Tommi Raunio

Visibility on the Web: Search Results and Spam

Pro Gradu Thesis in
Information Systems Science
11.05.2008

University of Jyväskylä
Department of Computer Science and Information Systems
Jyväskylä

ABSTRACT

Raunio, Tommi Petteri

Visibility on the Web: Search results and Spam
Jyväskylä: University of Jyväskylä
100 pages
Pro Gradu Thesis

Visibility on the Web is important for many people and commercial organisations as for many it has become one of the key issues for successful business. Most of the visibility is obtained through top rankings in search engine results. Simultaneously the end users have begun to rely more and more on search engines to locate the content they seek as finding information otherwise on the Web has become notoriously difficult. These issues have lead to an increased interest on search engine results as the top results gain more visitors than others. Due to importance of visibility, many companies and individuals try manipulate the search engine ranking algorithm for their benefit, so that their site would appear among the top results. This deteriorates the search engine functionality and has forced the search engine developers to develop new ranking algorithms and anti-spam filters which in turn the spammers try to counter. The race for top rankings has lead to a new digital arms race on the Web.

This study investigates the algorithms and methods that are used by each of the two sides and also explores the adversarial relationship that has developed between them. The research method is theoretic conceptual analysis based on latest academic literature and empiric evidence. The research result is a description about the methods that each of the involved parties use. Furthermore, the current balance between the two adversary sides is explained in detail. This study indicates that Web spam has a strong effect on search engine development, and it directly affects the structure and development of the Web. Also, although both sides have developed innovative methods in turn, neither of them seem to have a clear upper hand which indicates that the struggle for the top rankings is likely to go on.

KEYWORDS: Web, Search, Optimization, Spam

TABLE OF CONTENTS

1 INTRODUCTION

The World Wide Web today is huge, dynamic, self-organized, and strongly interlinked. The massive size of the Web has made finding information by surfing nearly impossible, and so the users have started relying more and more on search engines to locate the content they seek. Due to this popularity, search engines, such as Google, Yahoo and MS Live, have become sort of gateways into the Web which route Internet traffic through their search results. This in turn has lead to a massive interest towards them, both of commercial and political motives, as the top results gain more visitors than their lower ranking counter-parts. A whole new commercial industry has been born, referred as *Search Engine Optimization* (SEO), which sole aim is to influence search engine results. SEO-companies work with loose co-operation with search engines to provide the optimum results for their clients. Unfortunately, there are also some who try to abuse the ranking algorithm for their benefit. This kind of activity is the focus of this thesis and is referred as *Web spam*.

SEO-companies restructure the client website so that it ranks as high as possible with current content. They choose optimal keywords, construct optimized in-site link-structures, and overall make the site search engine friendly. Search engines approve all these activities. On the other hand, spammers do everything they can to get their page rank high in search results. For example, they construct spam blogs, generate garbage comments on forums, and construct link farms that span across thousands of sites. All this for the top rankings in search results. (Gyöngyi, Garcia-Molina, Berkhin & Pedersen 2006) When earlier spammers operated single-handed, today link exchange programs are everyday, and even legitimate advertisers work with spammers through go-betweens (Wang, Ma, Niu & Chen 2006).

Web spam (also often referred as *Search Engine Spam*) is commonly considered unethical as spammers derive profits by boosting certain pages up in the results and so doing deteriorate the search engine ranking algorithm. Search engine developers' aim is to develop algorithms that rank relevant pages ordered by global importance or popularity, overall so that the users find the results useful. Spammers on the other hand try to manipulate the ranking algorithm, so that their sites would appear among the first results, and so they undo what the search engine developers are trying to achieve. In response to this, search engines implement anti-spam measures which in turn the spammers try to bypass. This cycle has lead to a sort of everlasting arms race. Web spam has always been a nuisance to search engines, but during last years it has become one of the biggest challenges (Henzinger, Motwani & Silverstein 2002). In fact, it has become one of the most important factors for developing new ranking methods (Langville & Meyer 2006).

Commonly Web spam (or SEO for that matter) isn't very well known. This can be determined just by the fact that the end users still consider the organic search results to be relatively reliable. But Web spam does exist as evidence can be found by just using the search engines. One example of political motives is the ”miserable failure”-query which produced as top results the biographies of George W. Bush, Tony Blair or other well known political figure, depending on whose “supporters” had been the most active (BBC 2003). At the time of writing this thesis some search engines still produce this result. A well known example of commercial motives is the case from spring 2006 when Google gave BMW's homepage PageRank zero (*PR0*). The reason was that BMW had used methods that were disapproved by Google. As a result, BMW was dropped from the top results temporarily. (BBC 2006a) (Sobek 2003) Also it is worth to notice that Web spam is a bigger problem in big languages as then more pages tend to target the same keyword. All these manipulation attempts tend do raise concerns especially for the information that is used in decision making processes (Gori & Witten 2005).

It is possible to buy query keywords from the search engine companies in which case the desired link will appear among the sponsored results. Although being easier for the client, this doesn't have the same impact as being placed high in *organic results*. For the end-user the organic results carry more weight as they are considered to be globally objective. (Jansen & Resnick 2005) Furthermore, research show that it is essential to be placed within the top five in organic results as already the fourth gets less attention than the best three (Eyetools 2007).

Measures suggest that Web spam is a growing problem. In 2002 Fetterly, Manasse, & Najork found 8.1% of their dataset of 150 million URLs to be spam. In 2004 Gyöngyi & Garcia-Molina found 18% of their dataset to be spam and estimated that probably 10 – 15% of all content on the Web is spam. The latest figures show the same trend. In 2006 Benczúr, Csalogány & Sarlós found 20,9% of hosts in .uk domain to be spam. However, it is worth to mention that same authors only found 3.6% spam from Swiss evaluation sample which suggests that spam levels indeed differ from domain to domain. To mention some concrete numbers, in 2006 Webb, Caverlee & Pu alone found 350000 spam pages for their spam page database. Although most of the above is separate research, it is safe to assume that the level of spam in the Web is growing.

“Spam, spam, spam, spam, spam, spam, lovely spam! Wonderful spam!”. This phrase is from a Monty Python sketch from 1970. In the sketch Vikings repeat the same chorus over and over again and so doing irritate everyone else in the room. The sketch is funny, but unfortunately it appears that since then the Vikings have invaded the World Wide Web. Although there is no direct link, it is likely that the modern meaning of the word at least in some part comes from this sketch. It is rumoured that in modern sense the term first appeared in MUDs (Multi player dungeons) in the late 1980s when somebody literally spammed the game's user interface (Southwick & Falk 1998). However, the fist message to be called *spam* was the automated message sent by ARMM, a misfunctional program, which accidentally posted hundreds of messages to a news group. The messages got consequently named as spam. Soon after more such multi postings got named the same way. Currently e-mail spam is the most known form of spam as it is evident in almost every mailbox (Gyöngyi & Garcia-Molina 2005a). But also other forms of spam exist, such as *spim* as spam by instant messaging and *spit* as spam by internet telephony. (Becchetti, Castillo, Donato, Leonardi & Baeza-Yates 2006).

The precise definition for Web Spam remains elusive although by now hundreds of papers about it has been written. I believe the most accurate definition so far to be the one by Gyöngyi & Garcia-Molina (2004): “We use the term spamming (also, spamdexing) to refer to any deliberate human action that is meant to trigger unjustifiably favorable relevance or importance for some web page, considering the page's true value". This definition is quite strict as it classifies as spam "all types of actions intended to boost ranking, without improving the true value of a page". Also, this definition makes it difficult to distinguish good page design from spam. For example, is it spam, or just good design, to choose keywords that attract visitors. However, this seems to be the first credible definition for Web spam, and so it is also used in this thesis. Further, according to various articles this definition seems to be commonly accepted and is referred by for example Wu & Davison (2005) and Chellapilla & Chickering (2006).

Although Web spam is now defined, there's still a large grey area between the spam methods and the search engine approved SEO methods (Becchetti, Castillo, Donato, Leonardi & Baeza-Yates 2006). So far it has been left up for the search engines to decide what methods are acceptable and what are not. This adds to their influence as can be seen from the case Google versus BMW (BBC 2006). Commonly they have the moral support as it is their algorithm the spammers are deteriorating. However, it should be noted that even human users have sometimes difficulties in distinguishing what is spam and what is not. In fact, sometimes the only difference is how the page links to other pages.

1.1 Research problem

In this thesis I intend to answer the following questions:

1. How are search results ranked on the Web?
2. What spam techniques exist to improve visibility in Web search engines?
3. What impact does Web spam has on the development of Web search engines?

To put it more freely, I intend to find out what methods each of the two adversary side use and how these methods interact with each other. Furthermore, the relationship between the two adversary sides is interesting as they have now co-existed and interacted with each other for over ten years. Their development is now tied together more tightly than ever before. I intend to open this issue, so that the relationship between Web spam and search engine development can be discussed. First research question will be answered in chapters 2 and 3. Second question will be answered in chapters 4, 5, and 6 and the last research question is discussed in chapter 7.

1.2 Research objectives

Search engine persuasion, which includes Web spam and Search Engine Optimization, is relatively new area of research. Of Search Engine Optimization there are many marketing books, but very few texts that discuss the subject objectively. Of Web spam there aren't any academic books although many works mention the subject in passing. In fact, at time of making this thesis, this very text seems to be the longest text which directly discusses Web spam and its' effects. This is also the objective of this study, to make the first extensive conceptual work on Web spam. Also the objective is to collect the ranking algorithms, spam methods, anti-spam filters, and their interaction within one lids. Although this has been already shortly done by for example Gyöngyi & Garcia-Molina (2004) and Jones (2005), this work aims to discuss these subjects more throughoutly. A personal motivation for this thesis comes from the last years of my studies, from one project to be exact, where the assignment was to construct a Web search engine. Although it was little more than a prototype, it did raise some questions which in the end resulted into this thesis.

After reading this thesis, the reader should have developed a deep knowledge on how results are ranked on the Web, and by what techniques the spammers are trying to get their pages into the top results. Also the technical interaction between the two will be explained. But perhaps even more importantly, the reader will gain understanding on what is the motive behind Web spam and also what kind of effect does spam has on the Web and Web search industry. Most IT professionals are likely to find most of this thesis as new information as often Web search seems to be overlooked in education programs. Also in working life relatively few companies deal with search engines in other context than to use them for search. Even the people who have expert knowledge of Web search might find something new from the last chapters of this thesis which focus on the consequences of Web spam.

1.3 Research method and research material

The research method is theoretic conceptual analysis based on literature available to the subject. This kind of research is essential to the field in question as first, there are none or few collective researches done on this field and second, while the definitions of many essential concepts are still at large, there seems to be need for this kind of abstract research. Although this study is mainly abstract, many examples are used to clarify the theories and also to show how Web spam appears on the Web. These examples are either derived from the theories or are examples of spam pages that are can be found from the Web.

The challenge in doing research of Web spam, or with other commercially and politically infused subject for that matter, is that necessarily some of the informative material is of questionable nature and therefore unfit to be used as reference. This issue further demonstrated the immature nature of the field as sometimes trustworthy information backed with research was difficult to find. On the other hand, this same issue further motivated this research. Also perhaps due to the subject of this thesis, quite a lot of references to Web sources are used. Although academically these aren't considered to be the most reliable references, the latest information does tend to be found through the Web, especially in younger fields of research.

1.4 Research limitations

The research is limited to conceptual analysis of Web spam, search engines, and their relationship. The biggest limitation is the exclusion of empiric research as no larger empiric part is included except for the examples. The research is focused on ranking algorithms, Web spam, and on how these two interact. Several side topics are touched briefly of which the search engine architecture is the largest. The aspect of this study is mostly technical. This leaves out for example the economical and sociological implications which are bound to exist when few single sites route the most of Web traffic.

1.5 Thesis structure

The rest of the thesis is organized as follows. The next chapter deepens the understanding about the environment in which the search engines operate and also describes the common search engine architecture. Chapter three goes through in detail the ranking systems that are likely used in current search engines. These two chapters together describe how the modern search engines work. Chapters four, five, and six discuss the different techniques that are used on both sides. I.e. they describe what kind of techniques are thrown against the search engines and also how search engines repel these attacks. Chapter seven is essentially the new contribution of this thesis as it is based mostly on the previous chapters. Chapter eight offers the conclusions and is also the final chapter of this thesis.

2 WEB SEARCH ENGINES

The purpose of this chapter is to describe what are Web search engines, how do they work, and to further illustrate the challenges the search engines face today. First subchapter offers a general overview, second chapter describes the challenges in modern Web search, third chapter discusses the difference between anti-spam work and censorship, and the last subchapter briefly describes the most common search engine architecture.

2.1 Overview

Perkins (2001) defines Search engine as “A system that uses automated techniques, such as robots (a.k.a. spiders) and indexers, to create indexes of the Web, allows those indexes to be searched according to certain search criteria, and delivers a set of results ordered by relevancy to those search criteria.“ The term “Search engine” is usually associated exclusively with Web search as the term appeared with the introduction of WWW. Search tools in traditional collections are usually referred as “Information Retrieval Systems”. (Crowdhury 2004) According to the above definition, the first Web search engine was the Web crawler which was launched in April 1994 (Pinkerton 1994). Web crawler was the first to introduce results ordered by relevancy, therefore also being the first to fulfil the above definition. From 1990 onwards there were search services that did some part of the above definition, but none of them were really complete solutions. (Mauldin 1997)

Currently globally the most popular and also the most powerful search engine is Google. Google's success derives from powerful search algorithm combined to successful, if not brilliant business strategy (Page & Brin 1998). In ten years Google has grown from a research project into a major player in the whole field of information technology. Currently in the U.S. nearly half of the searches are done with Google, followed by Yahoo with a 28% share (Sullivan 2006) (comScore 2007). Also in Europe Google's share is major. However, although Google is strong in Western countries, it is worth to notice that it's not the number one everywhere in the World. In Russia Google holds only a residue of searches as the market there is dominated by several local Internet companies (Pfanner 2006). Globally significant search engines are also Live Search, which is owned by Microsoft, and Yahoo! which in turn is one of the oldest search engines still in use.

Web search has its' roots in traditional Information Retrieval industry which is concerned in finding information in traditional document collections such as library and decision support systems. Even the first methods to derive information from the Web were identical to the ones used in these traditional systems. (Pinkerton 1994) However, traditional IR methods won't work properly in Web environment as there are major differences between the Web and traditional IR environments (TABLE 1).

TABLE 1. Differences between traditional document collections and the Web (Chowdhury 2004) (Gulli & Signorini, 2005) (Bergman 2001)

Distribution	Usually one or few servers which route the queries and the traffic	Dynamic. No central control.
Size and growth rate	Size and growth rate both controllable	Size being counted in billions of pages and constantly growing
The surface Web vs. Deep Web	Documents are usually accessible through database server	Surface Web, which contains the static Web pages and. The larger part, Deep Web, which contains database driven pages and documents
Variety of formats	Controllable, usually specified	Uncontrollable
Quality of information	Usually it is possible to specify certain quality guidelines	Although it is possible to specify guidelines, there is nobody to enforce them
Frequency of changes	Although can be rapid, they are controlled, and changes can be tracked	Pages evolve from minutes to years
Ownership	Ownership of material easily determined	Owner rights difficult to determine and hence also copyrights
Variety of languages	Most of the collections are local, made in one or few languages	Search engines must be able to at least read all languages for global coverage
Search resource requirements	Controllable, new investments can be planned carefully and are needed rarely (on comparison)	Need for more efficient storage and computation constantly increase which makes the costs of hosting a global search engine phenomenal

2.2 Challenges

Managing a search engine in this environment is challenging at best. Table 2 puts together these challenges according to the frequently referred article by Henzinger, Motwani & Silverstein (2002).

TABLE 2. Challenges of modern search engines.

Natural language queries	As most users can't or won't use the query syntax correctly, therefore it would be more convenient if it would be possible to execute natural language queries
Multiple language support	To be a global player, support for multiple languages is essential. But how to achieve this.
Spam	Web spam is causing nuisance not only to users and companies that rely on Web traffic but also to search engine developers
Content quality	High quality Web pages should be valued more, but how to distinguish the high quality Web pages from others?
Quality evaluation	How to verify that the search results are useful to the end user as users are unwilling to give explicit feedback
Duplicate hosts	Finding duplicates from search engine index is well researched, but it would be optimal if the duplicate hosts were not crawled at all.
Web conventions	Most webmasters follow simple rules in the way they create the Web pages and search engines have become to rely on them. The challenge is to determine, what are these conventions and how to determine when they are being violated
Vaguely-Structured Data	HTML usually contains some layout data. It would be useful to use this in ranking algorithms as spam in layout is more difficult to achieve without breaking the whole page

A lot of issues actually get back to the dynamic nature of the Web as everybody is allowed to publish on the Web without any guidelines. However, search engines are trying to find information out of this jungle of information and at the same time act as police to the unethical policies. If common rules, or Web conventions, could be formalized, it would remove a lot of the grey area between SEO and spam. (Gori & Witten 2005)

2.3 Search engine policies: Censorship

This thesis often refers to “search engines” in general like they were alike. But they're not. While in the past University research projects could take on the whole industry, now it is dominated by international multimillion companies who have their own motives and agenda. The reason for this is that only these big companies have the resources to manage a global search engine.

This being said, it could be perceived that search engines order results according to their own agenda. But they restrict themselves from doing it as in the long term that would mean an end for their business. OpenText, a search engine from mid 90s, sold it's organic search results to companies, so that the end user couldn't differentiate organic results from sponsored results (CNet 1996). The company didn't last long as users lost their trust and moved to use other search services. The same case is still valid today: If the users perceive that they are receiving paid results, or the results are garbage otherwise, they simply change their search provider to another. This is why it's on search engine's own best interest to provide as objective search results as possible as this way they also retain more users.

Anti-spam work should not be mistaken for censorship. In censorship the search engine drops something out of the results that they don't want the end user to see. For example, this is something that Google did when they agreed to censor the search results for China (BBC 2006b). Fighting spam, on the other hand, is different. In anti-spam work the search engine either drops the spam pages out of the search engine index completely or they change the algorithm so that all spam pages drop to the rank they deserve. The informative content of the page is irrelevant as pages are just penalized for using methods that are not accepted by the search provider, whereas in censorship it's the informative content itself that is forbidden. Even the technological method between censorship and anti-spam work is different. In censorship the page is most likely stored in the index but is being filtered out when results are being sorted. In spam prevention the page is usually dropped out of the index completely as garbage or otherwise displayed far from the top results (in an optimal case).

2.4 Architecture

Typical architecture for search engine contains four major parts: a crawler module, page repository, indexer, and a ranking system (FIGURE 1). Due to the growth of the Web, scalability has become a major issue and must be taken account at every step (Page & Brin 1998). It should be noted that this chapter discusses only one possible search engine architecture. Current search engines are likely to use similar structures, but small differences are bound to be found. It's possible to target spam nearly against every component and therefore also anti-spam filters are implemented on every level (Svore, Wu, Burges & Raman 2007). The following subchapters discuss the different components in closer detail.

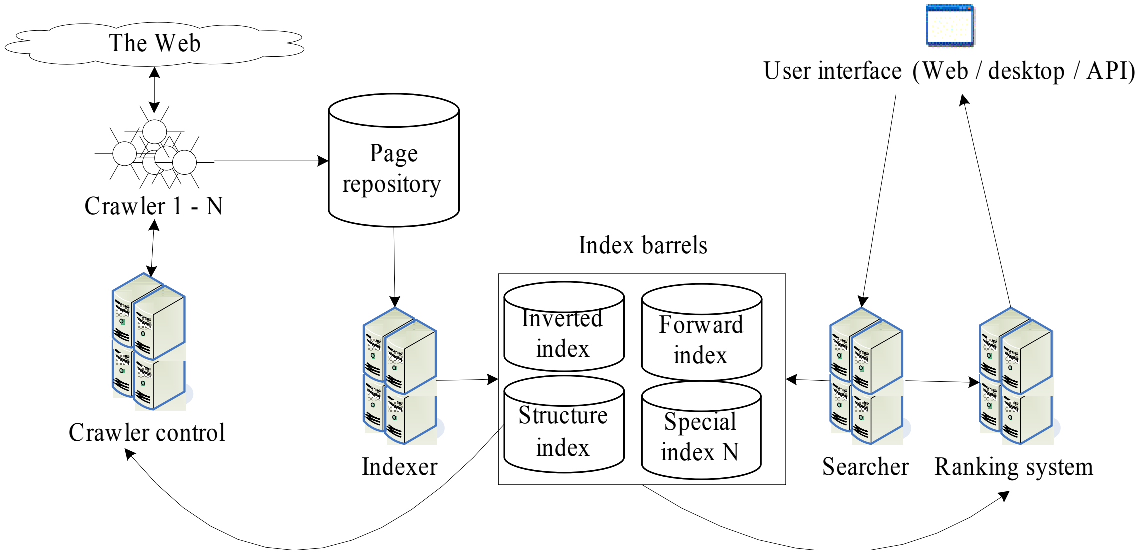


FIGURE 1. Common search engine architecture

2.4.1 Crawler

In order to perform searches in hundredth parts of second, the searched data needs to be stored locally first. As the Web is a globally distributed network, the data needs to be retrieved first by *crawlers* (often referred also as *spiders*). It is often said that the crawler “visits” the Web pages it reads which easily creates a misconception that crawlers are mysterious agents that roam on the Web. However, the everyday is much simpler as crawlers are just bits of computer code that run locally at search engine data centres and request pages over HTTP. (Nutch 2007)

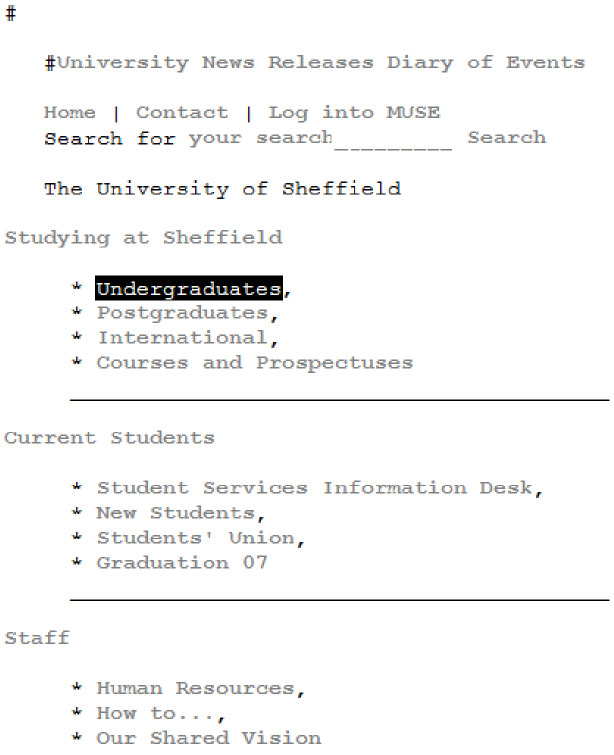
Usually the crawler module has various code threads that in turn run several crawlers to increase efficiency. The crawler main module parses links from crawled pages and adds them to unvisited list for future crawling. As Web pages constantly evolve, it is essential to revisit pages in order to read the current content. Crawlers store the Web pages they read into page repository. This is discussed in the next chapter. (Langville & Meyer 2006)

All legitimate crawlers follow the website policies which are defined in a text-file named robots.txt. Robots.txt is always placed on the domain root, so that all the crawlers are able to find it. By robots.txt, it's possible to ban all crawlers from that site or just forbid a few specific crawlers from accessing a selected set of pages (EXAMPLE 1). It should be noted that it is has been left up for the crawler to decide whether to read the robots.txt which also most likely means that hostile crawlers access all the pages, even the forbidden ones. These kind of crawlers have to be blocked otherwise, for example by their IP address.

```
User-agent: Googlebot
User-agent: MSNBot
Disallow: /Emails
```

EXAMPLE 1. An example robots.txt file that forbirds Google's and MS Live's crawlers from accessing the “Emails” folder.

