.

5. Tools for Detecting Plagiarism

We present an overview of a broad range of tools that are currently available. A detailed account of a number of the tools mentioned here can be found in a survey [Maurer, Kappe, Zaka 2006].

5.1 Usual Approach

The usual approach for detecting plagiarism splits a document into a (large) set of 'fingerprints'. A set of fingerprint contains pieces of text that may overlaps with one another. A fingerprint is then used as a query to search the web or a database, in order to estimate the degree of plagiarism. Most currently available software packages employ this technique. Variations between packages are only in the fingerprints used and search engines employed. The advantage of this method is that it is fairly stable against the re-arranging of text. It is however not stable against synonyms and translations.

5.2 Denial of Plagiarism

As mentioned in the previous section specific tools are required to address the denial of plagiarism. A Cloze procedure [Maurer & Zaka 2007] can be used to judge the likely owner of published works, when a dispute arises. Cloze works by concealing words in a document in a regular pattern. An individual is then required to fill in the blanks, with a word that he or she considers appropriate. It has been shown that the original author of a document is more likely to get words correctly, as compared to the one who copied [Maurer & Zaka 2007].

5.3 Stylometry

Stylometry is an attempt to analyse writing styles based on text similarity patterns. A particular text can be compared with the typical writing style of an individual based on his or her past works. Alternatively the text in a single paragraph can be compared with the overall style of writing as found throughout a paper. As opposed to the other methods mentioned, Stylometry is able detect plagiarism without the need for an external corpus of documents [Eissen & Stein 2006]. It can be applied to detect intrinsic patterns within documents that capture style parameters that include syntactic forms, text structure as well as the usage of key terms [Eissen & Stein 2006].

5.4 Manual Detection

This approach employs the manual selection of a phrase or one or more sentences representing a unique concept found in a text. This selected text is then used as a query to one or more search engines. An instructor or examiner may repeat this process a number of times, focusing and refining the query phrase in the process. Although this approach is simplistic, its potential in discovering plagiarism can be astonishing. The effectiveness of this approach depends mainly on the domain specific or contextual knowledge of a human expert in formulating meaningful queries. It is also possible that an expert knows exactly where a potential document may have been copied from, which could be used to narrow down search. The main difficulty in automating this approach fully is in having to determine the exact phrase or sentence to be used as query. Alternatively, partial automation or applying this approach in conjunction with other approaches may be explored.

6. Integrating Search Application Programmers Interface (API)

A home-grown plagiarism detection method [Maurer & Zaka 2007] built on top of Google's search API has surprisingly produced superior results as compared to leading software packages in the industry such as Turnitin® [Turnitin®] and Mydropbox® [Mydropbox®]. This is mainly due to Google's indexing of many more Web sites as compared to these plagiarism detection tools. Although there are many journals that are not freely available on the Web, most authors maintain a personal copy of their own publications on personal websites, which can then become indexed by search engines. The limitation of employing Google's free API, has however restricted their system to 1000 queries a day. As Google does not license their Search engine, they maintain the exclusive control to numerous potential applications. This will also increase the reliance of publishers on these global search engines.

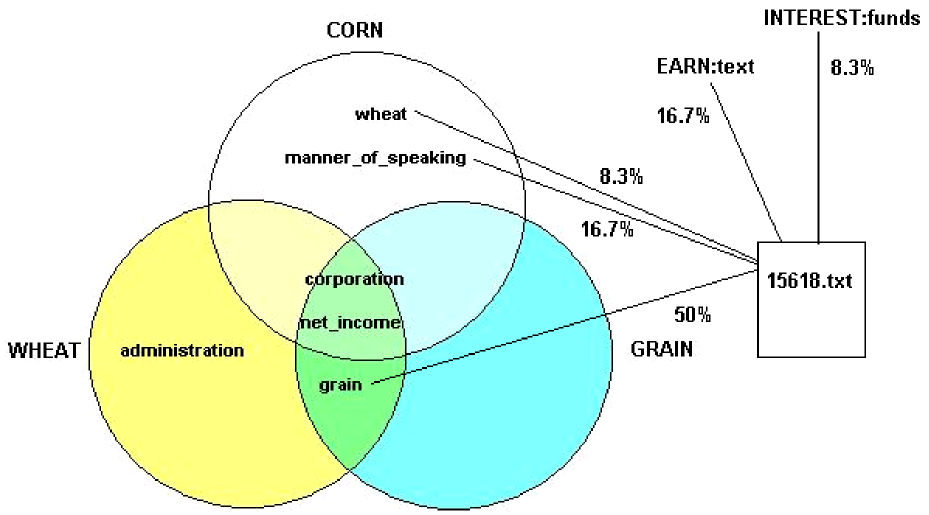
6.1 Advanced Plagiarism Detection

An essay grading system has been proposed employing Natural Language Processing techniques to build a propriety knowledge representation model [Dreher & Williams 2006]. Model answers of teachers are then compared with student answers to determine grade assignment. This form of text analysis attempts to serve plagiarism detection on the idea (conceptual) level. This approach has so far been proven to work well on a small database producing comparable performance with human experts.

Although current plagiarism tools appear adequate, they are not able to fight against a serious plagiarist, such as a core program plagiarist [Liu & al. 2006] The use of graphs has been proposed to illustrate data and control dependency in detecting plagiarism in software development projects. The use of such a representation of relationships makes it invariant to rearranging. This approach is analogous to one that employs conceptual dependencies in NLP applications.

A framework for the semantic characterisation of text has been proposed combining statistical analysis, machine learning together with semantic analysis[Ong & al. 2006]. Their conceptual schema extraction approach comprise of two phases: base-map learning and incremental learning. This approach can be applied for plagiarism detection whereby in the first phase, a semantic model is constructed for a set of documents that represent the base-knowledge to be modeled. A second level of learning is then applied to a document to be tested for copying (incremental learning). Figure 6 illustrates a test document [Ong & al. 2006] that depicts the associated degree of similarity for each concept in the model. This approach could also be used to narrow in on a smaller set of documents, for detecting plagiarism.

Figure 6: Illustrated Concept Level Plagiarism Detection



7. *What can we do?*

At the heart of a plagiarism detection capability resides a system to authenticate artefacts and determine the degree of similarity between established original works and potential infringements. The benefits of possessing this capability is extremely important as the (similarity) detection technology can be adapted to other areas, such as patent databases or automated question answering! It will further avoid the reliance of publishers on global search engines.

Plagiarism detection also requires the development of an enormous database of documents and texts. This database will be used to effectively protect intellectual property rights, as well as to prevent indiscriminate copying. We propose the development of a European Centre for Plagiarism and IPR violation Detection that will be responsible for the development and maintenance of such a database together with the similarity detection capability in providing a variety of services to preserve and protect intellectual property.

8. European Centre for Plagiarism and IPR violation Detection (ECPIRD[35])

There is an impending need to establish the ECPIRD in coping with the emerging developments in the global scene. We propose this Centre to be set up with an initial funding of € 20 million. It is expected to eventually become self-supporting, based on the revenues generated from services offered. In addressing the issues mentioned, a good search engine for plagiarism detection will need to be implemented. A sophisticated ranking capability as found in existing search engines may not be required. Nevertheless, the research and development carried out could become the basis for a competitive search engine.

The centre will thus maintain a workbench of detection tools that could be customised to suit a variety of applications. This workbench will include corpus independent tools, semi-automated support for manual detection, Optical Character Recognition (OCR), tools for denial of plagiarism, etc. It will be hosted on a Web Portal that serves as the front-end for the distributed plagiarism detection services offered by the ECPIRD.

This Centre would also need to be entrusted to come up with training modules to promote the scholarly and academic best-practises, to educate society to become more accountable and competent in undertaking knowledge-work. While efforts are being carried out to form the Centre, we will need to continue to use existing tools, and to carry out further research in plagiarism detection techniques such as Stylometry research. In the absence of sophisticated tools, the manual method serves to be an alternative. Apart from this we also propose to initiate a pilot project that will pave the way for a full-fledged plagiarism system.

8.1 Pilot project for Distributed Plagiarism Detection

Although, search engines are emerging as powerful plagiarism detection tools, their capability is limited to the shallow Web (visible). There is a huge number of sites that are currently locked up in the deep Web, beyond the reach of search engines. In order to address this, we propose the development of specialised domain-specific plagiarism detection systems. This entails a distributed approach where we will have separate facilities for plagiarism and IPR violation detection for each area of specialisation (i.e. Computer Science, Psychology). As an example, for the area of Computer Science, a server farm can be hosted at the Graz University of Technology, Austria. There could be similar facilities established in numerous localities throughout Europe and even across the world to effectively address multiple disciplines and languages.

Specialised ontology needs to be developed and maintained at each site for each area of specialisation. The availability of this ontology allows the capturing and encoding of domain-specific and context dependent knowledge that allows the n-depth addressing of plagiarism and IPR violation detection. This also enables the harnessing of domain knowledge to enhance the systems capability. Apart from that, the ontology provides a platform for the bottom-up construction of a multi-discipline Semantic Web. It will also facilitate the integration of distributed services towards the development of a

[35] Can almost be pronounced as EXPERT

consolidated front-end. A specialised ontology further supports the offering of personalised services based on specific needs of users. It also becomes a basis for the expansion of this system to incorporate other functionalities in future.

This proposal also ensures that no central agency will have an exclusive control over powerful technology and all resources. In order to ensure the neutrality of content, all such facilities will need to be managed by not-for-profit agencies such as universities and public libraries.

Taking a distributed approach in the development reduces the magnitude and cost of the project to a great extent. Addressing the deep web within a particular domain helps to simplify its design. The development of a complex system becomes more sustainable as individual agencies become responsible for the maintenance of each individual distributed facility. Grants can also be distributed fairly across European countries for setting up various facilities. This pilot project has to be developed in parallel with the development of an effective software package for plagiarism and IPR violation detection.

8.2 Layered Plagiarism and IPR violation Detection

We proposed the development of a layered plagiarism detection software which would enable a structured handling of this complex problem. This system would first highlight a theme to be extracted from the class of documents to be checked. Based on the theme, the focus of the similarity detection would be narrowed to a much smaller space (e.g. exotic plants of Borneo). The domain of the search will be used as a means to direct the processing to the most appropriate facility.

A second level of dimensionality reduction will then be applied to narrow the document space even further by employing heuristic-selection criteria (e.g. focus on recent documents only or restricted to a particular database). A further level of document space reduction can also be employed via a kind of semantic modelling.

When a small enough subset is identified, elaborate mining can be employed to generate detailed descriptions of evidences to support a case of plagiarism, if one exists. We propose a combination of plagiarism detection methods (in the workbench) to be employed together to achieve an effective performance.

In order to address the handling of synonyms and multilingual texts we propose an approach where each arbitrary document in Natural Language (e.g. German or Italian) is reduced into a set of a “normalized English version” terms.20 Pre-processing is first performed to remove stop words and commonly used terms. The list of terms is then reduced to set of derived standard root forms employing basic tools such as a lexicon and a Part-of-Speech (POS) tagger. Subsequently, sense disambiguation is performed to reduce the terms to singular senses, by employing a standard corpus as a reference. The derived normalised form is represented in English language and will be used as a base representation to index all documents. The advantage of employing such an approach is that publishers are more willing to supply such as “signature”, in return for an acknowledgement and assurance of a plagiarism free collection. Based on the compilation of normalised forms of documents, powerful plagiarism detection tools can be developed. Another benefit of offering their normalised collections is the additional traffic publishers will gain on their for-pay sites when such a system is in place and suspected plagiarized documents are being checked against originals.

Printed books and all text in stored images should be converted into digital images using OCR technology. This will also involve the digitisation and archiving of large collections of valuable content currently available only in printed form.

9. Conclusion

This paper has illustrated severe implications of plagiarism and IPR violation. Potential dangers of not addressing the issues mentioned leads to the lost of revenues due to inability to contain Intellectual property, the degradation of scientific culture and the loss of control over powerful technology.

As such plagiarism and IPR violation detection technology becomes absolutely essential. The establishment of a European Centre for Plagiarism and IPR violation detection ensures a balanced, sustainable growth and distribution of economic and social resources. This centre will be able to pool together resources towards the development of universal Plagiarism and IPR violation detection tools for preserving and protecting both textual and non-textual resources. Revolutionary measures will then need to be designed to protect us against the impending information explosion.

