What do the Neighbours Think?
Computing Web Page Reputations

Alberto O. Mendelzon
Department of Computer Science
University of Toronto
mendel@cs.toronto.edu

Davood Rafiei
Department of Computing Science
University of Alberta
drafiei@cs.ualberta.ca

Abstract

The textual content of the Web enriched with the hyperlink structure surrounding it can be a useful source of information for querying and searching. This paper presents a search process where the input is the URL of a page, and the output is a ranked set of topics on which the page has a reputation. For example, if the input is www.gamelan.com, then a possible output is "Java." We describe a simple formulation of the notion of reputation of a page on a topic, and report some experiences in the use of this formulation.

1 Introduction

The idea of exploiting the "reputation" of a Web page when searching has attracted research attention recently and even been incorporated into some search engines [16, 6, 11, 2, 3]. The idea is that pages with good reputations should be given preferential treatment when reporting the results of a search; and that link structure can be mined to extract such reputation measures, on the assumption that a link from page a to page b is, to some degree, an endorsement of the contents of b by the creator of a. The question that needs to be answered to use page reputation in Web search is: given a topic, what pages have the highest reputation on this topic? We consider a different question in this paper: given a page (or a Web site), on what topics is this page considered an authority by the Web community?

There are many potential applications for such computations. For example, a company may wish to know how its Web site is categorized by other pages that point to it, for several reasons: to assess its popularity, to check whether it is projecting the right image, or to detect problems signaled by unflattering links. Statistics of web access, such as the unique monthly visitor counts maintained by Web ranking services, are notoriously controversial [18] and do not provide insight on the topics that a Web site is perceived to be relevant to. A link-based reputation measuring service could be used not only by the site owners, but by anyone who needs to evaluate a site before using it as a source of information, or before transacting business with it.

Another example is the use of the reputation measure of a Web site listing a researcher's publications to help assess the impact of the researcher's work, with the obvious caveats against depending too heavily on any one measure of impact– caveats that, for that matter, also apply to more traditional methods such as print or online citation indexes.

Bulletin of the IEEE Computer Society Technical Committee on Data Engineering

There are difficulties in formalizing the notion of "reputation" effectively. The assumption that links are endorsements suggests that the number of incoming links of a page indicates its reputation. But in practice, links represent a wide variety of relationships such as navigation, subsumption, relatedness, refutation, justification, etc. In addition, we are interested not just in the overall reputation of a page, but in its reputation on specific topics.

In this paper, we describe a search process where the input is the URL of a page, and the output is a ranked set of *topics* on which the page has a reputation. (For our purposes, a topic is simply a term or a phrase, an admittedly simplistic definition.) For example, if the input is `www.informatik.uni-trier.de/~ley/db`, then among the outputs we would like to see with high rank topics such as "DBLP Bibliography", "database systems," etc. We present a simple formulation of the notion of reputation with the following two goals in mind: (1) it must be easy to compute in the setting of the Web; (2) it must be effective in measuring the reputations of pages. We point out that this formulation is a rough approximation to a more complex measure, based on random walk models of Web browsing behaviour. Finally, we report some test results on the effectiveness of our computations.

1.1 Related Work

Recent analyses of the linkage structure of the Web suggest that hyperlinks between pages often represent relevance [16, 6] or endorse some authority [11, 2, 3].

In a method incorporated into the Google search engine, Brin and Page [3] compute the importance of a Web page as the sum of the importance of its incoming links. The computation simulates the behaviour of a "random surfer" who either selects an outgoing link uniformly at random, or jumps to a new page chosen uniformly at random from the entire collection of pages. The PageRank of a page corresponds to the number of visits the "random surfer" makes to the page.

Kleinberg [11] proposes an algorithm that, given a topic, finds pages that are considered authorities on that topic. The algorithm, known as HITS, is based on the hypothesis that for broad topics, authority is conferred by a set of *hub pages*, which are recursively defined as a set of pages with a large number of links to many relevant authorities.