A common example to illustrate how the crawler sees the Web is to compare it to Lynx. Lynx is a text based Web browser from Unix and displays all the Web pages as text, much so how the crawler sees the pages when it visits them (EXAMPLE 2).



```
#

   #University News Releases Diary of Events

   Home | Contact | Log into MUSE
   Search for your search_________ Search

   The University of Sheffield

Studying at Sheffield

     * Undergraduates,
     * Postgraduates,
     * International,
     * Courses and Prospectuses
     ________________________________________

Current Students

     * Student Services Information Desk,
     * New Students,
     * Students' Union,
     * Graduation 07
     ________________________________________

Staff

     * Human Resources,
     * How to...,
     * Our Shared Vision
```

EXAMPLE 2. The homepage of the University of Sheffield (www.shef.ac.uk) in Lynx. Much how the crawler “sees” pages.

2.4.2 Page Repository

The first search engines contented only to store the Web index (discussed in the next chapter), but the modern search engines store also the whole textual content of the page. This enables the user to access the page even if in reality the page is temporarily or permanently removed from its' original location because even then the page is still available from the search engine cache. Google was the first search engine to create this kind of page repository and soon others followed (Page & Brin 1998). Due to the number of pages that these repositories hold, highly efficient storage systems and algorithms are needed.

2.4.3 Indexer

Modern search engines use a variety of indices to organize and accelerate the search. Langville & Meyer (2006) divide them to three types: content indices, structure indices, and special-purpose indices. Content indices can be of forward or inverted structure. Forward indices (EXAMPLE 3) are ordered by page ids, and each row contains the words that were on that page. Inverted index works just the opposite. They are ordered by word IDs and rows contain pointers to those pages where this word is found (Page & Brin 1998). (Arasu, Cho, Garcia-Molina, Paepcke & Raghavan 2001)

Page ID	Word ID
1	31, 23, 180...
2	1804, 13, 4
3	63, 1, 455
...	

Word ID	Page ID
1	1031, 4241, 4563...
2	4, 64664, 234...
3	67, 14, 6568
...	

EXAMPLE 3. Forward index on the left and invert index on the right

Structure index contains the Hyperlink structure of the Web and is also sometimes read by the crawler module to find new URLs to crawl. This index is also vital for the calculation of Hyperlink related ranking scores which depend on the Hyperlink structures. Special-purpose indices in turn contain more specific information to accelerate and enable certain types of queries. For example they can contain information about images or pdf-files. The indexes are accessed by the searcher. (Langville & Meyer 2006)

2.4.4 Ranking system

Ranking system is logically speaking the final component in the chain. Although it is only one component in the system, it is still perhaps the most vital one as it massively affects the traffic flow of the Web. In the worst case it has also the power to render the rest of the search engine useless as happened in the late 90s. In the very first search engines the ranking was not so essential as there wasn't so much information on the Web and certainly no spam. However, since than the situation has changed. Now the focus is to produce accurate results rather than to gather more data (Barabási 2002).

Broder (2002) divides the search engines into three generations according to the ranking system they use. First generation engines ranked pages according to on-page data, i.e. textual content and format. They resembled very much the search systems from traditional IR collections. These search engines are also often referred as TF-IDF search engines, according to the term weight scheme that they used. Second generation engines used off-page data for search such as the web graph. Google was the first of this generation. The third, and so far the latest, generation engines blend data from multiple algorithms and sources, so that user can for example receive images as a result. Most of the current search engines are of this generation. (Broder 2002) (Metaxas & DeStefano 2005)

Modern search engines use a variety of factors in their ranking algorithms to answer more accurately to users needs and also to obscure their specific ranking algorithm from the spammers (FIGURE 2). Ranking factors have different weights, so that for example words in meta-tags are considered more important than words in the 3rd header. Until recently, most ranking algorithms have originally been public academic work which has allowed the community to participate in their development (Page & Brin 1998) (Kleinberg 1999). For the same reason they have been also easily reverse-engineered. Because of this, the current search engines constantly adjust their ranking algorithms to keep the specifics of their algorithm hidden (Ridings, Shishigin & Whalen 2002). Today the ranking algorithms are highly guarded secrets as they are a critical success factors in the long term operation of any search engine. The current situation has also raised a question if public ranking algorithms are possible anymore due to the ease of which they can be reverse-engineered.

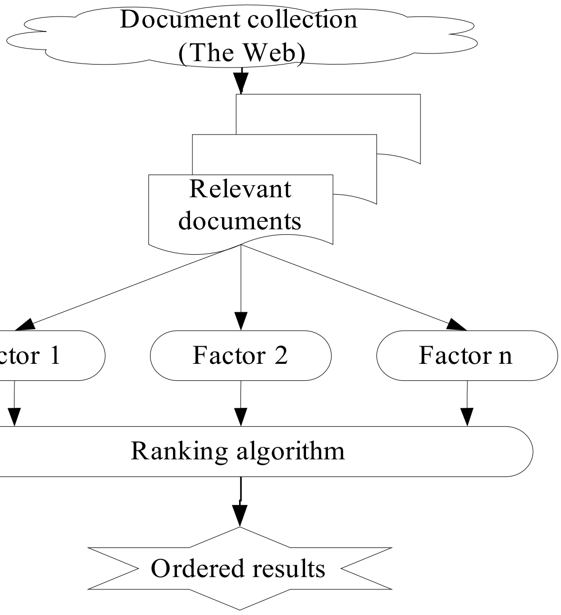


FIGURE 2. Logic behind the current ranking algorithms. There can be as many as 100 factors in the ranking algorithm.

3 RANKING ALGORITHMS

This chapter describes the different methods that have been used to rank results in Web search engines throughout their short history. In latter chapters it is essential to perceive the basic differences between the different ranking schemes because spam techniques are always targeted against specific family of ranking algorithms. The first subchapter describes the classic Information Retrieval models on top of which most early search engines based their ranking algorithms. Second subchapter describes ranking by page content, the most common method of ranking in the 90's. Currently the most popular ranking algorithms, Hyperlink algorithms, are described in chapter 3.3, 3.4, and 3.5. The last chapter is about ranking algorithms under development.

3.1 Classic Information Retrieval Models

Classic information retrieval models (further referred as IR models) are introduced at this point as they were the basis for early Web ranking algorithms. In the earliest Web ranking schemes they were used as such, and even the current Web engines probably apply them to some extent (Pinkerton 1994). Baeza-Yates and Ribeiro-Neto (1998) describe IR models as a tool for predicting what documents are relevant and what are not, i.e. as a notion of relevancy implemented by the system. The IR models should not be considered to be ranking algorithms by themselves as they just produce one of the inputs for the ranking algorithm (FIGURE 3), the document's IR or relevancy score. I.e, IR models resolve whether a document is relevant to the query or not. The higher the document's IR score, the more relevant the document.

Of the classic IR models, vector space model or its' modifications are the most used although also boolean model received some attention especially in the early days of the Web. However, boolean model has two major drawbacks. It produces too much results, especially in Web environment, and it produces no relevance score as the document simply is relevant or is not. On the other hand, Vector model does produce a relevancy score based on how well the query and the document match each other (Baeza-Yates & Ribeiro-Neto, 1998)

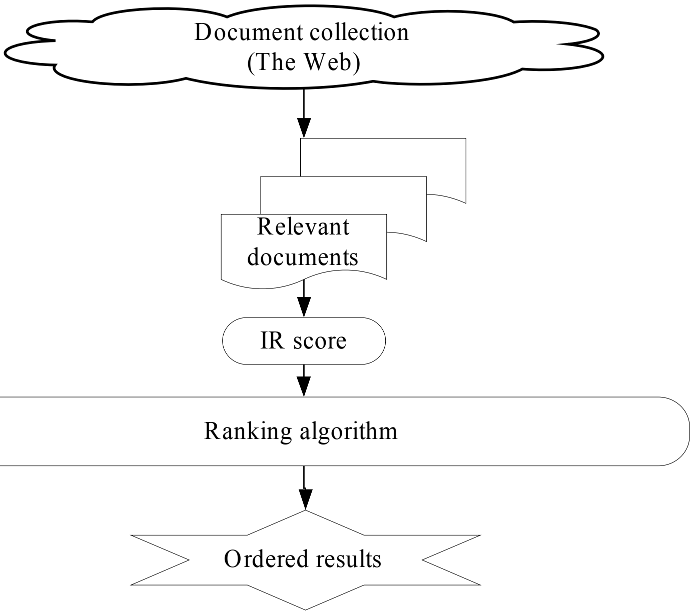


FIGURE 3. Ranking according to classic IR models

Vector (space) model forms the relevance score by comparing the query vector and the document vector which again are formed according to term weights. The degree of similarity between the document and the query is the angle between their corresponding vectors. In other words, the term weights (i.e. importance of a certain term) form up the vectors that in turn form up the relevancy score. The best known method for calculating term weights is known as TF-IDF (*term frequency – invert document frequency*). (Salton, Wong & Yang 1975)

Term frequency part of TF-IDF defines how often a specific word appears in the document. For example if the word "car" appears 8 times in a 40 word document, then tf("car") is 8/40 = 0.26. Inverted document frequency in turn is related to the number of documents in the collection that the word appears in. For example if the word "car" appears in 7 documents out of 100, then idf("car") is 100 / 7 = 14.3. (Baeza-Yates & Ribeiro-Neto 1998)

The purpose of this chapter was to describe what is the basic method for forming up the relevance score. Until 1997 the relevance score was nearly the only input that the ranking algorithms used. Even with current search engines the relevance score is usually formed by using some kind of variation of the vector space model (Page & Brin 1998). The early search engines that used this kind of relevance score as their sole input for ranking are often referred as "TF-IDF" engines according to the term weighting scheme they used. Their era lasted from about 1993 to 1998 after which the first practical Hyperlink models appeared. The next chapter, however, is about what factors were considered important in ranking algorithms that were created on top of these IR models.

3.2 Ranking by page content

Ranking results solely based on page content was the first method of ranking on the Web pioneered by RBSE spider and WebCrawler in 1993 and 1994 (Mauldin 1997). Although even the most modern ranking algorithms analyse and use page content as one factor, they are no longer used as the main factor to rank results on the Web. This is mainly for two reasons. First, they don't produce accurate results in huge collections such as the Web in the first place since it's nearly impossible to get accurate ranking with thousands of pages by using only the information that is supplied with the page. Second, as page content is directly controlled by the page owner, it is also easily manipulated (Ridings, Shishigin & Whalen 2002).

Commonly search engines consider their ranking algorithms to be critical business secrets as often their long term operations directly depend on it. As they guard the algorithm closely, it means that there is relatively little information about the specifics of commercial ranking algorithms. However, not only are the criteria for ranking by page content easy to determine, but one of the largest search engines back then, Lycos, published their criteria already in 1997 (Mauldin 1997). Also approximately year later, Pringle, Allison & Dowe (1998) published their work in which they had reverse-engineered the ranking criteria of four major search engines. However, the academic research lagged behind as spam had become a problem already by 1997 (Marchiori 1997b).

Mauldin published the specific ranking criteria of Lycos in 1997. Lycos was one of the first Web search engines and also from 1994 to 1997 it had one of the biggest indexes on the Web. The following factors were identified:

- Number of times the query term appears in the document
- Frequency of query term in the document
- Proximity of query terms in the document (i.e. how close the terms are to each other)
- Position where the query term appears in the document
- How closely the query matched individual words

The whole ranking algorithm functioned on basis of the IR score although the method for forming the IR score had been further developed from those presented in the previous chapter.

In 1998 Pringle, Dowe & Allison did more wide scale work. They reverse-engineered the ranking criteria of four major search engines by representing the ranking algorithms as decision trees. The factors that they considered to be significant were:

- Number of times the keyword occurs in URL
- Number of times the keyword occurs in the document title
- Number of words in document title
- Number of times the keyword occurs in meta fields
- Number of times the keyword appears occurs in the first header tag
- Number of words in the first heading tag
- Total number of times the keyword occurs in the document including title, meta, etc.
- Length of the document

It is worth noticing that the above points are not the absolute cut of factors that were used in this era, but rather items that Pringle, Allison and Dowe chose to test. They concluded in that the total number of matches on the keyword was important to at least three search engines. Also matches in headers, title, and meta tags were found to be essential for at least two engines. While this was not the absolute cut of all ranking algorithms, it does give some direction into what was the regular criteria for ranking before 1998. (Pringle, Dowe & Allison 1998)

Zhang & Dimitroff (2005) studied what on-site factors are considered significant by the modern search engines. They constructed a test-page which was then submitted to 19 search engines. Then they modified the test page and tracked its' position in search engine rankings. They concluded in the following items:

1. The number of duplicated keywords in title-tag increase visibility up to three duplications. Four duplicated keywords were found to be the point of diminishing returns
2. The number of duplicated keywords in text body increase visibility without any diminishing returns
3. Web pages that had keywords both in the title and text body received better visibility than pages with keyword just in their body or just in their title
4. Font colour, font case, font size, or similar features had no effect.
5. Web pages with meta data elements (for instance `<meta name="keywords" content="Tommi"/>)` received better visibility than pages without any meta data elements.
6. Web pages containing three meta data fields (title, subject, and description) received better visibility than other meta data field combinations
7. Of the meta data fields, subject (`<meta name="subject" content=""/>`) was found to be the most important
8. Meta data should contain only those keywords that also appear in title and body text. I.e. the meta data keywords should be always picked from the page.

The significance of the meta data fields is interesting, for it is commonly believed in the industry that meta data fields were spammed to death in the late 90's (Ridings, Shishigin & Whalen 2002). Also interesting is the finding that repetition does raise visibility to certain extent.

Nearly everything on the page can have an affect on its' ranking as search engines are usually reluctant to rely on single factors, due to the ease that they could be manipulated. These items can include for example page refresh rate, last refresh date (newer is better), how much does the content change per refresh, and how much the users click on the page on search results. On top of these there are also the Hyperlink features that are discussed in the next chapters. All these factors together make the ranking algorithm.

Although by now other kind of ranking algorithms have been developed, page content still has a strong position in determining the relevancy score of the page. As in most cases relevancy score and the authority score (or similar) form up the total score of the page, page content does still affect the overall page ranking. However, Ridings, Shishigin & Whalen (2002) argued that the relevancy score has a top limit, and it cannot be increased over a certain score. This means that webmasters can affect their Website's visibility to certain extent just by optimizing the page content, but to go over this limit they need the off-site factors, such as more off-site Hyperlinks linking to their page (in this thesis these links are further referred as *backlinks*). The effect of Hyperlinks is further discussed in the next chapters.

This chapter offered some insight to how page content was used in the ranking algorithms. As mentioned in the first paragraph, webmasters are able to manipulate the page content easily and this possibility is also exploited widely. So it is no surprise that in the last years of the TF-IDF ranking scheme the whole Web search industry was struggling under spam (Marchiori 1997b). For example, in 1997 only few search engines could “find themselves”. Searching “Altavista” in Altavista didn't produce the search engine as a result as in fact the search engine didn't even make it to first page on its' own results. However, around 1998 a new ranking scheme appeared. This is introduced in the next chapter.

3.3 Early Hyperlink algorithms

By 1998 the Web search industry was buckling under the sheer size of the Web and was in the death grip of the spammers (Langville & Meyer 2006). Into this situation came the ranking methods that derived their ranking criteria from the graph structure of the Web (FIGURE 4). This kind of ranking methods were developed from 1996 of which PageRank is clearly the most significant one. But contrary to common belief, PageRank wasn't the first Hyperlink algorithm to be developed. The first evidence of Hyperlink models range from 1996. These models are briefly discussed here.

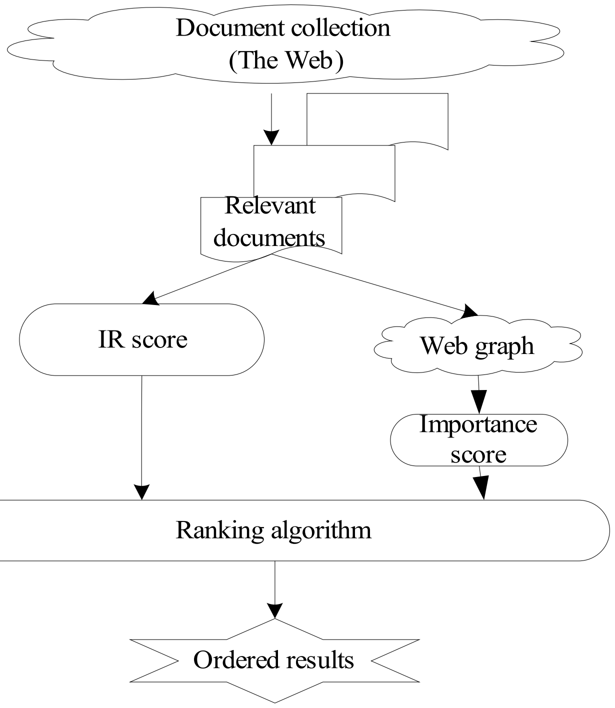


FIGURE 4. Graph structure of the Web is now being exploited in most of the ranking algorithms

Marchiori (1997a) developed one of the first Hyperlink algorithms. His algorithm considered the links on-site and gave higher rankings to pages that had linked to other sites. This kind of scoring function works just the opposite of current Hyperlink algorithms in a way that the current algorithms score backlinks, not just outlinks. Primarily the goal was to develop a spam resistant algorithm and secondarily to support the connectivity of the Web community. The idea was that the webmasters had to choose between two evils: linking to a competitor or facing a lower search engine visibility. The algorithm was designed to be implemented as a pre- or post-processor to the existing ranking algorithms.

Carrière & Kazman (1997) developed an algorithm that derived its' ranking both from page content and Hyperlink structure. The basic idea was that more connected pages should be ranked better. Their Hyperlink algorithm filtered out results that were returned by the IR model, but weren't connected enough as these nodes were considered to be “uninteresting. This kind of complete filtering is a bit questionable as also these pages could be of interest to the end user.

Credit of the first academically published Hyperlink algorithm probably belongs to Yuwono & Lee (1996). But while their algorithms were rather introductory in nature, the two algorithms presented in the above paragraphs were the first serious attempts to use the graph structure of the Web. Either of them would have been an improvement as spam pages back then had little connectivity and therefore would have ranked badly in these algorithms. However, neither of these algorithms got any air beneath their wings. The breakthrough came a year later.

3.4 PageRank

PageRank by Brin, Motwani, Page & Winogard (1998) is really a success. From its' publication in 1998 it has been the focus of search industry professionals, academics, and also naturally spammers. It is also definitely one of the decisive factors behind Google's success as it enabled Google to produce accurate results right from its' launch. Combined with successful business strategy, PageRank has made Google the definite number one in Web search industry in both numbers (Sullivan 2006) and prestige.

PageRank algorithm is about measuring the global importance of a webpages. It calculates a global importance score for each page on its' repository. The importance in

PageRank is realized by the number and quality of backlinks (i.e. links that link to the page A from elsewhere on the Web). The following subchapter describes the algorithm in closer detail.

3.4.1 PageRank formula

The simplified formula for calculating PageRank for webpage u is (Brin, Motwani, Page & Winogard 1998):

$$PR(u) = \sum_{v \in B_u} \frac{PR(v)}{N_v}$$

PR(u): PageRank of page u

B_u: Group of pages that link to page u

PR(v): PageRank of page v

N_v: Number of links on page v

Or as defined otherwise (Page & Brin 1998):

PR(A) = (1-d) + d (PR(T1)/C(T1) + ... + PR(Tn)/C(Tn))

PR(T1 – Tn): PageRanks of pages that have links to page A

C(T1 – Tn): Number of links on pages that have links to page A

d: Dampening factor. Usually set to 0.85

In plain words, the PageRank of webpage A is the sum of PageRanks of webpages linking to page A divided by the number of links on each page. I.e. it is as if every page casts a popularity vote but can divide it among many pages. Further, votes of pages that have high PageRank are considered to be more important. I.e. if there is a link from www.google.com to a page A, then also A's PageRank will be high in result of a vote of an important page. (Ridings, Shishigin & Whalen 2002)

The previous formula computes the PageRanks one page at a time. In industrial use the calculations are done in a matrix in which PageRanks for all pages can be calculated in one go. A huge Hyperlink matrix is formed where the webpages are placed on columns and rows. The entries, i.e. individual cells, represent the Hyperlinks between the pages. The formula above also shows that calculating PageRanks is an iterative process as page's PageRank depends on PageRanks of pages linking to it. For PageRank scores to converge, roughly 52 iterations are needed. (Langville & Meyer 2006)

As understanding how PageRank works is essential in latter chapters, I'll provide an example which further illustrates the mechanics behind PageRank. Lets calculate the first two iterations of PageRank calculation for the following graph:

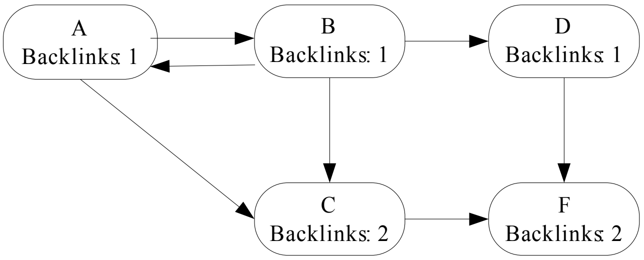


FIGURE 5. Hyperlink structure used in the example. Arrows represent Hyperlinks.

If we apply the simplified PageRank formula for this graph, we get the following calculations:

TABLE 3. PageRank values after first iteration

Iteration 1

Page	PR before	(1-d) + d (PR(T1)/C(T1) + ... + PR(Tn)/C(Tn))	PR after
A	1	1 - 0,85 + 0,85*(1 / 3)	0,43
B	1	1 - 0,85 + 0,85*(1 / 2)	0,58
C	1	1 - 0,85 + 0,85*(1 / 2 + 1 / 3)	0,86
D	1	1 - 0,85 + 0,85*(1 / 3)	0,43
F	1	1 - 0,85 + 0,85*(1 / 1 + 1 / 1)	1,85

TABLE 4. PageRank values after second iteration

Iteration 2

Page	PR before	(1-d) + d (PR(T1)/C(T1) + ... + PR(Tn)/C(Tn))	PR after
A	0,43	1 - 0,85 + 0,85*(0,58 / 3)	0,31
B	0,58	1 - 0,85 + 0,85*(0,43 / 2)	0,33
C	0,86	1 - 0,85 + 0,85*(0,43 / 2 + 0,58 / 3)	0,5
D	0,43	1 - 0,85 + 0,85*(0,58 / 3)	0,31
F	1,85	1 - 0,85 + 0,85*(0,43 / 1 + 0,86 / 1)	1,25

If continued for a few more loops, the values would eventually the values converge to a certain value.

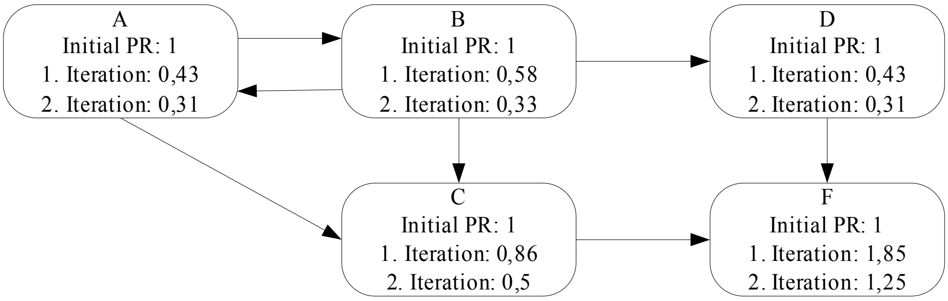


FIGURE 6. Web graph with PageRank values after the two first iterations. Assuming that these pages would be included in the same query, page F would rank highest.

