Google-bombing - Manipulating the PageRank Algorithm

Peter A. Hamilton
CMSC 676 - Information Retrieval

1. Introduction

With the growth of the Internet, the field of Information Retrieval (IR) has gained increasing importance. Quick, easy, and accurate information access is the deciding factor between successful search companies and their rivals. Likewise, the manipulation of IR systems for ulterior motives, known as adversarial IR, is just as important, as it can turn the successes of a search strategy against itself, at the expense of the general user. In this paper, a specific form of adversarial IR known as Google-bombing is examined, regarding its implementation and the different means by which it can be combatted.

To lay the foundation for Google-bombing, Section 2 discusses background information including the concepts of link analysis, spamdexing, and the general idea behind the PageRank algorithm. Section 3 discusses the history and implementation of a Google-bomb, providing several examples. Section 4 examines link analysis methods that could be applied to PageRank to “diffuse” Google-bombs, along with relevance feedback tactics that could also be employed.

2. Background

To understand the ideas and concepts utilized and manipulated by a Google-bomb, it is necessary to have a general understanding of several different topics in IR.

2.1 Link Analysis

Traditional IR systems focus primarily on the actual content of the documents in the corpus, utilizing TF-IDF and cosine-similarity scoring (or a similar approach) to compute similarity scores that indicate the documents which are “most similar” to the search query. Link analysis is an alternative method used to rank pages in light of a specific search query [5]. Link analysis selects specific document content (i.e. hyperlinks) and tries to use link text and frequency of page linking to augment document similarity scores. Combining document scoring with link analysis generally returns documents that both match the provided query and which are linked to by many other documents, indicating indirectly that the returned documents are more likely to be relevant, or that they were at least relevant to other website authors. Conversely, documents with few or no “incoming” links will appear lower in the result list.

2.2 Spamdexing

The phrase “spamdexing” combines the terms “spam” and “indexing”, indicating an application of spamming techniques to the indexing process [4]. Spamdexing is a general term referring to various techniques that take advantage of different features in an indexing system, allowing malicious and non-malicious users alike to artificially inflate the ranking of specific documents in the indexing process. These techniques range from stuffing a document with certain query terms, thereby causing
it to score higher for those terms in light of TF-IDF, to setting up local groupings of pages that all heavily reference each other to increase the ranking of the overall group.

2.3 The PageRank Algorithm

The PageRank algorithm is a specific algorithm for conducting link analysis [3]. Originally developed in the 1990s as a research project for Stanford University, it was later licensed exclusively to Google as a core part of its Internet indexing system. PageRank analyzes the links between web pages and uses them to recursively score each page, assigning it a page rank.

PageRank can be viewed in two ways: (1) a form of voting among web pages for other pages they are linked to in the page network, and (2) a model of the likelihood a web surfer will end up on a specific page given a random “walk” through the page network [5]. In the voting model, a page’s page rank is defined by the number of links to it from disparate pages. In turn, the higher page rank a page has, the more important are its links to other pages. A single page with a few links from highly ranked pages will therefore score better than a page with lots of links from poorly ranked pages. Figure 1 shows a visual representation of a page network after PageRank has assigned specific ranks to each page, demonstrating the effect of reputable page relationships.

![Figure 1](image)

Figure 1: This is an example network provided as input to the PageRank algorithm. Node size correlates to the labeled ranking of the document, or node. The rank of Node C demonstrates the importance of being linked to other important pages. Image taken from [2].

From the view of the random-walk model, each page rank indicates the general probability a web surfer has of visiting the page given a random walk through the network. If a page has lots of incoming links, it is more likely to be visited by the web surfer. If a page has lots of incoming links from other pages with lots of incoming links, it is even more likely to be visited by the web surfer. If a page is a sink in the page graph (e.g. Node A in Figure 1), it is assumed that the surfer randomly jumps to another page in the network.

In addition to link mapping, PageRank also associates links with the link text used and adds this information to the general query information related to a page. If a majority of links use a specific
phrase when linking to a target page, that phrase, when queried, will include the target page in the results list.

PageRank essentially models the flow of navigation and information through the page network, thereby modeling the likely flow of web surfer traffic. It is capable of finding pages that will be visited by any generic surfer and is incredibly useful in determining which pages are important to the overall structure of the network. It uses the summary information provided by link text and augments the original term-document matrix with this information to further ensure reliable query results.

3. Google-bombing

According to the New Oxford American Dictionary as of 2005, the term “Google-bombing” is defined as “the activity of designing Internet links that will bias search engine results so as to create an inaccurate impression of the search target” [6]. A Google-bomb is the result of an intentional set of actions whereby a target page is linked to by many different pages with the same link text, or key phrase, thereby associating the target with the key phrase in Google’s PageRank algorithm. This association is almost always erroneous or misleading, thereby disrupting the accuracy of the indexing system and possibly facilitating the spread of disinformation.

3.1 Examples

The first known Google-bomb occurred in 1999, when it was discovered that the search query “more evil than satan himself” returned the Microsoft homepage as its top result. Perhaps the most famous incident of Google-bombing occurred when the phrase “miserable failure” was tied to then President George W. Bush’s biography on the White House website in 2003.