In earlier work [17], we developed two formulations of the notion of reputation based on random walk models of simple Web browsing behaviours. The first model, based on the idea of one-level influence propagation, generalizes PageRank; the main difference is that the ranking is performed with respect to a specific topic instead of computing a universal rank for each page. The second model is a probabilistic formulation of a model similar to Hubs and Authorities; this formulation allows us to invert the search process, i.e., given the URL of a page, we can find the topics the page is an authority on. For differences between HITS and probabilistic approaches, see the work by Lempel and Moran [15].

The notion of adjusting link weights for HITS was studied by Chakrabarti et al. [4] and Bharat and Henzinger [2]. Based on the hub-and-authority structure of a community, Kumar et al. [13] show that a large number of such communities can be identified from their signatures in the form of complete bipartite subgraphs of the Web. Dean and Henzinger [6] suggest algorithms to find related pages of a given page solely based on the linkage structure around the page. Finally, Henzinger et al. [9, 10] use random walks on the Web to do URL sampling and also to measure the quality of pages stored in an index.

2 Our Approach

Given a page, we want to identify a ranked list of topics the page has a reputation on. This requires a ranking function and a collection (of pages and topics) over which the ranks can be computed.

We first define two ratios that relate a page p and a topic t. The *penetration* of page p on topic t, $P_p(t)$, is the fraction of pages on topic t that point to page p. (For our purposes, a page is *on* topic t simply when it contains the term or phrase *t*.) The *focus* of page *p* on topic *t*, $F_t(p)$, is the fraction of pages pointing to *p* that are on topic *t*. That is, if $I(p,t)$ is the number of pages that contain *t* and point to *p*, $In(p)$ is the number of pages that point to *p*, and *N(t)* is the number of pages that contain *t*, then

$$P_p(t) = I(p,t)/N(t)$$

and

$$F_t(p) = I(p,t)/In(p).$$

Note that these quantities can be interpreted as conditional probabilities: $P_p(T)$ is the conditional probability that a page points to *p*, given that it contains *t*, and $F_t(p)$ is the conditional probability that a page contains *t*, given that it points to *p*.

Using one of these ratios as a measure of the reputation of *p* on *t* is problematic. For example, $P_p(t)$ may overestimate the reputation on any topic of pages that have large in-degrees, while $F_t(p)$ may overestimate the reputation of those whose incoming links are narrowly concentrated on topic *t*. It is more appropriate to consider, not just the conditional probability that a page points to *p* given that it contains *t*, but how much larger (or smaller) is this probability than the unconditional probability that an arbitrary page points to *p*. Note that the latter probability is given by $L(p) = In(p)/N_w$, where N_w is the number of pages on the web. Let us define the *reputation measure* of *p* on *t*, $RM(p,t)$, as

$$RM(p,t) = \ (P_p(t) - L(p))/L(p).$$

We could equally define $RM(p,t)$ in terms of $F_t(p)$, letting $M(t) = N(t)/N_w$ be the probability that an arbitrary page contains term *t*:

$$RM(p,t) = \ (F_t(p) - M(t))/M(t) = \ (P_p(t) - L(p))/L(p).$$

If we now define $LM(t,p)$ as the probability that a page contains term *t* and points to page *p*, it is easily seen that

$$RM(p,t) = \ [LM(t,p)/(L(p) \times M(t))] \ - \ 1.$$

That is, $RM(p,t)$ measures how far from independent are the events "a page contains *t*" and "a page points to *p*." Brin *et al.* [1] propose a similar measure, called a *dependence rule*, as an alternative to association rules (which use confidence ratios similar to $P_p(t)$ and $F_t(p)$), and show how to use standard statistical tests to evaluate its significance; we do not pursue the latter topic in this paper.

Given a search engine that can compute estimates of $I(p,t)$, $N(t)$ and N_w, $RM(P,t)$ can be readily estimated as

$$RM(p,t) = \ (N_w I(p,t)/N(t) In(p)) - \mathit{1}.$$

We next show that the proposed ranking provides both an effective and easy to compute reputation measure.