3.4.2 Random surfer analogy

Brin, Motwani, Page & Winogard (1998) themselves use the analogy of random surfer to explain the basic premise behind PageRank: A surfer surfs on the Web by randomly selecting a new link from each page, but never clicking the back-button. If the surfer gets caught in a page with no Hyperlinks, then he randomly jumps to another page on the Web. As links are always randomly selected, those pages with high amount of backlinks get visited more often. Hence the probability that the user visits a certain page is the page's PageRank.

Actually, the random surfer analogy being now explained, also the factor d in PageRank formula can be further opened. The factor d is the probability the random surfer will get “bored” and randomly jumps to another location on the Web (Brin & Page 1998). This factor being usually set to 0.85, there is a 0.15 probability the user will request a new, random page. This modification is also the one that allows the random surfer to escape pages with no outlinks (Langville & Meyer 2006).

3.4.3 Discussion

In 2008 PageRank has been in industrial use for nine years and it still appears to be a central component in Google's ranking system (Google 2007a). However, it is likely that its' weight in the ranking system has been reduced just due the amount of market pressure it has to bear. Further, PageRank's basic premise doesn't hold any more as today it is the PageRank-algorithm itself that is the reason behind complex link structures (Gyöngyi & Garcia-Molina 2005b). However, only the engineers at Google know what exactly is the current PageRank algorithm, since it is very likely that the algorithm has been altered a great deal from its' launch.

PageRank along with Google was a welcomed change in 1998 as back then most search engines were infested with spam and were further developed as overly complex portals as was the trend at the time. Into this situation came Google which had a very easy-to-use interface and on top of that produced very accurate results. The success was quick to follow, and after well used opportunities Google made itself the strongest player in the search market.

It would be interesting to know, just how much PageRank has affected the shape Web just by existing. The effect can be perceived at least in the form of spam as link farms can be found all over the Web (Gyöngyi & Garcia-Molina 2005b). PageRank seems to be generating links by its' own and therefore affecting the very thing it is trying to measure. From the measures done in 1996 it can be seen that before PageRank the Web wasn't very interlinked as most of the pages contained no Hyperlinks at all (Bray 1996). This effect on Web's web graph is further opened in chapter 7.3.

One of the motives for the development of PageRank was the scale of spam in the late 90's as it was all but disabling the existing search engines. PageRank calculation does have a natural resilience to spam as backlinks are off-site factors and therefore out of direct reach to the page owner. (Brin, Motwani, Page & Winogard 1998) PageRank did tip the scales in favour of search engines for a while, but spammers learned to adapt within a few years. Also, because PageRank algorithm is public academic work, the exact principles and even calculations were already available.

Also, the way PageRank understands “importance” should be discussed. PageRank predicts that the more pages link to page A, the more important the page must be (Brin, Motwani, Page & Winogard 1998). But then what about:

- Reciprocal links: “Link to me and I'll link you”
- Link requirements: “Using our script requires you to put a link to our page”
- Friends and Family
- Free page add-ons such as counters
- Link spam

And as was determined above, PageRank itself is affecting the structure of the Web which further obscures the notion of “importance”. Because of these issues, PageRank can't be argued to be the sole conveyor of importance any more, but what other terms are there? Another term would be “popularity” as referred in Langville & Meyer (2006), but it faces the same credibility problems as a term as “importance”. PageRank has become a sort of standard of its' own which is not necessarily accurate regarding “importance”, but it is still an algorithm that is a standard for other search engines.

3.5 HITS

HITS, as in Hypertext Induced Topic Search, is a search system that is based on algorithm that was developed by Jon M. Kleinberg in 1998 (Gibson, Kleinberg & Raghavan 1998). In many aspects HITS is similar to PageRank. Both calculations are done in a matrix, both produce importance scores, and both are based on Hyperlink structures. But there are also major differences. Where PageRank produces only one score, HITS produces two, and also where PageRank calculation is done query independent, the HITS calculation is query dependent. Kleinberg's algorithm (further simply referred HITS) was used by the search engine Teoma. Whether it is used by Teoma's current owner, Ask.com, is unknown.

HITS divides webpages into two categories: authorities and hubs. Authority pages are pages that the end-users wish to find and hubs are pages that link to these authorities. I.e hub is nexus which has links to many authorities or other hubs. It is expected that the authorative pages have a high overlap in the hubs that link to them as the authorative pages are expected to be on the same topic (FIGURE 7). A good hub points to many good authorities and respectively a good authority is linked by many good hubs. This principle is called the "mutually reinforcing relationship". The "goodness" of hubs and authorities is measured by the hub and authority score which are assigned for every webpage. Also, the calculation is done at query time because the scores are calculated for relevant pages only, i.e. for a relatively small cluster of pages. This cluster is referred as the "neighbourhood graph". (Kleinberg 1999)

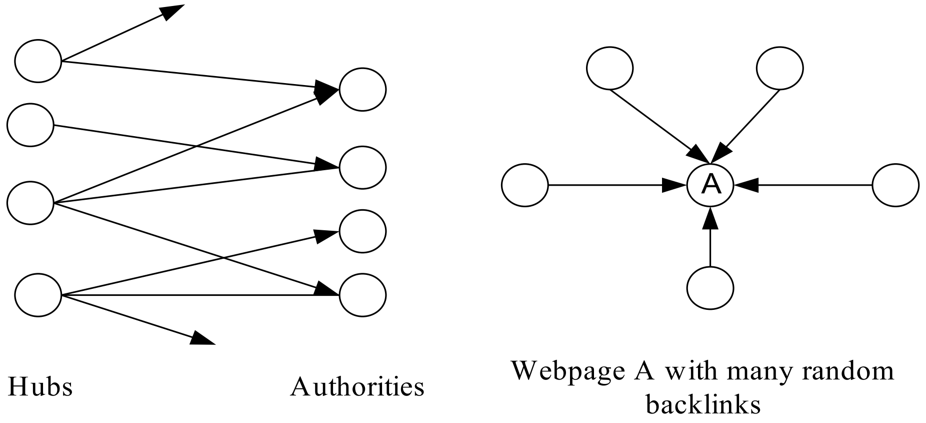


FIGURE 7. The difference between authorities and a normal webpage with many random backlinks (Kleinberg 1999).

Unfortunately, while being otherwise an efficent ranking algorithm, HITS doesn't manage very well with spam. HITS is especially vulnerable to situations where a page has few backlinks, but a large number of outlinks. After several iterations of HITS calculation this results in a situation where the page is assigned with a top hub score and respectively the pages it links to are assigned with a top authority score. If the page has a link pointing to another hub, this hub is also assigned with a top hub score. This formation results in a "*Tightly Knit Community*" against which HITS is specially vulnerable to (Wu B & Davison B 2005) (Lempel & Moran 2000). Tightly Knit Communities can be legitimite page collections, but often they are link farms created to mislead search engine ranking algorithms such as PageRank or HITS. (Li, Shang & Zhang 2002)

While commercially PageRank is the most important algorithm, as it is used by the superior market leader, HITS seems to be enjoying the favour of the academic community as a large number of papers have been published which introduce improvements to the to the basic algorithm. Henzinger & Bharat (1998) developed the BHITS algorithm which aimed to improve the vulnerability of HITS to special link patterns. Li, Shang & Zhang (2002) focused on the same issue and revised the algorithm to take account the "small-in-large-out" pages, but both of previous algorithms were still vulnerable to link farms. Also SALSA by Lempel & Moran (2000) was motivated by the HITS algorithm. Finally, as HITS really is vulnerable to link farms, it is also an efficent way to flush them out as done by Benczúr, Csalogány & Sarlós (2006).

It should be noted that when further the term "authority score" is used in this thesis, it refers not only to the HITS authority score but also to PageRank score. Although these numbers are calculated differently, the purpose is the same. This is why it was decided to use the single term "authority score" when referring both to the HITS authority score and the PageRank score.

3.6 Ranking algorithms under development

The development of early Web ranking algorithms was motivated by the desire to make accurate algorithms that would in turn produce accurate results. While this is still the final goal, now the driving force behind new algorithms seems to be spam (Langville & Meyer 2006). If before it was possible to build algorithms just to produce results in an optimal environment, now it is not possible to build any algorithm without taking Web spam into account. However, this chapter describes a few new ranking algorithms that both address the subject differently and are also naturally spam resistant.

Personalization through PageRank is suggested by Haveliwala (2003). The basic premise is simple: Instead of calculating just one PageRank value, one is calculated for each interest category. According to query context, i.e. user interests, the matching PageRank vectors are combined query time. Query context is determined from the query by an algorithm that tries to find three most matching categories. If the query is done by highlighting a word on a webpage, words near the query term can be used to determine the context. Easiest way would be to ask the user of what's he's interested of, but so far the users have been reluctant to supply any explicit information their interests. Another option would be to use an algorithm to predict user preferences.

As users are reluctant to give explicit feedback on search results, many automated measures are being researched. One of these is the use of machine learning. The learning algorithms would be there to help determine, which returned search results were actually useful. For example, it can be assumed that the end-users usually pick the first returned result. If on some queries the first results is rarely picked, then the first results is not probably relevant for this query and its' rank should be lowered. The difficulty is, however, how to extract this information, and further, how to use it. (Agichtein, Brill, Dumais & Ragno 2006)

The Internet is often referred to be the “Highway of Information”. By using this mindset, it is also possible to construct a ranking method. If the World Wide Web is the highway, then the surfers are cars in a highway network, and like real highways, some virtual highways have more traffic than others. (Langville & Meyer 2006). This traffic can be used as a factor in ranking algorithms by favouring those pages that have more traffic. Naturally counting surfers on every page is impossible, but estimates based on Hyperlink structures are possible. One of this kind of algorithms, TrafficRank, was introduced by Tomlin (2003).

NEC research from 1997 to 1999 showed that search engines couldn't keep up with the growth rate of the Web. As Web today is likely to be many times larger, it would suggest that no single search engine can cover it fully. Some, or perhaps the most, of search engine coverage is overlapping, but there are also segments of the Web that only a few single search engines cover. (Barabási 2002) (Signorini & Gulli 2005*). Meta-search* engines use other search engines to locate the content they seek. As the search engines have different a coverage of the Web, meta-search engines could be used to search the Web more extensively with one query. Usually all the major search engines offer some kind of limited API to their services, through which other services can access the search engine functionality. The limitation can be for example that only certain type and certain number of queries are allowed. The first generation of meta-search engines had their moment of sunshine in the end of 90's, but they nearly disappeared after newer, accurate search engines appeared. Now there seems to be some window for them again. (Langville & Meyer 2006).

In the past most ranking algorithms, such as PageRank and HITS, have been public work. While new ranking algorithms are still being developed open source, it is unknown to what extent these are used in search engines as for example MS Live search just settles to state that their algorithm is complex, and it uses factors such as page content and link structures. As commercial search engines consider their ranking algorithm to be a highly guarded business secret, it does raise a question whether open source ranking algorithms are possible any more due to the ease they can be reverse-engineered.

The emergence of Hyperlink algorithms in the end of 90's has been the biggest development so far. But that is not to say that the development has stopped there. Next big steps could be the emergence of personalized results and learning algorithms. Their emergence would also most likely start a new surge on the spammers side as they are completely different from current algorithms which measure the rest of the Web, not so the behaviour of search engine users.

For the popularity of search engines, there seems to be no end in sight as no other feasible alternatives exist. For example Web rings, which were used to form a Hyperlink ring of similar Web pages, have all but disappeared. This indicates that while the search engine operations continue as profitable as they are, we're about to see more developments due to the resources that the big search companies are putting on research and development currently.

3.7 Conclusions

This chapter introduced all the major ranking algorithms that has been in use so far. In a relatively short time, under ten years, the focus has shifted from page content into ranking by link structures, to the point that now most ranking algorithms are based on the latter. However, page content is still important as relevancy is usually determined from the page content, so keywords, headers, and meta-tags are still relevant and should be considered when optimizing the page for search engines. Current topics, on top of secrecy, seem to be optimizations to PageRank and HITS, personalized search, machine learning, incorporation of multimedia, and spam resistant algorithms. The driving force behind new ranking algorithms seems to be spam, and further algorithm development without taking spam into account seems now to be an unrealistic option.

4 CONTENT SPAM

This chapter is the first chapter to move to the other focus of this thesis, Web spam. Roughly spam methods can be divided into three main category: content spam, link spam, and hiding techniques (Gyöngyi & Garcia-Molina 2004). This chapter is about content spam, methods that were chronically speaking the first form of spam on the Web. These methods were also one of the initiatives for the development of Hyperlink algorithms as these spam methods had all but disabled the search engines back then.

4.1 Target algorithms

Chapter 3.1 introduced vector space models that used TF-IDF word weight formula. As most of the ranking techniques in the 90s were based on this model, it was an easy target. The spammers couldn't affect the IDF-part of the algorithm (inverted document frequency) as they don't have any control over search engine index size. This leaves the term frequency part which again is a on-site factor and easily modifiable. TF can be affected in two ways. The spammer can make the page very relevant for one query (i.e. TF(“word”) = 50/100 or make the page somewhat relevant for many queries (i.e. large page with many hit words). (Gyöngyi & Garcia-Molina 2004) As already mentioned in previous chapter, these methods don't have any direct effect on the Hyperlink algorithms which are mostly concerned in calculation of authority scores. However, they can have an affect to whether a certain page is relevant to the certain query and therefore also make an unrelevant page appear on results.

4.2 Techniques

All content spam techniques are based on the textual content of the page. The most basic method is to repeat keywords so that the search engine is mislead to believe that the page is really about that specific topic. As nonsense repetition and inconsistent sentences are easily detected, often various hiding techniques are also used in unison.

The following paragraphs describe all the content spam techniques according to the article by Gyöngyi & Garcia-Molina from 2004. Hiding techniques, which all closely relate to these methods, are discussed in chapter 6.

The most basic content spam methods is word repetition in different parts of an HTML-page. All the different components of an HTML-page can be spammed: header, title, meta, and body tags. As discussed in 3.2, search engines analyse these components differently and also do cross-checks, so that for example words in the title should also appear in page body. As in the end of 90s most spam pages contained dozens instances of the same word, search engines implemented rules that would penalize too many repetitions.

Simple sequential repetition is only one possible technique. Others include *dumping* in which the spammer copies a dictionary on to his page. This results the page to rank well on many queries. As on some the competition is signifigantly lower, the page is likely to rank well on these queries. In *weawing* the spammer copies for instance a news article on to his page and then inserts spam words randomly in the article. Although this also lowers the term tf-score, it also misleads anti-spam filters as otherwise the word distribution seems normal. In *pharse stitching* the spammer mixes sentences from different sources. As sometimes search queries are full sentences and as the spam page contain phrases from multiple sources, the page will do well on these queries. Some search engines also break down the URLs to further determine the relevancy of a page. This is why spam page URLs are often long and contain keywords (EXAMPLE 4).

- **`http://top-free-ringtones.isgreat.org/download-free-ringtones-for-cell-phone.html`**
- **`http://top1000-home-insurance.isgreat.org/best-home-owners-insurance-company%5C'.html`**

EXAMPLE 4. Example of spam URLs

4.3 Counters

Natural counter to content spam is to flush the extra terms out by statistical analysis. In 3.2 it was already suggested that in page title four repetitions were the point for diminishing returns although repetitions in page body increased visibility endlessly. Also, in general factors such as page evolution rate, components of the URL name, word distribution per sentence, replication of content, and link in-degree and out-degree can be measured. Outliers in the distribution are likely to be spam. (Fetterly, Manasse & Najork 2004)

One issue in spam detection is how to handle the detected pages. First answer would be to set the page score to zero, so that it would drop to bottom in search results. However, the spammer reaction would be to modify the page so that it would stay just in the allowed zone, i.e. for instance so that the keyword in the title is repeated only twice. This would result the page to have the highest possible (allowed) score. When more spammers do the same, a *flattening effect* occurs: Many pages receive the same top score, which eventually makes ranking by this factor impossible. (Marchiori 1997b) Another option would be just to ignore the extra spam factors, i.e. take count only the first two repetitions. However, for example Google just settles to drop the spam page out of its' index, permanently. When gotten into this black list, getting out is notoriously difficult. (Google 2006)

This chapter was a short description about the spam methods that were used in the 90's. They were a serious threat before the introduction of Hyperlink algorithms, and they nearly deteriorated the whole search industry. But as the old ranking algorithms, which were based on page content, got obsolete, so did some of the methods that were introduced in this chapter. However, a new threat appeared right after. This is discussed in the next chapter.

5 LINK SPAM

The development of Hyperlink algorithms quite effectively disarmed the most of content spam which turned the scales temporarily in search engine developers favour. However, it was soon discovered that also Hyperlink algorithms were susceptible to spam. This chapter is about *link spam*, currently the most wide spread form of spam as the current ranking algorithms are still vulnerable to it. Link spam is both difficult to detect and to rebuff as will be demonstrated in this chapter. Also, all spam methods are summarized in appendix 2.

5.1 Target algorithms

HITS and PageRank and their successors are rather easy to reverse-engineer as they have been originally public academic work. This is why spammers were able to form optimal link structures rather easily and even test them by going through the algorithm calculations, similar to what we did in chapter three. There's big market pressure against these algorithms as they were published over ten years ago, and by now their strengths and weaknesses are commonly known (Langville & Meyer 2006). It is likely that Google bears the most market pressure as it is the current market leader in most parts of the World (comScore 2007). As Google's competitors cover only a residue demand, the biggest global visibility is gained by concentrating efforts on Google. But also as the search engine ranking algorithms have similarities, ranking high in one search engine also often means high visibility on other engines. (Xing & Lin 2004)

Spoken language does matter. There is big difference in spam from Finnish (five million speakers) to English (average of 1.8 billion speakers) as spam pages are always targeted for certain keywords. Naturally brand names are the same in most languages, but search queries are also usually limited by localized words. This would indicate that competition within keywords is more intensive within big languages which further suggests that also spam is a bigger problem in those languages.

5.2 Outlinks

Adding outlinks is the easiest form of link spam as it is done directly with page under spammers control. Usually the purpose is to increase the hub score in HITS calculation, or to use the page as part of a link farm. If the spam page is assigned with a top hub score, the result is that the pages it links to receive a top authority score. This group of pages form up a tightly knit community, and as HITS is vulnerable to them, this group dominates the top results with the algorithm in question. (Wu & Davison 2005)

The easiest way to achieve a large number of outlinks quickly is to clone a web directory. Web directories, such as DMOZ Open Directory and Yahoo! Directory, contain thousands of links that are relatively easy to extract and copy (Gyöngyi & Garcia-Molina 2004). This can result in a spam page that has a its' own look and feel, but where the content is the same as in the original directory

From PageRank point of view, many outgoing off-site links can be a negative issue for other pages on the same site. There is a phenomenon called *PageRank leak* in which PageRank score is leaked out of the current website. For example if we consider the following two cases (FIGURE 8):

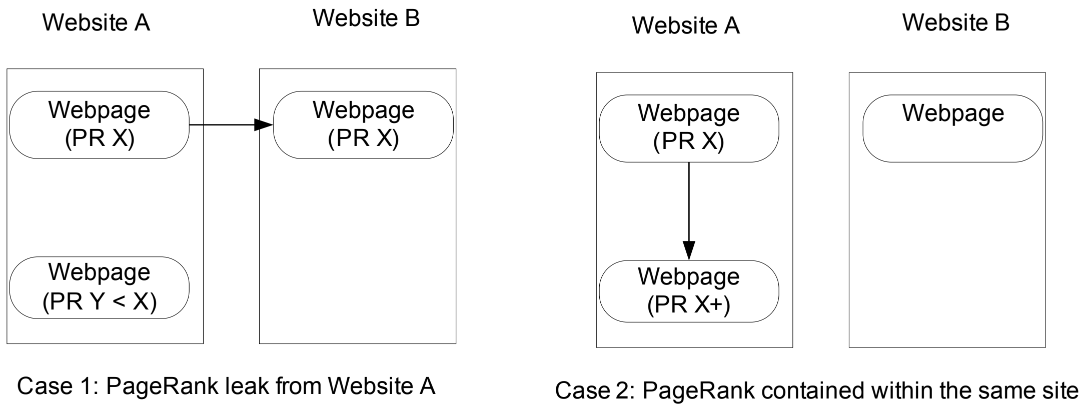


Figure 8. PageRank leak.

1. The only link from page in site A links to site B. The page in site B receives the full PageRank of the page on site A.

2. The only link from page A links to another page within the same site. The target page on site A receives the full PageRank of the source page. This way PageRank of the source page is used within the same website.

On case 1 PageRank leak occurs because site B receives the PageRanks score which could have been used also on site A. It should be noted that the PageRank value of the link page on site A doesn't decrease in any case, it is just the case of which page it gives its' “vote”. It's a missed chance to increase the PageRank of other pages on the same site. But then, PageRank leak doesn't matter if 1) Only hub score, which is not PageRank related, is desired or 2) There are no other pages on the website or 3) the PageRank score of other pages is irrelevant. This is not to say that linking outside the site is absolutely a negative issue in terms of visibility since it is possible that search engines have implemented measures that impose penalty for linking just within the same site. It is possible to construct websites so that the PR leak is minimized, by for example adding in-site links to pages that have off-site links (FIGURE 9). The PR leak for site A in FIGURE 9 is minimized since the vote of the link page is divided between site A and site B, so that the same site (A) gets the most of PR score. (Ridings, Shishigin & Whalen 2002)

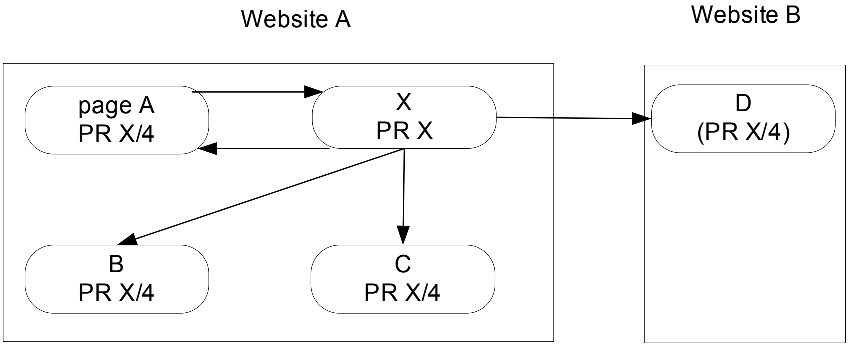


FIGURE 9. Site structure that gives only 1/4 PageRank to the other site. Note that PageRanks on the figure are only those that Page X gives to the other pages, not absolute PageRank values.

5.3 Backlinks

Backlinks are more difficult for spammers to control since they are out of their direct control. However, there are other ways to acquire backlinks as will be described in this chapter.