References

[Associated Press] Associated Press, College president cited for plagiarism , March 17,2004, http://www.msnbc.msn.com/id/4551042/

[Curtis 2004] Curtis, P., Quarter of students 'plagiarise essays', June 30, 2004 http://education.guardian.co.uk/students/work/story/0,,1250786,00.html

[Dawkins 2005] Dawkins, S,. Stealing work common, not limited to students, January 19, 2005 http://www.cw.ua.edu/vnews/display.v/ART/2005/01/19/41ee0c05c597c

[Dreher & Williams 2006] Dreher, H. ,Williams,R., Assisted Query Formulation Using Normalised Word Vector and Dynamic Ontological Filtering: Flexible Query Answering Systems: 7th International Conference, FQAS 2006, Milan, Italy, June 7-10, 2006 pp. 282 –294

[Eissen & Stein 2006] Eissen, S., Stein, B. Intrinsic Plagiarism Detection. In: *Proceedings of the 28th European Conference on Information Retrieval;* Lecture Notes in Computer Science, 2006, vol. 3936, Springer Pub. Co., 565-569.

[Foxnews] Foxnews, Star Trib Plagiarism Probe Clears Writer, Dec 16, 2006, http://www.foxnews.com/wires/2006Dec16/0,4670,StarTribunePlagiarismProbe,00.html

[Fredericks 2002] Fredericks,M.A., Cheating, the copy-and-paste way ,The Star Online Exclusive Report 2002 http://202.186.86.35/special/online/plagiarism/mike_cutpaste.html

[Guardian 2004] Guardian Unlimited, Plagiarising student sues university for negligence, May 27, 2004 http://education.guardian.co.uk/higher/news/story/0,,1226148,00.html

[Jacobs 2004] Jacobs,J., History Without History, Spelling Without Spelling, June 04, 2004 http://www.foxnews.com/story/0,2933,121840,00.html

[Jacobs 2005] Jacobs,J., The Under-Graduate, Cheating Confessions, April 11, 2005 http://www.foxnews.com/story/0,2933,152879,00.html

[Kulathuramaiyer & Balke 2007] Kulathuramaiyer N., Balke, W.T. "Restricting the View and Connecting the Dots - Dangers of a Web Search Engine Monopoly" Journal of Universal Computer Science, 2007, Vol. 12, No. 12, pp. 1731-1740.http://www.jucs.org/jucs_12_12/restricting_the_view_and

[Leach 2005] Leach,S.L. , Learning, Profs who plagiarize: how often?, The Christian Science Monitor, April 27, 2005 http://www.csmonitor.com/2005/0427/p15s01-legn.html

[Liu & al. 2006] Liu,C., Chen, C., Han,J., and Yu,P.S., GPLAG: Detection of Software Plagiarism by Program Dependence Graph Analysis, the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2006, pp. 872-881, Philadelphia, USA, http://www.ews.uiuc.edu/~chaoliu/papers/kdd06liu.pdf

[Maurer, Kappe, Zaka 2006] Maurer, H., Kappe, F., Zaka, B. Plagiarism- a Survey. *Journal of Universal Computer Science, 2006* 12, 8, 1050-1084.

[Maurer & Zaka 2007] Maurer, H., Zaka, B. Plagiarism- a Problem and How to Fight It, 2007, Proc. of ED-MEDIA 2007, AACE, Chesapeake, VA (2007), 4451-4458; http://www.iicm.tugraz.at/iicm_papers/plagiarism_ED-MEDIA.doc

[Mydropbox®] http://www.mydropbox.com

[NPR News] NPR News, Profile: Prime Minister Tony Blair Under Fire for Intelligence Report That Appears to be Plagiarized, Feb 8, 2003, http://www.npr.org/programs/atc/transcripts/2003/feb/030208.raz.html,

[Ong & al. 2006] Ong, S.C.,Kulathuramaiyer, N., Yeo, A.W., Automatic Discovery of Concepts from Text, Proceedings of the IEEE/ACM/WIC Conference on Web Intelligence 2006: pp.1046-1049

[Paper Mills] Paper Mills, http://www.coastal.edu/library/presentations/mills2.html

[Smith 2006] Smith,A., Plagiarism 'rife' at Oxford, March 15, 2006 http://education.guardian.co.uk/higher/news/story/0,,1731423,00.html

[Smithers 2005] Smithers, R, Crackdown urged on web exam plagiarism, November 22, 2005 http://education.guardian.co.uk/gcses/story/0,,1648106,00.html

[Turnitin®] http://www.turnitin.com

[Weber 2006] Weber, S., Das Google-Copy-Paste-Syndrom, Wie Netzplagiate Ausbildung und Wissen gefährden, Heise, Hannover, 2006

[Wikipedia – Deep Web] Wikipedia, Deep Web, http://en.wikipedia.org/wiki/Deep_web

[Wikepedia – Peary] Wikipedia, Peary,http://en.wikipedia.org/wiki/Robert_Peary

[Witten, Gori, Numerico 2007] Witten, I.H., Gori, M., Numerico, T., Web Dragons, Inside the Myths of Search Engine Technology, Morgan Kaufmann, San Francisco, 2007

Appendix 4: Restricting the View and Connecting the Dots – Dangers of a Web Search Engine Monopoly

Narayanan Kulathuramaiyer
Institute for Information Systems and New Media
Graz University ot Technology
nara@iicm.edu

Wolf-Tilo Balke
L3S Research Center and University of Hannover, Germany
balke@l3s.de

Abstract: Everyone realizes how powerful the few big Web search engine companies have become, both in terms of financial resources due to soaring stock quotes and in terms of the still hidden value of the wealth of information available to them. Following the common belief that "information is power" the implications of what the data collection of a de-facto monopolist in the field like Google could be used for should be obvious. However, user studies show that the real implications of what a company like Google can do, is already doing, and might do in a not too distant future, are not explicitly clear to most people.

Based on billions of daily queries and an estimated share of about 49% of the total Web queries [Colburn, 2007], allows predicting with astonishing accuracy what is going to happen in a number of areas of economic importance. Hence, based on a broad information base and having the means to shift public awareness such a company could for instance predict and influence the success of products in the market place beyond conventional advertising or play the stock market in an unprecedented way far beyond mere time series analysis. But not only the mining of information is an interesting feature; with additional services such as Google Mail and on-line communities, user behavior can be analyzed on a very personal level. Thus, individual persons can be targeted for scrutiny and manipulation with high accuracy resulting in severe privacy concerns.

All this is compounded by two facts: First, Google's initial strategy of ranking documents in a fair and objective way (depending on IR techniques and link structures) has been replaced by deliberatively supporting or ignoring sites as economic or political issues are demanding [Google Policy: Censor, 2007]. Second, Google's acquisition of technologies and communities together with its massive digitization projects such as [Burright**,** 2006**]** [Google Books Library, Project, 2006] enable it to combine information on issues and persons in a still more dramatic way. Note that search engines companies are not breaking any laws, but are just acting on the powers they have to increase shareholder value. The reason for this is that there are currently no laws to constrain data mining in any way. We contend that suitable internationally accepted laws are necessary. In their absence, mechanisms are necessary to explicitly ensure web content neutrality (which goes beyond the net neutrality of [Berners-Lee, 2006]) and a balanced distribution of symbolic power [see Couldry, 2003]. In this paper we point to a few of the most sensitive issues and present concrete case studies to support our point. We need to raise awareness to the threat that a Web search engine monopoly poses and as a community start to discuss the implications and possible remedies to the complex problem.

1. Introduction

Google has emerged as the undisputed leader in the arena of Web search. It has become the gateway to the world for many people, as it is the first point of reference for all sources of information. It has also successfully transformed the way we live our lives today in a number of way. At the strokes of the keyboard, it is now possible to gain access to vast reservoirs of information and knowledge presented by the search engine. But of course also our perception is shaped by what we see or fail to see. The situation is aptly characterized by the statement "Mankind is in the process of constructing reality by googeling" [Weber, 2006].

Moreover, with respect to the quality of the results gained by search engines, users have shown to be overly trusting and often rather naïve. Recent user behavior shows that the simple and efficient search facilitated by search engines is more and more preferred to tedious searches through libraries or other media. However, the results delivered are hardly questioned and a survey in the Pew Internet &

American Life Project come to the result that over 68% of users think that search engines are a fair and unbiased source of information: “While most consumers could easily identify the difference between TV’s regular programming and its infomercials, or newspapers’ or magazines’ reported stories and their advertorials, only a little more than a third of search engine users are aware of the analogous sets of content commonly presented by search engines, the paid or sponsored results and the unpaid or “organic” results. Overall, only about 1 in 6 searchers say they can consistently distinguish between paid and unpaid results.” [Fallows, 2005]

Taking the idea of personalized information access seriously indeed involves the restriction of the possible information sources by focusing the user’s view to relevant sites only. Google started business a decade ago with the lofty aim to develop the perfect search engine. According to Google’s co-founder Larry Page: “The perfect search engine would understand exactly what you mean and give back exactly what you want.” [Google Corporate Philosophy, 2007]. As knowledge of the world and the Web are interconnected and entwined, most search engine builders have grown to realize that they need to have “all knowledge of everything that existed before, exists now or will eventually exist” in order to build the envisioned perfect search engine. The supremacy of Google’s search engine is acknowledged [Skrenta, 2007b] even by its competitors [Olssen, Mills, 2005]. Google’s larger collection of indexed Web pages coupled with its powerful search engine enables it to simply provide the best search results.

In this paper we want to analyse the evident dangers that are in store for Web users and the community at large. We need to become aware of the silent revolution that is taking place. As a de-facto search engine monopolist Google may become the leading global player having the power and control to drastically affect public and private life. Its information power has already changed our lives in many ways. Having the power to restrict and manipulate users’ perception of reality will result in the power to influence our life further [Tatum, 2005]. We present concrete anchor points in this document to highlight the potential implications of a Web search engine monopoly.

2. *Connecting the Dots and the Value of Data Mining*

The real implications of what Google can do, is already doing or will do are not explicitly clear to most people. This section will provide insights into the extra-ordinary development of Google as a monopoly, providing evidences as to why this is a major concern.

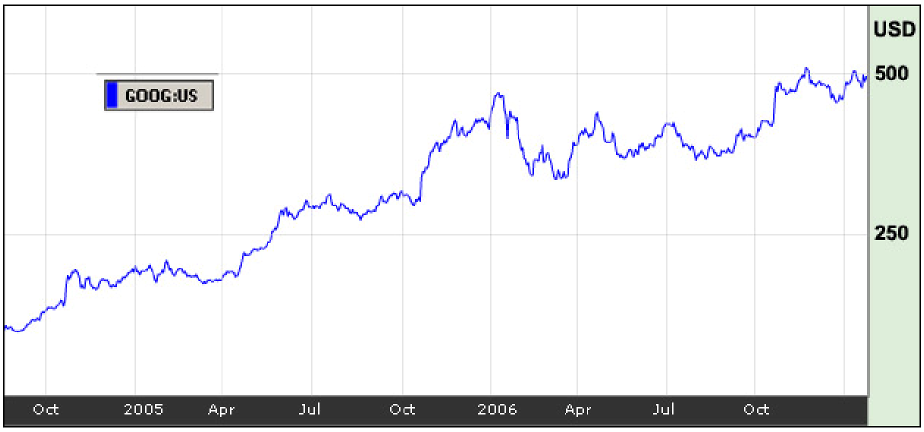
2.1 Unprecedented Growth

Google’s ability to continuously redefine the way individuals, businesses and technologists view the Web has given them the leadership position. Despite its current leadership position, Google aspires to provide a much higher level of service to all those who seek information, no matter where they are. Google's innovations have gone beyond desktop computers, as search results are now accessible even through portable devices. It currently provides wireless technology to numerous market leaders including AT&T Wireless, Sprint PCS and Vodafone.

Over time they have expanded the range of services offered to cover the ability to search an ever-increasing range of data sources about people, places, books, products, best deals, timely information, among many other things. Search results are also no longer restricted to text documents. They include phone contacts, street addresses, news feeds, dynamic Web content, images, video and audio streams, speech, library collections, artefacts, etc.