2.1 Ranking Effectiveness

We now provide some justifications on the effectiveness of the proposed ranking. When a page is created, it has no incoming links (except possibly some links from the same site, which we ignore in our computation). As other users become aware of the page, based on their judgments of the content of the page and its relevance to their topics of interest, they start including links to the page within the pages they create or maintain. After a while, if a large fraction of pages on a specific topic point to the page, it is natural to expect that the page has secured a reputation on that topic.

A similar interesting phenomenon can be seen in link navigation. Users frequently search the Web by alternating between the two modes of (1) searching for pages that contain some terms (using a search engine), and (2) following outgoing links from those pages. If a large fraction of pages on a specific topic points to a page, the page will be most likely visited by the Web surfers searching for pages on that topic. Specifically, for a given page *p* and topic *t*, $RM(p,t)$ is, to some degree, a rough approximation to the number of visits made to page *p* by a "random surfer" who wanders the Web by following the links and searching for pages on topic *t* (see [17] for details).

2.2 Rank Computations

Consider the rank computation in the simple case where both a page *p* and a word or phrase *t* are given. We can use a search engine to estimate the values of $In(p)$, $I(p,t)$ and $N(t)$, and then plug these estimates in to compute $RM(p,t)$. For example, we can estimate $In(p)$, $I(p,t)$ and $N(t)$ by respectively sending queries "+link:p," "+link:p +t" and "+t" to AltaVista (*www.altavista.com*) and retrieving the counts returned by the engine. The value of N_w, a constant which doesn't affect the ordering, can be estimated by the number of pages in the search engine collection; this is often publicly announced (e.g., [14]).

However, we are often interested in the case where the topics that a page has a reputation on are not known in advance. Thus the problem is: how to compute for a given page *p* the set of topics *t* such that $RM(p,t)$ is highest? A solution is to compute $RM(p,t)$ for every word or phrase that appears in page p. Although this is easy to compute, it is not good enough because, as pointed out in the literature, often a page is an authority on some term that is not mentioned in the page. For example, the IBM Almaden Research Center (*www.almaden.ibm.com*) has a reputation on "data mining," but this phrase does not appear anywhere in its home page. [1]

Another solution is to compute $RM(p,t)$ for every possible word or phrase that appears in a page that points to p. (We call such pages "incoming links" of p.) In general, this is infeasible in the setting of the Web since there can be tens or hundreds of thousands of incoming links, and examining all those links is a cumbersome process.

We now describe the solution adopted by TOPIC [19], a prototype developed at the University of Toronto for computing Web page reputations. Given a page, the system compiles a set of incoming links of the page. This is currently done using AltaVista and Lycos, but it can be equally done using other search engines. The size of the result set is limited by the maximum number of entries returned by the search engine; search engines often return no more than 1000 entries. Clearly the accuracy of the ranks will depend on the fraction of incoming links returned. Then, for each incoming link, the system extracts words and phrases from the "snippet" returned by the search engine, rather than the page itself. This avoids the additional overhead of downloading the page, under the expectation that the snippet of a page, to some degree, is representative of the topic of the page.

There are other issues which need to be dealt with. First, links within the same site are often created for navigation purposes; as mentioned above, these links are ignored. Second, it is quite likely to find one or more near-duplicate copies of the same document in the search engine collection, even though search engines usually try to avoid storing duplicates. To address this problem, duplicate snippets are removed. Third, stop words such as "the," "for," "in," etc. and rare terms such as "BBAAA" usually convey no specific meanings (for the purpose of computing page reputations) and are removed.

3 Examples

In this section, we report our experiences with TOPIC, a prototype that uses the proposed method for reputation measurements. The input to the prototype is the URL of a page, and the output is a ranked list of topics. The default value for the number of links to download is 300, but the user can change it to a smaller number (to get better speed) or to a larger number (to get better precision). The user can enter an optional term or phrase, in which case the reputation of the page is measured only on that particular topic, and the result is displayed within a list of top 10 authorities identified by Google [8] for the same topic, providing a comparative ranking. The

[1] Disclaimer: all the examples are current as of 00/08/28, but things change fast on the Web.

default search engine for downloading the incoming links and estimating the query counts is AltaVista, but the user can change it to another engine (currently Lycos being the only alternative).

