GOOGLE SEARCH ALGORITHM UPDATES AGAINST WEB SPAM

Ashish Chandra, Mohammad Suaib, and Dr. Rizwan Beg

Department of Computer Science & Engineering, Integral University, Lucknow, India

ABSTRACT

With the search engines' increasing importance in people's life, there are more and more attempts to illegitimately influence page ranking by means of web spam. Web spam detection is becoming a major challenge for internet search providers. The Web contains a huge number of profit-seeking ventures that are attracted by the prospect of reaching millions of users at a very low cost. There is an economic incentive for manipulating search engine's listings by creating otherwise useless pages that score high ranking in the search results. Such manipulation is widespread in the industry. There is a large gray area between the Ethical SEO and the Unethical SEO i.e. spam industry. SEO services range from making web pages indexed, to the creation of millions of fake web pages to deceive search engine ranking algorithms. Today's search engines need to adapt their ranking algorithms continuously to mitigate the effect of spamming tactics on their search results. Search engine companies keep their search ranking algorithms and ranking features secret to protect their ranking system from gaming by spam tactics. We tried to collect here all the Google's search algorithm updates from different sources which are against web spam.

KEYWORDS

Google, web spam, spam, search engine, search engine ranking, spamdexing, search algorithm updates, Google Panda, Google Penguin.

1. INTRODUCTION

The Internet has become a major channel for people to get information, run business, connecting with each other and for the purpose of entertainment and education. Search Engines are the preferred gateway for the web for recent years.

Web spam is one of the major challenges for search engine results. Web spam (also known as spamdexing) is a collection of techniques used for the sole purpose of getting undeserved boosted ranking in search result pages. With the widespread user generated content in web 2.0 sites (like blogs, forums, social media, video sharing sites etc.), spam is rapidly increasing and becoming a medium of scams and malware also.

When a web user submits a query to search engine, relevant web pages are retrieved. The search engine ranks the result on the basis of relevancy (i.e. Dynamic ranking) and authority score (i. e. Static ranking) of the page. For this purpose it uses the page content information, link structure of the page, temporal features, usage data (i.e. Wisdom of Crowd) etc [1]. After this process search engine sorts the list of these pages according to the score thus calculated and returns the result to user.

DOI : 10.5121/ieij.2015.3101

Traditional Information Retrieval methods assumes IR system as a controlled collection of information in which the authors of the documents being indexed and retrieved had no knowledge of the IR system and no intention of manipulating it. But in case of Web-IR, these assumptions are no longer valid. Almost every IR algorithm is prone to manipulation in its pure form.

A ranking system which is purely based on the vector space model can easily be manipulated by inserting many keywords in the document, whereas a ranking system purely based on counting citations can be manipulated by creating many fake pages pointing to a target page, and so on. Detecting spam is a challenging web mining task. The search engine companies always try to stay ahead of the spammers in terms of the ranking algorithms and their spam detection methods. Fortunately, from the point of view of the search engines, the target is just to adjust the economic balance for the prospective spammers, and not necessarily detecting 100% of the web spam. The web spam is essentially an economical phenomenon where the amount of spam depends on the efficiency and the cost of different spam generating techniques. If the search engine can maintain the costs for the spammers consistently above their expected gain from manipulating the ranking, it can keep web spam at low level.

Many existing heuristics for web spam detection are generally specific to a specific type of web spam and cannot be used if a new spamming technique appears. Due to the enormous business opportunities brought by popular web pages, many spam tactics have been used to affect the search engine ranking. New spam tactics emerge time to time, and spammers use different tricks for different types of pages. These tricks varies from violating recommended practices (such as keyword stuffing [2], cloaking and redirection [3] etc) to violating laws (such as compromising web sites to poison search results [4], [5] etc.). After a heuristic for web spam detection is developed, the bubble of Web visibility tends to resurface somewhere else. We are in need to develop models that are able to learn to detect any type of web spam and that are able to be adapted quickly to new unknown spam techniques. Machine learning methods are the key to achieve this goal. It will be not wrong to say that web spam is the greatest threat to modern search engines.

1.1 Term and Definitions

We have listed below some key terms and definitions in context with the topic for better understanding of this paper.

1.1.1 Spamming

Spamming refers to any deliberate action which is performed for the sole purpose of boosting page's position in search engine result.

1.1.2 Keyword Stuffing

Keyword stuffing refers to loading of a page with keywords (excessive repetition of some words or phrases) for boosting page's ranking in search engine result. It makes the text appearing as unnatural.