After going public in August 2004 the stock price recently reached a high of more than five times of the original issue price [see figure 7]. The rise in valuation was so steep that Google quickly exceeded the market capitalization of Ford and General Motors combined. M. Cusumano of MIT Sloan School of Management deduces that “Investors in Google’s stock are either momentum speculators (buying the stock because it is hot and going up) or they believe Google is a winner-takes-all kind of phenomenon, fed by early-mover advantage and positive feedback with increasing returns to scale.“ [Cusumano, 2005]

Figure 7: Development of the Google Stock (extracted from Bloomberg.com)



Google’s main source of income has been through its targeted advertisement that has been placed beside its search results as sponsored links. Their non-obtrusive, inconspicuous text-based advertisements that is dependent and related to search results, has made it into a billion-dollar company. The company is now poised to expand their advertisements even further to cover audio and video transmissions [Google Video Ads, 2006], [Rodgers, Z, 2006]. According to [Skrenta, 2007a], Google’s stake of the search market is actually around 70%, based on their analysis of web traffic of medium and large scale Web sites.

Besides this, Google has been quite successful in acquiring the best brains in the world to realize its vision by stimulating a rapid and explosive technological growth. Innumerable commercial possibilities have arisen from the creative energy and the supporting environment of Google. Google has been recognized as the top of the 100 best companies to work for in 2007, by Fortune Magazine. [Fortune Magazine, 2007] In evaluating and screening the large number of job applications they receive, Google’s encompassing mining capability is already being applied [Lenssen, 2007].

2.2 Technology Acquisition

Google has been aggressively buying up technology companies with a clear vision of buying into large user communities. Recently Google paid 1.5 billion for YouTube which has a massive community base. YouTube was reported to have 23 million unique visitors with 1.5 billion page views in U.S. alone, in October 2006. Apart from this Google has recently bought leading community software such as Orkut and Jot.

Google’s ability to integrate acquired technologies into an expanded portfolio distinguishes it from its competitors. The acquisition of a digital mapping company, Keyhole has brought about Google Earth, which enables the navigation through space, zooming in on specific locations, and visualising the real world in sharp focus. Google Earth provides the basis of building an enormous geographical information system, to provide targeted context-specific information based on physical locations. The databases that they have constructed provide a plethora of services to make them knowledgeable on a broad range of areas in a sense that is beyond the imagination of most people.

Google’s acquisition of Urchin analytics software established Google Analytics, which provides it the capability to analyse large amounts of network data. Link and traffic data analysis have been shown to reveal social and market patterns that includes unemployment and property market trends [see Trancer, 2007]. Google Analytics together with its Financial Trends analysis tool opens up an unprecedented level of discovery capabilities. Currently there are no laws that restrict data mining in any way at this moment, in contrast with telecommunication laws that prevent e.g. the taping of phone conversations. The rapid expansion of Google’s business scope which now has blurred boundaries raises the danger of them crossing over into everybody’s business.

2.3 Responsibility to Shareholders After Going Public

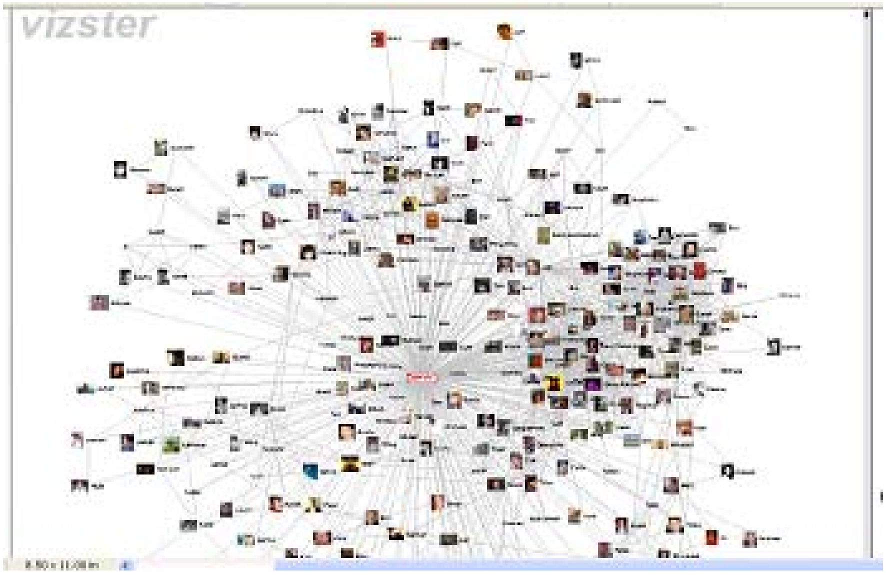
After going public Google's prime concern has to lie with their shareholders who can hold Google's management responsible for all decisions, also with respect to missed opportunities. Hence, what started as a quest for the best search engine respecting the user might turn into directly exploiting users by mining information, as well as shaping their view to increase revenues. "The world's biggest, best-loved search engine owes its success to supreme technology and a simple rule: Don't be evil. Now the geek icon is finding that moral compromise is just the cost of doing big business." [McHugh, 2003]

3. Data Mining and the Preservation of Privacy

Google has realized search has to cover all aspects of our life. Based on Google community management tools and the analytical capability, it will also be able to visualize and track social behavioral patterns based on user networks [see Figure 8]. The ability to link such patterns with other analysis highlights the danger of Google becoming the 'Big Brother'. Privacy and abuse of personalized information for commercial purposes will become a major concern. To make things worse, there are also currently no restrictions of what can be discovered and to whom it may be passed on to (for reasons such as tracking terrorism).

It has been shown that even in anonymized data individuals can be singled out by just a small number of features. For instance, persons can quite reliably be identified by records listing solely e.g., their birth date, gender or ZIP code [Lipson, Dietrich, 2004]. Therefore, only recently the release of a large anonymized data set by the internet portal provider AOL to the information retrieval research community, raised some severe concerns [Hafner, 2006]. It included 20 million Web queries from 650,000 AOL users. Basically the data consisted of all searches from these users for a three month period this year, as well as whether they clicked on a result, what that result was and where it appeared on the result page. Shortly after the release New York Times reporters were indeed able to connect real life people with some of the queries.

Figure 8: Social Network Visualisation (extracted from Heer et al, 2007)



4. Shaping the View of the World

4.1 Restricting Access According to Political Viewpoints

By adapting their index, search engines are in control to authoritatively determine what is findable, and what is kept outside the view of Web users. There is a major concern that search engines become gatekeepers regarding the control of information. As the information presented to users also shapes the worldviews of users, search engines face challenges in maintaining a fair and democratic access.

As with Google's digitization project there are already concerns about the bias in the information store, which mainly contains American-skewed resources [Goth G, 2005]. Other concerns stem from the control of information access as regulated by governments and are already heavily discussed in the community. As gatekeeper of information repositories, Google has for instance recently made allowances to freedom of access and accuracy as required by the Chinese government. [Goth G, 2005]. The policy of Google with regards to oppressive regimes is clearly highlighted by their censored version of Web search. [Wakefield, 2006]

4.2 Objectivity of Ranking Strategy and Product Bundling

Google's initial strategy of ranking documents in a fair and objective way (depending on link structures) has been replaced by its deliberatively supporting or ignoring sites as economic or political issues are demanding. It has been shown that Google's page ranking algorithm is biased towards the bigger companies and technology companies. [Upstil et al, 2003a]. [Upstil et al, 2003b] further indicates that the page ranks made available to public by Google, might not be the same as the actually used internal ranking.

A blog posting by Blake Ross, [Ross, 2007] reported that, Google has been displaying 'tips' that point searchers to Google's own product such as Calendar, Blogger and Picasa for any search phrase that includes words 'calendar' (e.g. Yahoo calendar), 'blog' and 'photo sharing', respectively (see Figure 9). He further added that, "In many ways, Google's new age 'bundling' is far worse than anything Microsoft did or even could do." As compared to Microsoft, Google has enough knowledge of what users want and can thus discreetly recommend its products at the right time. Paired with the Google business model of offering advertisement-supported services free to end users, this forms an explosive combination. If such bundling is not checked, a large number of companies could become sidelined and be forced into financial difficulties.

In order to illustrate the power of product bundling, Google's calendar service increased its market share by 333%, from June 2006 to December 2006. In the process it has overtaking MSN Calendar and is fast approaching Yahoo! Calendar in market share of US visits. As opposed to Yahoo and Microsoft, whose traffic mainly comes from their own mail users, Google's traffic however largely comes from their Search engine users [Prescott, 2007].

Figure 9: View of Google's Home Page (extracted from Ross, 2007)

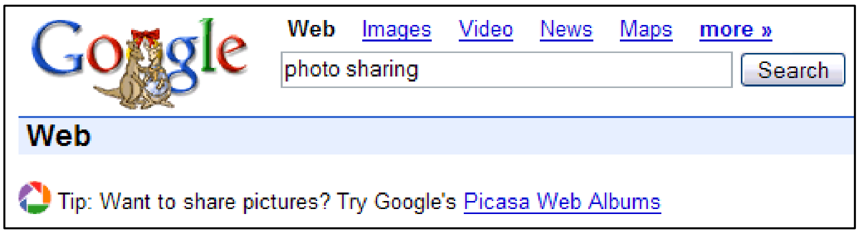


photo sharing

Tip: Want to share pictures? Try Google's Picasa Web Albums

4.3 Symbolic Power and Exclusive Control to the Most Powerful Search Engine Technology

As people become more and more dependent on the Web and become fully trusting to whatever it says, large search engines will then have the absolute power to influence the views of millions. This form of power is referred to as symbolic power [Couldry, 2003], which describes the ability to manipulate symbols to influence individual life. Web Mining has thus put in the hands of a few large companies the power to affect the lives of millions by their control over the universe of information. They have the power to alter the recording of historical events [Witten et al, 2007]. They also have the ability to decide on the 'account of truth' which could well be restricted or product-biased. The full potential of their symbolic powers is however yet to be seen.

The paper by [Maurer and Zaka, 2006] has revealed the exceptional ability of Google Search in detecting document similarity in plagiarism detection. Their results were superior to that of even established Plagiarism detection systems. As Google does not license its search technology to institutions, they maintain the exclusive control over a powerful search capability that could well be adapted to a wide range of applications developed in a variety of fields in future.

4.4 Monopoly over Networked Operating System

Google freely provides an expanding list of services that goes beyond search to cover numerous collaborative personal and community management tools such as shared document, and spreadsheets, Google Mail, Google Calendar, Desktop Search and Google Groups, Google Talk and Google Reader. These applications will drive users to get accustomed with integrated collaborative applications built on top of a Networked Operating system as opposed to Desktop operating systems. The emergence of a participative Web [see Maurer, Kolbitsch, 2006] together with an application development paradigm, mashups [see Kulathuramaiyer, Maurer, 2007] is further driving more and more developers to build integrated Web applications on the networked operating system. Google's firm control over its integrated hardware and software platform will enable it to dominate over a network operating system. According to a quote in a blog entry by [Skrenta, 2007b]: "Google is not the competitor, Google is the environment."

5. *Conclusions*

We have argued that a Web search engine monopolist has the power to develop numerous applications taking advantage of their comprehensive information base in connection with their data mining and similarity detection ability. This ranges from intellectual property violations to the personal identification of legal and medical cases. Currently Google is the most promising contender for a factual Web search engine monopoly. The obvious conclusion is that the non-constrained scope of Google's business will make it very difficult for competitors to match or contain their explosive expansion.

As the Web is a people-oriented platform, a consolidated community effort stands out as a neutralizing factor for the ensuing imbalance economical and social imbalance. Still the ranking mechanisms of leading search engines are predominantly based on popularity of sites. In this sense, 'netizens' thus hold the power, in determining the course and the future of the Web. Community-driven initiatives would be able to impose change and could even possibly call for a paradigm-shift. A good example are so-called Google-bombs [See Wikipedia, Google Bombs, 2007], which are a form of community influence on search result visibility. In 2005 community actions by political parties were able link the homepage of George W. Bush directly to the search phrase 'miserable failure' [Tatum, 2005]. The opposition party retaliated by also enlisting names of other leaders to the same phrase. [Tatum, 2006] highlights an incident where Google was forced to remove a top-ranked link from its search, as a result of community action. Prior to the removal, concerted community activity had managed to shift the poll positions of results.

