An Autonomous Page Ranking Method for Metasearch Engines

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

1 Introduction

Ordering search results collected from multiple sources is a challenge to metasearch engines. We present an "autonomous" ranking method, meaning one that does not depend on the individual rankings returned by the engines participating in the search. Instead, it applies our *TOPIC* method for evaluating the *reputation* of a web page on a topic [7, 6] to the problem of ranking the results to a query. (TOPIC is available at `www.cs.toronto.edu/db/topic`).

Methods for ranking query result pages based on some notion of relatedness or authoritativeness have been studied in the literature [5, 1] and implemented in commercial systems [3]. There is also work on incorporating user intentions (eg. 'current events', 'research papers' or 'individual home pages') to a metasearch engine [4].

TOPIC computes, given a page, those topics on which the page has highest reputation, by a combination of link and content analysis. In this paper, the topics are derived from the user's query; the reputation of each result page on the query topic is computed, and the value used to rank the result pages across all participating search engines, without biasing the ranking towards any of the sources.

A comparative study of link analysis algorithms has been published by Borodin et al. [2]. Our work differs from Kleinberg's approach [5] in that there is no concept of a base set that can affect the final ranking. It also differs from Page Rank [3], the ranking function used in Google, since it does not require storing a large collection of pages.

2 Proposed Method

We use two ratios for measuring the rank of a page on a search query. The *penetration* of page *p* on topic *t*, $P_p(t)$, is the fraction of pages on topic *t* that point to page *p*. It is easily estimated by dividing *I(p,t)*, the number of web pages on topic *t* that point to page *p* by *N(t)*, the number of web pages on topic *t*. (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*. It is similarly estimated by dividing *I(p,t)* by *In(p)*, the in-degree of page *p*. These quantities can be interpreted as probabilities: $P_p(t)$ is the probability that an arbitrary page on topic *t* points to *p*, and $F_t(p)$ is the probability that an arbitrary page pointing to *p* is on topic *t*. If one interprets the incoming links of a page as a retrieved set from the target set of pages on the topic, definitions of penetration and focus respectively correspond to those of recall and precision.

Values of $In(p)$, $I(p,t)$ and $N(t)$ can be estimated, for example, by respectively sending queries "+ml:p," "+ml:p +t" and "+t" to Lycos (*lycospro.lycos.com*) and retrieving the counts returned by the engine (we call these *count* queries).

We use penetration and focus as follows. Results of a search query are ordered according to their penetration ratio on the query term (we assume the search query is a single term or a phrase; the discussion of more general forms of queries is outside of the scope of this paper). The focus ratio can be used to identify pages that are not closely relevant to the search query. This is used to filter out well-linked pages such as *www.infospace.com* and *www.yahoo.com* that may acquire high penetration ranks independently of the query term.

3 Experiments

We built a prototype metasearch engine, called *TOPICsearch*, (`www.cs.toronto.edu/db/topic/search.html`) that uses the penetration ratio for ordering the results. TOPICsearch sends the user's query to several search engines (currently Alta Vista, Lycos and Google) and collects the results. The penetration ratio is computed for each page using count values retrieved from an engine chosen independently by the user (currently Alta Vista and Lycos are supported). Results are ordered by their penetration ratios before being presented to the user.

In our experiments we did not use the metasearch capability: we simply sent a query term to a particular search engine, say *S*, retrieved its top k answers for various values of k, reordered them by their penetration ratio on the query topic, and compared the

resulting ranking with the ranking returned by *S*. We used a list of 471 frequently typed queries, provided to us by a major search engine company, and tested two highly popular search engines in the role of *S*. We manually examined some of the pages that were pushed up or down by 5 positions or more in our ranking relative to *S*'s ranking; the anecdotal evidence suggests that our results are often substantially better than the original ordering.

For example, the three top-ranked pages by our engine for query 'Century 21' were (1) the Century 21 home page [1], (2) their branch in Dover, New Hampshire [2], and (3) their branch in Phoenix, Arizona [3]. The same pages were respectively ranked 35th, 7th and 19th by *S*. The three top-ranked pages by *S* were instead an inaccessible page [4], the branch in Northhampton, Massachusetts [5], and the branch in Pawtucket, RI [6].

For the query 'clickable map,' our method placed 'clickable map of the world' [7], 'Alabama clickable map' [8] and 'USA climate page' [9] as the top three pages. These pages were ranked 8th, 4th and 18th by *S*. *S* placed as its top three: 'Texas clickable map' [10], 'Texas lakes clickable map' [11], and an inaccessible Japanese site [12].

The results of these experiments for one particular engine, including all 471 queries, with k set to 50, are available at `www.cs.toronto.edu/~ mendel/S.html`.

Acknowledgements

This research was supported by the Natural Sciences and Engineering Research Council of Canada.

References

[1] K. Bharat and M.R. Henzinger. Improved algorithms for topic distillation in hyperlinked environments. In *Proceedings of the 21st International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 104–111, 1998.

[2] A. Borodin, G.O. Roberts, J.S. Rosenthal, and P. Tsaparas. Finding authorities and hubs from link structures on the world wide web. In *Proceedings of the 10th International World Wide Web Conference*, pages 415–429, Hong Kong, May 2001. Elsevier Science.

[3] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. In *Proceedings of the 7th International World Wide Web Conference*, pages 107–117, Brisbane, Australia, April 1998. Elsevier Science.

[4] E.J. Glover, S. Lawrence, M.D. Gordon, W.P. Birmingham, and C.L. Giles. Web search - your way. *Communications of the ACM*, 44(12):97–102, 2001.

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

[6] A.O. Mendelzon and D. Rafiei. What do the neighbours think? Computing Web page reputations. *IEEE Data Engineering Bulletin*, 23(3):9–16, 2000.

[7] D. Rafiei and A.O. Mendelzon. What is this page known for? Computing Web page reputations. In *Proceedings of the 9th International World Wide Web Conference*, pages 823–835, Amsterdam, May 2000. Elsevier Science.

[1] www.ctc21directory.com
[2] www.century21hometeam.com
[3] www.century21metro.com
[4] www.century21pro.com
[5] www.century21pva.com
[6] www.c21bk.com
[7] southport.ipl.nasa.gov/imagemaps
[8] www.eng.auburn.edu/alabama/map.html
[9] www.cdc.noaa.gov/USclimate/states.fast.html
[10] www.tpwd.state.tx.us/maps/clicpark.html
[11] www.tpwd.state.tx.us/fish/infish/regions/instate.htm
[12] home.ntt.com/japan/map/