1.1.3 Link Farm

Link Farm refers to excessive link exchanges, large scale links creation campaign, buying and selling links, link creation using automated programs etc. just for the sake of increased PageRank.

1.1.4 Doorway Pages

Doorway Pages are typically a large collection of low quality content pages where each page is optimized to rank for a specific keyword. These pages ultimately drive users to a specific target page by funnelling the traffic.

1.1.5 Cloaking

Cloaking is a search engine optimization technique in which the search engine crawler is served different copy of the page than that served to the normal user's web browser. Cloaking is a form of Doorway Page technique. It can be achieved by malicious redirect of page.

1.1.6 Indexing

Indexing refers to collecting, parsing, storing content of web pages for fast and accurate accessing of information at searching time.

1.1.7 Search Engine Ranking

Search engines rank web pages according to two main features. (i) Relevancy of page with respect to query (Dynamic Ranking). (ii) Authoritativeness of the page (Static Ranking).

Dynamic Ranking is calculated at search time and depends on search query, user's location, location of page, day, time, query history etc.

Static Ranking uses hundreds of query independent features of the page like length of the page, frequency of keywords, number of images, compression ratio of text etc. It is pre-computed at the time of indexing [6].

1.2 Related Work

There are many surveys done on web spam and detection methods [1], [6]. Many modern techniques of spamming as analyzed by authors in their researches in [1], [2], [3], [4], [5], [6]. In our knowledge there is no scholarly paper which covers actual implementation of spam detection algorithms by today's search engines like Google. We are presenting this paper to fill this gap.

1.3 Structure of the Paper:

We have divided this paper in four sections. In the section 2, we have enumerated all important updates released by Google which are concerned with web spam detection and filtering from search results. In section 3 we have analyzed the findings of the paper. The section 4 contains the conclusion of the paper.

2. GOOGLE ALGORITHM UPDATES

2.1 PageRank

The Google's cofounder Larry Page invented a system known as Page Rank [7]. This system works on link structure of web pages to decide ranking of web pages.

PageRank counts the number and quality of links to a page to calculate a rough estimate of a website's global importance. It can be assumed that important websites are more likely to receive

high number of links from other websites. Initially Google's search engine was based on Page Rank and signals like title of page, anchor text and links etc.

PageRank is calculated as:

$$PR(p_i) = \frac{1-d}{N} + d \sum_{p_j \in M(p_i)} \frac{PR(p_j)}{L(p_j)}$$

where: $p_1, p_2, \ldots, p_N$ are the pages under consideration,

$M(p_i)$ is the set of pages that link to p_i,

$L(p_j)$ is the number of outbound links on page p_j,

N is the total number of pages.

Currently Google search engine uses more than 200 signals for ranking of web pages as well as to combat web spam. Google also uses the huge amount of usage data (consisting of query logs, browser logs, ad-click logs etc.) to interpret complex intent of cryptic queries and to provide relevant results to end user.

2.2 Google ToolBar

Google launched its Toolbar for the Internet Explorer web browser in year 2000 with a concept of ToolBar PageRank (TBPR).

In the year 2001 , Google's Core Search Quality Engineer, Amit Singhal revised Page and Brin's original algorithm completely by adding new signals. One of these signals is commercial or non-commercial nature of the page. Google Engineer Krishna Bharat [8], studied that links from recognized authorities should carry more weight and discovered a powerful signal that confers extra credibility to references from experts' sites.

In the year 2002 search engine giant Google gave a clear message to SEO (Search Engine Optimization) industry that Google do not require Search Engine Optimization at all. It started penalizing sites which try to manipulate Page-Rank and started rewarding informative non-commercial websites.

2.3 Boston

The Boston update was announced at SES-Boston in year 2003. It incorporated local connectivity analysis and it gave more weight to authoritative sites in indexing as well as on the search result page. [9]

2.4 Cassandra

The Cassandra update was launched in April 2003 to combat against basic link quality issues (such as massive linking farms, cross-linking by co-owned domains, multiple links from same site) and by taking into account factors like link text, navigation structure, page title, hidden text and hidden links. [10]

2.5 Dominic

The Dominic update was launched in May 2003. This update was related with basic link calculations.

2.6 Esmereldo

The Essmereldo update was an infrastructure change followed by Fritz update which enabled Google search engine to update its index daily rather than existing monthly complete overhauling of index.

