Web Spam Detection

Marc Najork
Microsoft Research, Mountain View, CA, USA

Synonyms

Spamdexing; Google bombing; Adversarial information retrieval

Definition

Web spam refers to a host of techniques to subvert the ranking algorithms of web search engines and cause them to rank search results higher than they would otherwise. Examples of such techniques include content spam (populating web pages with popular and often highly monetizable search terms), link spam (creating links to a page in order to increase its link-based score), and cloaking (serving different versions of a page to search engine crawlers than to human users). Web spam is annoying to search engine users and disruptive to search engines; therefore, most commercial search engines try to combat web spam. Combating web spam consists of identifying spam content with high probability and – depending on policy – downgrading it during ranking, eliminating it from the index, no longer crawling it, and tainting affiliated content. The first step – identifying likely spam pages – is a classification problem amenable to machine learning techniques. Spam classifiers take a large set of diverse features as input, including content-based features, link-based features, DNS and domain-registration features, and implicit user feedback. Commercial search engines treat their precise set of spam-prediction features as extremely proprietary, and features (as well as spamming techniques) evolve continuously as search engines and web spammers are engaged in a continuing "arms race."

Historical Background

Web spam is almost as old as commercial search engines. The first commercial search engine, Lycos, was incorporated in 1995 (after having been incubated for a year at CMU); and the first known reference to "spamdexing" (a combination of "spam" and "indexing") dates back to 1996. Commercial search engines began to combat spam shortly thereafter, increasing their efforts as it became more prevalent. Spam detection became a topic of academic discourse with Davison's paper on using machine learning techniques to identify "nepotistic links," i.e., link spam [4], and was further validated as one of the great challenges to commercial search engines by Henzinger et al. [9]. Since 2005, the workshop series on *Adversarial Information Retrieval on the Web* (AIRWeb) provides a venue for researchers interested in web spam.

Foundations

Given that the objective of web spam is to improve the ranking of select search results, web spamming techniques are tightly coupled to the ranking algorithms employed (or believed to be employed) by the major search engines. As ranking algorithms evolve, so will spamming techniques. For example, if web spammers were under the impression that a search engine would use click-through information of its search result pages as a feature in their ranking algorithms, then they would have an incentive to issue queries that bring up their target pages, and generate large numbers of clicks on these target pages. Furthermore, web spamming techniques evolve in response to countermeasures deployed by the search engines. For example, in the above scenario, a search engine might respond to facetious clicks by mining their query logs for many instances of identical queries from the same IP address and discounting these queries and their result click-throughs in their ranking computation. The spammer in turn might respond by varying the query (while still recalling the desired target result), and by using a "bot-net" (a network of third-party computers under the spammer's control) to issue the queries and the click-throughs on the target results.

Given that web spamming techniques are constantly evolving, any taxonomy of these techniques must necessarily be ephemeral, as will be any enumeration of spam detection heuristics. However, there are a few constants:

- Any successful web spamming technique targets one or more of the features used by the search engine's ranking algorithms.
- Web spam detection is a classification problem, and search engines use machine learning algorithms to decide whether or not a page is spam.
- In general, spam detection heuristics look for statistical anomalies in some of the features visible to the search engines.

Web Spam Detection as a Classification Problem

Web spam detection can be viewed as a binary classification problem, where a classifier is used to predict whether a given web page or entire web site is spam or not. The machine learning community has produced a large number of classification algorithms, several of which have been used in published research on web spam detection, including decision-tree based classifiers (e.g., C4.5), SVM-based classifiers, Bayesian classifiers, and logistic regression classifiers. While some classifiers perform better than others (and the spam detection community seems to favor decision-tree-based ones), most of the research focuses not on the classification algorithms, but rather on the features that are provided to them.

