Marco Gori and Ian Witten

The Bubble of Web Visibility

Promoting visibility as seen through the unique lens of search engines.

The Web seems like a Borgesian library with a huge amount of information. Access to this treasure is mediated by dragons in the guise of search engine operators who compete amongst themselves for dominance. The battleground is so hostile that few will survive—indeed, just one will likely achieve overall dominance in all but specialist corners of the library. The allegory is compelling—except that the treasure is not private property but a public good, and the dragons must serve their own business interests rather than the public cause to compete successfully.

A vital source of information in everyday life, all the content of the Web is readily available—in principle. But the dominant mode of access is through search engines, and—in practice—the view they offer is restricted. Typical queries match hundreds of thousands of documents, which search engines sort according to ranking heuristics. For pragmatic reasons users view only the highest ranked pages, which gives rise to the notion of Web page visibility as perceived through the lens provided by search engines.

The heuristics adopted by popular search engines favor the construction of artificial communities that are expressly designed to promote Web sites, a speculative process we call the "bubble of Web visibility." Moreover, conventional ranking schemes are biased and partial—by design—and although the principles are public, the precise details of the bias are closely guarded trade secrets.

$$\text{score of page } p = \sum_{q \text{ points to } p} \frac{\text{score of page } q}{\text{number of pages pointed to by } q}$$

Figure 1. Determining PageRank.

Promoting Visibility

Today's search engines, the guardians of the Web, measure Web visibility according to the "authority" of each page. Following general principles of information retrieval, the documents that match a query are sorted by an index that takes into account their similarity to the query and the absolute authority of the pages themselves. Typically, the notion of authority is independent of page content. For example, in the Google search engine [7], which we use for illustration because of its widespread popularity, the score of a page p is determined as shown in Figure 1. This score is referred to as PageRank.

Web page visibility becomes fully defined only at query time, when retrieved pages are sorted for presentation to the user. In practice, you can improve the likelihood that your Web site appears prominently in response to queries relating to its content

by taking specific actions to promote it. For example, as shown in Figure 2, you can capitalize on an understanding of the circuital propagation of PageRank to organize Web pages in a way that forces a given target page's rank to increase without bound; this is commonly called a "link farm."

A different way to boost visibility is based on the fact that Google matches the terms in a query not only against each retrieved page, but also against the "anchor text" that links other pages to it. Unlike the method of boosting PageRank, this technique acts at query time by increasing the number of matches with a target page. Related ways of boosting visibility apply to other important search engines.

Depending on the search engine, there are many other ways to promote visibility. When a page is created, words and phrases can be inserted that people are likely to search for. It is more effective to match phrases than individual words, and multiple synonyms dramatically increase one's chance of appearing near the top of the result set. Words and phrases inserted specifically to promote visibility can be hidden from readers using a variety of techniques, ranging from being tagged as "metadata" that does not form part of the visible content, to being written large but in the same color as the background. In a particularly insidious technique, "cloaking," the Web server delivers a special version of the page to particular named Web crawlers. This technique can reduce to zero the correlation between the Web as seen by the crawler and that experienced by regular users.

A continual battle of wits is waged between search engine operators and those who seek to enhance Web page visibility artificially. The situation resembles the computer virus-antivirus escalation war, except that the economic benefit of Web visibility is larger and far more direct. The net effect for the information consumer is to distort, in arbitrary and unpredictable ways, the criteria used to rank retrieved results. PageRank, for instance, is a purely social way of evaluating information: pages gain importance when important sites reference them explicitly. This mechanism provides a very effective practical solution for many topics on the Web but its potential manipulation by link farm raises serious concerns, especially for information that is relevant for decision processes and for shaping opinions. Worse still, the general trend of promoting visibility is forcing search engine operators to modify their ranking model in ad hoc and secret ways in an attempt to outwit those who try to exploit it. This is the lens through which society sees its Web-based information.

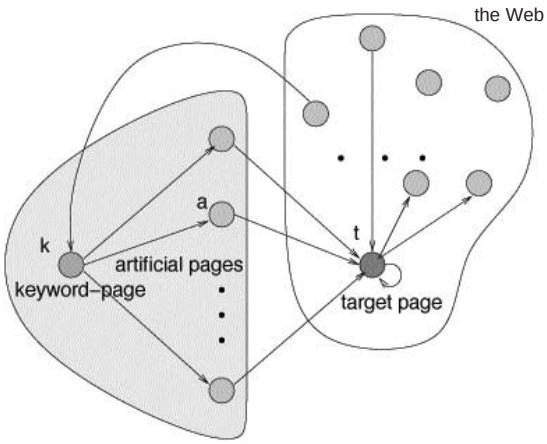

Figure 2. How to promote a Web page by playing on the number of input links. A keyword page is created that links to a set of artificial pages containing the selected keywords, which link to the target page.