5.3.1 Honey pot

The term *honey pot* refers to the case where the spammer's page is so attractive that other people naturally link to it. The information can be for example Unix-instructions, flight schedules, manuals, or whatever that would make the page moderately appealing. The contents is always copied elsewhere from the Web as the only reason for the page to exist is that it's participating in a link farm. The honey pot page functions as a proxy which collects legitimate backlinks from elsewhere the Web and passes their authority score onwards to other spam pages (FIGURE 10). Note that fake Web directories that were presented in the previous chapter could function as a honey pot. (Gyöngyi & Garcia-Molina 2004)

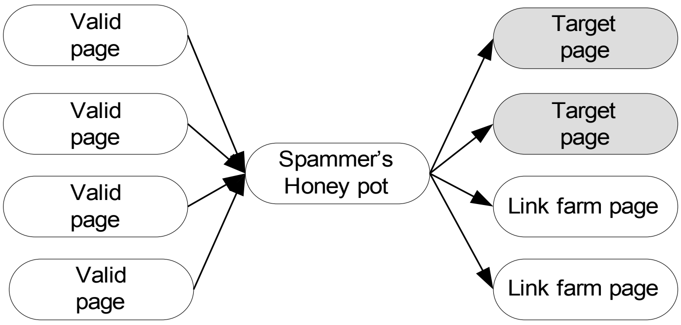


FIGURE 10. Honey pot

The purpose of honey pot is usually to increase the targets' PageRank or hub score, but they can also be used to deceive anti-spam measures as for anti-spam filters it is difficult to distinguish: The honey pot is linked from valid web pages and it can also contain links to other valid pages, among the spam links. The valid outlinks can be there to increase the hub score or just simply to mislead anti-spam algorithms. (Gyöngyi, Garcia-Molina & Pedersen 2004) It is possible to calculate the relation between good and bad outlinks, assuming that it's possible to separate them. But then again, if their relation nears one to one, is it a spam page at all?

5.3.2 Infiltrating a web directory

In this case the spammer is able to post his spam page to a legitimate Web directory, either because of the lack of control in the directory or by simply misleading the directory administrator. As Web directories usually have a high hub and PageRank score, this also results the spam page having a high authority score (FIGURE 11). Spam page again points to the target page which also can be in the directory. (Gyöngyi & Garcia-Molina 2004)

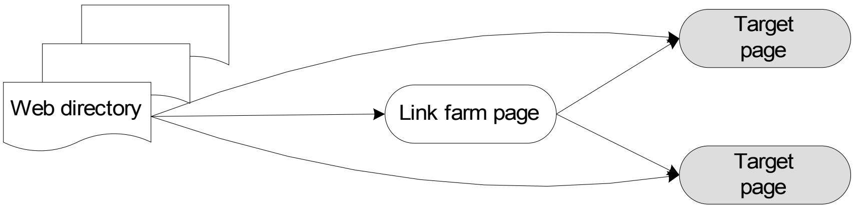


FIGURE 11. Web directory has links to the link farm page and the target pages. This way the targets receive the directory authority score twice, once through a legitimate link from a directory and another time through a spam proxy.

Wikipedia is rather special case of Web directory, or so at least the search engine algorithms see it. Wikipedia is a popular all-around dictionary which contains information, well, about everything. As a result, its' pages are also linked very often. Furthermore, as its' pages are very interlinked, nearly all the pages on Wikipedia have a high authority score (APPENDIX 1). This is why Wikipedia ranks on the first page on nearly every query, no matter how popular or commercial the query keyword might be. For example, with a highly commercial Google query “BMW” Wikipedia is already the fifth. But although Wikipedia also links to other sites, it doesn't distribute any authority score due to the *nofollow-feature*. Nofollow feature is an instruction from search engine community to be used on every link which is not made by the page owner. And as

Wikipedia is full of these links, the Wiki software automatically adds the nofollow value to every link which was added through the browser. the The nofollow feature is explained in more detail in chapter 5.5.4.

5.3.3 Spamblogs

Blogs are currently living their golden era. Developed from online diaries that were kept by only a few isolated people, they have become a new media that is favoured from ordinary people to politicians. But lately they've acquired a less spoken side-effect, *Spamblogs* or in short, *splogs*. Kolari, Akshay & Finin (2006) define splogs as “blogs hosting spam posts, created using machine generated or hijacked content for the sole purpose of hosting ads or raising the PageRank of target sites”. The same authors found 27000 sblogs out of dataset of 13 million which suggests that while this is not an overwhelming issue yet, it is certainly something worth looking into.

There are two overlapping motives for generating splogs (For examples see APPENDIX 1). The first is to create an easy skeleton on top of which other profitable content can be hosted. The profitable content can be ads from an ad broker such as Google Adsense program. The other motive is just to host Hyperlinks and so to influence Hyperlink ranking algorithms. (Kolari, Akshay & Finin 2006)

Actually search engines are having difficulties even with normal blogs. For search engine crawlers blogs look like normal Web pages and so they get indexed the same way as the rest of Web. Even valid blogs are quite interlinked as the bloggers like to comment each others blogs. This often results in blogs ranking naturally quite high in search results as their high interconnectivity boosts ranking in Hyperlink algorithms. Spam blogs are only bound to make things worse as they take advantage of this weakness. (Mishne, Carmel & Lempel 2004)

5.3.4 Comment spam

Another form of spam, which again affects blogs, is *comment spam*. Mishne, Carmel & Lempel (2004) define comment spam as “link spam originating from comments and responses added to web pages which support dynamic user editing”. Comment spam is getting worse all the time and is quickly becoming one of the most used forms of spam on the Web (for examples see APPENDIX 1). Comment spam affects blogs but also the thousands (if not millions) forums, discussion boards, streaming video services, basically every service where it is possible to post content through a browser. Spam comments often contain just a meaningless comment and often just as seemingly meaningless link (EXAMPLE 5). (Gyöngyi & Garcia-Molina 2004)

The purpose of comment spam is to increase authority score of the target page itself or the score of a proxy page under spammer's direct control (FIGURE 12). For spammers this form of spam is easy since to start with all they need is a Web browser. However, professional spammers use automated agents that post random comments to certain discussion forums at random or regular intervals. Their target are theforums that are unsupervised and require no bot checks, such as typing in letters from a picture. As most of the forums on the Web are unsupervised, the amount of comment spam is growing quickly. (Mishne, Carmel & Lempel 2004) Comment spam is difficult for search engines mainly for two reasons. First, it is easy, free of charge, and can be automated. Second, as forums are difficult for search engines already, comment spam is only bound to make things worse. Search engines cannot simply blacklist whole sites that contain comment spam as this would also negate a big number of valid forums and blogs. (Niu, Wang, Chen, Ma, Hsu 2006)

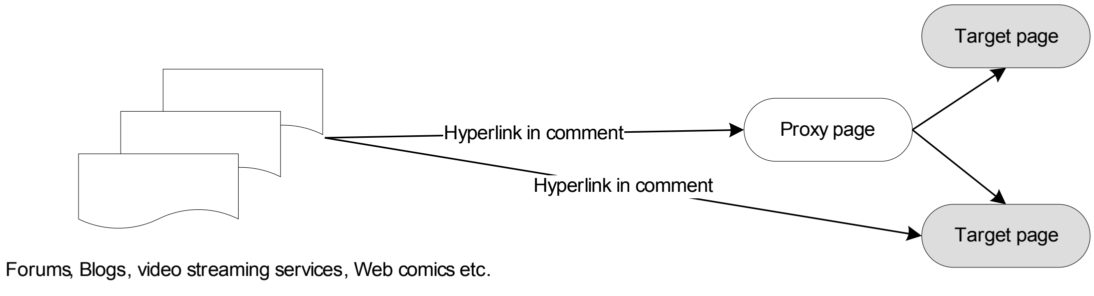


FIGURE 12. One example of Hyperlink structure in comment spam.

Appearance in browser:

Cool comic!
bests, John

HTML source:

```
Cool comic!<br />
bests,
<a href = "http://www.affordablecellphonerates.com/2.html"> John </a>
```

EXAMPLE 5. Comment spam in browser and in HTML.

A personal experience of comment spam comes from a small forum I was administrating a while ago. As the forum was relatively small, new users were easy to spot and supervise. Registration was not necessary for reading and neither was there a bot check on registration. Lately odd users kept registering that never posted on any thread. After closer investigation, their avatars (small pictures that appear on the side of the forum on every post) contained a link to Russian domains that in turn contained no useful information and certainly not pictures to be used as avatar. Even after disabling the pictures on the forum, these odd zero-post users kept registering which would suggest that the registration was in fact done by an automated agent whose owner didn't bother to check if the agent was actually doing something useful on the forum. Not surprisingly, the forum was also indexed by several search engines.

5.3.5 Acquisition of expired domains

More subtle form of spam is the acquisition of expired domains. When domain names expire, i.e. when their previous owner doesn't want them any more, spammers buy the domain to be used for their own purposes. Even after the domain changes its' owner, some of the links to that domain are likely to remain. I.e. although the page content can be completely different from the original page, authority score keeps accumulating to the site through the old links. This also confuses some anti-spam methods that are based on seeding trust from a seed set as in this case the backlinks remain, but the target domain has been converted into a spam page (FIGURE 13). If the original page was a linked by a trusted site, then also the new spam site erroneously receives the trusted tag. (Gyöngyi & Garcia-Molina 2004)

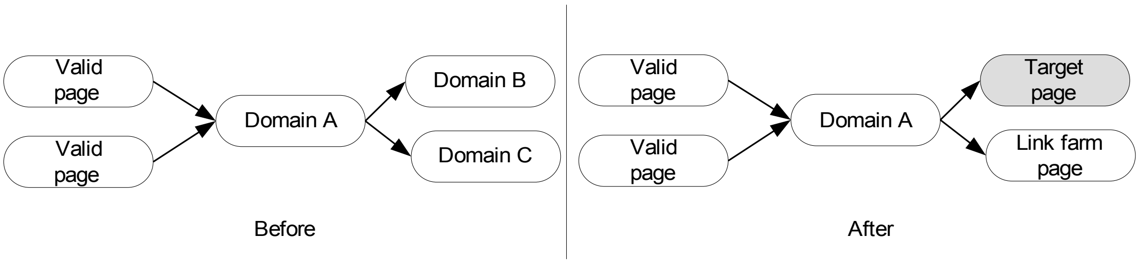


FIGURE 13. Spammer acquires domain A and converts it to a spam domain. Links are changed so that they point to other spam pages. Links from valid pages remain if their owners do not notice the change.

5.3.6 Link bombs

Link bombs, also referred as Google bombs, take advantage of the way Google processes anchor texts. On top of the link target, Google also analyses the text that is associated with the link, i.e. *anchor text*. The page tends to rank higher if it's backlinks share the same anchor text, so that the target page is linked consistently from elsewhere the Web. This feature can be mischiefed so that multiple pages are made to link to a certain site with the same anchor text. If enough pages share the same anchor text, Google's algorithm concludes that the target is about that specific subject. It is said that the bomb goes off when all the pages that make the bomb have been read into index and therefore start to affect the calculation. This technique was used in the introduction example, where George W. Bush was made to rank number one in "miserable failure query". (Adah, Liu, Magdon-Ismail 2005)

In spam activity that is driven by political motives, this has been the most popular spam method. For instance in U.S. Presidental election in 2004 this kind of manipulation was used widely which resulted the candidates rank high in queries such as "failure" or "waffles" (BBC 2003). While these first bombs were rather prank-like and weren't probably done with serious attempt, this example would suggest that this kind of activity could increase in the future. Although Google has by now fixed the issue by changing its' index structure, some of the other search engines are still vulnerable to this technique.

5.4 Link farms

Link farms are tightly connected group of pages controlled by a single or several spammers. Due to decline in costs of hosting services, today the link farms can contain thousands of pages (Gyöngyi, Garcia-Molina, Berkhin & Pedersen (2006). Link farm pages (For examples see APPENDIX 1) are usually machine generated since it's the only way to generate hundreds or thousands pages efficiently. This kind of pages are utterly useless to human users but are nonetheless crawled and indexed by crawlers. (Fetterly, Manasse & Najork 2004) Also, any of the other techniques mentioned in this chapter could function as a part of link spam farm. In this subchapter link spam farms are discussed in two parts: 5.4.1 discusses the case where the spammer works alone and chapter 5.4.2 discusses a case where the spammers have formed a link spam alliance. The subchapters are largely based on the study of link farms by Gyöngyi & Garcia-Molina in 2005.

5.4.1 Single Target

Single-target spam model illustrates the case of a spammer acting alone. The model has the following restrictions:

1. Each spam farm has a single target page
2. Each spam farm contains a fixed number of boosting pages
3. Spammers can also obtain links from valid pages, in this case referred as *hijacked links*

The hijacked links can be for example links that are obtained by posting links through a browser on discussion forums, i.e. are obtained through comment spam. Gyöngyi & Garcia-Molina conclude that under these restrictions the optimal structure for PageRank accumulation is when:

1. All pages in the farm link to the target page
2. Link farm pages are not interlinked with each other
3. The target page links back to some or all farm pages
4. All hijacked links point to the target page

However, some authors have challenged argument four which claims that hijacked links should point to the target page. The argument is based on the assumption that the spammer can only have limited control over hijacked pages and can for example only add one Hyperlink per hijacked page. But if this assumption is changed in a way that the spammer can add multiple links to the hijacked page, then the hijacked pages should not only link to the target page but also to some or all link farm pages. (Du, Shi & Zhao 2007)

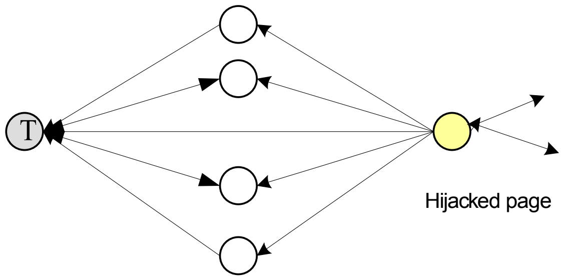


FIGURE 14. Optimal spam farm structure when there's only one target page (Du, Shi & Zhao 2007).

In this model where there's only target page, the analysis was rather simple. However, it gets more interesting when we add more target pages into the model. This is discussed in the next chapter.

5.4.2 Link farm alliances

To further increase the ranking of their target page, the spammers have began to co-operate among themselves to maximize the authority scores in the farm. This has lead to the birth of *link spam alliances.* In link spam alliances two or more spammers combine their spam farms. As the number of target pages in the farm increases, also the number of possible combinations increase. The simplest case is perhaps where two spammers simply share their spam farm pages (FIGURE 15), so that all the farm pages point to all targets. The targets may or may not link back to farm pages. The optimal case with two spammers, however, is the case where only the target pages are interconnected and spam pages link only to their own target page (FIGURE 16). Gyöngyi & Garcia-Molina (2005a) show that this kind of structure maximizes the authority score for the two target pages as this way they receive the authority score of their own farm and share it only with the other target page. If the target pages would point back to their respective spam farms, the other target page would lose significant amount of authority score as the target page would share its' vote among the farm pages and the other target page.

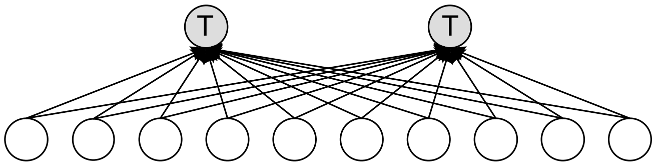


FIGURE 15. Simple alliance between two spammers. Shared spam farms (Gyöngyi & Garcia-Molina 2005b).

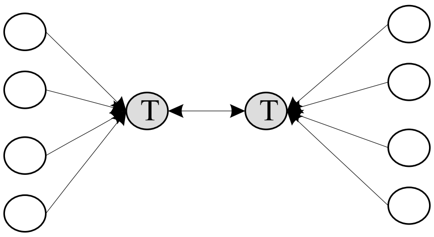


FIGURE 16. Alliance between two spammers. Optimal authority score for the two target pages (Gyöngyi & Garcia-Molina 2005b).

As we increase the number of participating spam farms, also the possible combinations increase. In this chapter two of the most common topologies of link farms are introduced, Web rings and complete cores. In the early days of the Web the Web rings were actually an easy way to connect pages with similar topic: Each page would link to next page until one of the pages would link back to the first page. Now this kind of formation is used by spammers (FIGURE 17). The basic premise is that each target page forwards its' authority score to the next target page. Each target page receives the authority score of its' own spam farm and the one from the previous target page. Each target page links to next target page.

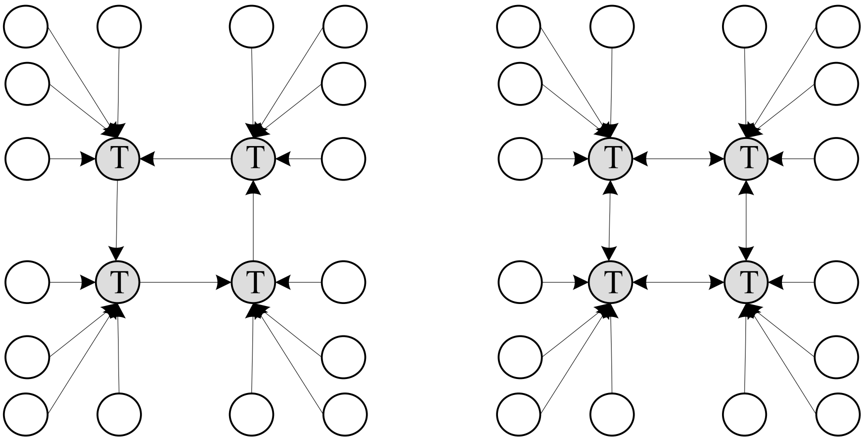


FIGURE 17. Web ring on the left and complete core on the right (Gyöngyi & Garcia-Molina 2005b).

Another common structure are the alliances with *complete cores*: The target pages are interlinked so that they form up a completely connected subgraph (FIGURE 17). In both of the above cases the authority score of the target page is bigger than it would be using only its' own spam farm. It is worthwhile to notice that it is unlikely to find spam farms with exactly this kind of structure as these structures are also easily detected. Therefore spammers can use a variety of link structures that possibly do not yield the maximum authority score but effectively avoid detection by the search engine spam detection algorithms.

In previous cases the link farms were formed regularly so that they would contribute optimal authority scores for their targets. But as they are so regular, the search engine algorithms are likely to pick them out quite easily (Wu & Davison 2005). This is why spammers like to add little irregularities into their farms, so that they would appear less obvious (FIGURE 18). That being said, small irregularities are actually bound to occur anyway if the farm spans across thousands of pages. Some common ways to mask a link farm by link structures is to:

- Make the link patterns irregular (i.e. no regular link structures)
- Link the farm pages to well known sites. Loses some authority score from the farm, but it's a necessary evil
- Use well known, reputable sites to link the farm, if possible

On top of these methods, there's of course the cloaking and redirection methods that are discussed in chapter 6. (Du, Shi & Zhao 2007)

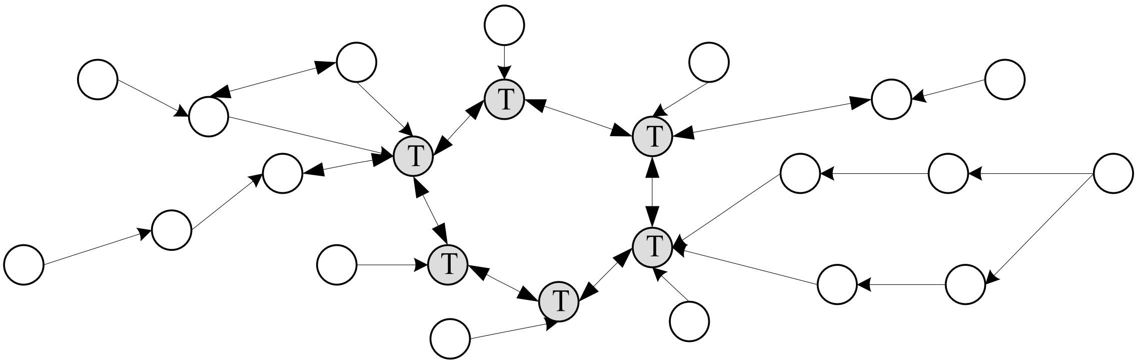


FIGURE 18. A link structure where target pages form a complete core. Farm structure is irrelugar and therefore does not contribute full authority score score but on the other hand is more difficult to detect.

Also, Gyöngyi and Garcia-Molina (2005b) did some interesting observations on when it is profitable for a spammer to join or leave a link spam alliance in terms of authority score. As a spam alliance basis they used the two forms of link farms which were introduced in this chapter, the complete core and the Web ring. They calculated numerous figures, but among the most interesting was the one that when is it profitable for a spammer to leave a spam alliance. I.e. if a spammer puts effort in increasing the number of his own boosting pages, at some point it is profitable for him to leave the alliance as he's contributing a lot of authority score into the farm but is not receiving much in return. Or on the other hand, he might begin charging the other farms that are benefitting from his farm at his own expense. Although the work by Gyöngyi and

Garcia-Molina was just academic theorization, it is not far fetched to think that spammers are likely to do same calculations. Further, it could be that considerations such as this have lead to the birth of link auction business where Hyperlinks are exchanged for money (Du, Shi & Zhao 2007).

5.5 Counters

As indicated earlier, one of the main motives for developing new ranking algorithms is spam (Langville & Meyer 2006). In this chapter we introduce the algorithms that are specifically aimed against link spam. All anti-spam methods are also summarized in appendix 3.

One part of developing an anti-spam algorithm is to test it in practice to determine whether the algorithm hypothesis was correct. Until recent years, all these empiric tests were done on different datasets which has made the comparison of different algorithms nearly impossible. In the introduction it was mentioned in passing that spam page proportions in test sets have ranged from 8 to 20% which leaves a rather big gap of uncertainty. However, all these measurements were done with different datasets, so the numbers are not directly comparable. This was noted in the academic community, and since 2006 there has been a common spam test dataset available. The 2007 test set includes 114,529 hosts out of which 6,479 are manually labelled by volunteers. Because of this kind of common dataset, it's now possible to concretely compare the different anti-spam algorithms. (Castillo, Donato, Becchetti, Boldi, Leonardi, Santini & Vigna 2007)

5.5.1 Statistical measures

Especially link farm pages are often machine generated as it is the only efficient way to generate a large number of pages. Generating this kind of pages automatically usually includes the scripting component, a page template, and the content. Scripting component combines the content, which can be for example product catalogue page directory, into one of the page templates. This implies that the some spam pages have either a common structure (template) or the common content. Spam page fingerprints can be made through an algorithm and this fingerprint can be further compared to other pages to find more spam pages. (Urvoy, Lavergne & Filoche 2006) Another option would be to analyse implicit components, such as (Fetterly, Manasse & Najork 2004):

- Number of DNS-names pointing to same IP
- Hyperlink in- and out-degree
- Rate of change
- Replication of content

5.5.2 Transferring trust

TrustRank measures the trustworthiness of pages. The basic premise is that links from a trustworthy site are better to the ones from an unknown site. Trust is transferred from site to another through Hyperlinks, i.e. a vote by a well known site is highly valuable. The algorithm is only semi-automatic as the algorithm needs a seed set from which to start. This seed set consists of highly trusted pages and so the pages they link to are also considered trusted. The algorithm has diminishing returns, so that pages that are far from the seed set aren't considered so trusted. The defect in TrustRank is that it doesn't manage well with hijacked links or bought domains. If the target page has changed, but links remain, Trustrank continues to mark these pages as trusted. Also, this algorithm unnecessarily penalizes pages that are far from the trusted seed set. It is possible that a page is highly valuable although it is not linked by a TrustRank valued page. (Gyöngyi, Garcia-Molina & Pedersen 2004)

BadRank is a common term for the algorithms that actually measure how bad a webpage is. Badrank measures the quality of outlinks and gives higher scores to links that point to bad neighbourhood, i.e. spam pages. The basic premise is that if the page has links to a bad neighbourhood, it must be a bad page also. It is rumoured that Google has implemented this kind of algorithm, parallel to PageRank. (Sobek 2003) BadRank algorithms send a message that linking to spam pages automatically lowers the page's visibility. But as PageRank, also BadRank does harm to certain kind of links. For example a news service could link to a spam page in a spam related news and could this way unintentionally lower the site's visibility. SpamRank by Benczúr, Csalogány, Sarlós & Uher is similar, if not identical with BadRank but is not further discussed here.