We advocate that in the long run internationally accepted laws are necessary to both curtail virtual company expansion and to characterize the scope of data mining. In their absence, the monopoly of Google should be considered carefully. We feel that the community has to wake up to the threat that a Web search engine monopoly poses and discuss adequate action to deal with its implications.

Acknowledgement

The authors would like to thank Professor Hermann Maurer for his invaluable initial input and in pointing out the various developments.

References

[Battelle,J., 2005] Battelle,J., The Search- How Google and Its Rivals Rewrote the Rules of Business and Transformed our Culture, Porfolio, Penguin Group, New York, 2005

[Berners-Lee, 2006], Berners-Lee, T, Neutrality of the Net, http://dig.csail.mit.edu/breadcrumbs/blog/4, 2006

[Burright,2006] Burright, M, Database Reviews and Reports-Google Scholar -- Science & Technology , http://www.istl.org/06-winter/databases2.html, 2006

[Colburn, 2007] Colburn, M., comScore Releases December Search Rankings, http://battellemedia.com/archives/003270.php, 2007

[Couldry, 2003], Couldry, N., Media and Symbolic Power: Extending the Range of Bourdieu's Field Theory.' Media@lse Electronic Working Papers Department of Media and Communications, LSE No 2. http://www.lse.ac.uk/collections/media@lse/Default.htm, 2003

[Cusumano, 2005] Cusumano, M: Google: What it is and what it is not. CACM, Vol. 48(2), 2005

[Fallows,2005] Fallows,D., Search Engine Users. Report in The Pew Internet & American Life Project, http://www.pewinternet.org/pdfs/PIP_Searchengine_users.pdf, 2005

[Fortune 100, 2007] Fortune 100 Best Companies to work in 2007: http://money.cnn.com/ magazines/fortune/bestcompanies/2007/full_list/, Accessed 13 January 2007

[Google Corporate, Philosophy, 2007] Google Corporate: Philosophy Statement, 2007 Website: http://www.google.com/corporate/tenthings.html, Accessed 13 January 2007

[Google Corporate, Technology, 2007] Google Corporate: Technology, http://www.google.com/corporate/tech.html, Accessed 13 January 2007

[Google Books Library Project, 2006], Google Books Library Project, http://books.google.com/googlebooks/library.html , 2006

[Google Policy, 2007] Google Policy, http://www.google.com/support/bin/answer.py?answer=17795, accessed 30 January 2007

[Google Video Ads, 2006]A look inside Google AdSense, Introducing Video Ads, http://adsense.blogspot.com/2006/05/introducing-video-ads.html, 2006

[Goth G, 2005],Who and Where are the New Media Gatekeepers, IEEE Distributed Systems Online 1541-4922 2005,IEEE Computer Society, July 2005 Vol. 6, No. 7; http://ieeexplore.ieee.org/iel5/8968/32220/01501754.pdf?isnumber=&arnumber=1501754 July 2005

[Hafner, 2006] Hafner,K. , Tempting Data, Privacy Concerns; Researchers Yearn To Use AOL Logs, But They Hesitate. The New York Times, August 23, 2006

[Heer, Boyd, 2005] Heer,J., Boyd,D., Vizster: Visualising Online Social Networks, IEEE Symposium on Information Visualisation, http://www.danah.org/papers/InfoViz2005.pdf , 2005

[Kulathuramaiyer, Maurer, 2007], Kulathuramaiyer,N., Maurer, H. Current Development of Mashups in Shaping Web Applications, submitted to Ed-Media 2007

[Lenssen, 2007], Lenssen, P, Google’s Automated Resume Filter, Google Blogscopped, http://blog.outer-court.com/archive/2007-01-03-n81.html, 2007

[Lipson, Dietrich, 2004] Lipson, H. , Dietrich, S., 2004, Levels of Anonymity and Traceability (LEVANT). In Bergey, J.; Dietrich, S.; Firesmith, D.; Forrester, E.; Jordan, A.; Kazman, R.; Lewis, G.; Lipson, H.; Mead, N.; Morris, E.; O’Brien, L.; Siviy, J.; Smith, D.; & Woody, C. Results of SEI Independent Research and Development Projects and Report on Emerging Technologies and Technology Trends (CMU/SEI-2004-TR-018), pp. 4-12, http://www.sei.cmu.edu/publications/documents/04.reports/04tr018.html, 2004.

[Maurer,Kolbitsch, 2006] Maurer, H.; Kolbitsch, J., The Transformation of the Web: How Emerging Communities Shape the Information we Consume, J.UCS (Journal of Universal Computer Science) 12, 2 (2006), 187-213

[Maurer, Zaka 2006] Maurer,H., Zaka,B., Plagiarism- A Problem and How to Fight it, Website: http://www.iicm.tugraz.at/iicm_papers/plagiarism_ED-MEDIA.doc , 2006

[McHugh, 2003] McHugh, J., Google vs. Evil, Wired Magazine, Issue 11.01, 2003

[Olsen,Mills, 2005], Olsen, S. Mills, E., AOL to Stick with Google, http://news.com.com/AOL+to+stick+with+Google/2100-1030_3-5998600.html, 2005

[Prescott, 2007] Prescott,L, Google Calendar Up Threefold Since June, http://weblogs.hitwise.com/leeann-prescott/2007/01/google_calendar_up_threefold_s_1.html, 2007

[Rodgers, 2006] Rodgers, Z, Google Opens Audio Ads Beta, http://clickz.com/showPage.html?page=3624149, 2006

[Ross, 2007] Ross, B., Tip: Trust is hard to gain, easy to lose. http://www.blakeross.com/2006/12/25/google-tips/, Accessed 13 January 2007

[Skrenta, 2007a] Skrenta, R., Google's true search market share is 70%, Website: http://www.skrenta.com/2006/12/googles_true_search_market_sha.html, Accessed 17th January, 2007

[Skrenta, 2007b] Skrenta, R., Winner-Take-All: Google and the Third Age of Computing, Website: http://www.skrenta.com/2007/01/winnertakeall_google_and_the_t.html, Accessed 17 January, 2007

[Tatum, 2006], Tatum, C., Deconstructing Google Bombs-A breach of Symbolic Power or Just a Goofy Prank, http://www.firstmonday.org/issues/issue10_10/tatum/index.html, Accessed 31 January 2007

[Trancer, 2007] Trancer, B., July Unemployment Numbers (U.S.) - Calling All Economists, http://weblogs.hitwise.com/billtancer/2006/08/july_unemployment_numbers_us_c.html Accessed 17 January 2007

[Upstill, 2003a] Upstill, T, Crawell, N and Hawking, D., Predicting Fame and Fortune, Proceedings of the 8th Australasian Document Computing Symposium, Australia, http://cs.anu.edu.au/~Trystan.Upstill/pubs/pubs.html#adcs2003 , 2003

[Upstill, 2003b] Upstill, T, Craswell, N and Hawking, D, 2003b, Query-Independent Evidence in Home Page Finding , ACM Transactions on Information Systems, Vol. 21, No. 3, http://cs.anu.edu.au/~Trystan.Upstill/pubs/tois-may03-final.pdf, 2003

[Vise, Malseed, 2006] Vise, D.A., Malseed,M., The Google Story- Inside the Hottest Business, Media and Technology Success of our Time, Pan MacMillan Books, Great Britain, 2006

[Wakefield, 2006] Wakefield, J, Google faces China challenges , BBC News, 25 January http://news.bbc.co.uk/2/hi/technology/4647468.stm, 2006

[Weber, 2006] Weber, S., Das Google-Copy-Paste-Syndrom, Wie Netzplagiate Ausbildung und Wissen gefährden, Heise, Hannover, 2006

[Wikipedia, Google Bombs, 2007] Wikipedia, Google Bombs,
http://en.wikipedia.org/wiki/Google_bomb, Accessed, 30 January, 2007

[Witten & al, 2007] Witten, I.H., Gori, M., Numerico, T., Web Dragons, Inside the Myths of Search Engine Technology, Morgan Kaufmann, San Francisco, 2007

Appendix 5: Data Mining is becoming Extremely Powerful, but Dangerous

Working paper by N. Kulathuramaiyer and H. Maurer

Abstract: Data Mining describes a technology that discovers non-trivial hidden patterns in a large collection of data. Although this technology has a tremendous impact on our lives, the invaluable contributions of this invisible technology often go unnoticed. This paper discusses advances in data mining while focusing on the emerging data mining capability. Such data mining applications perform multidimensional mining on a wide variety of heterogeneous data sources, providing solutions to many unresolved problems. This paper also highlights the advantages and disadvantages arising from the ever-expanding scope of data mining. Data Mining augments human intelligence by equipping us with a wealth of knowledge and by empowering us to perform our daily task better. As the mining scope and capacity increases, users and organisations become more willing to compromise privacy. The huge data stores of the 'master miners' allow them to gain deep insights into individual lifestyles and their social and behavioural patterns. Data integration and analysis capability of combining business and financial trends together with the ability to deterministically track market changes will drastically affect our lives.

1. Introduction

As we become overwhelmed by an influx of data, Data Mining presents a refreshing means to deal with this onslaught. Data Mining thus holds the key to many unresolved age-old problems. Having access to data thus becomes a powerful capability which can be effectively harnessed by sophisticated mining software.

According to [Han, Kamber, 2006] data mining is defined as the extraction of interesting (non trivial, implicit, previously unknown and potentially useful) information or patterns from data in large databases. We take a broad view of data mining, where we also include other related machine based discoveries such as deductive query processing and visual data mining. Databases may include both structured data (in relational databases), semi structured data (e.g. metadata in XML documents) as well as unstructured documents such as text documents and multimedia content.

Data mining has been widely employed for the learning of consumer behaviour based on historical data of purchases made at retail outlets. Demographic data as collected from loyalty cards is combined with behavioural patterns of buyers to enable retailers in designing promotional programmes for specific customer segments. Similarly, credit card companies use data mining to discover deviations in spending patterns of customers to overcome fraud. Through this, these companies are able to guarantee the highest quality of service to their customers.

Despite these success stories in areas such as customer relationship modelling, fraud detection, banking, [KDD, 2005], the majority of applications tend to employ generic approaches and lack due integration with workflow systems. As such, Data Mining is currently seen as being at a chasm state and has yet to become widely adopted by the large majority [Han, Kamber, 2006].

2. Overview of Data Mining Process

Having access to data thus becomes a powerful capability which can then effectively be harnessed by sophisticated mining software. The statement by O'Reilly [O'Reilly, 2006] that 'Data is the Next Intel Inside' illustrates its hidden potency. Data at the hands of credit card companies, will allow them to profile customers according to lifestyles, spending pattern and brand loyalty. Political parties are now able to predict with reasonable accuracy how voters are likely to vote [Rash, 2006].

Data Mining involves the extraction of patterns from a collection of data via the use of machine learning algorithms. Sophisticated mining technologies of today integrate multiple machine learning algorithms to perform one or more of the following functions:

- construct an aggregated or personalised predictive model of systems , events and individuals being studied and supporting decision making by employing these models in a number of ways (extraction of classification patterns)
- identify similarity/dissimilarity in terms of distributional patterns of data items and their relationships with associated entities (clustering)
- uncover associational and behavioral patterns based on relationship drawn from transactional data (associational pattern mining)
- determine trends highlighting both the norm as well as deviations based on observable patterns. (e.g mathematical data modelling)
- Determine sequential patterns amongst a number events or state of data objects to depict behavioural patterns (sequential pattern mining)

Data Mining typically describes an automated acquisition of knowledge from a large collection of data. This traditional view corresponds to the knowledge creation phase in knowledge management. Current developments of data mining have expanded this to also cover support for knowledge organization and consolidation with respect to existing domain knowledge, and data visualization for iterative learning. Data Mining can thus be seen as complementing and supplementing knowledge management in a variety of phases.

3. Data Mining Process

In order to describe the processes involved in performing data mining we divide it into 3 phases: domain focussing, model construction (actual mining using machine learning algorithms), and decision making (applying the model to unseen instances).

Data focusing phase involves the application of some form of clustering or may incorporate intensive knowledge engineering for complex applications. At the model construction phase, a model of generalized patterns is constructed to capture the intrinsic patterns stored in the data.