Not all Google-bombing incidents are done for personal or “immoral” reasons. As of 2005, the top result of the query “Jew” was an anti-semetic website due to the query’s derogatory nature. A Google-bombing campaign was started by a Jewish blogger to change that result to the Wikipedia article on Jewish culture [8]. As of early 2006, the Google-bomb has been a success, despite a counter-bomb devised by supporters of the original page.

A less serious Google-bomb that has not been fixed and is still active targets the “French military victories” query. When Googled, the top result is a fake Google error page that states “Your search - french military victories - did not match any documents. Did you mean: french military defeats?” [1].

3.2 Relevance

Google-bombing is interesting because it reveals one of the primary weaknesses of the composite indexing approach used by Google. Note that no actual page content of the target page is ever manipulated when targeted by a Google-bomb. It ranks the same when compared to all other search queries. However, since link analysis is used to associate additional query terms or phrases with the document that do not appear in the document, it is possible to artificially inflate the page’s relationship with the Google-bomb key phrase and therefore indirectly change the term-document matrix.
4. Diffusing a Google-bomb

The rise of Google-bombing was detected throughout the early 2000s and, for a while, Google refused to intervene [7]. While initially considered a prank, Google-bombing has found serious applications in political and commercial circles, associating competitors, rivals, and enemies with negative or derogatory terms. As of 2007, Google has neutralized many of the known Google-bombing key phrases by modifying the PageRank algorithm. Searching for most Google-bomb key phrases now yields pages describing the Google-bombing phenomenon and the history of the term. This does not stop new Google-bombs from being created. Despite Google’s secrecy, the details of the PageRank algorithm allow us to reasonably guess at some of the algorithmic and manual changes adopted to diffuse Google-bombs.

4.1 Linker Reputability

Google-bombing is successful because it exploits the assumption that website authors link to pages that they find particularly important. In this model, the very existence of a link adds weight to the ranking of the target page, increasing its rating for given queries. However, when an author repeatedly links to the same page, or when a multitude of authors do so in a deceptive manner, the rank for the target still increases, regardless of whether the page is relevant. This general problem can be addressed by taking into account the reliability, or reputability, of the linker.

Generally, given certain domains or servers, certain websites could be given default reputability rankings (i.e. government (.gov) and education (.edu) sites might be more reliable than a .com page). Reputability can also be defined by a high page rank threshold, where only the most trafficked, and therefore the most popular, pages should be considered reliable in that most users visit them and rely on them for information. As a web crawler scans the Internet, it can recompute page ranks and shift page rankings that are no longer “on the front page” by analyzing the presence or location of links, allowing new and more relevant material to percolate to the top of the results list.

4.2 Link Text Analysis

Google-bombing also takes advantage of the assumption that link text is an accurate description of the page being linked to. If a large percentage of incoming links contain the same general terms, it is likely that a query containing those terms should yield the target page in its results list. As with linker reputability, this assumption relies on the honesty of web authors. If a majority of incoming links have the same text but that text is in no way related to the page or page content, the authors can artificially increase the rating of the page given queries that contain the link terms.

There are several ways to combat the abuse of link text. First, check the target page for the terms in the link text. If the target page contains none of the link terms, it is likely the information in the link has nothing to do with the target page and therefore the link should not be included in the link analysis for that page. Alternatively, a history of link text could be maintained, tracking the types of link terms used when linking to the target page. If a sudden shift is detected in both the number of links and the “average” link text, it may be that a Google-bomb is forming. The latter alternative, however, is far more costly in terms of time and space complexity and should only be used for a moderately sized corpus.

Finally, it is possible to combine both linker reputability and link text analysis, where authors who post uninformative links accrue a slight penalty on their page rank. This diminishes the influence of authors attempting to tie pages to meaningless query terms, but can also harm those authors who publish lots of links and who do not always check to ensure their links have meaningful text.
4.3 Human Intervention

While the previously listed methods can accurately detect and neutralize Google-bombs, they can also inadvertently change the ranking of pages that are not Google-bomb targets. The only current full-proof alternative to effectively detect and mitigate Google-bombing is to use human evaluation, which can be done in various ways using relevance feedback.

Web surfers should be allowed to rank the page results of their queries, especially if the page results are unsatisfactory. Additionally, web surfers might have the option to rank a page as wholly irrelevant, in that it contains no information concerning the original query. Given enough notices from various users, Google, or any generic search company, could then take appropriate investigative measures and manually disable the “bomb”, if it actually is one.

Note that even these “last resort” techniques can be abused by a large enough population of web surfers intent on hiding real Google-bombs or fabricating fake ones. Additionally, manual solutions are frowned upon due to the time and resources required to maintain them. Even though human “filtering” may work, the algorithmic approaches listed above are preferable.

5. Conclusion

The Google-bomb phenomenon is an interesting case study in the weaknesses of indexing systems, specifically applying to the PageRank algorithm. At a more fundamental level, Google-bombing reveals some of the basic nature of humanity in that individuals do not hesitate to abuse the rules of a bias-less system for personal gain, whether it be for humorous, monetary, political, or personal reasons. Like the war between cryptosystem developers and code breakers, search engine optimizers will have to take into account the actions of web authors looking to advance their own agendas via indexing systems in future search engine development.

References