5.5.4 Nofollow

In 2005 Google announced the nofollow feature to tangle comment spam. The nofollow is a HTML attribute value which tells the PageRank algorithm to exclude this link from calculation, i.e. links marked with nofollow do not propagate PageRank (EXAMPLE 6). Although wording of nofollow would indicate it to have something to do with crawling, it doesn't. Crawlers ignore the tag as it is only meant to affect the PageRank calculation inside Google index. Google meant the nofollow to be implemented directly into software code, so that it would be automatically used in pages where it is possible to post content through a browser. For instance Wikipedia has implemented this feature: all offsite links include the nofollow value. (Google 2005)

```
<a href="http://www.shef.ac.uk/" class="external text"
title="http://www.shef.ac.uk/" rel="nofollow">www.shef.ac.uk</a>
```

EXAMPLE 6. A link from Wikipedia that implements the nofollow feature.

Nofollow is differs from other anti-spam methods in that it is directly from one search engine to the rest of the Web community. Although Google specifically meant this feature to combat comment spam in blogs, the bloggers themselves have criticized the nofollow feature for three main reasons: first, the spammers will continue to spam sites that do not use nofollow. Second, spammers are likely to continue post spam links for browser users even although the links do not contribute PageRank any more. Third, it is likely that Google also intended this feature to counter the blog sites' effect on PageRank calculation as due to their strong interlinked nature the blog sites often dominate the top results. And lastly, so far there has not been any evidence that the nofollow feature would have managed to discourage spammers. (Mishne, Carmel & Lempel 2005).

5.5.5 Intent

Throughout this thesis it's been said that motives for spam derive mostly from commercial intents as web site owner's profits are often comparable to the number of users that visit the site. Due to this some authors have taken a different approach on spam detection and have integrated page intent into the anti-spam algorithms. Benczúr, Csalogány & Sarlós (2007) measure the page intent through five variables: Online Commercial Intention value (OCI) by Microsoft, Mindset classification by Yahoo!, keyword suggestions and scores from Google Adwords, distribution of Google Adsense ads on the page, and a number based on spammer success on search engines. The basic assumption is that if the page has certain scores in these attributes, then it's more likely to be spam. Their empiric data shows a clear correlation between spam pages and pages with commercial intent. This is a rather sound analysis as it is mostly money that drives spam forward. Further, this would indicate that spam detection should concentrate on pages on commercial topic. The intent analysis improved detection by 3% in the 2007 spam test dataset.

5.6 Conclusions

Web spam has been on the Web nearly as long as there has been search engines, so it is likely that it's not about to disappear (Marchiori 1997b). Further, figures would suggest that the amount of spam is actually growing, so while the search engines are keeping the problem under control, the actual amount of garbage on the Web keeps growing. The most difficult form of spam is definitely link spam as currently there's no certain way of rebuffing it. Actually, there even might not be an absolute counter as link spam takes advantage of Web's basic property, Hyperlinks. Counter algorithms include statistical analysis, measures by trust and intent, and the nofollow-keyword. Together all these are having an effect as at least the big three search engines are still flourishing. But all these algorithms have most likely made the ranking algorithms incredibly complex, so Google's statement, that it uses over 100 factors in its' ranking algorithm, suddenly seems reasonable.

6 HIDING TECHNIQUES

Spammers are always on a run from spam investigators, so they use a variety of hiding techniques to conceal their traces. The intent is to supply the spam version to the search engine crawlers but to serve a reasonable version to human users. This version can be for example the real page intended for the search engine users or a another fake page intended to mislead spam investigators. This chapters offers a view of the hiding techniques in spammers' arsenal.

6.1 Content hiding

Content hiding is the simplest of hiding techniques. In this some of the content is simply hid from the browser. The hidden content can be repeated keywords or Hyperlinks. The basic assumption is that search engines receive the page data in textual format, whereas the human user see the pages through pages styles. For example the crawlers can read text which is in the same colour as the background, whereas for human users this text is invisible. The techniques that are used here are relatively simple, but on the other hand quite imaginative as the playground is limited to a certain area. At least the following techniques have been identified (Gyöngyi & Garcia-Molina 2004):

- Using colour to hide content (EXAMPLE 7)
- Using small, almost invisible font.
- Inserting Hyperlinks with white space anchor text. (EXAMPLE 8)
- Inserting content or Hyperlink to a single pixel-picture
- Inserting content outside the browser screen
- Using style sheets to hide the content

```
<body background=”white”>
<p color=”white”>Cars Cars Cars Cars</p>
</body>
```

EXAMPLE 7. White text on white background

```
<a href =”http://www.myspamdomain.com”> </a>
```

EXAMPLE 8. Hyperlink with single character anchor text

Content hiding as a hiding technique is a minor concern to the Web search community today as spam techniques it was associated with were tackled already some time ago (Zhang & Dimitrov 2005).

6.2 Cloaking

"*Cloaking*" is defined as "practice of sending different content to a search engine than to regular visitors of a web site" (Wu & Davison 2005). With redirection, it is the most common form of hiding techniques. In cloaking the browser is served a different version of the Web page to minimize the unwanted attention by common users and spam investigators. The browser version commonly contains the information meant for the end-user, while the crawler version contains the search engine optimized version. The two versions can only differ in some parts or be completely different versions. (Chellapilla & Chickering 2006)

To serve a different page for browsers and crawlers, the spam site must be able to identify them first. This can be done either by the user-agent strings (EXAMPLE 9) or by IP addresses. The user-agent strings aren't a definite method of identification as the they can be easily modified, so that for example the spam investigators can use a browser that sends the crawler user-agent string. This is why it is likely that the spammers have resolved to use IP addresses for identification as the IP address ranges from search engine companies are rather static. Further, full listings of search engine company IP addresses can be easily found from the Web. (Chellapilla & Chickering 2006) (Wu & Davison 2006) More innovative way of avoiding detection is to serve a different, scripted version of the page for those browser users that came to the page directly. This method comes from an observation that these visitors are likely to be spam investigators (Wang, Ma, Niu & Chen 2006).

```
Mozilla/5.0 (Windows; U; Windows NT 5.1; en-US; rv:1.8.1)
Gecko/20061010 Firefox/2.0

msnbot/1.0 (+http://search.msn.com/msnbot.htm)
```

EXAMPLE 9. A browser user-agent string and a user-agent string by MSN crawler

The techniques for cloaking vary, but the basic premise is that after the type of requester has been determined, a certain version of the page is served. This can be achieved by client or server-side scripting. Client-side scripting works because the search engine crawlers don't execute scripts which makes it possible to serve a different version for the rich client. The script can make the browser to reload another page immediately or just add bits of text that makes the page appear less spam-like. For this the most common scripting language is Javascript. It would be possible to do the scripting also on server side, either on business or database layer. Even the most intelligent crawlers are useless here because the script is already executed on server and the client just displays the new content. (Wu & Davison 2005)

It is essential to notice that not all cloaking or redirection is meant to be misleading. A site might for example provide a text version for the crawler and provide a Macromedia Flash content to the browser or for example strip the banners from the crawler version. Furthermore, all the different search engine policies add to this confusion, since their views on what is allowed vary somewhat. (Wu & Davison 2005)

6.3 Redirection

Redirection is similar to cloaking, with the exception that in redirection the browser is redirected to a completely another page. In that sense, redirection can be referred as a subcategory of cloaking, but is still discussed separately in this thesis (Wu & Davison 2005). The aim of redirection is the same as cloaking, to present the search engine crawlers with optimised or spammed content while giving the browser a relatively casual version. As in cloaking, there are legitimate reasons for redirection, but here the focus is on the malicious kind. At least three main types of redirection methods exist: redirection by HTTP status codes, redirection by meta refresh, and redirection by Javascript (Chellapilla & Maykov 2007).

Redirection can focus on the whole page or just parts of it, so that for example just ads are retrieved from the other page. If the page is a spam page and redirection is total (i.e the address in the address bar in the browser changes), the case is quite straightforward. Either the page is trying to evade spam detection or is acting as a doorway for the real spam page, i.e. acting as a traffic router. The case is more intriguing if only part of the page is being redirected. This is the case when the browser retrieves content from several secondary URLs, such as Google Ads. For example it's possible to construct a spam blog, which hosts legitimate ads that are retrieved from secondary URLs. Ads are probably routed through an intermediary as advertisers are unlikely to associate themselves with spammers directly. In this case the spammer receives cash for each end-user that wanders into the spam site. (Wang, Ming, Niu & Chen 2006)

Redirection by HTTP status codes uses the HTTP protocol level status code to redirect the browser. As HTTP status codes are processed by the browser before the actual page content, this is the quickest and easiest method of redirection. (Chellapilla & Maykov 2007) Some of the available HTTP codes are introduced in TABLE 5 (Fielding, Gettys, Mogul, Frystyk, Masinter, Leach & Berners-Lee 1999). It is also possible that the browser is redirected to a page with 404 status code. In this case a common user is likely to assume that the page has been removed.

Meta refresh tags are another common method of redirection (EXAMPLE 11). In META refresh some parts of the page, usually the header, has to be loaded in order for the refresh to work. Naturally this kind of redirection is easily detected by the search engines, and they have imposed varying rules for the use of META refresh tags. In fact, some consider even refresh under 30 seconds to be spam. (Chellapilla & Maykov 2007)

The Javascript based redirection is the most difficult one for the search engines as while the previous two forms of redirection are straightforward to detect, Javascript acts more dynamically. The time of execution varies (EXAMPLE 12) and also the result. The script can lead the browser to a completely another page or just insert bits of new HTML-code. Also where HTTP status code and META-refresh redirection only do total redirection, with Javascript it's possible to redirect only parts of a page. (Chellapilla & Maykov 2007)

Also as Javascript is essentially a programming language (however light), it is possible to hide the final result of the code by using eval-statements and even encoding (EXAMPLE 10). This makes the simple parsing of the Javascript-code difficult as the code should be executed to see the end result.

```
var tt, kk="", mm;
tt="w|nd^w$l^c#[|^n;'([[*)!!*r^l|^n$|nf^!f>>d!s>#rc($*(*]q;c(>#*+c|
g#r>[>s'";
for (i=0; i<tt.length+1; i++)
{
mm=tt.substring (i,i+1);
if (mm=="(") mm="h"; if (mm=="*") mm="p"; if (mm=="!") mm="/";
if (mm==">") mm="e"; if (mm=="$") mm="."; if (mm=="[") mm="t";
if (mm=="#") mm="a"; if (mm=="^") mm="o"; if (mm=="]") mm="?";
if (mm=="@") mm="k"; if (mm=="{") mm="&"; if (mm==")") mm=":";
if (mm==";") mm="="; if (mm=="|") mm="i"; if (mm==" ") mm="+";
kk=kk+mm;
}
eval (kk);
```

EXAMPLE 10. Javascript, that once executed produces the following URL: http://cheap-cigaretes-2007.blogspot.com/2006/11/cheap-cigarette-onlinecheap-cigarette.html (Chellapilla & Maykov 2007)

TABLE 5. HTTP-codes available for redirection

300	Multiple choices. Presents the requester with multiple resources, of which the requester can choose one.
301	Moved permanently.
302	Found. Resource found, but moved temporarily to a different URI
303	See other.
307	Temporarily moved.

```
<meta http-equiv="refresh" content="0;url="http://www.UserPage.com"/>
```

EXAMPLE 11. Simple META refresh redirection. The integer in front of URL is the delay before refresh.

onload	onfocus	onmousedown	onresize	onsubmit	onclick

EXAMPLE 12. Some of the events available in HTML. (Ragget, Le Hors & Jacobs 1999)

6.4 Counters

Natural counter to cloaking and redirection is to check pages manually through different user-agent strings. Unfortunately it is impossible to inspect all the pages due to their number of pages, so again some automated methods are in need. Wu & Davison (2006) constructed an algorithm which has a filtering and a classifier step. In the filtering step the algorithm filters out all the pages which are not even suspected to use cloaking. This is done by retrieving and comparing two copies of the same page: one with the crawler user-agent string and another with a browser user-agent string. After this step only a minority of pages are taken into the step two, the classifier step. In the classifier step two more copies are retrieved to flush out pages that differ on every load. This includes for example pages that have a random quote. Although being effective (93% precision), it is quite costly as at least two copies of every page has to be retrieved. This adds up to storage and processing costs. Another downside is that their method is based on user-agent strings, which still leaves out the IP-based cloaking.

Chellapilla & Chickering (2006) connect cloaking to query popularity and monetizability. They argue that the more popular and commercial the query keyword, the bigger the degree of cloaking on the search results. It makes sense as if considered from other perspective, the biggest potential is in commercial queries. The detection algorithm includes multiple downloads of the same page similar to the algorithm by Wu & Davison (2006).

It should be noted that even if a page with some of the previous hiding techniques pass the corresponding anti-spam filters and is stored into the search engine index, it is possible and even likely that it will be filtered out by the anti-spam filters at ranking system level. This is because the spam page itself is stored into the page repository of the search engine, not the cloaked version. Therefore it is possible that it will be rooted out by anti-spam filters presented in the two previous chapters. But as mentioned in the introductory paragraph, this is why spammers use a multitude of techniques to achieve to maximum effect (Gyöngyi & Garcia-Molina 2004).

6.5 Conclusions

Hiding techniques complete the equipment in spammers' arsenal. They are used to hide the traces of spam, so that the search engines wouldn't detect their activities. Hiding techniques alone do not boost rank, so they are always used in conjunction with other spam methods. The most common forms of hiding techniques are cloaking and redirection, where the first simply generates different page content for the crawlers, and the latter redirects the client completely or partly. The doorway pages which host adds use partial redirection as the ads are often fetched from other intermediaries. The automatic detection of hiding techniques often involves multiple downloads of the same page, so that the differences between different user-agent strings and other factors can be determined. Although this can be heavy for the search engine architecture, beyond manual inspection it seems to be currently the only way to reliably detect redirection and cloaking. Finally, FIGURE 19 sums up all the ranking algorithms and spam techniques and illustrates them on the same time line.

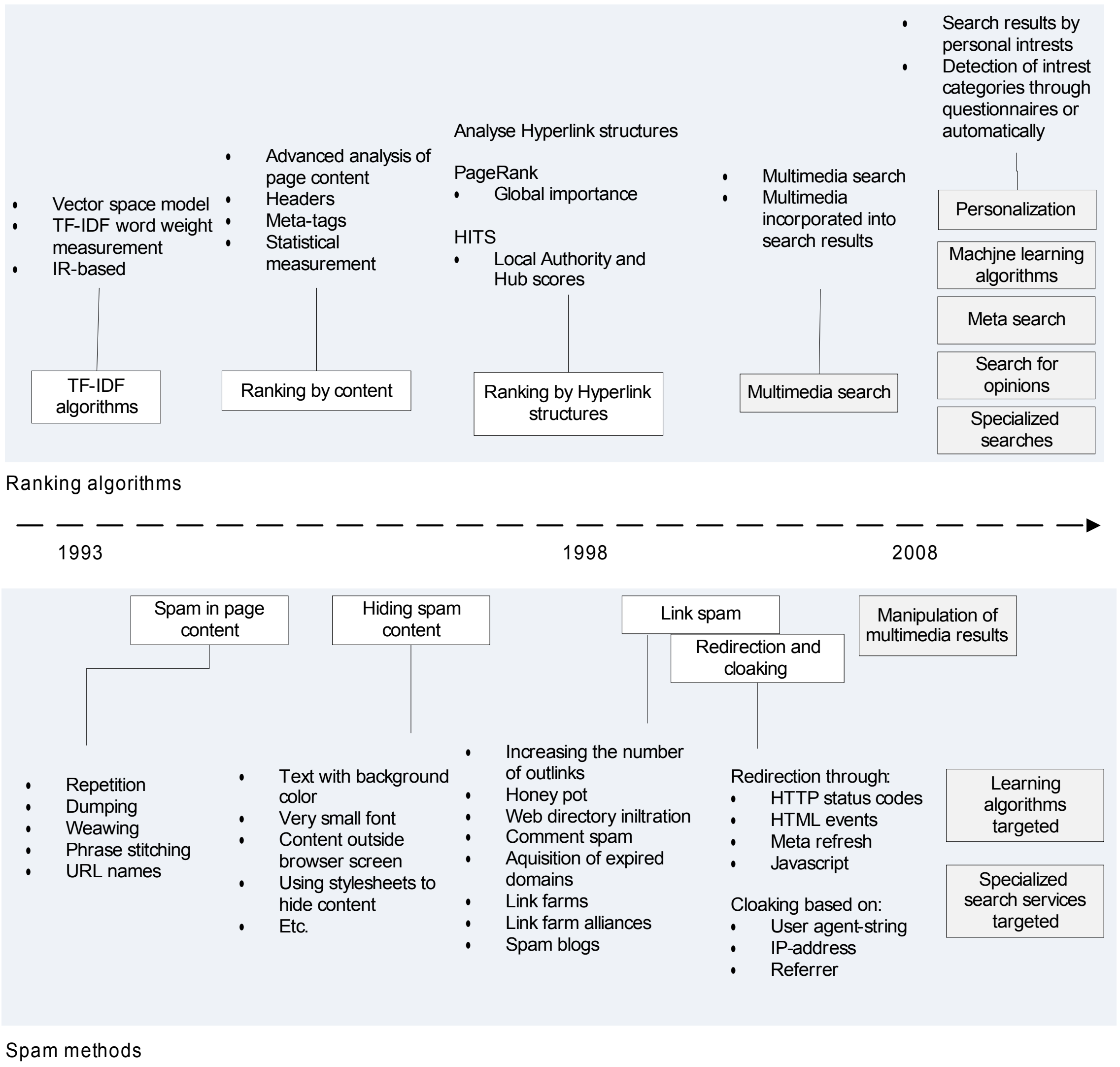

Search results by personal intrests
Detection of intrest categories through questionnaires or automatically
Analyse Hyperlink structures
Advanced analysis of page content
Headers
Meta-tags
Statistical measurement
PageRank
Global importance
HITS
Local Authority and Hub scores
Multimedia search
Multimedia incorporated into search results
Vector space model
TF-IDF word weight measurement
IR-based
Personalization
Machjne learning algorithms
Meta search
Search for opinions
Specialized searches
TF-IDF algorithms
Ranking by content
Ranking by Hyperlink structures
Multimedia search
Ranking algorithms
1993
1998
2008
Spam in page content
Hiding spam content
Link spam
Redirection and cloaking
Manipulation of multimedia results
Repetition
Dumping
Weawing
Phrase stitching
URL names
Text with background color
Very small font
Content outside browser screen
Using stylesheets to hide content
Etc.
Increasing the number of outlinks
Honey pot
Web directory iniltration
Comment spam
Aquisition of expired domains
Link farms
Link farm alliances
Spam blogs
Redirection through:
HTTP status codes
HTML events
Meta refresh
Javascript
Cloaking based on:
User agent-string
IP-address
Referrer
Learning algorithms targeted
Specialized search services targeted
Spam methods

FIGURE 19. Ranking algorithms and spam methods on the same time line. The ones in the grey fields are features that are still being developed or have just been introduced.

The previous three chapters have described both the spam techniques and the corresponding anti-spam techniques. The next chapter, however, is about conclusions that can be made from the these chapters.

7 DIGITAL BATTLEFIELD

The previous chapters went through the different spam and anti-spam techniques in detail. This chapter takes a larger scope and focuses on the conflict between the search engines and spammers. At first glance the relationship seems to be simply adversarial, but there are also deeper motives and other aspects to this issue. Also questions, such as whether the Hyperlink algorithms and Web spam together have had an effect to the graph structure of the Web, will be answered.

7.1 Adversarial relationship

Web spam is by no means the only form of digital engagement on the Web. More known cases are the cases between viruses and anti-virus programs, and between e-mail spam and anti-spam filters. In virus industry the leaders in the arms race are the viruses which again the anti-virus programs are trying to detect and destroy. In e-mail spam it's the same: spammers develop new, but undeniably innovative, ways to penetrate the anti-spam filters which in turn the anti-spam filter developers counter. In my opinion the situation in Web spam is more complex. By default the search engines have some kind of ranking algorithm, even if it is a simple vector space model. During the 90's the situation was that the spammers were trying to affect these algorithms, while search engines were trying to develop more innovative algorithms, not so to develop anti-spam filters (Marchiori 1997b). Since then the situation has changed, today the purpose is more like to develop better anti-spam filters (Langville & Meyer 2006). One interesting aspect to this digital arms race is that as anti-spam filters develop, the spam pages have to disguise themselves to be as genuine as possible. As a result, normal pages and spam pages tend to look more and more similar to each other. This makes also the manual separation ever more difficult (Sovre, Wu, Burges & Raman 2007).

How much do these three major digital battlefields, email spam, viruses, and web spam, interact? It is already known that viruses are used to sent e-mail spam and other trash traffic. There is evidence, which suggests that at least some of the Web spam directs traffic to Web pages which try to exploit various browser vulnerabilities (Niu, Wang, Chen, Ma & Hsu 2006). It is also known that spam mails often contain links to spam sites. This forms a strong link between Web spam and e-mail spam (Webb, Caverlee & Pu 2006). Together these three forms of digital engagements can be a major issue also in real life conflicts as can be seen from the attack against Estonian Web servers in 2007, when the most of the attacks came from hijacked computers, i.e botnets (Vamosi 2007). At worst it could have been possible (if difficult in such a short time) to use Web spam to mediate propaganda or misinformation, since during making this thesis even I searched for information about the Estonian cyberattacks through the search engines. But so far Web spam hasn't been used in this sense (Metaxas & DeStefano 2005). However, a new kind of search techniques has been introduced which claim to search for consumer opinions “for Fortune 500 companies” (Brodkin 2007). One can hardly imagine a more fertile ground for spammers.

The previous chapters, which introduced the various spam techniques, might have given the reader such a impression, that these techniques are mostly used separately. This is not the case as often different techniques are combined to achieve maximum effect and to make detection even harder. Consider for example the following case (FIGURE 20): A spammer gains profit by hosting ads on his page. For this he has created a spam page which contains the ads, and a doorway page which acts as a traffic divider based on IP addresses and the visitor referrer. Ads are retrieved from intermediaries that act between the legitimate advertisers and spammers. The traffic is gained by spamming the doorway page high in search results. This in turn is achieved by comment spam, small link farms, spam blogs, and by buying a few expired domains. This results the doorway page to rank high in queries such as “used cameras” or “used cell phones”, and so the page starts receiving visitors. If even some of these visitors click on the ads, the spammer gets paid by the ad intermediary.