2.7 Florida

The Florida update was released in November, 2003. This update severely hit the low value SEO tactics like keyword stuffing by adding new factors to calculate search ranking. some of the factors are:

- repetitive in-bound anchor text with little diversity,
- heavy repetition of keyword phrases in title and body,
- lack of related/supportive vocabulary in the page.

2.8 Austin

The Austin [11] update was launched in January 2004. This update impacted on-page spam techniques like invisible text and links, META tag stuffing (abnormally long META tag of HTML), link exchange with off-topic sites. It is speculated that this update included **Hilltop** [12] algorithm.

The Hilltop algorithm is a topic sensitive approach to find documents relevant to the specific keyword or topic. When a user enters a query or a keyword into the Google search, this algorithm tries to find relevant keywords whose results are more informative about that query or keyword.

During this update Google started giving more value to the restricted top level domain websites such as educational (.edu, .ac), military (.mil) and Government websites (.gov) [13] because it is very difficult for spammers to have controlling access to these websites.

2.9 Brandy

The Brandy update was launched in February 2004. This algorithm update was a massive index update to add a lot more authoritative sites. This update incorporated Latent Semantic Indexing (LSI) to add capability of understanding synonyms for enhanced keyword analysis. LSI is based on the assumption that the words that are used in the same contexts are likely to have similar meanings.

This update increased attention to anchor text relevance and added concept of **link neighbourhood.** Link Neighbourhood refers to who is linking to your site. Links must be from relevant topic sites.

The Brandy update also added new factors of content and link quality and slightly reduced importance of PageRank. It used outbound links to calculate authority of the page. This feature is similar to the hub score in HITS algorithm [14].

According to Surgey Brin (Cofounder of Google), over-optimized use of title, h1, h2, bold, italic are no longer important features for ranking.

2.10 NoFollow

The "nofollow" [15] is a value of rel attribute of href tag in HTML. This value was proposed in January 2005 collectively by three top search provider companies of world (Google, Yahoo and Microsoft) to combat spam in blog comments.

An example of nofollow link is as following:

<a href="http://msn.com/about.html" rel="nofollow">

If Google sees 'nofollow' in a link then it will:

- not follow through to the target page,
- not count the link for calculating Page Rank,
- not consider the anchor text in determining the term relevancy of target page.

2.11 Jagger

The Jagger update was released in October 2005. The Jagger update targeted low quality links such as link farms, paid links, reciprocal links etc.

2.12 Vince

The Vince update was launched in February 2009. Vince strongly favors big brands as they are trusted sources.

2.13 May Day

The May Day update was released in May 2010. This update targeted sites with thin content optimized for long tail keywords.

2.14 Caffeine

The Caffeine update was launched in June 2010. This update was an infrastructural change. It was to speed up searching and crawling which resulted 50% fresher index. It revamped entire indexing system to make it more easier to add new signals such as heavier keyword weighting and importance of domain age.

2.15 Negative Review

The Negative Review update was launched in December, 2010. The Negative Review update targeted sites ranking due to negative reviews. It incorporated sentiment analysis.

2.16 Social Signals

The Social Signals update was launched in December 2010. This update added social signals to determine ranking of pages. Social Signals include data from social networking sites like Twitter and FaceBook. This update added concept like Social Rank or Author Rank.

According to Google CEO Eric Schmidt, in the year 2010, 516 updates were made in ranking system.

2.17 Attribution

The Attribution update was released in January, 2011. This update impacted duplicate/scrapped content websites by penalizing only copying site and not the original content websites. According to the Google's web spam engineer Matt Cutts [16]: "The net effect is that searchers are more likely to see the sites that wrote the original content rather than a site that scrapped or copied the original site's content".

2.18 Panda

The Panda algorithm update was first release in February 2011. The Panda update went global in April, 2011. This update aimed to lower rank of low quality websites and increased ranking of news and social networking sites. Panda is the filter to down rank sites with thin content, content farms, doorway pages, affiliates websites, sites with high ads-to-content ratio and number of other quality issues [17].

Panda update affects ranking of entire website rather than individual page. It includes new signals like data about the site users blocked via search engine result page directly or via the chrome browser [18]. Panda has improved scrapper detection. Its algorithm requires huge computing power to analyze pages so it is run periodically. The latest version of Panda (4.1) was released in September 2014. Panda update is named on Google's Search Engineer Navneet Panda with a patent named on him. [19].