The model generated is then employed for decision-making. Simplistic applications of data mining tend to merely employ the model to predict the likelihood of events and occurrences, based largely on past patterns. Amazon, for example, is able to recommend books according to a user’s profile. Similarly, network operators are able to track fraudulent activities in the usage of phone lines by tracking deviation patterns as compared to standard usage characterization. Figure 10 compares the elaborate mining tasks performed in an emerging application as opposed to a traditional data mining system.

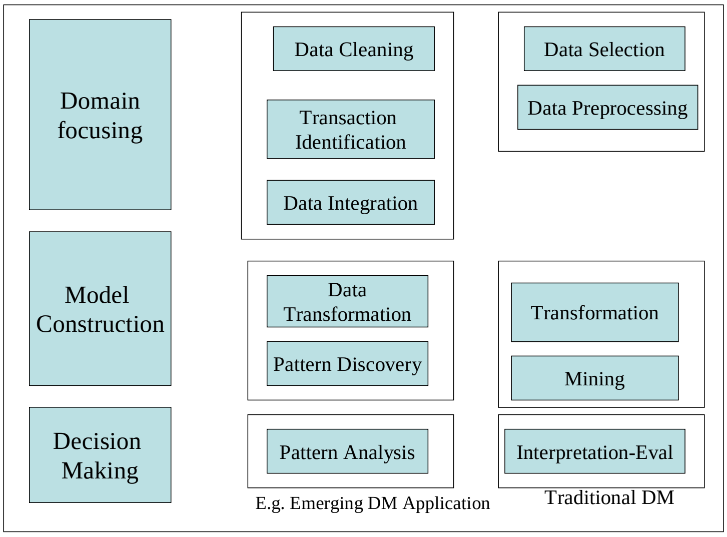


Figure 10: Illustration of steps involved in a DM Process

In emerging applications, domain focusing will be concerned with the discovery of causal relationships (e.g. using Bayesian networks) as a modeling step. Multiple sources of data need to be incorporated in the discovery of likely causal relationship patterns. A complex form of data mining is required even at the phase of domain focusing. This will involve an iterative process whereby hypothesis generation could be employed to narrow the scope of the problem to allow for a constrained but meaningful data collection. For complex domains such as this, domain focusing can also provide insights on constraining and refining the data collection process. Domain focusing would thus perform problem detection, finding deterministic factors and to hypothesize relationships that will be applied in the model [Beulens et al, 2006].

The model construction phase will then employ a variety of learning algorithms, to profile events or entities being modeled. As this stage may negate model relationships, domain focusing will need to be repeated and iteratively performed. The model construction phase will allow the incremental development of a model, based on a complex representation of the causal networks [Beulens et al, 2006].

Mining methods (e.g. clustering, associational rule mining, Bayesian or neural network classifiers) could be used to verify the validity of causal associations. Once a potential causal link is hypothesized, verification can be done via the application of data mining methods. A combination of approaches could be employed to include deviation detection, classification, dependence model and causal model generation [Beulens et al, 2006].

The Decision Making phase will subsequently apply the validated causal relationship model in exploring life case studies. An environment for an interactive explorative visual domain focusing is crucial, to highlight directions for further research. Data mining could serve as a means of characterization of profiles for both areas prone to disasters or those that are safe. The results of the data mining process would need to be seen merely as recommendations or suggested relationship in complex applications. It is thus typical to engage the community in verifying the results in an on-the-job scenario (as described in[Mobasher, 2005]).

4. Applications of Data Mining

Emerging forms of data mining applications deal with complex-structured domain problems e.g. environmental modelling and medical data mining. The lack of awareness of the structure of complex domain problems reveals the importance of data collection presenting a need for discovery social behavioural patterns. Social relationship discovery will serve as a means of providing a deeper understanding for many of such domain problems.

4.1 Environmental Modeling Application

For complex-structured problems, data mining could be used to provide answers by uncovering patterns hidden beneath layers of data. In many cases, domain focusing has in the past has been the biggest challenge. Data mining could be employed for the modeling of environmental conditions in the development of an early warning system to address a wide range of natural disasters such as avalanches, landslides, tsunami and other environment events such as global warming. The main challenge in addressing such problems is in the lack of understanding of structural patterns characterizing various parameters.

As highlighted by [Maurer et al, 2007], although a large variety of computer-based methods have been used for the prediction of natural disasters, the ideal instrument for forecasting has not been found yet. As highlighted in their paper, there are also situations whereby novel techniques have been employed but only to a narrow domain of limited circumstances.

Integration of multiple databases and the compilation of new sources of data are required in the development of full-scale environmental solutions. As advances in technology allow the construction of massive databases through the availability of new modes of input such as multimedia data and other forms of sensory data, data mining could well provide a solution. In order to shed insights on a complex problem such as this, massive databases that were not previously available need to be constructed e.g. data about after event situations of the past [Maurer et al, 2007]. Such data on past events could be useful in highlighting patterns related to potentially in-danger sites. Data to be employed in this mining will thus comprise of both of weather and terrestrial parameters together with other human induced parameters such as vegetation or deforestation over a period of time [Maurer et al, 2007]. As such the mining of social and behavioural patterns will also play an important role in such complex applications.

4.2 Medical Application

In the medical domain, data mining can be applied to discover unknown causes to diseases such as 'sudden death' syndrome or heart attacks which remain unresolved in the medical domain. The main difficulty in performing such discoveries is also in collecting the data necessary to make rational judgments. Large databases need to be developed to provide the modeling capabilities. These databases will comprise of clinical data on patients found to have the disease, and those who are free of it. Additionally non-traditional data such as including retail sales data to determine the purchase of drugs, and calls to emergency rooms together with auxiliary data such as micro array data in genomic databases and environmental data would also be required. [Li, 2007]

Non traditional data could also incorporate major emotional states of patients by analyzing and clustering the magnetic field of human brains which can be measured non invasively using electrodes to a persons' heads. [Maurer et al, 2007] Social patterns can also be determined through profile mining as described in the previous section to augment the findings of this system. The power of considering social patterns in gaining deeper insights on focused individual is explain in the next section.

The development of large databases for medical explorations will also open possibilities for other discoveries such as mining family medical history and survival analysis to predict life spans. [Han and Kamber, 2006] The process of data mining in such domains will involve numerous iterations in data focusing and scooping before data mining validates relationships.

5. Web Mining and Search as a Social Data Mining Application

Web Mining can typically be divided into Web Page content mining, Web structure mining and Web log mining (including search log). Traditional search engines basically utilised web content only for building their index of the Web. Web structure has however become important in current search engines which employs web structure patterns to determine popularity of websites (i.e. PageRank algorithm). Web log mining is also becoming an important source of data mining for providing an analysis of customer relationship modeling and other form of user profiling. Global search engines of today combine these three forms of mining to provide results that is able to meet users needs better. As the semantic Web emerges, content mining of semantic relationships will also begin to be explored. As such travel agents could mine Web contents to populate their semantic web of flight schedules, or hotel price listings. We will now focus on Web Search as a complex form data mining at both a collective level and at a personal level, in revealing the true potential of unlimited data mining.

Search engines have turned the Web into a massive data warehouse as well as a playground for automated discovery of hidden treasures. Web Search is thus viewed as an extensive form of multidimensional heterogeneous mining of a largely unstructured database for uncovering an unlimited number of mind-boggling facts. The strength of search engines stems from its absolute control over vast collections of data and the various analytical tools it has. These tools include text processing and mining systems, translators, content aggregators, data visualisation tools, data integration tools, data traffic analyzer, financial trends analyzer, context aware systems, etc. The analytical tools provide alternative mining resources for enhancing the quality of discoveries.

The data at their disposal include, but are not limited to web pages, email collections, discussion groups and forums, images, video, books, financial data, news, desktop content, scholarly papers, patents, geographical data, chronological data, community generated dynamic tagged content (video, music, writings), product offerings, local business data, shared documents, and user profiles. The expanse of data resources is effectively exploited in their ability to support users' decision-making process, as well as in providing alternative channels for further investigation. The scale of data available is in the range of peta bytes, and it much greater than the terra bytes of data available at the hands of large global corporations such as Walmart.

Search engines can either simultaneously or incrementally mine these datasets to provide a variety of search results which include phone contacts, street addresses, news feeds, dynamic web content, images, video, audio, speech, books, artifacts.

The distinguishing feature of the mining performed is seen in domain focusing. A model allowing the characterization of aggregated user search behaviour is constructed [Coole et al, 1997]. This phase may also involve associational subject link analysis, requiring a context-sensitive domain analysis (as done for mobile users). This Mining phase involves the derivation of aggregated usage profiles based on a multidimensional mining of usage patterns according to clustered characterization. Figure 11 illustrates the scope and extent of mining performed by search engines.

By analyzing search history over a period of time, search engines have access to a great deal of insights into lives of presumably 'anonymous' searchers. A search query can indicate the intent of a user to acquire particular information to accomplish a task at hand. Search engines can thus track patterns and drifts in global user intentions, moods, and thoughts.

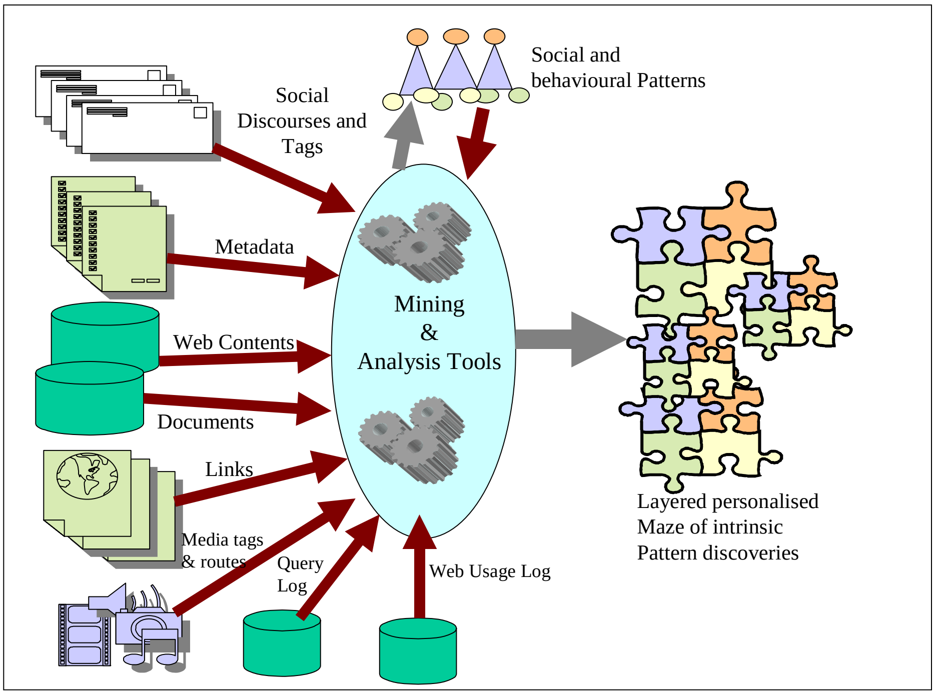

Social and behavioural Patterns
Social Discourses and Tags
Metadata
Web Contents
Documents
Links
Media tags & routes
Query Log
Web Usage Log
Mining & Analysis Tools
Layered personalised Maze of intrinsic Pattern discoveries

Figure 11: Extensive Mining of Heterogeneous Data sources

Search traffic patterns is another data source that can be applied to highlight relationships between search terms and events. For instance the number of searches for "Christmas presents" peaks in the early part of the month of December [Hopkins, 2007]. Search traffic data analysis have also been shown to reveal social and market patterns such as unemployment and property market trends (see [Trancer, 2007]). Apart from that the intentions of global users can be modelled by terms employed in search. An illustration of intent revealing search queries can be discovered via the AOL search database [Zhao, Sapp, 2006] (which allows the search results leaked by AOL [Kantor, 2006] to be used for academic purposes). The intent logs make the intentions of users absolutely obvious to global search companies. Examples of such search queries include 'finance major vs. accounting major', 'shoes from India', 'what does vintage mean', 'baby looney tunes baby shower party invitations', 'how many web pages are on the internet', 'feline heartworm remedies', 'salaries for forensic accountants', 'hamilton county chinese stores that sell chinese candy white rabbit', 'how do you read the stock market or invest in stock', 'on the basis of the settlement of unfortunate past and the outstanding issues of concern', 'riverview towers apartment information in Pittsburg'.
A sudden burst of search term frequency have been observed seeking quick answers to questions posed in reality shows, such as "Who wants to be a Millionaire" [Witten, 2007]. An emerging paradigm, mashups (see [Kulathuramaiyer, Maurer, 2007]) together with mobile web services further allows the discovery of localised contextual profiles.

