Journal of Universal Computer Science, vol. 12, no. 12 (2006), 1731-1740
submitted: 3/12/06, accepted: 27/12/06, appeared: 28/12/06 © J.UCS

Restricting the View and Connecting the Dots – Dangers of a Web Search Engine Monopoly

Narayanan Kulathuramaiyer
Faculty of Computer Science and Information Technology,
University Malaysia Sarawak, Malaysia
nara@fit.unimas.my

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initial strategy of ranking documents in a fair and objective way (depending on IR techniques and link structures) has been replaced by deliberatively supporting or ignoring sites as economic or political issues are demanding [Google Policy: Censor, 2007].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.

Keywords: Web Mining, Search Engines and Information Retrieval, Social Issues
Categories: H.3.0, I.2.6, K.4.2, K.5.2

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 1]. 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 1: 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]

2.4 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 2]. 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).

Figure 2: Social Network Visualisation (extracted from Heer et al, 2007)

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.

3 Shaping the View of the World

3.1 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]

3.2 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 3). 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].

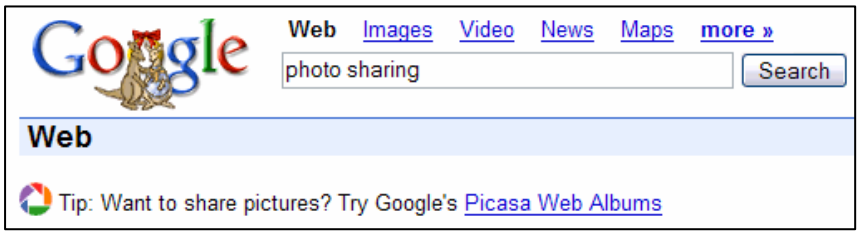

Figure 3: View of Google's Home Page (extracted from Ross, 2007)

3.3 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.

3.4 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."

4 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, 2006]. 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 for 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

[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

[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

[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

[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

[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

[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, 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

[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

[McHugh, 2003] McHugh, J., Google vs. Evil, Wired Magazine, Issue 11.01, 2003

[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

[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

[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 et al, 2007] Witten, I.H., Gori, M., Numerico, T., Web Dragons, Inside the Myths of Search Engine Technology, Morgan Kaufmann, San Francisco, 2007