3.1 Known Authoritative Pages

For our first experiment, we selected the home pages of a set of U.S. database research groups, whose reputations we knew, from the DBLP bibliography [5]. [2] As shown in Figure 1, the results look quite reasonable. Note the high reputation of the *Microsoft Research* home page on the phrase "Data Engineering Bulletin," due to the fact that the site hosts the online version of this Bulletin.

URL : www-db.research.bell-labs.com 192 links examined (out of 193 available) **Topics**: Database Systems, Data Mining, ACM, Databases, Computer Science
URL : www.research.microsoft.com/research/db 212 links examined (out of 349 available) **Topics**: Technical Committee on Data Engineering, IEEE Data, Active Databases, Data Engineering Bulletin, Database Research, SIGMOD, SIGMOD Record, DBLP, VLDB, Database Systems
URL : www.almaden.ibm.com 999 links examined (out of 7815 available) **Topics**: IBM Almaden Research Center, Search Engines, Data Mining, Microscopy, Visualization
URL : www-db.stanford.edu 1000 links examined (out of 5637 available) **Topics**: Database research, Data Warehousing, Database Systems, Data Mining, Stanford
URL : db.cs.berkeley.edu 73 links examined (out of 130 available) **Topics**: Computer Science, Berkeley, Database, Research
URL : www.db.ucsd.edu 43 links examined (out of 78 available) **Topics**: XML, Database, Research, Project, Information

Figure 1: Selected database research group home pages and their reputations

3.2 Unregulated Web Sites

For our second experiment, we selected the home pages of a number of Canadian departments of Computer Science. The main characteristic of these sites is that they are unregulated, in the sense that users store any documents they desire in their own pages. Our goal was to find out how a site is perceived overall, without singling out specific pages stored on the site.

The results, as shown in Figure 2, can be surprising. The Computer Science Department at the University of Toronto has a high reputation on "Russian history" and "travel" mainly because a Russian graduate student of the department has put online a large collection of resources on Russia, and many pages on Russia link to it. The high reputation on "hockey" is due to a former student who used to play on the Canadian national women's hockey team. The Department of Computer Science at the University of British Columbia has a high reputation on "periodic table" because the site keeps an online version of the periodic table of chemical elements. The site also has a high reputation on "Anime" and "Manga," Japanese animation and comic art, because a staff member has put online a collection of pages on the subject and many other pages link to this collection. The reputation

[2] Disclaimer: the selection is arbitrary and not intended to be a complete list of groups of high reputation.

of the Department of Computing Science at the University of Alberta on “virtual reality” and “chess” and the reputation of the Department of Computing Science at Simon Fraser University on “data mining” and “reasoning” are to be expected. The high reputation of the CS Department at Simon Fraser University on “whales” is due to a 3D animation project being carried out on whales.

The results reported here, even though some are surprising, turn out to be consistent with other data once we examine the pages in question. For example, the Russian page at the University of Toronto reported over one million hits since 1997 for a period of two years; the number of unique visits (visits with different host names) to the page was slightly over 333,000. Similarly, Andria Hunter’s hockey page, for the period of two weeks starting from July 19, 2000 showed 2,700 average daily visits. The total number of visits since the creation of the page in March 1995 was about 1.8 million.