Targeted advertisements based on keyword-bidding is currently employed by search engines. In the near future, complex mining capabilities will provide personalised context specific suggestions which will intrude into our daily life. E.g. It would be possible via RFID technology, for a user passing by an intelligent electronic billboard (as described in [NewScientistTech, 2007]), to encounter a highly personalized messages such as 'Nara, you have not purchased your airline ticket yet, you have only 2 weeks for your intended flight. I do know of a discount you can't refuse'. Such a level of user profiling could easily be achieved by merely connecting shopping cart analysis, together with cookies and calendar entries. Figure 12 illustrates the layered mining that could be employed to facilitate such a discovery. This is described by [Kulathuramaiyer, Balke, 2006] as connecting the dots, to illustrate the ability to extract and harness knowledge from massive databases at an unprecedented level.

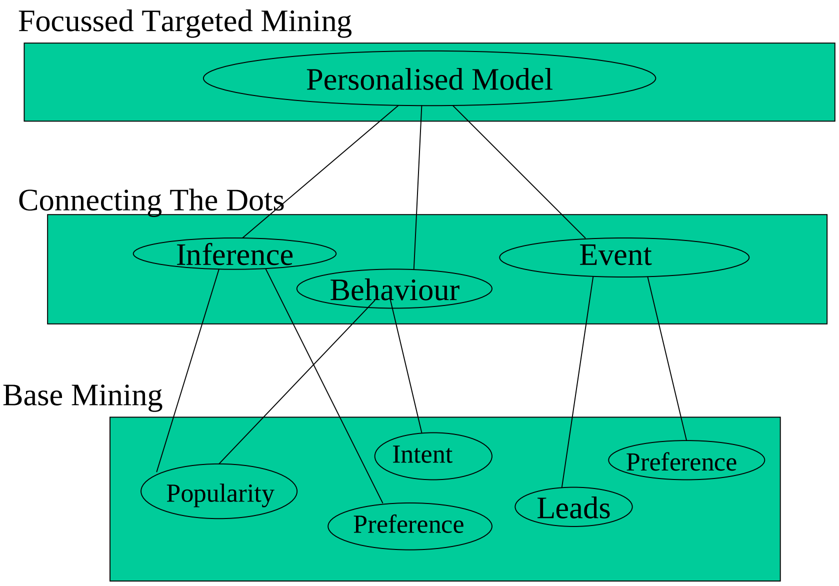


Figure 12: Insightful Connected Discoveries

Figure 13 then illustrates the amount of knowledge about anonymous users that could be established by global search engines, via the connection of dots (see [Kulathuramaiyer, Balke 2006]). We will now concentrate on the implications of this technology on our lives.

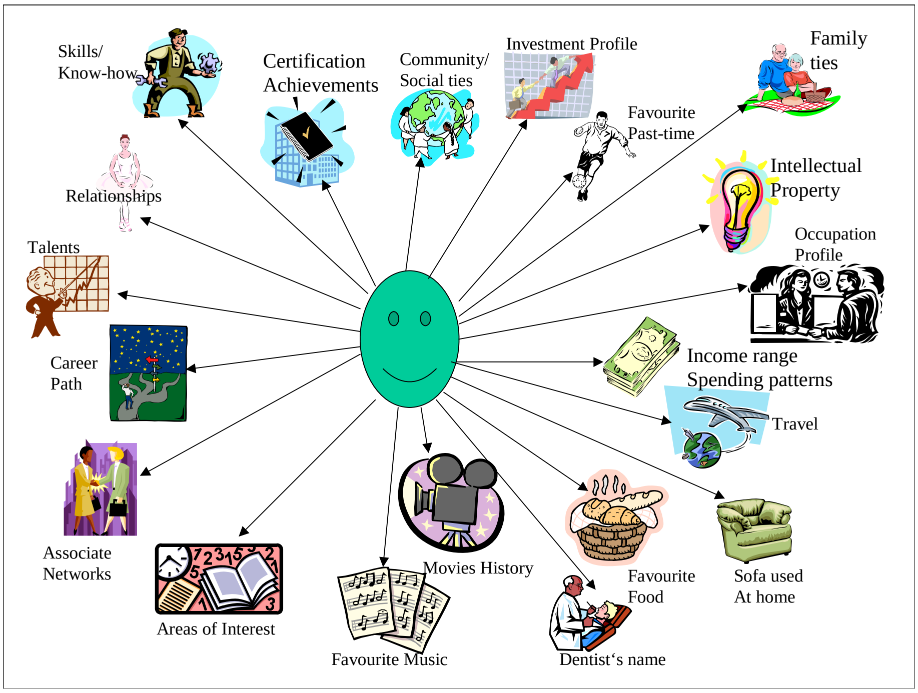


Figure 13: Search and Social Interaction History can reveal a great deal of information about users

6. Implications of Data Mining

6.1 The Advantages of Data Mining

Data mining has crept into our lives in a variety of forms. It has empowered individuals across the world to vastly improve the capacity of decision making in focussed areas. Powerful mining tools are going to become available for a large number of people in the near future.

The benefits of data mining will include preserving domestic security through a number of surveillance systems, providing better health through medical mining applications, protection against many other forms of intriguing dangers, and access to just-in-time technology to address specific needs. E.g. Police in the United States are using a software called CopLink that is able to discover relationships leading to identification of suspects by connecting data from multiple sources [Goldman, 2003]. Data Mining will also provide companies effective means of managing and utilizing resources. People and organizations will acquire the ability to perform well-informed (and possibly well-researched) decision-making. Data mining also provides answers through sifting through multiple sources of information which were never known to exist, or could not be conceivably acquired to provide enlightening answers. Data Mining could be combined with collaborative tools to further facilitate and enhance decision-making in a variety of ways. Data mining is thus able to explicate personal or organizational knowledge which may be locked in the heads of individuals (tacit knowledge) or in legacy databases, to become available. Many more new benefits will emerge as technology advances.

6.2. Disadvantages of Data Mining

A great deal of knowledge about users is also being maintained by governments, airlines, medical profiles or shopping consortiums. Mining applications with dramatic privacy infringement implications include search history, real-time outbreak and disease surveillance program, early warning for bio-terrorism [Spice, 2005] and Total Information Awareness program [Anderson, 2004]. For example, search history data represents an extremely personal flow of thought patterns of users that reflects ones quest for knowledge, curiosity, desires, aspirations, as well as social inclinations and tendencies. Such logs can reveal a large amount of psychographic data such as user's attitudes towards topics, interests, lifestyles, intents and beliefs. A valid concern would be that the slightest leak could be disastrous. The extent of possible discoveries has been clearly illustrated by the incidence where AOL released personal data of 658,000 subscribers [Jones, 2006], [Kantor, 2006].

Another common danger is profiling where there is a possibility of drastic implications such as a conviction being made based on the incriminating evidences of mining results. There is also a danger of over-generalization based on factors such as race, ethnicity, or gender. This could result in false positives, where an entirely innocent individual or group is targeted for investigation based on a poor decision making process. For the domestic security application, a reasonable data mining success rate of 80% implies that 20% of all US citizens (or 48 million people) would be considered false positives [Manjoo, 2002].

Data mining will further empower mining kingpins to be able to go beyond the ability to PREDICT what is going to happen in a number of areas of economic importance, but actually have the power to KNOW what will happen. It is possible to detect trends and social behavioural patterns (e.g. stock purchase, property acquisition) by analyzing traffic data and data logs of millions of users. Could such a capability not be applied to exploit the stock market in an unprecedented way?

By connected domain focusing, they will also have the capacity to make judgments on issues and persons with scary accuracy. There has been instances when the data archived by global search engines have been used as incriminating evidences in criminal proceedings which have led to convictions.

7. *What can we do?*

Most related works are concerned more about privacy, but it is no longer the main issue of concern. [Kovatcheva, 2002] has a proposed a means of protecting anonymity by the use of anonymity agents and pseudonym agents to avoid users from being identified. Their paper also proposed the use of negotiation and trust agents to assist users in reviewing a request from a service before making a rational decision of allowing the use of personal data.

A similar agent-based approach is described by [Taipale, 2003] via rule-based processing. An "intelligent agent" is used for dispatching a query to distributed databases. The agent will negotiate access and permitted uses for each database. Data items are labeled with meta-data describing how that item must be processed. Thus, a data item is protected as it retains relevant rules by which it describes the way it has to be processed. The main challenge then lies in coming up with guidelines and rules such that site administrators or software agents can use to direct various analyses on data without compromising the identity of an individual user. This approach however is not applicable for areas such as Web search, where a standard framework for conscientious mining is far from sight.

Furthermore, the concern with the emergence of extensive mining is no longer solved by addressing privacy issues only. As more and more people are willing to compromise privacy, the questions that we pose are: Who do we trust as the gatekeeper of all our data? Do we then trust all our private data at the hands of a commercial global company?

One approach to overcome the concerns mentioned above is by employing a distributed data mining approach, where separate agencies will maintain and become responsible and accountable for different (independent segments of) data repositories. This proposal further ensures that no central agency will have an exclusive control over the powerful mining technology and all resources. [Kulathuramaiyer, Maurer, 2007] In order to realize this solution, governments have to start playing a more proactive role in maintaining and preserving national or regional data sources. If such initiatives can be put in place, then it will be possible to also explore the sensitivities associated with data control even at a applications design phase. The research by [Friedman, et. al., 2003] can be applied in considering data sensitivities at the design process.

We are also beginning to see partnerships between global search engines and governments in this respect. Such partnerships should however be built upon a balanced distribution of earnings. Such a solution can be complemented by the existence of a large number of autonomous context-specific or task specific miners.

8. *Conclusion*

As data mining matures and becomes widely deployed in even more encompassing ways, we need to learn to effectively apply it to enrich our lives. At the same time, the dangers associated with this technology needs to be minimized by deliberate efforts on the part of the enforcement agencies, data mining agencies and the users of the system. There should be strict regulations to prevent the abuse or misuse of data. Users should also be made aware of the privacy policies in order to make an informed decision about revealing their personal data. The success of such regulations and guidelines can only be guaranteed if they are backed up by a legal framework.

References

[Anderson, 2004], Anderson, S.R., Total Information Awareness and Beyond, Bill of Rights Defense Committee; White paper. The Dangers of Using Data Mining Technology to Prevent Terrorism, July 2004

[Battelle, 2005] Battelle, J., The Search- How Google and Its Rivals Rewrote the Rules of Business and Transformed our Culture, Porfolio, Penguin Group, New York, 2005

[Beulens et al, 2006] Beulens, A., Li, Y., Kramer, M., van der Vorst, J., Possibilities for applying data mining for early Warning in Food Supply Networks, CSM'06, 20thWorkshop on Methodologies and Tools for Complex System Modeling and Integrated Policy Assessment, August, 2006 http://www.iiasa.ac.at/~marek/ftppub/Pubs/csm06/beulens_pap.pdf

[Coole et al, 1997] Coole, R. Mobasher, B., Srivastava, J., Grouping Web Page References into Transactions for Mining World Wide Web Browsing Patterns, Proceedings of the 1997 IEEE Knowledge and Data Engineering Exchange Workshop Page: 2, 1997 ISBN:0-8186-8230-2 IEEE Computer Society

[EFF, 2006] American Travelers to Get Secret 'Risk Assessment' Scores, Electronic Frontier Foundation (EFF), November 30, 2006, http://www.eff.org/news/archives/2006_11.php

[Friedman, et. al., 2003] Friedman, B., Kahn, P. H., Borning, A, Value Sensitive Design: Theory and Methods, June 2003, http://www.ischool.washington.edu/vsd/vsd-theory-methods-draft-june2003.pdf

[Goldman, 2003] Goldman, J., Google for Cops: Revolutionary software helps cops bust criminals, TechTV, April 12, 2003, http://www.techtv.com/news/scitech/story/0,24195,3424108,00.html

[Graham, et. al., 2003] Graham, J., Page, C. D., Kamal, A., Accelerating the Drug Design Process through Parallel Inductive Logic Programming Data Mining, Proceedings of the Computational Systems Bioinformatics IEEE Computer Society, 2003 http://ieeexplore.ieee.org/iel5/8699/27543/01227345.pdf

[Han and Kamber, 2006] Han, J., and Kamber, M., Data Mining: Concepts and Techniques, 2nd ed., The Morgan Kaufmann Series in Data Management Systems, Jim Gray, Series Editor , Morgan Kaufmann Publishers, March 2006.

