

# Both Sides of the Digital Battle for a High Rank from a Search Engine

Timothy Jones

Department of Computer Science

University of Otago

Dunedin

New Zealand

timothy.jones@stonebow.otago.ac.nz

## Abstract:

Because of the financial gain in achieving a high search engine rank, modifying a web page to unfairly alter its ranking has become a common practice. Techniques to achieve this are generally known as *web spam* because of the adverse affect on the relevance of the results returned by the search engine. This paper contains an overview of the techniques used to create web spam and the defenses available to search engines. In addition, we comment on why further research into combating web spam is required. This paper provides insights into the state-of-the-art of both sides of the battle to achieve an unfairly high search engine placement for a page.

**Keywords:** Adversarial information retrieval, Web spam, Spamdexing, Search engine spam, Information Retrieval

## 1. Introduction

Over the last few years, search engines have become the primary access point for many users of the web. The first two results returned by a search engine are followed far more than any other result, and users frequently only view the first page of results (Joachims, Granka *et al.* 2005). This means a high ranking within the results of a search engine is critical for good online business (Totty and Mangalindan 2003).

Because of the potential financial gain in achieving a high search engine ranking for a page, services such as ‘Search Engine Optimizers’ (SEOs) have become popular. Some of these services simply offer advice on structuring pages with relevant terms and headings. The major search engines accept this kind of behavior as legitimate (Google 2005). Unfortunately, some have ventured beyond this by attempting to ‘second guess’ the ranking functions, creating pages and/or content with no purpose other than to alter the rank of a target page. The search engines frown upon this because it affects their control over the relevance of the information returned. If advertising companies completely controlled the results of search engines it would become difficult to find information that was not connected with a product of some kind. Consequently, if a page exploits the ranking algorithms of a search engine, it may be struck from the index (Totty and Mangalindan 2003; Google 2005).

Unfairly altering the results of a search engine query has become known as *web spam* (Gröngyi and Garcia-Molina 2005; Joachims, Granka *et al.* 2005; Metaxas and DeStefano 2005), *search engine spam* (Wu and Davison 2005) and *spamdexing* (Gröngyi and Garcia-Molina 2005). Throughout this paper we shall use the term *web spam* or simply *spam*. This is not to be confused with email spam (which is unsolicited email).

The precise definition of web spam is contentious. Gröngyi *et al.* (2005) define web spam as “*any deliberate human action that is meant to trigger an unjustifiably favorable relevance or importance for some web page, considering the page’s true value*”. We prefer this definition because it is quite general and will cover the discovery of new techniques. Throughout this paper we shall also use the term *target page* or simply *target* to refer to the page whose ranking is being altered.

The rest of this paper is organized as follows: Section 2 is a short history of search engine ranking algorithms with respect to web spam. Section 3 describes the various techniques in a web spammer’s toolbox, while Section 4 describes the current state-of-the-art in defending against such techniques. Section 5 concludes the paper. This paper is intended to give the reader an understanding of the state-of-the-art of web spam creation and detection, as well as an insight into why further research into combating web spam is required.

## 2. History of Target Algorithms

### 2.1 TF.IDF

The first generation of search engines emerged when the web was gaining popularity in the early 1990s. Most of these used a ranking scheme called *TF.IDF* (Harman 1992). *TF.IDF* ranks documents by similarity to the query. Unfortunately this term-based ranking is especially vulnerable to manipulation by including many repetitions of expected keywords.

This may result in a high ranking, but often unfortunately also results in meaningless pages. In order to combat these ‘visit traps’ the search engines changed their ranking schemes.

## 2.2 The Link Voting Principle

In 1996, Lycos pioneered the first ‘implicit voting’ scheme. This scheme treated links to a site as a ‘vote’ for that site. Sites with many votes would be ranked highly. Realizing this, spammers created *link farms* (large networks of interlinked pages) to artificially boost a site’s popularity. Fortunately, the cost of maintaining these link farms was prohibitive in 1996.

## 2.3 PageRank

In 1998, Page *et al.* (1998) created PageRank. PageRank is a link-voting scheme that is superior to previous web ranking functions because the ‘reputation’ of a page contributes to the power of its link vote. A page’s reputation can be thought of as the probability that a user would come across that page by following a random path across the web. Each page begins with a default reputation which represents the probability of jumping to that location without a hyperlink (sometimes called the ‘teleportation’ value). Additional reputation comes in to a page through incoming links, and flows out through the outgoing links. Thus, a page with many incoming links passes on a large reputation to pages it links to. Pages are ranked by their final reputation score.