URL : www.cs.toronto.edu 1000 links examined (out of 8400 available) **Topics**: Russian History, Neural, Travel, Hockey
URL : www.cs.ualberta.ca 1000 links examined (out of 10557 available) **Topics**: University of Alberta, Virtual Reality, Language, Chess, Artificial
URL : www.cs.ubc.ca 999 links examined (out of 17958 available) **Topics**: Confocal, Periodic Table, Anime, Computer Science, Manga, Mathematics
URL : www.cs.sfu.ca 963 links examined (out of 2055 available) **Topics**: Whales, Simon Fraser University, Data Mining, Reasoning

Figure 2: Selected Computer Science Department home pages and their reputations

3.3 News Sites

So far, we have only ranked topics for a given site. For this purpose it does not matter whether we use $RM(p,t)$ or the penetration $P_p(t)$, since they produce the same ordering of topics for a fixed p. In this experiment, we evaluated a set of sites, all of them news providers, on a predetermined set of topics. For a fixed topic t, ranking a set of sites by their value of $RM(p,t)$ amounts to ordering them by their focus $F_t(p)$. Since we are interested in comparing both within and across sites, we show in Figure 3, both the penetration and focus for each combination of site and topic, revealing interesting patterns. For example, *CNN* has the largest penetration of any site on every topic, but relatively low focus, showing that it is well-known on all the topics but not specifically known for any single one. On the other hand, *wired.com*, while ranking a close second in penetration to *CNN* on “technology,” has substantially higher focus on this topic than any other site.

3.4 Personal Home Pages

In another experiment, we selected a set of personal home pages and used our system to find the reputation topics for each page. We expected this to describe in some way the reputation of the owner of the page. The results, as shown in Figure 4, can be revealing, but need to be interpreted with some care. Tim Berners-Lee’s reputation on the “History of the Internet,” Don Knuth’s fame on “TeX” and “Latex,” Jeff Ullman’s reputation on “database systems” and “programming languages” and Jim Gray’s reputation on “database,” “research” and “systems” are to be expected. The Comprehensive TeX Archive Network (CTAN) is frequently cocited with Don Knuth’s home page mainly within TeX information pages. Alberto Mendelzon’s high reputation on “data warehousing” and

	CNN (www.cnn.com)	BBC (www.bbc.co.uk)	ABC (www.abcnews.go.com)	Wired.com
International News	[0.0277 , 0.0170]	[0.0121 , 0.0201]	[0.0021 , 0.0279]	[0.0057 , 0.0090]
Weather	[0.0096 , 0.2228]	[0.0029 , 0.1845]	[0.0005 , 0.2338]	[0.0013 , 0.0744]
Sports	[0.0043 , 0.1724]	[0.0012 , 0.1265]	[0.0003 , 0.2531]	[0.0017 , 0.1698]
Entertainment	[0.0074 , 0.1632]	[0.0049 , 0.2913]	[0.0003 , 0.1516]	[0.0023 , 0.1294]
Travel	[0.0040 , 0.1421]	[0.0016 , 0.1548]	[0.0003 , 0.2264]	[0.0012 , 0.1082]
Technology	[0.0041 , 0.1957]	[0.0011 , 0.1390]	[0.0002 , 0.2234]	[0.0038 , 0.4560]
Business	[0.0034 , 0.2831]	[0.0018 , 0.3946]	[0.0002 , 0.3399]	[0.0022 , 0.4635]

Figure 3: [P , F] ranks of news provider sites on a set of topics

"OLAP," on the other hand, is due to an online research bibliography he maintains on these topics, and not to any merits of his own.

URL : www.w3.org/People/Berners-Lee   789 links examined (out of 1334 available) **Topics**: Tim Berners-Lee, History Of The Internet, Hypertext, Pioneers, Brief, W3C
URL : www-cs-faculty.stanford.edu/˜ knuth   1000 links examined (out of 1543 available) **Topics**: Don Knuth, TeX Users, LaTex, Linux, CTAN, Donald, Computer Science
URL : www-db.stanford.edu/˜ ullman   294 links examined (out of 423 available) **Topics**: Ullman, Database Management Systems, Database Systems, Database Design, Data Mining, Programming Languages
URL : www.research.microsoft.com/˜ gray   57 links examined (out of 74 available) **Topics**: Database, Research, Systems, Information
URL : www.cs.toronto.edu/˜ mendel   161 links examined (out of 162 available) **Topics**: Data Warehousing, OLAP, Data Mining, Bibliography, Computer Science, Database, Research