2.19 Ad-above-the-fold

The Ads-above-the-fold algorithm update was released in January 2012. This update was to devaluate sites with too much advertisements. This is an improvement to page layout algorithm.

2.20 Penguin

The Penguin update was released in April 2012. This update is purely web spam algorithm update [20]. It adjusts a number of spam factors including keyword stuffing, in-links coming from spam pages, anchor text/link relevance. Penguin detects over optimization of tags and internal links, bad neighborhood, bad ownership etc. The latest version of Penguin (3.0) was released in October 2014.

2.21 Exact Match Domain (EMD)

The Exact Match Domain algorithm update was released in September 2012. The EMD update devaluates domain names which are exactly matching its keywords if content quality of the pages is not good.

2.22 Phantom

The Phantom update was released in May 2013. This update detects unnatural link patterns, cross-linking (like network) between two or more sites with large percentage of links between them, heavy use of exact match anchor text etc.

2.23 Payday Loan

The Payday Loan update was released in June 2013 and May 2014. This algorithm update targets heavily spam industry queries like payday loans, mortgage rate trends, pornography, cheap apartments etc.

2.24 Pigeon

The Pigeon update was released by Google in July, 2014. This update was released to provide more accurate results by taking into account the distance and location of local brands from the searcher. The Pigeon update aims to serve more relevant results. It was against the spammers promoting local brands in global search results and it was in favor of average honest business.

2.25 HTTPS / SSL

The HTTPS/SSL update was released in August 2014 by Google. This update gives preference to secure websites that adds encryption for data transfer between web server and web browser. Google says that initially this boost is slight but it may be increased if this update gives a positive effect in search results [21].

HTTPS is the Secure Hypertext Transfer Protocol which uses encryption while transmission of data. The SSL stands for Secure Socket Layer. SSL is a Transport Layer Security protocol for secure data transfer over Internet. SLL certificate is a trust certificate issued be trust authorities. The SSL certificate requirement makes creating link farms economically infeasible for spammers. This update is an attempt to make Internet more secure for end users as well as for their sensitive data.

2.26 Pirate 2.0

The Pirate 2.0 update was released by Google in October, 2014. This update targeted websites that serves illegal and pirated content. The mostly hit web sites are the torrent web sites which offer illegal downloads of pirated software and pirated copyrighted digital media content [22].

3. ANALYTICAL SUMMARY

We observed some key points while writing this paper. These key points are enumerated as following:

- Almost all updates are based on text, there is little work done on multimedia content such as images, video, audio etc.
- The Google's search algorithm updates show a shift from Page Rank to the Quality of Page Content.
- Google focuses on the safety and security of data as well as end users, by promoting secure websites that provide encryption and demoting sites that provide malware, pirated content and scams.
- Google gives more importance to the websites which provide information over the commercial website which try to sell something to end users.
- Google gives more importance to the trusted authority web sites such as big brand websites, websites having TLDs reserved for statutory authorities such as .gov, .edu, .mil etc.

- Google search quality team makes around 350 to 500 changes in the search ranking system every year. In the year 2010, the number of updates was 516.
- Google gives low priority to the websites which are optimized for generally spam industry queries and keyword such as 'Cheap loans', 'Pharmacy' etc.
- Websites which are highly optimized for ranking are discouraged by Google's search system.

We have summarized the Google's search algorithm updates in the Table 1. These updates are categorized according to the website features they deal with and the year when they were launched.

Table I. Algorithm Updates According to Page Feature

Feature Type	**Google Algorithm Update Name**	**Year**
Page Content Quality, Plagiarism	May Day	2010
	Panda	2011 - 2014
	Attribution	2011
Link Structure, Link Farms	Cassandra	2003
	Dominic	2003
	No Follow	2005
	Jagger	2009
	Penguin	2012-2014
	Phantom	2013
Website Authority	Boston	2003
	Vince	2009
	Social Signals	2010
	HTTPS / SSL	2014
	Pirate	2014
Keyword Stuffing	Florida	2003
	EMD	2012
	Penguin	2012-2014
Topic Relevancy, Sentiment Analysis	Austin	2004
	Brandy	2004
	Pigeon	2014
	Negative Review	2010
Monetization	Ad-above-the-fold	2012
	Pay Day Loan	2013-2014
Cloaking, Redirection	Penguin	2012-2014