Because of the flow-through nature of PageRank, isolated interconnected link farms do not gain reputation (because reputation flows in and out equally). However, there are designs for link farms that can be used against PageRank (discussed in Section 3).

## 2.4 HITS

HITS (Klienberg 1999) was introduced in 1999 to find authoritative pages on a particular topic, but it is often applied to the whole web. HITS assigns ‘hub’ and ‘authority’ scores to each page. A *hub* is a page that links to many authorities, and an *authority* is a page that links to many hubs. Sites often linked to (such as *cnn.com*) are likely to have high authority scores, while sites with often linked from (such as *slashdot.org*) are likely to have high hub scores. Pages are ranked by a combination of their hub and authority scores. Spammers can increase hub scores on a link farm by linking to many known authorities, and then link their (now important) hubs to a target page to make it an authority. As a link based scheme, HITS is also vulnerable to well designed link farms. A link farm optimized for HITS is not optimized for PageRank, and vice-versa. HITS and PageRank are the main targets for today’s web spammers.

## 3. The Spammer’s Toolbox

In order to fully understand the techniques for defeating web spam, we must first understand the techniques used to create it. Almost all spam techniques attack one of the algorithms outlined in Section 2. An extensive list of examples is outlined in Gröngyi *et al.* (2005), but we include several here for completeness. For further information see their excellent paper. We will use their categorization of *boosting techniques* for web spam techniques designed to increase ranking, and *hiding techniques* designed to hide the use of boosting from the user.

There are three classes of web page relevant to the web spammer: *Inaccessible* pages are pages the spammer has no control over, and represent the majority of the web. *Accessible* pages are the pages a spammer can add information to, but not control (such as blog comments). The third category is the spammer’s *own* pages – the pages that the spammer can completely control. These are often employed in some sort of link farm, and (with today’s hosting prices), can be high in number. The target page is assumed to belong to this third category.

### 3.1 Boosting Techniques

- *Body Spam*: Adding extra terms to the body of a page in order to boost the target’s TF.IDF score. Usually these are not arranged in a meaningful order, eg: “resort ski holiday tickets camping”.
- *Title Spam*: Some search engines give a higher weight to terms in the title of a page. Adding extra terms to the title can also increase the TF.IDF score.
- *Anchor Text Spam*: Search engines often index the text of outgoing links as part of the destination page, because the link often describes the destination page (for example, a link pointing to a casino page might contain the text “Las Vegas, Gambling, Millionaire”). Using anchor text spam does not require altering the target page. In fact, it can be done without the cooperation of the target page owner. Using this technique without cooperation (usually to rank the page highly for rude or humorously unrelated queries) is called *Google Bombing* and has recently received media attention (BBC News 2005). While anchor text spam mainly targets TF.IDF-like ranking schemes, it is often seen with *link spam* (discussed below).

- *URL Spam*: Search engines often index the URL of a page as part of the page text. Registering many domain names can be expensive, so some spammers will add spam terms to the hostname, relying on a customized DNS server to resolve any hostname within their domain to their page. Again, this targets TF.IDF-like schemes, and is also often seen with link spam.
- *Link Spam*: Link spam is one of the most common spamming techniques. This is likely due to the popularity of algorithms like PageRank and HITS. Link spam is the creation of hyperlinks for no purpose other than to alter the rank of a page. These links need not necessarily be added to the target page itself; they may exist solely as link farms. Link spam usually works by adding incoming links to a page solely to alter its PageRank reputation, or increase its HITS 'authority'. There are many different varieties of link spam, including:
  - *Outgoing Link Spam*: An unusual application of link spam is to add a large number of outgoing links from the target page to well-known pages. This increases the HITS hub score of the target. This can be done quickly and easily by copying any one of many online directories (such as Yahoo!). This technique will often decrease the target page's PageRank score (because reputation will be flowing out of the target page).
  - *Incoming Link Honey Pot*: Some spammers clone useful information (possibly also a web directory) in the hope that users of the web will stumble across their page and link to it. This is a useful technique to fool owners of otherwise inaccessible pages to link to the target page.
  - *Crawling to Post Links (Comment Spam)*: Spammers often crawl and post links to pages that are only 'accessible' to them, such as blogs or guest books. Many of these accessible pages are not moderated, and even if they are the link can be cleverly hidden. Some guest books allow users to give a home URL, which could easily be spammed. To avoid detection the message body may be something generic like "Great page – thanks!"
  - *Link Farms*: While the introduction of PageRank curbed the effectiveness of isolated link farms, it is relatively easy to build a link farm that is not isolated (using comment spam or content cloning). There are a number of different link farm strategies, many of which are under academic investigation. It has been shown that link farms provide little improvement of the PageRank of an already highly ranked page (Baeza-Yates, Castillo *et al.* 2005). This means the use of a link farm may not move a page that is already in the top few results, but it doesn't stop link farms from promoting a page from near the bottom of the list. The monetary cost of manipulating PageRank with link farms has also been investigated (Clausen 2004).
  - *Expired Domain Purchasing*: When a domain expires, the content previously stored there immediately disappears, but links to that domain do not. Expired domains with a high PageRank or HITS score can be purchased and used in a link farm or as target page locations.