Figure 4: Selected personal home pages and their reputations

4 Conclusion

We have presented a method for evaluating the reputation of a page which is both easy to compute and, according to our preliminary tests, effective. There are some limitations to our method that should be mentioned. First, we don't exploit relationships among terms such as synonyms, generalization, specialization, etc. Second, our computation is affected by the fraction of incoming links returned by search engines and the unpredictable order in which they are returned: this affects both the choice of relevant topics and the estimation of the ratios. We are currently working on refinements of the method to overcome these limitations, as well as on more systematic evaluation of the method's effectiveness.

Acknowledgements

This research was supported by the Natural Sciences and Engineering Research Council Canada and Communications Information Technology Ontario.

References

[1] S. Brin, R. Motwani, and C. Silverstein, Beyond market baskets: generalizing association rules to dependence rules. *Data Mining and Knowledge Discovery* 2(1), pages 39–68, 1998.

[2] K. Bharat and M. R. Henzinger. Improved algorithms for topic distillation in hyperlinked environments. In *Proc. of the ACM SIGIR Conference*, pages 104–111, 1998.

[3] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. In *Proc. of the WWW7 Conference*, pages 107–117, Brisbane, Australia, April 1998. Elsevier Science.

[4] S. Chakrabarti, B. Dom, P. Raghavan, S. Rajagopalan, D. Gibson, and J. Kleinberg. Automatic resource compilation by analyzing hyperlink structure and associated text. In *Proc. of the WWW7 Conference*, Brisbane, Australia, April 1998. Elsevier Science.

[5] DBLP Bibliography. www.informatik.uni-trier.de/~ley/db.

[6] J. Dean and M. R. Henzinger. Finding related pages on the Web. In *Proc. of the WWW8 Conference*, pages 389–401, Toronto, Canada, May 1999. Elsevier Science.

[7] D. Florescu, A. Levy, and A. Mendelzon. Database techniques for the World Wide Web : a survey. *ACM SIGMOD Record*, 27(3):59–74, September 1998.

[8] Google. www.google.com.

[9] M. R. Henzinger, A. Heydon, M. Mitzenmacher, and M. Najork. Measuring index quality using random walks on the Web. In *Proc. of the WWW8 Conference*, pages 213–225, Toronto, Canada, May 1999. Elsevier Science.

[10] M. R. Henzinger, A. Heydon, M. Mitzenmacher, and M. Najork. On near-uniform url sampling. In *Proc. of the WWW9 Conference*, pages 295–308, Amsterdam, May 2000. Elsevier Science.

[11] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. In *Proc. of ACM-SIAM Symposium on Discrete Algorithms*, pages 668–677, January 1998.

[12] R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. Extracting large-scale knowledge bases from the Web. In *Proc. of the VLDB Conference*, pages 639–650, September 1999.

[13] R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. Trawling the Web for emerging cyber-communities. In *Proc. of the WWW8 Conference*, pages 403–415, Toronto, May 1999. Elsevier Science.

[14] S. Lawrence and C.L. Giles. Accessibility of information on the Web. *Nature*, 400:107–109, 1999.

[15] R. Lempel and S. Moran. The stochastic approach for link-structure analysis (SALSA) and the TKC effect. In *Proc. of the WWW9 Conference*, pages 387–401, Amsterdam, May 2000. Elsevier Science.

[16] Netscape Communications Corporation. What's related. Web page, www.netscape.com/escapes/related/faq.html.

[17] D. Rafiei and A. O. Mendelzon. What is this page known for? Computing Web page reputations. In *Proc. of the WWW9 Conference*, pages 823–835, Amsterdam, May 2000. Elsevier Science.

[18] Jon Swartz. Net rankings vex dot-coms. *USA Today*, June 2000. www.usatoday.com/life/cyber/invest/in794.htm.

[19] TOPIC. www.cs.toronto.edu/db/topic.