4. CONCLUSIONS

Identifying and detecting web spam is an on-going battle between search engines and spammers which is going on since search engines allowed searching of the web. In this paper we have studied and analyzed algorithm updates made by Google to combat spamdexing in their search result. After studying these algorithm updates, we can say that Google has radically improved the ability to detect low quality websites which provide no useful information to users. Google search quality team makes around 350 to 500 changes in the search ranking system every year to mitigate chances of spammers who try to play with the ranking system of Google. But due to rapidly changing technology and open nature of the web, spammers may invent new tactics to manipulate the ranking system. We believe the war between Google and spammers will go on in future years also. We hope that in near future Google will release updates to analyze digital media formats also (such as images, video, audio etc.) to check the quality of content of the web pages.
We also believe that web spam is a socio-economic phenomenon which can be dealt with up to some extent if end-users are aware of it and there are such preventive measures that add extra cost to web spam generation.

REFERENCES

[1] Adversarial Web Search, Carlos Castillo and Brian D. Davison, Foundations and Trends in Information Retrieval Vol. 4, No. 5 (2010) 377–486

[2] T. Moore, N. Leontiadis, and N. Christin. "Fashion Crimes: Trending Term Exploitation on the Web". In Proceedings of the ACM CCS Conference, October 2011.

[3] D. Y. Wang, S. Savage, and G. M. Voelker. "Cloak and Dagger: Dynamics of Web Search Cloaking". In Proceedings of the ACM CCS Conference, October 2011.

[4] L. Lu, R. Perdisci, and W. Lee. SURF: "Detecting and Measuring Search Poisoning". In Proceedings of the ACM CCS Conference, October 2011.

[5] D. Y. Wang, S. Savage, and G. M. Voelker. "Juice: A Longitudinal Study of an SEO Campaign". In Proceedings of the NDSS Symposium, February 2013.

[6] Chandra, Ashish, and Mohammad Suaib. "A Survey on Web Spam and Spam 2.0." International Journal of Advanced Computer Research, Volume-4 Number-2 Issue-15 pp. 635-644, June-2014.

[7] L. Page, S. Brin, R. Motwani, and T. Winograd. "The pagerank citation ranking: Bringing order to the web", 1998.

[8] Krishna Bharat, Bay-Wei Chang, Monika Henzinger, Matthias Ruhl, "Who Links to Whom: Mining Linkage between Web Sites", In Proceedings of the IEEE International Conference on Data Mining (ICDM '01), San Jose, CA (2001).

[9] How Google's Algorithm Rules the web, Steven Levy 23 Feb 2010, http://www.wired.com/2010/02/ff-google-algorithm/all/1.

[10] Google Cassandra update, http://level343.com, Blog Archive 14 March, 2011.

[11] Austin, http://www.searchenginejournal.com/the-latest-on-update-austin-googles-january-update/237.

[12] Bharat K. and Mihaila G.A., Hilltop: "A Search Engine: Based on Expert Documents", Technical Report, University of Toronto (1999).

[13] Zhu V., Wu G. and Yunfeg M., "Research and Analysis of Search Engine Optimization Factors Based on Reverse Engineering", In Proceedings of the 3rd International Conference on Multimedia Information Networking and Security, 225-228 (2011).

[14] M. J. Kleinberg, "Authoritative sources in a hyperlinked environment," Journal of the ACM, vol. 46, no. 5, pp. 604–632, 1999.

[15] nofollow http://en.wikipedia.org/wiki/Nofollow.

[16] Blog Post : Algorithm Change Launched http://www.mattcutts.com/blog/algorithm-change-launched/ .

[17] TED 2011: The 'Panda' That Hates Farms: A Q&A With Google's Top Search Engineers http://www.wired.com/2011/03/the-panda-that-hates-farms/.

[18] Blog Post by Amit Singhal, Search Quality Engineer, Google. http://googlewebmastercentral.blogspot.in/2011/04/high-quality-sites-algorithm-goes.html.

[19] Panda, Navneet. "US Patent 8,682,892". USPTO. Retrieved 31 March 2014.

[20] Another step to reward high-quality sites http://googlewebmastercentral.blogspot.in/2012/04/another-step-to-reward-high-quality.html.

[21] Google Web Master Central: Official news on crawling and indexing sites for the Google index, August 6, 2014, http://googlewebmastercentral.blogspot.in/2014/08/https-as-ranking-signal.html.

[22] Google's New Search Down Ranking Hits Torrent Sites Hard, October 23, 2014, http://torrentfreak.com/googles-new-downranking-hits-pirate-sites-hard-141023/.