Taxonomy of Web Spam Techniques

Content spam refers to any web spam technique that tries to improve the likelihood that a page is returned as a search result and to improve its ranking by populating the page with salient keywords. Populating a page with words that are popular query terms will cause that page to be part of the result set for those queries; choosing good combinations of query terms will increase the portion of the relevance score that is based on textual features. Naïve spammers might perform content spam by stringing together a wide array of popular query terms. Search engines can counter this by employing language modeling techniques, since web pages that contain many topically unrelated keywords or that are grammatically ill-formed will exhibit statistical differences from normal web pages [11]. More sophisticated spammers might generate not a few, but rather millions of target web pages, each page augmented with just one or a few popular query terms. The remainder of the page may be entirely machine-generated (which might exhibit statistical anomalies that can be detected by the search engine), entirely copied from a human-authored web site such as Wikipedia (which can be detected by using near-duplicate detection algorithms), or stitched together from fragments of several human-authored web sites (which is much harder, but not impossible to detect).

Link spam refers to any web spam technique that tries to increase the link-based score of a target web page by creating lots of hyperlinks pointing to it. The hyperlinks may originate from web pages owned and controlled by the spammer (generically called a *link farm*), they may originate from partner web sites (a technique known as *link exchange*), or they may originate from unaffiliated (and sometimes unknowing) third parties, for example web-based discussion forums or in blogs that allow comments to be posted (a phenomenon called *blog spam*). Search engines can respond to link spam by mining the web graph for anomalous components, by propagating distrust from spam pages backwards through the web graph, and by using content-based features to identify spam postings to a blog [10]. Many link spam techniques specifically target Google's PageRank algorithm, which not only counts the number of hyperlinks referring to a web page, but also takes the PageRank of the referring page into account. In order to increase the PageRank of a target page, spammers should create links on sites that have high PageRanks, and for this reason, there is a marketplace for expired domains with high PageRank, and numerous brokerages reselling them. Search engines can respond by temporarily dampening the endorsement power of domains that underwent a change in ownership.

Click spam refers to the technique of submitting queries to search engines that retrieve target result pages and then to "click" on these pages in order to simulate user interest in the result. The result pages returned by the leading search engines contain client-side scripts that report clicks on result URLs to the engine, which can then use this implicit relevance feedback in subsequent rankings. Click spam is similar in method to *click fraud*, but different in objective. The goal of click spam is to boost the ranking of a page, while the goal of click fraud (generating a large number of clicks on search engine advertisements) is to spend the budget associated with a particular advertisement (to hurt the competitor who has placed the ad or simply to lower the auction price of said ad, which will drop once the budget of the winning bidder has been exhausted). In a variant of click fraud, the spammer targets ads delivered to his own web by an ad-network such as Google AdSense and obtains a revenue share from the ad-network. Both click fraud and click spam are trivial to detect if launched from a single machine, and hard to detect if launched from a bot-net consisting of tens of thousands of machines [3]. Search engines tackle the problem by mining their click logs for statistical anomalies, but very little is known about their algorithms.

Cloaking refers to a host of techniques aimed at delivering (apparently) different content to search engines than to human users. Cloaking is typically used in conjunction with content spam, by serving a page containing popular query terms to the search engine (thereby increasing the likelihood that the page will be returned as the result of a search), and presenting the human user with a different page. Cloaking can be achieved using many different techniques: by literally serving different content to search engines than to ordinary users (based for example on the well-known IP addresses of the major search engine crawlers), by rendering certain parts of the page invisible (say by setting the font to the same color as the background), by using client-side scripting to rewrite the page after it has been delivered (relying on the observation that search engine crawlers typically do not execute scripts), and finally by serving a page that immediately redirects the user's browser to a different page (either via client-side scripting or the HTML "meta-redirect" tag). Each variant of cloaking calls for a different defense. Search engines can guard against different versions of the same page by probing the page from unaffiliated IP addresses [13]; they can detect invisible content by rendering the page; and they can detect page modifications and script-driven redirections by executing client-side scripts [12].

Key Applications

Web spam detection is used primarily by advertisement-financed general-purpose consumer search engines. Web spam is not an issue for enterprise search engines, where the content providers, the search engine operator and the users are all part of the same organization and have shared goals. However, web spam is bound to become a problem in any setting where these three parties – content providers, searchers, and search engines – have different objectives. Examples of such settings include vertical search services, such as product search engines, company search engines, people search engines, or even scholarly literature search engines. Many of the basic concepts described above are applicable to these domains as well; the precise set of features useful for spam detection will depend on the ranking algorithms used by these vertical search engines.