The Bubble

Page promotion is becoming popular because the cost of sustaining the growth of Web visibility is affordable. The resulting distortion is creating a "bubble" just like those experienced in the stock market. Johansen et al. [5] point out that the eight most significant financial crashes from 1929 to 1998 share common preconditions. In particular, economic bubbles are caused by local, self-reinforcing, imitative behavior between traders. Holland's infamous tulip craze in the late-sixteenth century, when tulip bulbs exceeded gold in value, is an early example of the havoc that irrationality can play with valuations.

The general trend of promoting visibility is forcing search engine operators to modify their ranking model in ad hoc and secret ways in an attempt to outwit those who try to exploit it.

The essential ingredient of imitation is fundamental to the growth of economic bubbles. Consider the set of pages that match a given query. If all page owners want their wares to appear high in the result list,[1] the competition becomes an unstable process in which page owners continually buy additional visibility. This is local self-reinforcing imitation, just as in the stock market.

Recently, occasional criticisms of search engine operators have begun to emerge in the press: This is just a bellwether for the speculative process outlined here.

Conclusion

These observations call for a serious reconsideration of Web searching—not just a series of "peaceful interludes" but a paradigm shift, an "intellectually violent revolution" [6]. One might try to address speculative Web visibility scams individually (as search engine companies are no doubt doing); however, the bubble is likely to reappear in other guises. Moreover, issues of equality of access to information, transparency, and fairness can only be addressed within new paradigms; new, personalized views of the Web that supplement today's horizontal search services. Different users may merit different answers to the same query, and the technology must guarantee that such personal views cannot be manipulated.

Sophisticated automatic analysis of Web page content and the knowledge structures they represent may stem the tide of manipulative growth of visibility. Automatic analysis of page content is gaining central importance in both crawling [2, 4] and page scoring [3], and can help detect artificial communities conceived solely for page promotion. Perhaps the Semantic Web [1], by couching Web content in a form that is meaningful to computers, will help defeat attempts to misrepresent its importance.

Web page visibility was invented by the dragons to protect the treasure and to make it accessible. And indeed they have provided a great service to the Web community. But current notions of visibility are unlikely to be sustainable, at least for a large and important portion of the Web. The Web is a public good and biased views will become the target of sustained protest. The dragons will better serve users, and hence their own business, by anticipating the protest. They need to strengthen links with the research community and devise new models of access, new ways to evaluate the treasure. C

[1]Many Web page promotion companies guarantee that their clients' pages will appear in the top 10 positions for a given set of keywords.

References

1. Berners-Lee, T., Hendler, J., and Lassila, O. The Semantic Web. *Scientific American 284*, 5 (May 2001), 28–37.
2. Chakrabarti, S., van den Berg, M., and Dom, B. Focused crawling: A new approach to topic-specific Web resource discovery. In *Proceedings of the 8th International World Wide Web Conference*. ACM Press, 1999.
3. Diligenti, M., Gori, M., and Maggini, M. A unified probabilistic framework for Web page scoring systems. In *IEEE Transactions on Knowledge and Data Engineering 16*, 1 (Jan. 2004).
4. Diligenti, M., Coetzee, F., Lawrence, S., Giles, C., and Gori, M. Focused crawling using context graphs. In *Proceedings of the 26th International Conference on Very Large Databases, VLDB 2000,* May 2000.
5. Johansen, A., Sornette, D., and Ledoit, O. Predicting financial crashes using discrete scale invariance. *Journal of Risk 1*, 4 (1999).
6. Kuhn, T. *The Structure of Scientific Revolutions* (second edition). The University of Chicago Press, 1970.
7. Page, L., Brin, S., Motwani, R., and Winograd, T. *The Pagerank Citation Ranking: Bringing Order to the Web*. Technical Report, Stanford University, 1998.

Marco Gori (marco@dii.unisi.it) is a professor of computer science at the University of Siena, Italy.

Ian Witten (ihw@cs.waikato.ac.nz) is a professor of computer science at the University of Waikato, New Zealand.