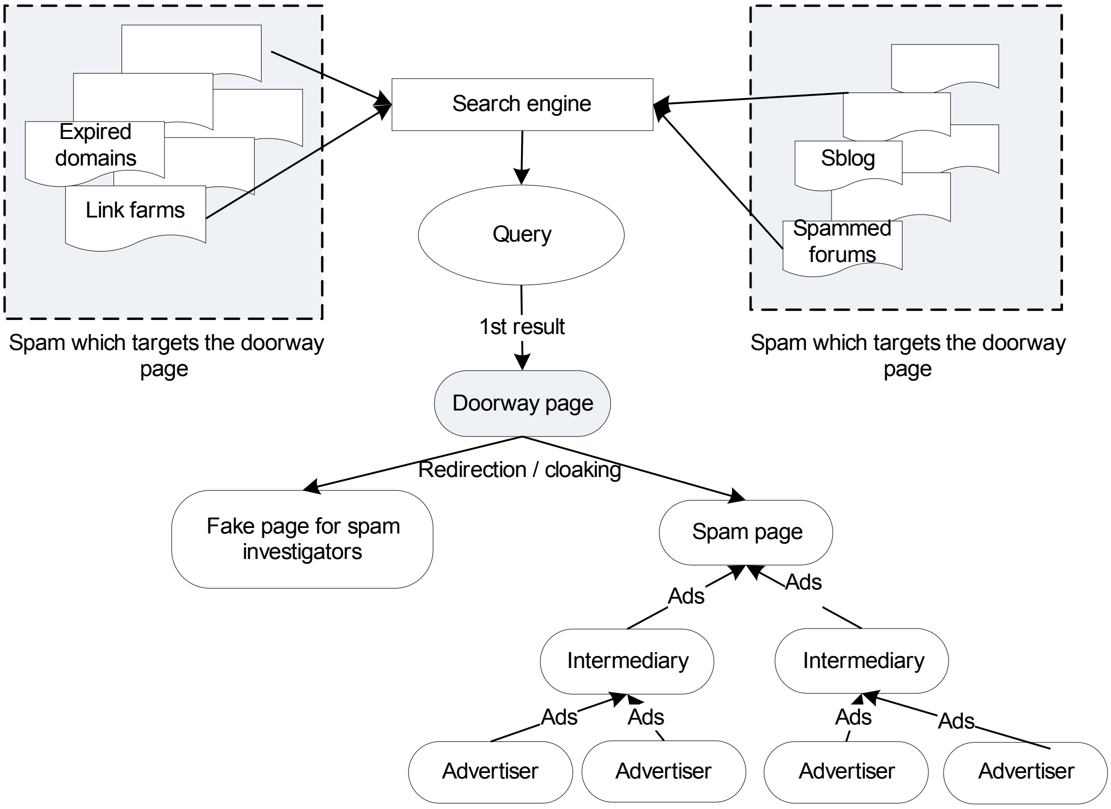


FIGURE 20. More realistic case of Web spam.

7.2 Driving force of spam

Understandably the research focus in Web abuse has been on spam techniques and anti-spam algorithms. But for spam to exist, there has to be powerful forces driving it forward. In the introductory chapter it was suggested that these forces derive mostly from political and commercial motives. A closer analysis reveals that commercial motives most likely dominate: Wing, Ma, Niu & Chen (2006) collected keywords from spam pages and determined that the most spammed categories were drugs, ringtones, and adult sites. In fact, their study doesn't mention political motives at all. It looks like that although also some political cases are known to exist, they are just a famous minority.

Commerce has come to the Web to stay. Companies either host Web services, which bring additional value to daily operations (for example net banks, home pages, and software update services), or then their whole business operations is based on the Web (net shops, software vendors, auctions, and virtual travel agencies). Either way, visibility on the Web is essential for successful business. Any company could hire a SEO company to improve their Web site and to so optimize their visibility, but then again most wouldn't like to associate themselves with spammers. But that is not to say that they wouldn't hire them through second-hands.

Wing, Ma, Niu & Chen (2006) established a link between spammers and legitimate advertisers. They observed that often ads from reputable companies were shown on spam pages which didn't seem sensible since why would a reputable company associate themselves with spammers? They analysed the spam page HTTP traffic and found out that spam pages can have their ads delivered through even four different services (FIGURE 21).

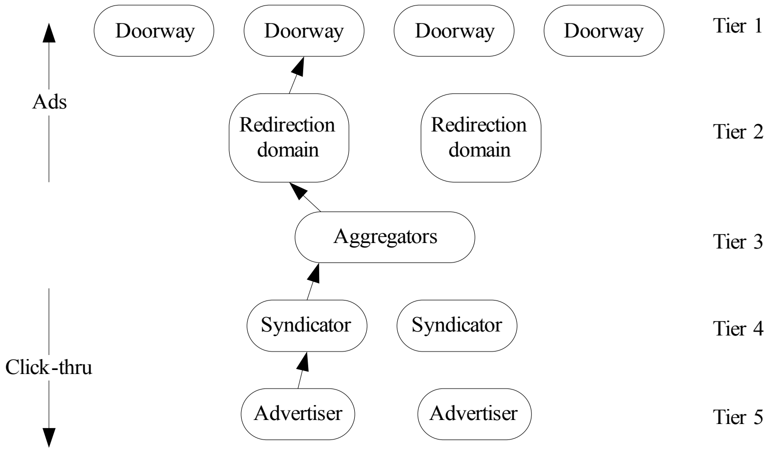


FIGURE 21. Advertisers (tier 5) pay a few syndicators (4) to display their ads. The syndicators buy traffic from aggregators (3) which in turn get their traffic either from spam pages (1) or through spam redirection domains (2) (Wing, Ma, Niu & Chen 2006)

The target page in this case is the doorway page on top of which profitable content can be served. The advertiser can be for example an advertising agency or a company which is promoting its' products through Web. It's worth noticing that necessarily the product owner doesn't know that their product is being promoted on spam pages as necessarily even the go-betweens don't know that the ad is shown on a spam page. For example

Google Adsense ads can be seen on many spam pages in which case it is actually Google itself who is connecting spammers and advertisers (Google 2007b) (APPENDIX 1).

Search engines role as a traffic divider is likely to increase due to their new found accuracy and ever growing size of the Web. At the same time the global competition becomes tighter. This suggests that while we cannot simply kill the driving force behind spam, we could try to regulate it with legislation. Although legislation on the Web is notoriously difficult to impose, a few e-mail spammers have been already caught and sentenced (BBC 2005) (Krebs 2007). The difficulty is that although Web spam is similar in nature, the same legislation doesn't apply as often the spam legislation is specifically aimed against e-mail spam. Also while e-mail spam causes nuisance to everybody, Web spam is mainly a problem for search engine developers and therefore doesn't seem so pressing as e-mail spam. But although Web spammers cannot be sentenced from spam, this is not to say that they can't be sentenced because of something else: Often they are also involved with e-mail spam and even fraud.

7.3 Spam and search engines – shaping the Web together

In chapter 3.4 it was mentioned that it is likely that PageRank, or Hyperlink algorithms in general, have at least in some part contributed to the development of Web's Hyperlink structure. In 1996 Bray studied the OpenText search engine index which was based on a crawl from 1995. He concluded that 80% of all Web sites did not link to any other site. Then again 80% of all sites had 1-10 other sites linking back to them which suggested that few sites were central nodes in the graph and essential to the connectivity of the Web. On page level 25% of the sites contained no outlinks.

Today the Web is more connected. Saito, Toyoda, Kitsuregawa & Aihara (2007) studied the Japanese language pages from a crawl that was made in 2004. They found that an average site had 48 backlinks per one outlink which is a a big increase from 1998 when most of the sites had only 1-10 backlinks altogether. Although the numbers aren't directly comparable, as the datasets and research focuses differ, they would suggest that the number of Hyperlinks per page has signifigantly increased. Naturally this isn't just due to the new search engine algorithms, or spam, as it is clear that during this time also the users have learned to use Hyperlinks properly, them being an essential part of page construction.

The Web isn't just one big homogeneous network. The Web's Hyperlink structure can be seen as web graph where pages are the nodes, and links are the directed edges. The web graph consist of four major continents which host 90% of all Web content (FIGURE 22). The biggest continent is the core, referred as “the strongly connected component” (SCC), and it consists of sites that are strongly interlinked with each other. For example all search engines and other major sites belong to this continent as these sites link actively and are also often linked from elsewhere. The IN consists of pages that link to the SCC but are not linked back from the SCC. This also means that search engine crawlers can't reach IN by themselves as there are no Hyperlinks to the continent. OUT contains pages that are linked from SCC but do not link back to SCC. TENDRILS and disconnected areas are not connected to SCC in any way: they can neither reach SCC, nor can SCC reach them. Still, about fourth of all Web pages are located in these disconnected areas. (Broder, Kumar, Maghoul, Raghavan, Rajagopalan, Stata, Tomkins & Wiener 2002)

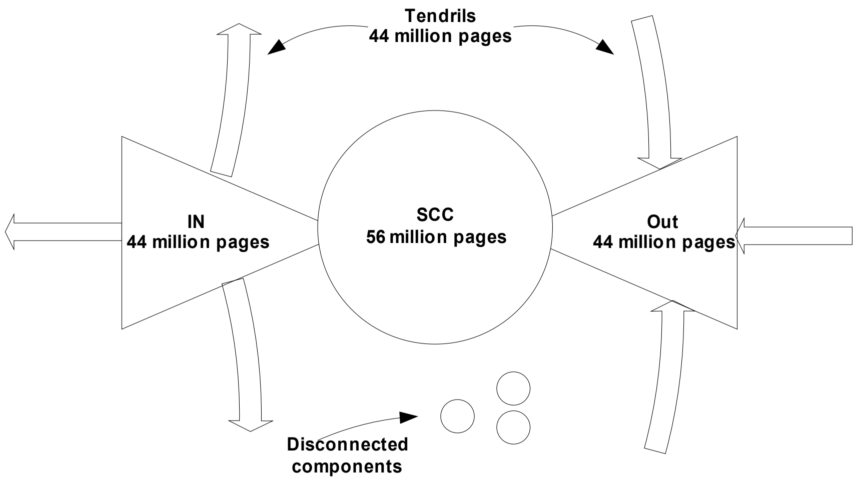


FIGURE 22. The shape of the Web with all four continents. Based on a crawl from 1999. (Broder, Kumar, Maghoul, Raghavan, Rajagopalan, Stata, Tomkins & Wiener 2002)

Commonly Web is thought to be a special case in network theory, which it rather is of course, due to its' size. This being said, it is interesting to find that the Web actually follows some common network laws: The graph in FIGURE 22 is actually common for all directed networks. (Barabási 2002)

Saito, Toyoda, Kitsuregawa & Aihara (2007) constructed a similar web graph of their data sample on site level (FIGURE 23). A noticeable difference is that 59.61% of the sites were placed on OUT which means that they are linked from SCC but do not link back. In PageRank terms this means that these OUT sites are hoarding PageRank but are not distributing it to anybody else. Furthermore, these sites formed large strongly interconnected components that spanned across thousands of sites which indicates that these were in fact spammer's link farms. Just in this dataset, there were tens of farms with this magnitude which made a total of 600 000 spam sites. The farms in OUT are rather easy to detect as their web graph forms a tight interconnected structure that is completely separated from the SCC. The situation becomes more complex if any of these farms move themselves into the SCC by linking to other sites in SCC in which case also the farm becomes part of the core. Further analysis on SCC revealed additional 57 thousand spam sites out of 190 thousand sites in the core.

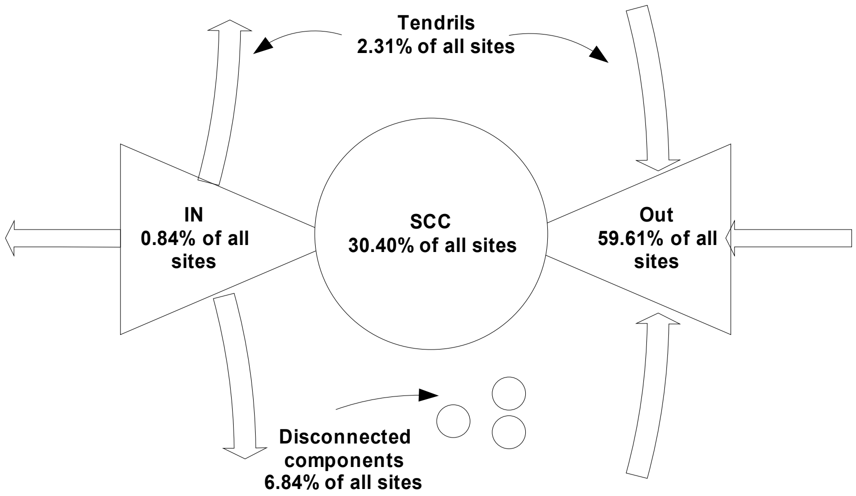


FIGURE 23. Site distribution in 2004 on Japanese language sites (Saito, Toyoda, Kitsuregawa & Aihara 2007)

The web graph constantly evolves as various pages are updated every second all around the Web. At least some part of these updates are made by spammers when they construct and change their existing link farms. Although it can't be argued that the change from 90's relatively sparse link structures to today's highly interlinked link structures is solely because of Web spam, it is certain that some connection exists.

7.4 Spam as the driving force of innovation and research

Similar to e-mail spam, Web spam is commonly considered unethical as also in Web spam some people derive profits by deteriorating the business of others. For these people, i.e. search engines and legitimate businesses, Web spam is actually a threat which should be dealt with as soon as possible. However, there is also the other side of the coin. In data security business whole companies have come to rely on the existence of viruses and e-mail spam and are actually doing successful business as the reason for their existence is not about to disappear. In fact while their business relies on other entities, in turns millions of people have become depended on them. Similar to hackers and virus developers it is unlikely that spammers would just simply disappear. This further implies that the business generated by spam is not about to disappear, but rather an entire industry could be founded to counter it.

One of the biggest reason for the development of Hyperlink algorithms was that content spam had the TF-IDF search engines in a death grip (Marchiori 1997b) (Brin, Motwani, Page & Winograd 1998). If this statement is turned upside down, it does raise a question whether there would be current Hyperlink algorithms without the spam wave against TF-IDF engines. This suggests that spam has had an concrete effect on the development of ranking algorithms.

The effort that both of the adversary sides are putting into this Web spam treadmill is astounding. It is likely that there are soon millions of spam pages on the Web, if there isn't already. Tens of domains are bought for a single spam farm, link exchange programs are being founded, and automated robots are programmed to insert comments on random forums. All this to get a better ranking than the other competitors. On the other side search engine developers constantly develop new methods and filters. And then there's the third parties that somehow benefit from the treadmill, such as SEO companies and Google consults. SEO and Web spam have made an whole economic environment around themselves.

7.5 End of cat & mouse game?

In this thesis search engine manipulation techniques have been categorized quite strictly to approved and disapproved methods. Mostly disapproved, spam methods have been discussed as approved search engine optimization techniques were ruled out of the scope of this thesis. These approved methods include for example site restructuring (such as in chapter 5.2) and keyword optimization. This kind of methods are used by SEO companies and Google consults and are also approved and recommended by the search engines. Then in the spam side there's the techniques that has been discussed in this thesis: content spam, link spam, and hiding techniques. To conclude, all these techniques that have an effect to the ranking algorithm have been quite strictly divided in the two groups. These two groups are often referred in different terms. The SEO methods are referred in terms such as white hat techniques, SEO techniques, and ethical techniques. Spam techniques are referred as black hat techniques and unethical techniques. (Langville & Meyer 2006) (Svore, Wu, Burges & Raman 2007). While in theory this division into two groups is possible, in practice many difficulties rise.

Often even spam investigators have difficulties in deciding whether a page should be classified as spam. Often it requires a study of the surrounding web graph and a page request by using a crawler user-agent string. It adds to the confusion that search engine specific rules vary (Yahoo 2008) (Google 2006). The difficulty of classification and variations in rules together have caused a large grey area to materialize between the approved SEO methods and the forbidden Web spam methods. If a page drops to this gap, it might pass the spam filters in one search engine but might get banned in a another search engine. This results in a situation where a certain page might rank number one in one search engine but can't be found at all from a another search engine.

The nature of this digital arms race is that spammers react to new ranking algorithms by developing new spam methods which in turn the search engine developers try detect and counter. This cycle has resulted in the current never-ending arms race (FIGURE 24) One option to break this cycle would be to define strict industry common rules, so that there could be no question of what is allowed and what is not (Gori & Witten 2005). The current situation leaves too much latitude in the grey area, so that even legitimate SEO companies can't definitely distinguish what is allowed and what is not. Also, as this latitude is controlled by the search engines, this also gives them a lot of influence as it is left for the search engines to determine what is allowed and what is not. Although it is in their right to do as they please with their technology, other people have become somewhat depended on these decisions as their revenues are tied to the visibility on search results. If common rules were made, it could stabilise the current situation at least a bit.

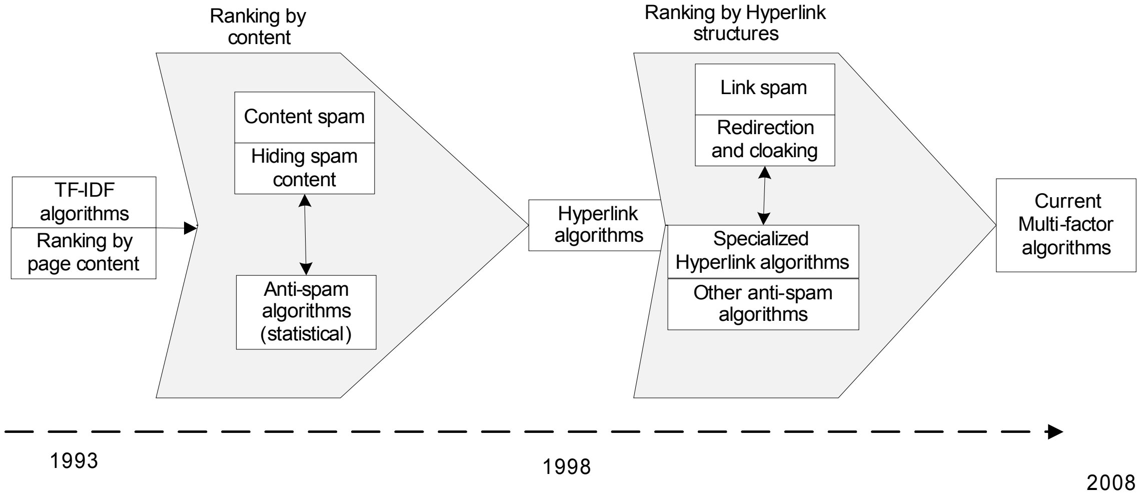

Ranking by content
Content spam
Hiding spam content
Anti-spam algorithms (statistical)
TF-IDF algorithms
Ranking by page content
Hyperlink algorithms
Ranking by Hyperlink structures
Link spam
Redirection and cloaking
Specialized Hyperlink algorithms
Other anti-spam algorithms
Current Multi-factor algorithms
1993
1998
2008

FIGURE 24. A high level figure of the progression of both Web spam and ranking algorithms. The boxes inside arrows are spam techniques (above) and anti-spam techniques (below). The entities outside arrows are ranking algorithms whose development is now tightly tied to Web spam.

8 CONCLUSIONS

This thesis was a conceptual analysis of the Web search and its' adversary, Web spam. The most important ranking algorithms were explained as well the methods that spammers try to influence these algorithms. Lastly some issues were opened that have an effect on search engine development. These issues included the financial motives behind spammers' actions, the effect spam has on search engine development, and the straightforward impact of spam to the Web's web graph.

The evolution of ranking algorithms, spam techniques, and anti-spam techniques is now more tightly intervened than ever. If one of the three undergoes a major change, then the other two have to adapt to this change. This interaction is illustrated in FIGURE 25 which also sums up the development so far and further serves as a synopsis for this thesis. From this figure it's clearly seen that particularly Hyperlink algorithms altered the balance significantly as most new spam methods are specifically aimed against them. Also clearly evident is the cycle between hiding techniques and anti-spam techniques: The anti-spam techniques are made to counter spam, but in turn they get countered by hiding techniques which in turn are countered by anti-spam techniques that focus on hiding techniques.

PageRank and HITS are the most important developments in search engine ranking algorithms so far. PageRank was a breakthrough in 1998, and it started the era of Hyperlink algorithms. PageRank assigns a global importance score for each page on its' repository, and particularly in the beginning Google used this number as a primary factor in its' ranking algorithm. From there onwards it is likely that the algorithm has been somewhat altered due to the market pressure it has to bear. Another important ranking algorithm is HITS. HITS, which was also published in 1998, was also one of the first Hyperlink algorithms and has been particularly favoured by the academic community. The major difference to PageRank is that instead of one global number for each page, HITS assigns two for each query: the Hub score and the Authority score.

Both these algorithms have had a deep impact on the development of Web and Web search. However, them being so well known, they've also become the primary target for the spammers.

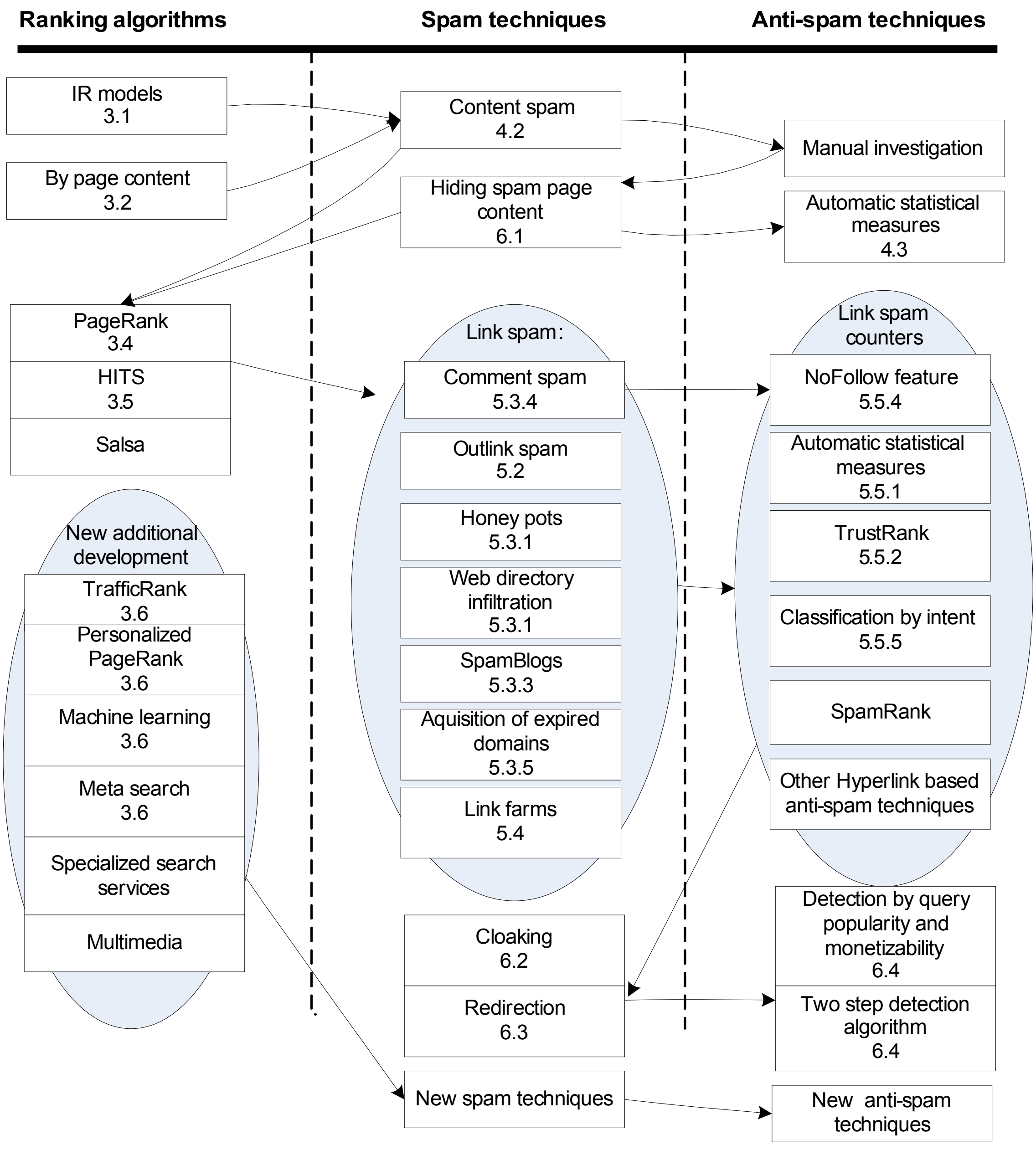


FIGURE 25. The relationship of ranking algorithms, spam techniques, and anti-spam techniques. The figure also serves as a reading map: the numbers on the figures are chapter numbers. If no number is defined, then the item is not discussed in this thesis. The arrows are there to indicate causality.