### 3.2 Hiding Techniques

Techniques that directly alter a target page are usually hidden from regular users as they decrease the credibility of the page by reducing the attractiveness (Bailey, Gurak *et al.* 2001). Consequently, techniques for hiding the alterations have been invented.

- *HTML or web scripting trickery*: Techniques for preventing browsers from rendering links or paragraphs of text fall in this category. These range from simple HTML illusions (white text on a white background) to scripts to set the 'visible' attribute of page elements to false upon page load. These techniques are often visible in the source of the web page.
- *Cloaking*: The practice of serving different content to a search engine crawler than a legitimate visitor. This can be achieved either by detecting the 'user agent' supplied to the web server (Google's spider currently identifies itself as GoogleBot/2.1) or by detecting that the request came from an IP address known to belong to a search spider. Cloaking can be done legitimately in order to assist the spider (for example providing a page free of formatting and images in order to reduce file size) or illegitimately (for example serving a completely different page to the spider, perhaps laden with anticipated query terms only). Wu and Davison (2005) discuss the difficulty defining and identifying legitimate cloaking.
- *Redirection*: Hiding the spam content of a page by instructing the browser to redirect to an alternative page before the page has finished loading. In the simplest case this is achieved with a meta 'refresh' tag. More sophisticated approaches include redirects in javascript or some other scripting language, as search spiders do not currently process scripts.

## 4. Combating Web Spam

Because of the recent uprising of web spam, several researchers have turned their attention to identifying it. Most of the current research focuses on defeating techniques that affect PageRank and HITS. This is probably because these algorithms form the basis for most currently popular search engines.

### 4.1 Statistical Detection

Fetterly *et al.* (2004) provide a list of attributes that are often present in spam pages. This list was generated by an extensive statistical analysis of two large web data sets (several million pages each). Their extensive analysis found the following were good indicators of a spam page:

- Large numbers of dots, dashes and digits in the URL.
- Large numbers of hostnames resolving to a single IP address.
- Disproportionately high ratio of incoming to outgoing links to a page.
- Large sets of pages with little variance in content.
- Content that appears radically different each time it is downloaded, indicating dynamic generation irrespective of page requested.

These results were used to generate the overall probability that a given page is spam. They estimate that they detect half the spam pages in their collection, with a 14% false positive rate. If detection by statistics gains popularity, we believe some of these predictors (such as URL construction) will become obsolete as spammers adapt.

### 4.2 Graph Based Detection

Wu and Davison (2005) suggest a technique to detect and penalize link farms. Their technique identifies areas of high interconnectedness in the web graph and reduces the weighting of link votes within them. They note that their technique is vulnerable to link farms with a large number of duplicate nodes all pointing to a target (link farms with no interconnection). This approach can also penalize some legitimate pages with high of interconnectedness (such as news networks). They suggest the use of a whitelist to accommodate such sites.

Gröngyi *et al.* (2004) suggest an opposite approach designed to identify trustworthy pages. They first ask humans to identify a small set of trustworthy pages, and then rank pages according to their distance from the trusted page, treating 'trust' like reputation in PageRank.