Future Directions

Search engines are increasingly leveraging human intelligence, namely the observable actions of their user base, in their relevance assessments; examples include click-stream analysis, toolbar data analysis, and analysis of traffic on affiliate networks (such as the Google AdSense network). It is likely that many of the future spam detection features will also be based on the behavior of the user base. In many respects, the distinction between computing features for ranking (promoting relevant documents) and spam detection (demoting facetious documents) is artificial, and the boundary between ranking and spam suppression is likely to blur as search engines evolve.

Experimental Results

Several studies have assessed the incidence of spam in large-scale web crawls at between 8% and 13% [11,5]; the percentage increases as more pages are crawled, since many spam sites serve a literally unbounded number of pages, and web crawlers tend to crawl high-quality human-authored content early on. Ntoulas et al. describe a set of content-based features for spam detection; these features, when combined using a decision-tree-based classifier, resulted in an overall spam prediction accuracy of 97% [11].

Data Sets

Castillo et al. have compiled the WEBSPAM-UK2006 data set [2], a collection of web pages annotated by human judges as to whether or not they are spam. This data set has become a reference collection to the field, and has been used to evaluate many of the more recent web spam detection techniques.

Cross-references

► Indexing the Web
► Web Page Quality Metrics
► Web Search Relevance Feedback
► Web Search Relevance Ranking

Recommended Reading

1. Becchetti L., Castillo C., Donato D., Leonardi S., and Baeza-Yates R. Using rank propagation and probabilistic counting for link-based spam detection. In Proc. KDD Workshop on Web Mining and Web Usage Analysis, 2006.
2. Castillo C., Donato D., Becchetti L., Boldi P., Leonardi S., Santini M., and Vigna S. A reference collection for Web spam. ACM SIGIR Forum, 40(2):11–24, 2006.
3. Daswani N. and Stoppelman M. and the Google Click Quality and Security Teams. The anatomy of clickbot.A. In Proc. 1st Workshop on Hot Topics in Understanding Botnets, 2007.
4. Davison B.D. Recognizing nepotistic links on the web. In Proc. AAAI Workshop on Artificial Intelligence for Web Search, 2000.
5. Fetterly D., Manasse M., and Najork M. Spam, damn spam and statistics. In Proc. 7th Int. Workshop on the Web and Databases, 2004, pp. 1–6.
6. Gyöngyi Z., and Garcia-Molina H. Spam: its not just for Inboxes anymore. IEEE Comput., 38(10):28–34, 2005.
7. Gyöngyi Z. and Garcia-Molina H. Web Spam Taxonomy. In Proc. 1st Int. Workshop on Adversarial Information Retrieval on the Web, 2005, pp. 39–47.
8. Gyöngyi Z., Garcia-Molina H., and Pedersen J. Combating Web spam with TrustRank. In Proc. 30th Int. Conf. on Very Large Data Bases, 2004, pp. 576–587.
9. Henzinger M., Motwani R., and Silverstein C. Challenges in web search engines. ACM SIGIR Forum 36(2):11–22, 2002.
10. Mishne G., Carmel D., and Lempel R. Blocking blog spam with language model disagreement. In Proc. 1st Int. Workshop on Adversarial Information Retrieval on the Web, 2005, pp. 1–6.

11. Ntoulas A., Najork M., Manasse M., and Fetterly D. Detecting spam web pages through content analysis. In Proc. 15th Int. World Wide Web Conference, 2006, pp. 83–92.

12. Wang Y.M., Ma M., Niu Y., and Chen H. Spam double-funnel: connecting Web spammers with advertisers. In Proc. 16th Int. World Wide Web Conference, 2007, pp. 291–300.

13. Wu B. and Davison B. Detecting semantic cloaking on the web. In Proc. 15th Int. World Wide Web Conference, 2006, pp. 819–828.