[Hofgesang, Kowalczyk, 2005] Hofgesang, P. I., and Kowalczyk, W., Analysing Clickstream Data: From Anomaly Detection to Visitor Profiling, ECML/PKDD Discovery Challenge 2005 http://www.cs.vu.nl/ci/DataMine/DIANA/papers/hofgesang05pkdd.pdf

[Hopkins, 2005] Hopkins, H., Poker & Fantasy Football - Lessons on Finding Affiliate Partnerships, Hitwise Weblogs, 22 Jan 2005 http://weblogs.hitwise.com/heather-hopkins/2005/11/

[Jenssen, 2002] Jenssen, D., Data mining in networks. Invited talk to the Roundtable on Social and Behavior Sciences and Terrorism. National Research Council, Division of Behavioral and Social Sciences and Education, Committee on Law and Justice. Washington, DC. December 11, 2002

[Jones, 2007] Jones, K .C., Fallout From AOL's Data Leak Is Just Beginning , http://www.informationweek.com/news/showArticle.jhtml?articleID=191900935, accessed 2007

[Kantor, 2006] Kantor, A., AOL search data release reveals a great deal, USA Today, August, 17, 2006, http://usatoday.com/tech/columnist/andrewkantor/2006-08-17-aol-data_x.htm

[KDD, 2005] KDDnuggets : Polls: Successful Data Mining Applications, July 2005 http://www.kdnuggets.com/polls/2005/successful_data_mining_applications.htm

[Kovatcheva, 2002] Kovatcheva, E., Tadinen ,H., The technological and social aspects of data mining by means of web server access logs http://www.pafis.shh.fi/~elikov02/SFISWS2/SFIS2.html 18 January 2002

[Kulathuramaiyer, Balke, 2006] Kulathuramaiyer, N., Balke, W.-T., Restricting the View and Connecting the Dots — Dangers of a Web Search Engine Monopoly, J,UCS Vol. 12 , Issue 12, pp.1731 – 1740, 2006

[Kulathuramaiyer N., Maurer, H., 2007], Kulathuramaiyer, N., Maurer, H., "Why is Fighting Plagiarism and IPR Violation Suddenly of Paramount Importance?" Proc. of International Conference on Knowledge Management, Vienna, August 2007.

[Lane 2003] Lane,M., How Terror Talk is Tracked, BBC News Online Wednesday, 21 May, 2003, http://news.bbc.co.uk/2/hi/uk_news/3041151.stm

[Li, 2007] Li, C. S., Survey of Early Warning Systems for Environmental and Public Health Applications, in Wong, ,S., Li,, C. S.,(eds.), Life Science Data Mining, Science, Engineering, and Biology Informatics- Vol. 2, 2007 http://www.worldscibooks.com/compsci/etextbook/6268/6268_chap01.pdf

[Manjoo, 2002] Manjoo, F., Is Big Brother Our Only Hope Against Bin Laden?, Dec. 3, 2002 http://www.salon.com/tech/feature/2002/12/03/tia/index_np.html

[Maurer et al., 2007] Maurer, L., Klingler, C., Pachauri, R. K., Tochtermann, K,. Data Mining as Tool for Protection against Avalanches and Landslides, Proc. Environmental Informatics Conference, Warsaw, 2007

[Milne, 2000] Milne, G. R., Privacy and ethical issues in database/interactive marketing and public policy: A research framework and overview of the special issue, Journal of Public Policy & Marketing, Spring 2000

[Mobasher, 2005] Mobasher, B., Web Usage Mining and Personalisation, in Singh, M. P. (ed.) Practical Handbook of Internet Computing, Chapman & Hall/ CRC Press, 2005 http://maya.cs.depaul.edu/~mobasher/papers/IC-Handbook-04.pdf

[Mobasher, 2006] Mobasher, B., Data Mining for Personalization. In The Adaptive Web: Methods and Strategies of Web Personalization, Brusilovsky, P., Kobsa, A., Nejdl, W. (eds.). Lecture Notes in Computer Science, Vol. 4321. Springer-Verlag, Berlin Heidelberg, 2006, http://maya.cs.depaul.edu/~mobasher/papers/aw06-mobasher.pdf

[NewScientistTech, 2007] Street advertising gets local-stock-savvy, NewScientist.com, January 10, 2007 http://technology.newscientist.com/article/mg19325854.900-street-advertising-gets-localstocksavvy.html

[Rash, 2006] Rash, W., Political Parties Reap Data Mining Benefits ,eWeek.com enterprise News and reviews, November 16, 2006; http://www.eweek.com/article2/0,1895,2060543,00.asp

[Shermach, 2006] Shermach, K. Data Mining: where legality and ethics rarely meet, http://www.crmbuyer.com/story/52616.html, 25th January 2006

[Singel, 2003] Singel,R., Sep, 18, 2003 http://www.wired.com/news/privacy/0,1848,60489,00.html

[Spice, 2003] Spice, B., Privacy in age of data mining topic of workshop at CMU, March 28, 2003 http://www.post-gazette.com/nation/20030328snoopingnat4p4.asp

[Taipale, 2003] Taipale, K.A. "Data Mining and Domestic Security: Connecting the Dots to Make Sense of Data". Colum. Sci. & Tech. L. Rev. 5 (2). SSRN 546782 / OCLC 45263753, December 15, 2003.

Tanasa,D., and Trousse,B., Advanced Data Preprocessing for Intersites Web Usage Mining, AxIS Project Team, INRIA Sophia Antipolis Published by the IEEE Computer Society MARCH/APRIL 2004 http://ieeexplore.ieee.org/iel5/9670/28523/01274912.pdf?arnumber=1274912

[Trancer, 2007] Trancer, B. 2007, July Unemployment Numbers (U.S.) - Calling All Economists http://weblogs.hitwise.com/bill-tancer/2006/08/july_unemployment_numbers_us_c.html Accessed 17 January 2007

[Vaughan-Nichols, 2006] Vaughan-Nichols, S. J., Researchers make Make Search more intelligent, Industry Trends in Computer, (Eds.) Lee Garber, IEEE Computer Society, December 2006

Vise, D.A., Malseed,M., 2006, The Google Story- Inside the Hottest Business, Media and Technology Success of our Time, Pan MacMillan Books, Great Britain, 2006

Witten, I.H., Gori, M., Numerico, T., Web Dragons, 2007, Inside the Myths of Search Engine Technology, Morgan Kaufmann, San Francisco, 2007

[Zhao, Sapp, 2006], Zao, D., Sapp, T., AOL Search Database, http://www.aolsearchdatabase.com/

Appendix 6: Emerging Data Mining Applications: Advantages and Threats

N. Kulathuramaiyer, H.Maurer

Abstract: Data Mining describes a technology that discovers non-trivial hidden patterns in a large collection of data. Although this technology has a tremendous impact on our lives, the invaluable contributions of this invisible technology often go unnoticed. This paper addresses the various forms of data mining while providing insights into its expanding role in enriching our life. Emerging forms of data mining are able to perform multidimensional mining on a wide variety of heterogeneous data sources, providing solutions to many problems. This paper highlights the advantages and disadvantages arising from the ever-expanding scope of data mining. Data Mining augments human intelligence by equipping us with a wealth of knowledge and by empowering us to perform our daily task better. As the mining scope and capacity increases, users and organisations become more willing to compromise privacy. The huge data stores of the 'master miners' allow them to gain deep insights into individual lifestyles and their social and behavioural patterns. The data on business and financial trends together with the ability to deterministically track market changes will allow an unprecedented manipulation of the stock market. Is it then possible to constrain the scope of mining while delivering the promise of better life?

1. Introduction

As we become overwhelmed by an influx of data, Data Mining presents a refreshing means to deal with this onslaught. Data Mining thus holds the key to many unresolved age-old problems. Having access to data thus becomes a powerful capability which can be effectively be harnessed by sophisticated mining software. Data at the hands of credit card companies will allow them to profile customers according to lifestyles, spending patterns and brand loyalty. Political parties on the other hand are able to predict with reasonable accuracy how voters are likely to vote. [Rash, 2006]

According to [Han and Kamber, 2006] data mining is defined as the extraction of interesting (non trivial, implicit, previously unknown and potentially useful) information or patterns from data in large databases. We take a broad view of data mining, where we also include other related machine based discoveries such as deductive query processing and visual data mining. Databases may include both structured data (in relational databases), semi structured data (e.g. metadata in XML documents) as well as unstructured documents such as text documents and multimedia content.

Despite the success stories in areas such as customer relationship modelling, fraud detection, banking, [KDD, 2005], the majority of applications tend to employ generic approaches and lack due integration with workflow systems. As such, Data Mining is currently at a chasm state and has yet to become widely adopted by the large majority [Han and Kamber, 2006].

2. Data Mining Process

Data Mining typically describes an automated acquisition of knowledge from a large collection of data. This traditional view corresponds to the knowledge creation phase in knowledge management. Current developments of data mining have expanded this to also cover support for knowledge organization and consolidation with respect to existing domain knowledge, and data visualization for iterative learning. Data Mining can thus be seen as complementing and supplementing knowledge management in a variety of phases.

In order to describe the processes involved in performing data mining we divide it into 3 phases: domain focusing, model construction (actual mining using machine learning algorithms), and decision making (applying the model to unseen instances).

Data focusing phase involves the application of some form of clustering or may incorporate intensive knowledge engineering for complex applications. At the model construction phase, a model of generalized patterns is constructed to capture the intrinsic patterns stored in the data.

The model generated is then employed for decision-making. Simplistic applications of data mining tend to merely employ the model to predict the likelihood of events and occurrences, based largely on past patterns. Amazon, for example, is able to recommend books according to a user's profile. Similarly, network operators are able to track fraudulent activities in the usage of phone lines by tracking deviation patterns as compared to standard usage characterization.

3. Applications of Data Mining

3.1 Web Search As Data Mining

Search engines have turned the Web into a massive data warehouse as well as a playground for automated discovery of hidden treasures. Web Search is thus viewed as an extensive form of multidimensional heterogeneous mining of a largely unstructured database for uncovering an unlimited number of mind-boggling facts. The scale of data available is in the range of peta bytes, and it much greater than the terra bytes of data available at the hands of large global corporations such as Walmart.

Search engines can either simultaneously or incrementally mine these datasets to provide a variety of search results which include phone contacts, street addresses, news feeds, dynamic web content, images, video, audio, speech, books, artifacts. In performing domain focusing, a model allowing to characterize aggregated user search behaviour is used [Coole et al, 1997]. This phase could involve associational subject link analysis, requiring a contextual domain analysis (for mobile users). This Mining phase involves the derivation of aggregated usage profiles based on a multidimensional mining of usage patterns according to clustered characterization. By analyzing search history over a period of time, search engines have access to a great deal of insights into lives of presumably 'anonymous' searchers. A search query can indicate the intent of a user to acquire particular information to accomplish a task at hand. Search engines can thus track patterns and drifts in global user intentions, moods, and thoughts.

3.2 Environmental Modelling Application

There are complex problems for which data mining could be used to provide answers by uncovering patterns hidden beneath layers of data. In many cases, domain focusing has in the past has been the biggest challenge. Data mining could be employed for the modeling of environmental conditions in the development of an early warning system to address a wide range of natural disasters such as avalanches, landslides, tsunami and other environment events such as global warming. The main challenge in addressing such a problem is in the lack of understanding of structural patterns characterizing various parameters which may currently not be known.

As highlighted by [Maurer et al], although a large variety of computer-based methods have been used for the prediction of natural disasters, the ideal instrument for forecasting has not been found yet. As highlighted in their paper, there are also situations whereby novel techniques have been employed but only to a narrow domain of limited circumstances.