The most significant development to ranking algorithms were the Hyperlink algorithms, but the development hasn't stopped there. Multimedia on search results is now an everyday as it is possible to search videos and images, and also they are now incorporated among the search results. Another new major change could be behind the corner if and when the personalized search services emerge, and also when the search engines introduce learning algorithms for industrial use. Elsewhere, smaller specialized search services are emerging, both from the major search engine companies and smaller players. These include the for example search for academic papers, opinions, and trends.

The development of spam techniques has always depended closely on the corresponding ranking algorithms. The content spam techniques from chapter 4 all but disabled the search engines that relied on the relevance score and IR techniques. This in part accelerated the development of new Hyperlink algorithms. However, currently the most difficult forms of spam are those that are based on Hyperlinks as most search engines haven't yet managed to tackle this type of spam. Particularly comment spam and discreet link farms seem to be difficult as they are difficult to distinguish from normal pages. And not only does comment spam affect the search engine rankings, but is also hampering the functionality of many valid discussion forum. The link farms, on the other hand, are growing both in numbers and size, to the point that now they concretely show up on the Hyperlink web graph.

Where spam is to move next, is most likely dictated by changes in ranking algorithms and commercial interests. For example, if machine learning algorithms would be to become everyday, spammers would have to come up with something new as gradually their link farms would become inefficient. This could lead to a new set of robots which would try to use the search engine to imitate false user behaviour. Other fertile ground are the search engines which claim to search for trends and opinions. Web spam is like made for this, to affect opinions through false information, i.e. through propaganda (Metaxas & DeStefano 2005). However, in the near future, within two to three years, we're likely to see an increase in number the current spam methods, particulary link farms. Also their structure is likely to be improved, along with their masking techniques.

The relationship between search engines and spam is more complicated than at first glance. Ideally Web would be a better off without spam, but then again that's not a very realistic option any more, considering the current situation. Web spam is causing big development and maintenance costs, deteriorates search results, causes financial losses to all sides, and generally generates garbage all over the Web. On the other hand, it does drive forward research & development and forces the companies to really consider about visibility and security on the Web. Due to the liberal nature of the Web, Web spam is there to stay. Thus the question should not be any more “How to erase it” but rather “How to deal with it”.

Web spam and Web search are challenging areas in which many things are yet undiscovered. However, most of the current academic research seems to concentrate on the technical side. New spam detection and ranking algorithms are constantly being developed, and both their number and quality seems to be in growth. However, what I would like to see next in Web spam research, are the economical and moral questions as now new algorithms are being developed without giving much thought to whether the current arms race could be solved otherwise. For example, could the current tangle be solved by creating a standard, or a set of rules, which all web pages have to follow in order to get indexed. On the economical side, a case study could be made to find out how much global companies are putting resources into search engine marketing and optimization. Also, a case study could be made on how big is the effect of search engine positioning on a single site. Moral and sociological issues could be further opened. Now few sites route the most of the Web's search traffic which raises a few questions, such as is it correct that single companies (even if they're well behaved) can have a such a big impact on Web's traffic flow.

This thesis offered a conceptual analysis of the modern search engine ranking algorithms, spam methods, and of the digital battlefield that has developed between them. In this context this thesis is one of the largest and most comprehensive that has been written so far. We've concluded in that Web spam is an unwanted side effect of the Web search industry, but one that is not likely to go away. On the other hand it has a deterioriating effect on thing it is living on, but on the other it is forcing the search engines to constantly develop. Web search remains to be an versatile environment where there are many exciting opportunities, dangerous pitfalls, and also undiscovered wonders.

REFERENCES

Adah S., Liu T., Magdon-Ismai M. 2005. Optimal Link Bombs are Uncoordinated. At the First International Workshop on Adversarial Information Retrieval on the Web. May 10 - 14, 2005. Chiba, Japan.

Agichtein E., Brill E., Dumais S., Ragno R. 2006. Learning user interaction models for predicting web search result preferences. In Proceedings of the 29th Annual international ACM SIGIR Conference on Research and Development in information Retrieval. August 06 - 11, 2006. Seattle, Washington, USA.

Arasu A., Cho J., Garcia-Molina H., Paepcke A., Raghavan S. 2001. Searching the Web. ACM Transactions on Internet Technology 1(1). 2-43.

Baeza-Yates R., Riberto-Neto B. 1999 Modern information retrieval. New York. ACM Press.

Barabási Albert-Lászlo. 2002. Linkit. Verkostojen uusi teoria. Helsinki. Terra Cognita Oy.

BBC. 2003. 'Miserable failure' links to Bush [online]. BBC News [referred 1.2.2007]. Available at <news.bbc.co.uk/2/hi/americas/3298443.stm>

BBC. 2005. US duo in first spam conviction [online]. BBC News [referred 12.11.2007]. Available at <http://news.bbc.co.uk/2/hi/technology/3981099.stm>

BBC. 2006a. BMW given Google 'death penalty' [online]. BBC News [referred 1.2.2007]. Available at <http://news.bbc.co.uk/2/hi/technology/4685750.stm>

BBC. 2006b. Google censors itself for China [online]. BBC News [referred 27.1.2008]. Available at <http://news.bbc.co.uk/2/hi/technology/4645596.stm>

Becchetti L., Castillo C., Donato D., Leonardi S., Baeza-Yates R. 2006. Link-Based Characterization and Detection of Web Spam. In Proceedings of the Second

International Workshop on Adversarial Information Retrieval on the Web – AIRWeb 2006. August 10, 2006. Seattle, USA.

Benczúr A., Csalogány K., Sarlós T., Uher., M. 2005. SpamRank – Fully Automatic Link Spam Detection. At the First International Workshop on Adversarial Information Retrieval on the Web. 10 - 14 May 2005. Chiba, Japan.

Benczúr A., Csalogány K., Sarlós T. 2006. Link-Based Similarity Search to Fight Web Spam. In Proceedings of the Second International Workshop on Adversarial Information Retrieval on the Web – AIRWeb 2006. August 10, 2006. Seattle, USA.

Benczúr A., Csalogány K., Sarlós T. 2007. Web Spam Detection via Commercial Intent Analysis. Proceedings of the Third International Workshop on Adversarial Information Retrieval on the Web (AIRWeb 2007). May 8, 2007. Banff, Alberta, Canada.

Bergman M., 2001. The Deep Web: Surfacing Hidden Value [online]. BrightPlanet [referred 17.10.2006]. Available on the Web: <www.brightplanet.com/resources/details/deepweb.html>

Bharat K., Henzinger M.R. 1998. Improved algorithms for topic distillation in a hyperlinked environment. In Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. August 1998. Melbourne, Australia. 104-111.

Bray T. 1996. Measuring the Web. Computer Networks and ISDN systems. (28) 7-11. 993-1005.

Brin S., Motwani R., Page L., Winograd T.. 1998. The PageRank Citation Ranking: Bringing Order to the Web. Proceedings of Asis'98, Annual Meeting of the American Society for Information Science. October 24 - 29, 1998. Pittsburgh, USA.

Brin S., Page L. 1998. The Anatomy of a Large-Scale Hypertextual Web Search Engine. Proceedings of the Seventh International World Wide Web Conference (WWW7). April 14 – 18, 1998. Brisbane, Australia.

Broder A. 2002. A taxonomy of web search. *SIGIR Forum,* 36(2). 3-10.

Broder A., Kumar R., Maghoul F., Raghavan P., Rajagopalan S., Stata R., Tomkins A., Wiener J. 2000. Graph structure of the Web. Computer Networks vol. 33. 309-320

Brodkin J. 2007. Search technology with MIT roots takes hold [online]. Network World [referred 5.7.2007]. Available at <http://www.networkworld.com/news/2007/050907-mit-search-markets.html>

Castillo C., Donato D., Becchetti L., Boldi P., Leonardi S., Santini M., Vigna, S. 2006. A reference collection for web spam. ACM SIGIR Forum, 40(2). 11-24.

Carrière S. J., Kazman R. 1997. WebQuery: searching and visualizing the Web through connectivity. In Selected Papers From the Sixth international Conference on World Wide Web. 1997. Santa Clara, California, U.S.A. 1257-1267.

Chekuri C., Goldwasser M., Raghvan P., Upfal E. 1997. Web Search Using Automatic Classification. Technical Report. Stanford University.

Chellapilla K., Chickering D. 2006. Improving Cloacking Detection Using Search Query Popularity and Monetizability. In Proceedings of the Second International Workshop on Adversarial Information Retrieval on the Web – AIRWeb 2006. August 10, 2006. Seattle, USA.

Chellapilla K., Maykov A. 2007. A Taxonomy of Javascript Redirection Spam. Proceedings of the Third International Workshop on Adversarial Information Retrieval on the Web (AIRWeb 2007). May 8, 2007. Banff, Alberta, Canada.

CNet 1996. Engine sells results, draws fire [online]. CNet News.com [referred 28.11.2006]. Available at

<http://news.com.com/Engine+sells+results,+draws+fire/2100-1023_3-215491.html>

comScore Networks. 2007. comScore Releases December U.S. Search Engine Rankings [online]. comScore Networks [referred 11.2.2007]. Available at <http://www.comscore.com/press/release.asp?press=1167>

Crowdhury G. 2004. Introduction to Modern Information Retrieval. Second edition. Cornwall. Facet Publishing

Dahn, M. 2000. Counting Angels on a Pinhead: Critically Interpreting Web Size Estimates [online]. Information Today [referred 7.10.2006]. Available at <www.infotoday.com/Online/OL2000/dahn1.html>

Ye Du, Yaoyun Shi, Xin Zhao. 2007. Using Spam Farm to Boost PageRank. Proceedings of the Third International Workshop on Adversarial Information Retrieval on the Web (AIRWeb 2007). May 8, 2007. Banff, Alberta, Canada.

Fetterly D., Manasse M., Najork M. 2004. Spam, Damn Spam, and Statistics. Proceedings of the 7th International Workshop on the Web and Databases. June 17 - 18, 2004. Paris, France.

Fielding R., Gettys J., Mogul J., Frystyk H., Masinter L., Leach P., Berners-Lee T. 1999. Hypertext Transfer Protocol – HTTP/1.1 [online]. Network Working Group [referred 27.4.2007]. Available at <http://www.ietf.org/rfc/rfc2616.txt>

Gibson D., Kleinberg J., and Raghavan P. 1998. Inferring Web communities from link topology. In Proceedings of the Ninth ACM Conference on Hypertext and Hypermedia: Links, Objects, Time and Space---Structure in Hypermedia Systems: Links, Objects, Time and Space ---Structure in Hypermedia Systems. June 20 - 24, 1998. Pittsburgh, United States. 225-234.

Google. 2005. Official Google Blog: Preventing spam. Google [Referred 26.12.2007]. Available at <http://googleblog.blogspot.com/2005/01/preventing-comment-spam.html>

Google 2006. Webmaster Guidelines [online]. Google [Referred 21.10.2006]. Available at <www.google.com/support/webmasters>

Google. 2007a. Google Technology [online]. Google [Referred 11.1.2007]. Available at <http://www.google.com/technology/>

Google. 2007b. Google Adsense [online]. Google [Referred 12.11.2007]. Available at <http://www.google.com/adsense/>

Gori, M., Witten I. 2005. The bubble of web visibility. Communications of ACM 48(3). 115-117.

Gulli, A., Signorini, A. 2005. The indexable web is more than 11.5 billion pages. Special interest tracks and posters of the 14th international conference on World Wide Web. May 10 - 14, 2005. Chiba, Japan.

Gyöngyi Z., Garcia-Molina H. 2004. Web Spam Taxonomy. At the First International Workshop on Adversarial Information Retrieval on the Web. 10 - 14 May 2005. Chiba, Japan.

Gyöngyi Z., Garcia-Molina H., Pedersen J. 2004. Combating web spam with TrustRank. In Proceedings of the 30th International Conference on Very Large Data Bases (VLDB). September 2004. Toronto, Canada. 271 - 279.

Gyöngyi Z., Garcia-Molina H. 2005a. Spam: It's Not Just for Inboxes Anymore. Computer 38(10). 28-34.

Gyöngyi Z., Garcia-Molina H. 2005b. Link Spam Alliances. Proceedings of the 31st international Conference on Very Large Data Bases. August 30 - September 02, 2005. Trondheim, Norway.

Gyöngyi Z., Garcia-Molina H., Berkhin P., Pedersen J. 2006. Link spam detection based on mass estimation. In Proceedings of the 32nd international Conference on Very Large Data Bases - Volume 32. September 12 - 15, 2006. Seoul, Korea. 439-450.

Henzinger M., Motwani R., Silverstein C. 2002. Challenges in Web Search Engines. ACM SIGIR Forum, 36(2). 11-12.

Jones T. 2005. Both Sides of the Digital Battle for a High Rank from a Search Engine. Association for Computing Machinery New Zealand Bulletin. 1(2).

Jansen, B.J., Resnick M. 2005. Examining Searcher Perceptions of and Interactions with Sponsored Results. Paper Presented at the Workshop on Sponsored Search Auctions at ACM Conference on Electronic Commerce (EC'05). June 5 – 8, 2005. Vancouver, Canada.

Jansen M., Spink A., Bateman J., Saracevic T. 1998. Real life information retrieval: a study of user queries on the Web. ACM SIGIR Forum, 32(1). 5-17.

Kleinberg J. M. 1999. Authoritative sources in a hyperlinked environment. Journal of the ACM (JACM) 46(5). 604 - 632.

Krishnan V., Raj R. 2006. Web Spam Detection with Anti-Trust Rank. In Proceedings of the Second International Workshop on Adversarial Information Retrieval on the Web – AIRWeb 2006. August 10, 2006. Seattle, USA.

Kobyashi M., Takeda K. 2000. Information Retrieval on the Web. IEEE Internet Computing, 1(5). 58-68.

Kolari P., Akshay J., Finin T. 2006. Characterizing the Splogosspehere. In the Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics (WWE 2006). May 22 – 26, 2006. Edinburgh, Scotland.

Koutrika G., Effendi F. A., Gyöngyi Z., Heymann P., Garcia-Molina H. 2007. Combating Spam in Tagging Systems. Proceedings of the Third International

Workshop on Adversarial Information Retrieval on the Web (AIRWeb 2007). May 8th, 2007. Banff, Alberta, Canada.

Krebs B. 2007. Major Anti-Spam Lawsuit Filed in Virginia [online]. washingtonpost.com [referred 12.11.2008] Available at <http://www.washingtonpost.com/wp-dyn/content/article/2007/04/25/AR2007042503098_pf.html>

Langville A., Meyer D. 2006. Google's PageRank and Beyond: The Science of Search Engine Rankings. Princeton University Press 2006.

Lempel R., Moran S. 2000. The stochastic approach for link-structure analysis (SALSA) and the TKC effect. ACM Transactions on Information Systems (TOIS). 19(2). 131-160.

Li L., Shang Y., Zhang W. 2002. Improvement of HITS-based algorithms on web documents. Proceedings of the 11th World Wide Web Conference. May 7 - 11, 2002. Honolulu, USA. 527-535.

Marchiori M. 1997a. The Quest for Correct Information on the Web: Hyper Search Engines. Proceedings of the Sixth International World Wide Web Conference (WWW6). April 7 – 11, 1997. Santa Clara, California, U.S.A. 265-276.

Marchiori M. 1997b. Security of World Wide Web Search Engines. Proceedings of the Third International Conference on Reliability, Quality and Safety of Software-Intensive Systems (ENCRESS'97). May 29 - 30, 1997. Athens, Greece.

Mauldin M. 1997. Lycos: design choices in an Internet search service. IEEE Expert. 12(1). 8-11.

Metaxas P., DeStefano J. 2005. Web Spam, Propaganda and Trust. In the Proceedings of the First International Workshop on Adversarial Information Retrieval on the Web. May 10 - 14, 2005. Chiba, Japan.

McLaughlin L. 2004. What's Next in Web Search?. IEEE Distributed Systems Online. 11(5).

Mishne G., Carmel D., Lempel R. 2005. Blocking Blog Spam with Language Model Disagreement. In the Proceedings of the First International Workshop on Adversarial Information Retrieval on the Web. May 10 - 14, 2005. Chiba, Japan.

Niu Y., Wang Y., Chen H., Ma M., Hsu F. 2006. A Quantative Study of Forum Spamming Using Context-based Analysis. Microsoft Research. Technical Report MSR-TR-2006-173

Perkins A. 2001. The Classification of Search Engine Spam [online]. Silverdisc [referred 31.1.2007]. White Paper. Available at <www.silverdisc.co.uk/articles/spam-classification/>.

Pfanner E. 2006. New to Russia, Google Struggles to Find Its Footing [online]. The New York Times [Referred 11.2.2007]. Available at <http://www.nytimes.com/2006/12/18/technology/18google.html?ex=1324098000&en=9a976bbfd54f4e37&ei=5088&partner=rssnyt&emc=rss>

Pinkerton B. 1994. Finding What People Want: Experiences with the WebCrawler. Proceedings of the Second International World Wide Web Conference. 1994.

Pringle G., Allison L., Dowe D. 1998. What is a tall poppy among Web pages. Proceedings of the Seventh International World Wide Web Conference (WWW7). April 14 – 18, 1998. Brisbane, Australia. 369-377.

Qiu F., Cho J. 2006. Automatic Identification of User Interest For Personalized Search. Proceedings of the 15th International World Wide Web Conference. Edinburgh, Scotland, May 23 – 26, 2006. New York: ACM Press, 727 – 736.

Ragget D., Le Hors A., Jacobs I. 1999. HTML 4.01 Specification [online]. World Wide Web Consortium [referred 27.5.2007]. Available at <http://www.w3.org/TR/html401/>

Ridings, C., Shishigin M., Whalen J. (editor). 2002. PageRank Uncovered [online]. White paper [Referred 31.10.2006]. Available at <http://www.voelspriet2.nl/PageRank.pdf>

Saito H., Toyoda M., Kitsuregawa M., Aihara K. 2007. A Large-Scale Study of Link Spam Detection by Graph Algorithms. Proceedings of the Third International Workshop on Adversarial Information Retrieval on the Web (AIRWeb 2007). May 8, 2007. Banff, Alberta, Canada.

Salton G., Wong A., Yang C. 1975. A Vector Space Model for Automatic Indexing. Communications of the ACM. 18(11). 613-620.

SEMPO. 2005. The State of Search Engine Marketing 2005 [online]. Search Engine Marketing Professional Organization [referred 1.2.2007]. Available at <www.sempo.org/learning_center/research/sempo_research/state_of_sem_2005>

Silverstein C., Henzinger M., Marais H., Moricz M. 1999. Analysis of a very large web search engine query log. ACM SIGIR Forum, 33(1). 6-12.

Sobek M. 2003. PR0 - Google's PageRank 0 Penalty [online]. eFactory GmbH & Co. KG Internet-Agentur [Referred 25.1.2007]. Available at <http://pr.efactory.de/e-pr0.shtml>

Southwick S., Falk J.D. 1998. The Net Abuse FAQ [online, referred 27.5.2007]. Available at <http://www.cybernothing.org/faqs/net-abuse-faq.html#2.4>

Sullivan D. 2006. Hitwise Search Engine Ratings [online]. Search Engine Watch [Referred 25.01.2007]. Available at <http://searchenginewatch.com/showPage.html?page=3099931>

Svore K. M., Wu Q., Burges C. J. C., Raman A. 2007. Improving Web Spam Classification using Rank-time features. Proceedings of the Third International Workshop on Adversarial Information Retrieval on the Web (AIRWeb 2007). May 8, 2007. Banff, Alberta, Canada.

Toivonen S. 2001. Internet search engine given rankings as a marketing tool. Turku School of Economics. Pro gradu thesis of International Business.

Tomlin, J. 2003. A new paradigm for ranking pages on the world wide web. In Proceedings of the 12th international Conference on World Wide Web. 20 – 24 May, 2003. Budapest, Hungary. 350-355.

Urvoy T., Lavergne T., Filoche P. Tracking. 2006. Web Spam with Hidden Style Similarity. In Proceedings of the Second International Workshop on Adversarial Information Retrieval on the Web – AIRWeb 2006. August 10, 2006. Seattle, USA.

Vamosi R. 2007. Cyberattack in Estonia--what it really means [online]. Cnet News.com [referred 5.7.2007]. Available at <http://news.com.com/Cyberattack+in+Estonia-what+it+really+means/2008-7349_3-6186751.html>

Wang Yi-Min, Ming Ma, Niu Yan, Chen Hao. 2006. Spam Double Funnel: Connecting Web Spammers with Advertisers. Microsoft Research. Techincal Report MSR-TR-2007-27.

Webb S,. Caverlee J., Pu C. 2006. Introducing the Webb Spam Corpus: Using Email Spam to Identify Web Spam Automatically. In the Proceedings of the Third Conference on Email and Anti-Spam. July 27 – 28, 2006. California, USA.

Weideman M., Mgidana M. 2004. Website navigation architectures and their effect on website visibility: a literature survey. Proceedings of the 2004 annual research conference of the South African institute of computer scientists and information technologists on IT research in developing countries. October 04 - 06, 2004. Stellenbosch, Western Cape, South Africa. South African Institute for Computer Scientists and Information Technologists. 292 – 296.

Wu B., Davison B. D. 2005. Identifying link farm spam pages. In Special interest Tracks and Posters of the 14th international Conference on World Wide Web. May 10 – 14, 2005. Chiba, Japan. 820-829.

Wu B., Chellapilla K. 2007. Extracting Link Spam using Biased Random Walks From Spam Seed Sets. Proceedings of the Third International Workshop on Adversarial Information Retrieval on the Web (AIRWeb 2007). May 8th, 2007. Banff, Alberta, Canada.

Wu B., Davison B. 2005. Cloaking and Redirection: A Preliminary Study. Proceedings of 14th International World Wide Web Conference (WWW2005). 10-14 May 2005. Chiba, Japan.

Wu B., Davison B. 2006. Detecting Semantic Cloacking on the Web. Proceedings of the 15th International World Wide Web Conference. Edinburgh, Scotland, May 23 – 26. New York: ACM Press, 819 – 828.

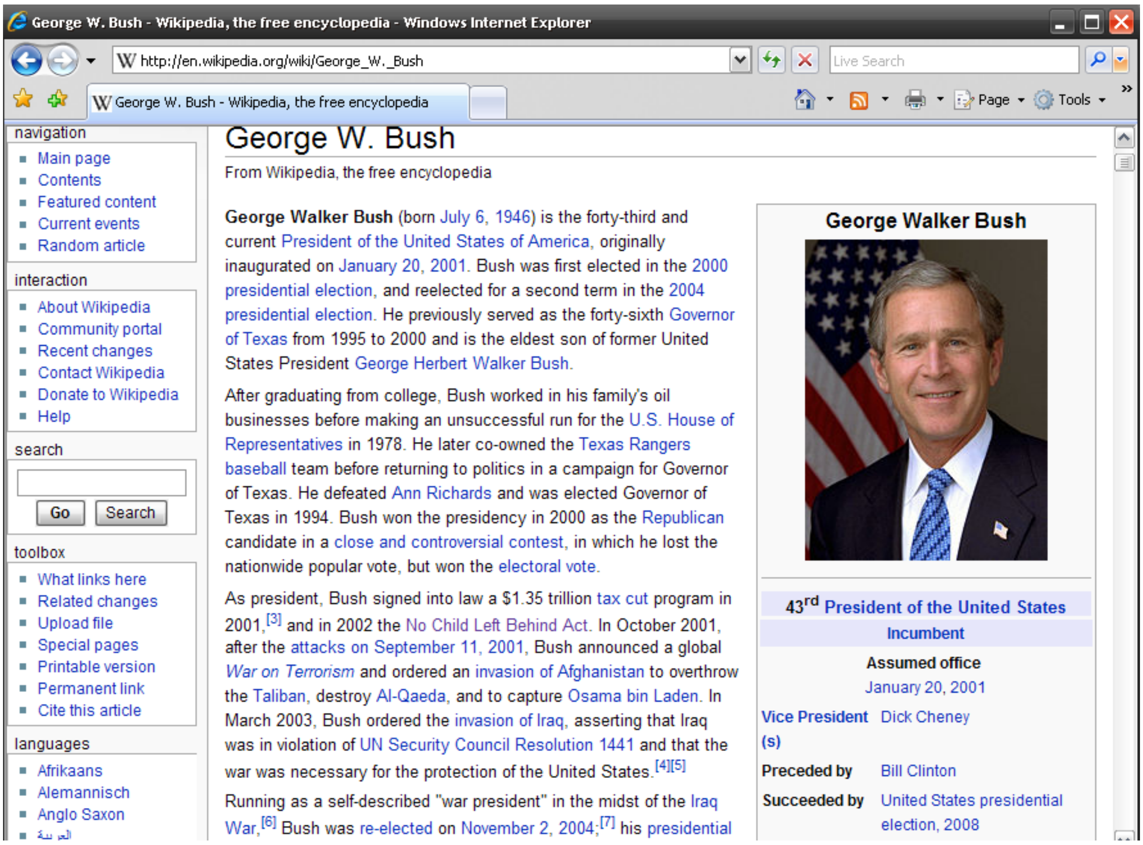
Xin B., Lin Z. 2004. The Impact of Search Engine Optimization on Online Advertising Market. In Proceedings of the 8th international Conference on Electronic Commerce: the New E-Commerce: innovations For Conquering Current Barriers, Obstacles and Limitations To Conducting Successful Business on the internet. August 13 - 16, 2006. Fredericton, New Brunswick, Canada. 519-529.

Yuwono B., Lee L. 1996. Search and Ranking Algorithms for Locating Resources on the World Wide Web. Proceedings of the Twelfth international Conference on Data Engineering. February 26 - March 01, 1996.164-171.

Zhang J., Dimitroff A. 2005. The impact of webpage content characteristics on webpage visibility in search engine results. Information Processing & Management. 41(3). 665-715.

Yahoo 2008. Yahoo! Search Content Quality Guidelines [online]. Yahoo! [Referred 27.1.2008]. Available at <http://help.yahoo.com/l/us/yahoo/search/basics/basics-18.html>

APPENDIX 1: EXAMPLES



George W. Bush

From Wikipedia, the free encyclopedia

George Walker Bush (born July 6, 1946) is the forty-third and current President of the United States of America, originally inaugurated on January 20, 2001. Bush was first elected in the 2000 presidential election, and reelected for a second term in the 2004 presidential election. He previously served as the forty-sixth Governor of Texas from 1995 to 2000 and is the eldest son of former United States President George Herbert Walker Bush.

After graduating from college, Bush worked in his family's oil businesses before making an unsuccessful run for the U.S. House of Representatives in 1978. He later co-owned the Texas Rangers baseball team before returning to politics in a campaign for Governor of Texas. He defeated Ann Richards and was elected Governor of Texas in 1994. Bush won the presidency in 2000 as the Republican candidate in a close and controversial contest, in which he lost the nationwide popular vote, but won the electoral vote.

As president, Bush signed into law a $1.35 trillion tax cut program in 2001,[3] and in 2002 the No Child Left Behind Act. In October 2001, after the attacks on September 11, 2001, Bush announced a global *War on Terrorism* and ordered an invasion of Afghanistan to overthrow the Taliban, destroy Al-Qaeda, and to capture Osama bin Laden. In March 2003, Bush ordered the invasion of Iraq, asserting that Iraq was in violation of UN Security Council Resolution 1441 and that the war was necessary for the protection of the United States.[4][5]

Running as a self-described "war president" in the midst of the Iraq War,[6] Bush was re-elected on November 2, 2004;[7] his presidential

George Walker Bush

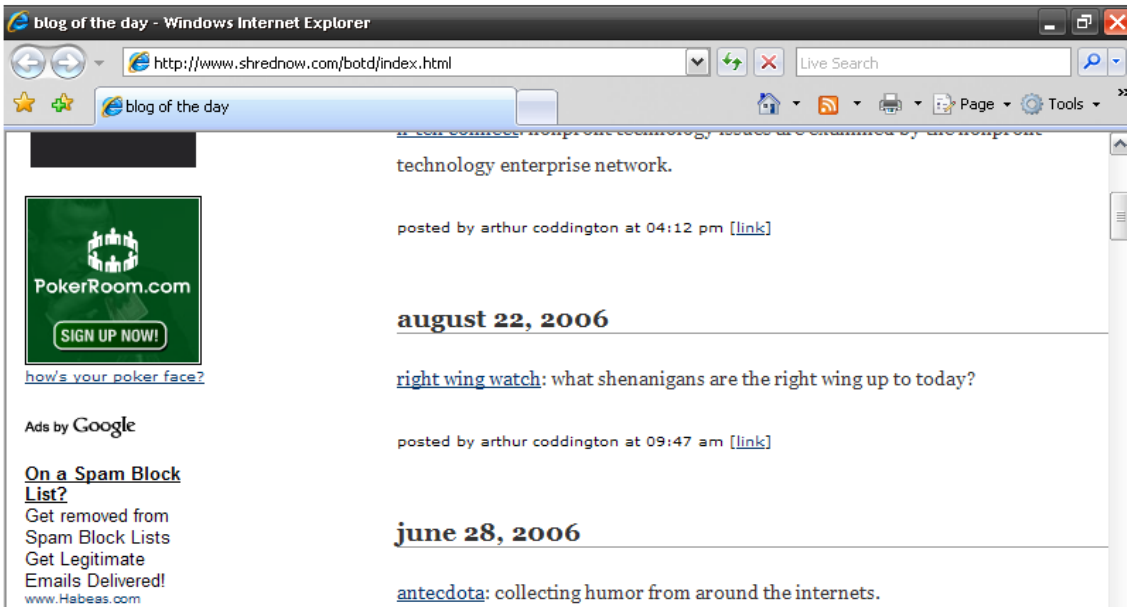
43rd President of the United States

Incumbent

Assumed office
January 20, 2001

Vice President(s)	Dick Cheney
Preceded by	Bill Clinton
Succeeded by	United States presidential election, 2008

EXAMPLE 1. The strongly interlinked Wikipedia. Already the three first paragraphs contain tens if not hundreds of links.


blog of the day - Windows Internet Explorer
http://www.shrednow.com/botd/index.html
Live Search
blog of the day
Page
Tools
PokerRoom.com
SIGN UP NOW!

how's your poker face?

Ads by Google

On a Spam Block List?
Get removed from Spam Block Lists Get Legitimate Emails Delivered!
www.Habeas.com

technology enterprise network.

posted by arthur coddington at 04:12 pm [link]

august 22, 2006

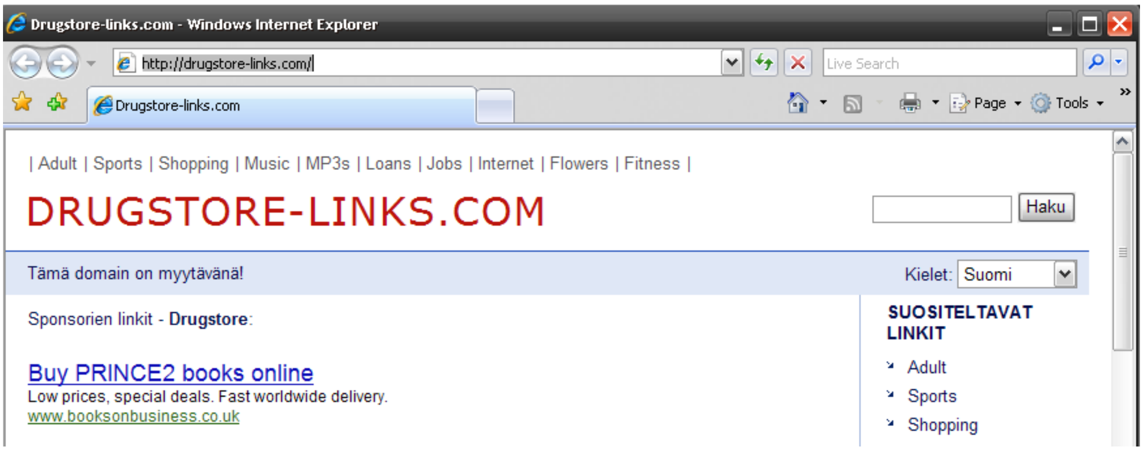
right wing watch: what shenanigans are the right wing up to today?

posted by arthur coddington at 09:47 am [link]

june 28, 2006

antecdota: collecting humor from around the internets.

EXAMPLE 2. A spam blog which is hosting ads


Drugstore-links.com - Windows Internet Explorer
http://drugstore-links.com/
Live Search
Drugstore-links.com
Page
Tools

| Adult | Sports | Shopping | Music | MP3s | Loans | Jobs | Internet | Flowers | Fitness |

DRUGSTORE-LINKS.COM

Haku

Tämä domain on myytävänä!

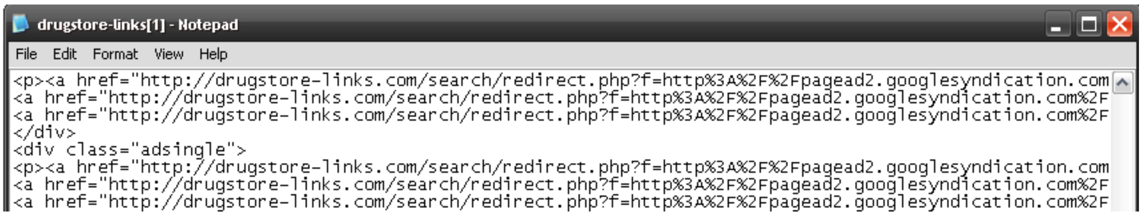
Kielet: Suomi

Sponsorien linkit - **Drugstore**:

Buy PRINCE2 books online
Low prices, special deals. Fast worldwide delivery.
www.booksonbusiness.co.uk

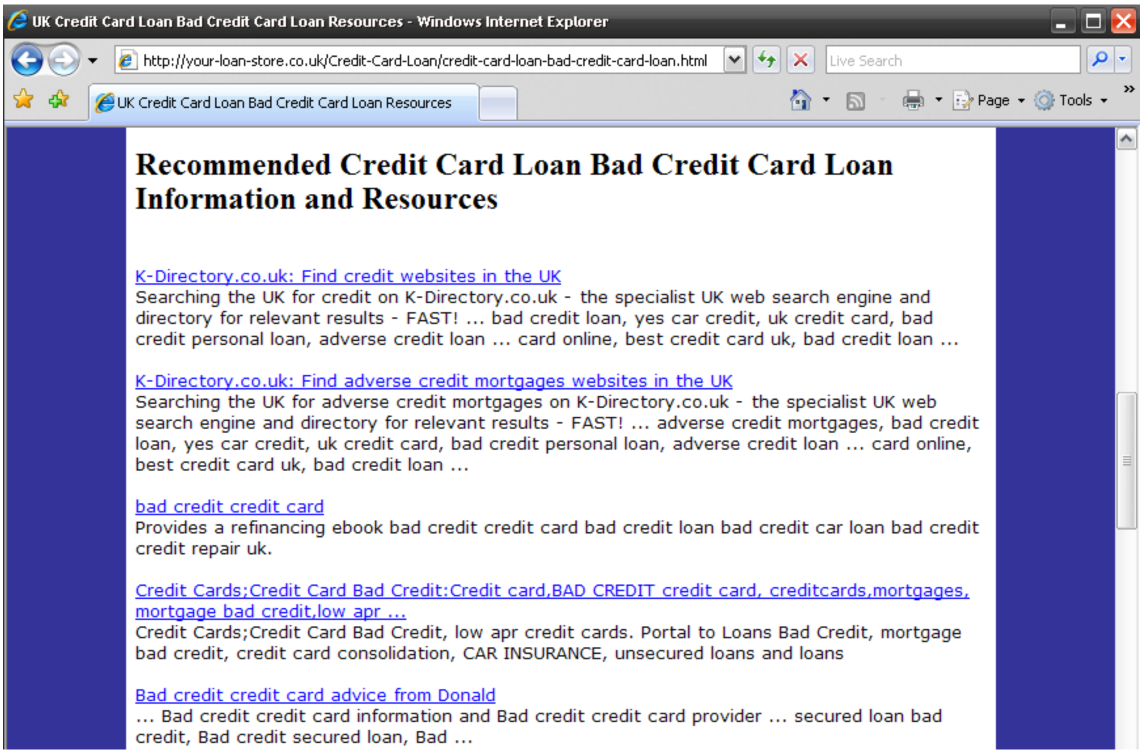
SUOSITELTAVAT LINKIT

- Adult
- Sports
- Shopping



```
<p><a href="http://drugstore-links.com/search/redirect.php?f=http%3A%2F%2Fpagead2.googlesyndication.com
<a href="http://drugstore-links.com/search/redirect.php?f=http%3A%2F%2Fpagead2.googlesyndication.com%2F
<a href="http://drugstore-links.com/search/redirect.php?f=http%3A%2F%2Fpagead2.googlesyndication.com%2F
</div>
<div class="adsingle">
<p><a href="http://drugstore-links.com/search/redirect.php?f=http%3A%2F%2Fpagead2.googlesyndication.com
<a href="http://drugstore-links.com/search/redirect.php?f=http%3A%2F%2Fpagead2.googlesyndication.com%2F
<a href="http://drugstore-links.com/search/redirect.php?f=http%3A%2F%2Fpagead2.googlesyndication.com%2F
```

EXAMPLE 3. Google Adsense Ads that are disguised as links


UK Credit Card Loan Bad Credit Card Loan Resources - Windows Internet Explorer
http://your-loan-store.co.uk/Credit-Card-Loan/credit-card-loan-bad-credit-card-loan.html
Live Search
UK Credit Card Loan Bad Credit Card Loan Resources
Page
Tools

Recommended Credit Card Loan Bad Credit Card Loan Information and Resources

K-Directory.co.uk: Find credit websites in the UK
Searching the UK for credit on K-Directory.co.uk - the specialist UK web search engine and directory for relevant results - FAST! ... bad credit loan, yes car credit, uk credit card, bad credit personal loan, adverse credit loan ... card online, best credit card uk, bad credit loan ...

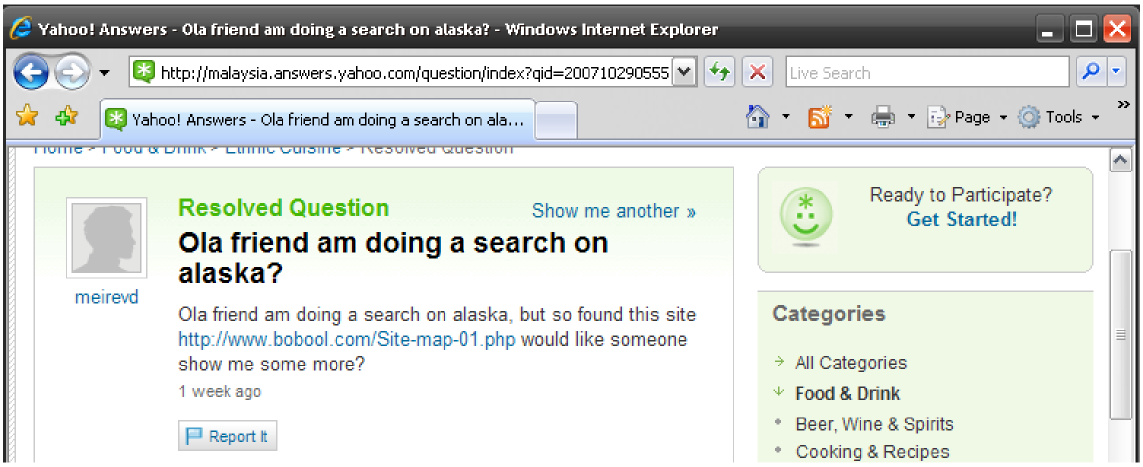
K-Directory.co.uk: Find adverse credit mortgages websites in the UK
Searching the UK for adverse credit mortgages on K-Directory.co.uk - the specialist UK web search engine and directory for relevant results - FAST! ... adverse credit mortgages, bad credit loan, yes car credit, uk credit card, bad credit personal loan, adverse credit loan ... card online, best credit card uk, bad credit loan ...

bad credit credit card
Provides a refinancing ebook bad credit credit card bad credit loan bad credit car loan bad credit credit repair uk.

Credit Cards;Credit Card Bad Credit:Credit card,BAD CREDIT credit card, creditcards,mortgages, mortgage bad credit,low apr ...
Credit Cards;Credit Card Bad Credit, low apr credit cards. Portal to Loans Bad Credit, mortgage bad credit, credit card consolidation, CAR INSURANCE, unsecured loans and loans

Bad credit credit card advice from Donald
... Bad credit credit card information and Bad credit credit card provider ... secured loan bad credit, Bad credit secured loan, Bad ...

EXAMPLE 4. Example of a spam page. The same site contains thousands of links.


Yahoo! Answers - Ola friend am doing a search on alaska? - Windows Internet Explorer
http://malaysia.answers.yahoo.com/question/index?qid=200710290555
Live Search
Yahoo! Answers - Ola friend am doing a search on ala...
Page
Tools
Home > Food & Drink > Ethnic Cuisine > Resolved Question
Show me another »
Ready to Participate?
Get Started!
Report It

Resolved Question

Ola friend am doing a search on alaska?

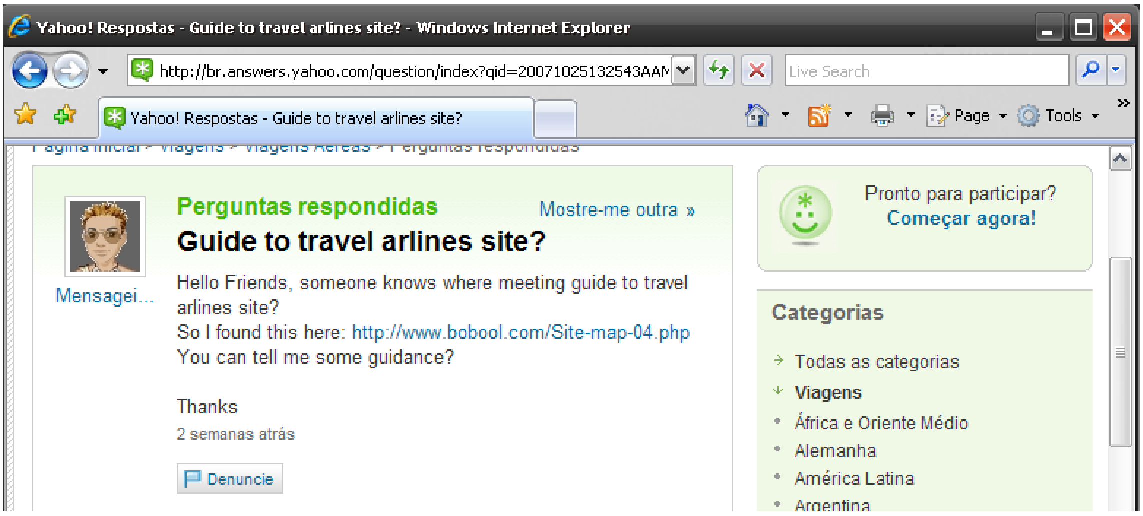
meirevd

Ola friend am doing a search on alaska, but so found this site http://www.bobool.com/Site-map-01.php would like someone show me some more?

1 week ago

Categories

- All Categories
- Food & Drink
- Beer, Wine & Spirits
- Cooking & Recipes


Yahoo! Respostas - Guide to travel arlines site? - Windows Internet Explorer
http://br.answers.yahoo.com/question/index?qid=20071025132543AAM
Live Search
Yahoo! Respostas - Guide to travel arlines site?
Page
Tools
Página inicial > Viagens > Viagens Aéreas > Perguntas respondidas
Mostre-me outra »
Pronto para participar?
Começar agora!
Denuncie

Perguntas respondidas

Guide to travel arlines site?

Mensagei...

Hello Friends, someone knows where meeting guide to travel arlines site?
So I found this here: http://www.bobool.com/Site-map-04.php
You can tell me some guidance?

Thanks

2 semanas atrás

Categorias

- Todas as categorias
- Viagens
- África e Oriente Médio
- Alemanha
- América Latina
- Argentina

EXAMPLE 5. Comment spam

APPENDIX 2: SUMMARY OF SPAM METHODS

	Target score type	**Method**	**Level of spammer's control**
Content spam	Relevancy	By choosing page keywords carefully spammers mislead search engine into believing that the page is about those topics	Full
URL spam	Relevancy	Some search engine algorithms also analyse URL composition. As URLs can be bought cheaply, spam words appear often in URLs.	Full
Honey pot	Authority	The spam page contains such information that others naturally link to it. The content is copied from other sources. Acquired authority score is forwarded to real targets	Partial
Infiltrate a Web directory	Authority	Add a spam page link into a Web directory	Partial
Spam blog	Authority	A blog on a blog site, which contains rubbish, or the content is copied from elsewhere. The purpose is to accumulate authority score or host profitable content	Full
Comment spam	Authority	Post spam URLs through a browser	Partial
Aquisition of expired domains	Authority	Spammer buys en expired domain, which is still linked elsewhere from the Web.	Partial
Link bombs	Relevancy / authority	Spammer links to the target with a misleading anchor text. Search engine algorithms then assign the page with the anchor text and so the page appears on these queries.	Full
Link farms: single target	Authority	A tight network of sites, which can span across thousands of pages. In this case there's only one target page, which most likely indicates that the farm is operated by a single spammer	Full
Link farms: alliances	Authority	A site structure, which is connected through a complex link structure. For example link exchange programs	Full
Cloaking	Hiding technique	Spam page to search engine crawler – descreet version to everybody else	Full
Redirection	Redirect traffic / hiding technique	Redirects the browser to other page. Crawler is served with the original version.	Full

APPENDIX 3: SUMMARY OF ANTI-SPAM METHODS

	Target	**Method**
Statistical analysis for page content	Content spam	Analyse number of repetitions, sentence structure, frequency of change, URL composition and so forth
PageRank	Content spam	One of the reasons for developing PageRank was to counter spam. The algorithm ignores content and emphasizes link structures
Statistical analysis for link structures	Link spam	Analyse both single pages (in and out degree) and link structures. Outliers in the distribution are likely to be spam
TrustRank	Link spam	Transfer trust onwards from a manually selected seed set. The basic premise is that good pages link to other good pages.
BadRank	Link spam	Works just the opposite of Trustrank. The algorithm transfers a distrust score from a spam page back to its' backlinkers
Intent	Link spam	Combine factors that measure page intent into the anti-spam algorithm. Pages with commercial intent are more likely to be spam.
NoFollow	Comment spam	An additional value to a-tag to indicate that this link has not been added by the page owner and should not be used in PageRank calculation
Detecting semantic cloaking	Cloaking	Two phase algorithm, where first phase filters out most of the valid pages and the second phase filters the rest. All that is left should be pages that use cloaking for suspicious purposes.
Query popularity and monetizability	Cloaking	An algorithm that starts from the assumption that popular and commercial searches contain more spam, i.e. this algorithm focuses on certain pages
SpamRank	Link spam	A three-phased algorithm that detects and penalizes sources of undedeserved PageRank scores