Metaxas and DeStefano (2005) liken web spam to social propaganda, and use techniques from social science (usually used to detect propaganda). Their algorithm requires a human to find an untrustworthy seed page. Then it traverses the incoming links to a page in order to discover which pages strongly support it. These pages are also marked as untrustworthy.

### 4.3 Comment Spam Detection

Popular blog sites attempt to stop comment spam at its source by preventing it from being posted. Such prevention techniques include requiring user registration, blacklist filtering (Allen 2005) and preventing HTML comments completely. These strategies are only temporary fixes, and often reduce the page's functionality.

Mishne *et al.* (2005) propose identifying comment spam by comparing the language models in the commented page and the comment itself. A language model models the probability of words occurring in a given text. Given the language model of the original page, the probability that the comment was created using the same language model can be calculated. If this probability is low, it is unlikely that the comment is related to the page. Their results look promising, though their test collection is small. They note that most of the incorrect classifications occur with very short comments. As a potential solution to this they suggest appending the linked page to the comment before language model generation.

### 4.4 Detecting Cloaking and Redirection

Cloaking detection is difficult. Even if we know that a page changes between loads it still may not be spam (it may be a dynamic page). Wu and Davison (2005) compare the pages returned by four separate crawls (two reporting to the server as a common web browser, two reporting as a search spider) using a thresholded difference calculation that takes into account unique term and link differences between the pages. Their algorithm seems quite effective, although the authors do note the difficulty in separating malicious and acceptable cloaking once detected.

They also examine the types of redirect present in their dataset. More pages in the browser dataset were redirected than pages in the crawler dataset, suggesting some pages employ cloaking to hide redirections. The authors conjecture that pages with an immediate redirection are more likely to be spam than pages with a time delayed one. Again, separating

malicious and innocuous redirection is problematic – some pages detect the browser type and then redirect harmlessly to a browser optimized page.

The authors outline future work in the area of detection of cloaking and redirection, particularly the observation that there may be sites that use IP based cloaking that would not have been picked up in their tests. Their study is a useful insight into the difficulties in the area.

#### 4.5 Secrecy

Search engines do not disclose the full details of their ranking algorithms, in order to protect their business and to prevent easy exploitation. Security by secrecy is a useful way to slow the progress of attack, but it is not a solution. Reverse engineering can be used on the results of a search engine; not only by the spammers, but also as a topic of academic interest (Bifet, Castillo *et al.* 2005). Because of the relative ease of black box reverse engineering, hiding a ranking algorithm by non-disclosure is a time limited defense.

### 5. Conclusion

Because of the popularity of link voting algorithms such as PageRank and HITS, most state of the art techniques for web spamming rely on link spamming, usually in the form of a link farm. Term spamming techniques are also seen, but these techniques only fine tune the ranking of a page instead of dramatically altering it.

For a link farm to function it must remain undetected, and it must retain incoming links. Therefore, defense against web spam has focused on the detection of link farms (so they can be ignored or penalized), or the detection (for removal) of comment spam. There is current research into the detection of cloaking and redirection, but this has proven problematic due to the difficulty of classifying cloaking and redirection.

The current defenses may fix many of today's problems, but (as the authors of the techniques often note) they are not invulnerable to further manipulation. Even a combined solution is unlikely to provide a permanent fix. The development of new defense techniques will likely encourage the development of new offense techniques, creating a cycle or a kind of spam 'arms race'. Much like email spam, it is unlikely that a perfect solution will ever be found.

Combating web spam is important because of the potential damage to the web if it were to remain unchecked. In the extreme case, all of the top results from a search engine could be altered. If a malicious party were able to completely manipulate the first page of results so that it contained a mixture of only credible resources for an idea or non-credible resources against the idea, users could be lead to a specific conclusion about a specific topic. This is particularly dangerous as typically users do not crosscheck information found on the internet (Graham and Metaxas 2003). The credible resources need not even be well-known – their credibility could be manufactured (Bailey, Gurak *et al.* 2001). It is clear that web spam must be kept under control.

There exist no standard test collections for web spam; individual authors gather their own. This makes it difficult to compare the performance of different algorithms. Such a set of test collections is required. An argument against such collections is that they may not represent the current state of the art of web spam. An argument for standardized test collections is that without them algorithms for spam detection cannot be accurately compared; even with the same test set gathering technique, web spam may have evolved between the testing of the techniques compared. Our view is that standardized test collections will enable more accurate comparison of detection algorithms, but frequent additions of new collections are required to keep up with web spam evolution.

We have provided an insight into both sides of the battle to achieve a high search engine ranking. The interested reader is referred to the recent *1<sup>st</sup> International Workshop on Adversarial Information Retrieval on the Web* (Davison 2005) as a starting point for further research.

### 6. Acknowledgements

Thanks to the numerous proofreaders who kept this paper on topic. Special thanks to Andrew Trotman for his constructive advice and contagious enthusiasm for Information Retrieval.

### 7. References

- Allen, J. (2005) *MT-Blacklist - A Movable Type Anti-spam Plugin*. retrieved September 2005, from <http://www.jayallen.org/projects/mt-blacklist/>.
- Baeza-Yates, R., C. Castillo, et al. (2005). PageRank Increase under Different Collusion Topologies. *Proceedings of the 1st International Workshop on Adversarial Information Retrieval on the Web (AIRWeb)*.
- Bailey, B., L. Gurak, et al. (2001). An Examination of Trust Production in Computer-Mediated Exchange. *Proceedings of Human Factors and the Web*.

- BBCNews. (2005). *'Miserable Faliure' links to Bush*. retrieved August 2005, from <http://news.bbc.co.uk/2/hi/americas/3298443.stm>.
- Bifet, B., C. Castillo, et al. (2005). An Analysis of Factors Used in Search Engine Ranking. *Proceedings of the 1st International Workshop on Adversarial Information Retrieval on the Web (AIRWeb)*.
- Clausen, A. (2004). The Cost of Attack of PageRank. *Proceedings of the International Conference on Agents, Web Technologies, and Internet Commerce (IAWTIC)*, Gold Coast, Australia.
- Davison, B. D. (2005). *AIRWeb '05: First International Workshop on Adversarial Information Retrieval on the Web*. retrieved September 2005, from <http://airweb.cse.lehigh.edu/>.
- Fetterly, D., M. Manasse, et al. (2004). Spam, damn spam, and statistics: using statistical analysis to locate spam web pages. *Proceedings of the 7th International Workshop on the Web and Databases (WebDB '04)*
- Google. (2005). *Google Information for Webmasters*. retrieved August 2005, from <http://www.google.com/webmasters/seo.html>.
- Graham, L. and P. T. Metaxas (2003). "Of course it's true, I saw it on the Internet!" Critical thinking in the internet era. *Communications of the ACM* 46(5), 70-75.
- Gröngyi, Z. and H. Garcia-Molina (2005). Web Spam Taxonomy. *Proceedings of the 1st International Workshop on Adversarial Information Retrieval on the Web (AIRWeb)*.
- Gröngyi, Z., H. Garcia-Molina, et al. (2004). Combating Web Spam With TrustRank. *Proceedings of the 30th International Conference on Very Large Data Bases (VLDB '04)*.
- Harman, D. (1992). Ranking Algorithms.in *Information Retrieval Data Structures & Algorithms*. W. B. Frakes and R. Baeza-Yates.
- Joachims, T., L. Granka, et al. (2005). Accurately Interpreting Clickthrough Data as Implicit Feedback. *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*.
- Klienbergl, J. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM* 46(5), 604-632.
- Metaxas, P. T. and J. DeStefano (2005). Web Spam, Propaganda and Trust. *Proceedings of the 1st International Workshop on Adversarial Information Retrieval on the Web (AIRWeb)*.
- Mishne, G., D. Carmel, et al. (2005). Blocking Blog Spam with Language Model Disagreement. *Proceedings of the 1st International Workshop on Adversarial Information Retrieval on the Web (AIRWeb)*.
- Page, L., S. Brin, et al. (1998). *The PageRank Citation Ranking: Bringing Order to the Web*. Technical report, Stanford University.
- Totty, M. and M. Mangalindan (2003). *Cat and Mouse: As Google becomes web's gatekeeper, sites fight to get in*. Wall Street Journal. CCXLI.
- Wu, B. and B. Davison (2005). Cloaking and Redirection: A Preliminary Study. *Proceedings of the 1st International Workshop on Adversarial Information Retrieval on the Web (AIRWeb)*.
- Wu, B. and B. Davison (2005). Identifying link farm spam pages. *Proceedings of the 14th International WWW Conference 2005*.