Integration of multiple databases and the compilation of new sources of data are required in the development of full-scale environmental solutions. As advances in technology allow the construction of massive databases through the availability of new modes of input such as multimedia data and other forms of sensory data, data mining could well provide a solution. In order to shed insights on a complex problem such as this, massive databases that were not previously available need to be constructed e.g. data about after event situations of the past [Maurer et al, 2007]. Such data on past events could be useful in highlighting patterns related to potentially in-danger sites. Data to be employed in this mining will thus comprise of both of weather and terrestrial parameters together with other human induced parameters such as vegetation or deforestation over a period of time. [Maurer et al, 2007]

Domain focusing will be concerned with discovery of causal relationships (e.g. using Bayes networks) as a modeling step. Multiple sources of data which include new sources of data need to be incorporated in the discovery of likely causal relationship patterns. A complex form of data mining is required even at the phase of domain focusing. This will involve an iterative process whereby hypothesis generation could be employed to narrow the scope of the problem to allow for a constrained but meaningful data collection. For complex domains such as this, unconstrained data collection may not always be the best solution. Domain focusing would thus perform problem detection, finding deterministic factors and to hypothesize relationships that will be applied in the model. [Beulens et al, 2006] describe a similarly complex representation for an early warning system for food supply networks.

Subsequently, the model construction phase will employ a variety of learning algorithms, to profile events or entities being modeled. As this stage may negate model relationships, domain focusing will need to be repeated and iteratively performed. The model construction phase will allow the incremental development of a model, based on a complex representation of the causal networks. [Beulens et al, 2006]

Mining methods such as clustering, associational rule mining, neural networks will be used to verify the validity of causal associations. Once a potential causal link is hypothesized, verification can be done via the application of data mining methods. [Beulens et al, 2006], have proposed a combination of approaches which include deviation detection, classification, dependence model and causal model generation.

The Decision Making phase will then apply the validated causal relationship model in exploring life case studies. An environment for an interactive explorative visual domain focusing is crucial, to highlight directions for further research. Data mining could serve as a means of characterization of profiles for both areas prone to disasters or those that are safe.

3.3 Medical Application

We will briefly discuss another form of mining that has a high impact. In the medical domain, data mining can be applied to discover unknown causes to diseases such as ‘sudden death’ syndrome or heart attacks which remain unresolved in the medical domain. The main difficulty in performing such discoveries is also in collecting the data necessary to make rational judgments. Large databases need to be developed to provide the modeling capabilities. These databases will comprise of clinical data on patients found to have the disease, and those who are free of it. Additionally non-traditional data such as including retail sales data to determine the purchase of drugs, and calls to emergency rooms together with auxiliary data such as micro array data in genomic databases and environmental data would also be required. [Li, 2007]

Non traditional data could also incorporate major emotional states of patients by analyzing and clustering the magnetic field of human brains which can be measured non invasively using electrodes to a persons’ heads. [Maurer et al, 2007] Social patterns can also be determined through profile mining as described in the previous section to augment the findings of this system. Findings of functional behavior of humans via the genomic database mining would also serve as a meaningful input.

The development of large databases for medical explorations will also open possibilities for other discoveries such as mining family medical history and survival analysis to predict life spans. [Han and Kamber, 2006]

4. The Advantages of Data Mining

Data mining has crept into our lives in a variety of forms. It has empowered individuals across the world to vastly improve the capacity of decision making in focussed areas. Powerful mining tools are going to become available for a large number of people in the near future.

The benefits of data mining will include preserving domestic security through a number of surveillance systems, providing better health through medical mining applications, protection against many other forms of intriguing dangers, and access to just-in-time technology to address specific needs. Mining will provide companies effective means of managing and utilizing resources. People and organizations will acquire the ability to perform well-informed (and possibly well-researched) decision-making. Data mining also provides answers through sifting through multiple sources of information which were never known to exist, or could not be conceivably acquired to provide enlightening answers. Data Mining could be combined with collaborative tools to further facilitate and enhance decision-making in a variety of ways. Data mining is thus able to explicate personal or organizational knowledge which may be locked in the heads of individuals (tacit knowledge) or in legacy databases, to become available. Many more new benefits will emerge as technology advances.

5. Disadvantages of Data Mining

A great deal of knowledge about users is also being maintained by governments, airlines, medical profiles or shopping consortiums. Mining applications with dramatic privacy infringement implications include search history, real-time outbreak and disease surveillance program, early warning for bio-terrorism [Spice, 2005] and Total Information Awareness program.[Anderson, 2004] For example, search history data represents an extremely personal flow of thought patterns of users that reflects ones quest for knowledge, curiosity, desires, aspirations, as well as social inclinations and tendencies. Such logs can reveal a large amount of psychographic data such as user’s attitudes towards topics, interests, lifestyles, intents and beliefs. A valid concern would be that the slightest leak could be disastrous. The extent of possible discoveries has been clearly illustrated by the incidence where AOL released personal data of 658,000 subscribers [Jones, 2006].

Another common danger is profiling where there is a possibility of drastic implications such as a conviction being made based on the incriminating evidences of mining results. There is also a danger of over-generalization based on factors such as race, ethnicity, or gender. This could result in false positives, where an entirely innocent individual or group is targeted for investigation based on a poor decision making process. For the domestic security application, a reasonable data mining success rate of 80% implies that 20% of all US citizens (or 48 million people) would be considered false positives [Manjoo, 2002].

Data mining will further empower mining kingpins to be able to go beyond the ability to PREDICT what is going to happen in a number of areas of economic importance, but actually have the power to KNOW what will happen, hence can e.g. exploiting the stock market in an unprecedented way. They also have the capacity to make judgments on issues and persons with scary accuracy.

6. What can we do?

A majority of related works are more concerned about privacy, but it is no longer the main issue of concern. [Kovatcheva, 2002] has a proposed a means of protecting the anonymity by the use of anonymity agents and pseudonym agents to avoid users from being identified. Their paper also proposed the use of negotiation and trust agents to assist users in reviewing a request from a service before making a rational decision of allowing the use of personal data.

A similar agent-based approach is described by [Taipale, 2003] via rule-based processing. An "intelligent agent" is used for dispatching a query to distributed databases. The agent will negotiate access and permitted uses for each database. Data items are labeled with meta-data describing how that item must be processed. Thus, a data item is protected as it retains relevant rules by which it describes the way it has to be processed. The main challenge then lies in coming up with guidelines and rules such that site administrators or software agents can use to direct various analyses on data without compromising the identity of an individual user. This approach however is not applicable for areas such as Web search, where a standard framework for conscientious mining is far from sight. Furthermore, the concern with the emergence of extensive mining is no longer solved by addressing privacy issues only. As more and more people are willing to compromise privacy, the questions that we pose are: Who do we trust as the gatekeeper of all our data? Do we then trust all our private data at the hands of a commercial global company?

One approach to overcome the concerns mentioned above is by employing a distributed data mining approach, where separate agencies will maintain and become responsible and accountable for different (independent segments of) data repositories. This proposal further ensures that no central agency will have an exclusive control over the powerful mining technology and all resources. [Kulathuramaiyer, Maurer, 2007] In order to realize this solution, governments have to start playing a more proactive role in maintaining and preserving national or regional data sources. We are also beginning to see partnerships between global search engines and governments in this respect. Such partnerships should however be built upon a balanced distribution of earnings. Such a solution can be complemented by the nurturing of a larger number of autonomous context-specific or task specific miners.

7. Conclusion

As data mining matures and becomes widely deployed in even more encompassing ways, we need to learn to effectively apply it to enrich our lives. At the same time, the dangers associated with this technology needs to be minimized by deliberate efforts on the part of the enforcement agencies, data mining agencies and the users of the system. There should be strict regulations to prevent the abuse or misuse of data. Users should also be made aware of the privacy policies in order to make an informed decision about revealing their personal data. The success of such regulations and guidelines can only be guaranteed if they are backed up by a legal framework

References

[Anderson, 2004], Anderson, S.R., Total Information Awareness and Beyond, Bill of Rights Defense Committee; White paper. The Dangers of Using Data Mining Technology to Prevent Terrorism, July 2004

[Beulens et al, 2006] Beulens, A., Li, Y., Kramer, M., van der Vorst, J., Possibilities for applying data mining for early Warning in Food Supply Networks, CSM'06, 20thWorkshop on Methodologies and Tools for Complex System Modeling and Integrated Policy Assessment, August, 2006 http://www.iiasa.ac.at/~marek/ftppub/Pubs/csm06/beulens_pap.pdf

[Coole et al, 1997] Coole, R. Mobasher, B., Srivastava, J., Grouping Web Page References into Transactions for Mining World Wide Web Browsing Patterns, Proceedings of the 1997 IEEE Knowledge and Data Engineering Exchange Workshop Page: 2, 1997 ISBN:0-8186-8230-2 IEEE Computer Society

[Han and Kamber, 2006] Han, J., and Kamber, M., Data Mining: Concepts and Techniques, 2nd ed., The Morgan Kaufmann Series in Data Management Systems, Jim Gray, Series Editor , Morgan Kaufmann Publishers, March 2006.

[Jenssen, 2002] Jenssen, D., Data mining in networks. Invited talk to the Roundtable on Social and Behavior Sciences and Terrorism. National Research Council, Division of Behavioral and Social Sciences and Education, Committee on Law and Justice. Washington, DC. December 11, 2002

[Jones, 2007] Jones, K .C., Fallout From AOL's Data Leak Is Just Beginning ,

http://www.informationweek.com/news/showArticle.jhtml?articleID=191900935, accessed 2007

[KDD, 2005] KDDnuggets : Polls: Successful Data Mining Applications, July 2005

http://www.kdnuggets.com/polls/2005/successful_data_mining_applications.htm

[Kovatcheva, 2002] Kovatcheva, E., Tadinen ,H., The technological and social aspects of data mining by means of web server access logs http://www.pafis.shh.fi/~elikov02/SFISWS2/SFIS2.html 18 January 2002

[Kulathuramaiyer, Balke, 2006]Kulathuramaiyer, N., Balke, W.-T., Restricting the View and Connecting the Dots — Dangers of a Web Search Engine Monopoly, J,UCS Vol. 12 , Issue 12, pp.1731 – 1740, 2006

[Kulathuramaiyer N., Maurer, H., 2007], Kulathuramaiyer, N., Maurer, H., "Why is Fighting Plagiarism and IPR Violation Suddenly of Paramount Importance?" Proc. of International Conference on Knowledge Management, Vienna, August 2007.

[Li, 2007] Li, C. S., Survey of Early Warning Systems for Environmental and Public Health Applications, in Wong, ,S., Li,, C. S.,(eds.), Life Science Data Mining, Science, Engineering, and Biology Informatics- Vol. 2, 2007 http://www.worldscibooks.com/compsci/etextbook/6268/6268_chap01.pdf

[Manjoo, 2002] Manjoo, F., Is Big Brother Our Only Hope Against Bin Laden?, Dec. 3, 2002 http://www.salon.com/tech/feature/2002/12/03/tia/index_np.html

[Maurer et al., 2007] Maurer, L., Klingler, C., Pachauri, R. K., and Tochtermann, K,. Data Mining as Tool for Protection against Avalanches and Landslides, Proc. Environmental Informatics Conference, Warsaw, 2007

[Milne, 2000] Milne, G. R., Privacy and ethical issues in database/interactive marketing and public policy: A research framework and overview of the special issue, Journal of Public Policy & Marketing, Spring 2000

[Mobasher, 2005] Mobasher, B., Web Usage Mining and Personalisation, in Singh, M. P. (ed.) Practical Handbook of Internet Computing, Chapman & Hall/ CRC Press, 2005 http://maya.cs.depaul.edu/~mobasher/papers/IC-Handbook-04.pdf

[Rash, 2006] Rash, W., Political Parties Reap Data Mining Benefits ,eWeek.com enterprise News and reviews, November 16, 2006
http://www.eweek.com/article2/0,1895,2060543,00.asp

[Spice, 2003] Spice, B., Privacy in age of data mining topic of workshop at CMU, March 28, 2003 http://www.post-gazette.com/nation/20030328snoopingnat4p4.asp

[Taipale, 2003] Taipale, K.A. "Data Mining and Domestic Security: Connecting the Dots to Make Sense of Data". Colum. Sci. & Tech. L. Rev. 5 (2). SSRN 546782 / OCLC 45263753, December 15, 2003.

List of Figures

Sections 1 - 16:

Appendices 1-6: