Searching the Web Through User Information Spaces

Athanasios Papagelis[1] and Christos Zaroliagis[1,2]

[1] Department of Computer Engineering and Informatics,
University of Patras, 26500 Patras, Greece
[2] Computer Technology Institute, P.O. Box 1122, 26110 Patras, Greece
papagel@ceid.upatras.gr, zaro@ceid.upatras.gr

During the last years web search engines have moved from the simple but inefficient syntactical analysis (first generation) to the more robust and usable web graph analysis (second generation). Much of the current research is focussed on the so-called *third generation* search engines that, in principle, inject "human characteristics" on how results are obtained and presented to the end user. Approaches exploited towards this direction include (among others): an alteration of PageRank [1] that takes into account user specific characteristics and bias the page ordering using the user preferences (an approach, though, that does not scale well with the number of users). The approach is further exploited in [3], where several PageRanks are computed for a given number of distinct search topics. A similar idea is used in [6], where the PageRank computation takes into account the content of the pages and the query terms the surfer is looking for. In [4], a decomposition of PageRank to basic components is suggested that may be able to scale the different PageRank computations to a bigger number of topics or even distinct users. Another approach to web search is presented in [2], where a rich extension of the web, called semantic web, and the application of searching over this new setting is described.

In this work we depart from the above lines of research and propose a new conceptual framework for representing the web and potentially improving search results. In particular, the new framework views the web as a collection of web-related data collected and semi-organized by individual users inside their information spaces. These data can be explicitly collected (e.g., bookmarks) or implicitly collected (e.g., web-browsing history). Our approach is based on the observation that users act as small crawlers seeking information on the web using various media (search engines, catalogs, word-of-mouth, hyperlinks, direct URL typing, etc). They tend to store and organize important-for-them pages in tree-like structures, referred to as *bookmark collections*, where the folder names act as tags over the collected URLs. This method of organizing data helps people to recall collected URLs faster, but can also be used as a kind of semantic tagging over the URLs (the path to the URL can be perceived as different ways to communicate the URL itself). This information constitutes part of the user's *personal information space* and it is indicative of his interests. One might argue that people do not collect bookmarks or that they do not organize them in any reasonable manner. However, as our experiments show, people indeed collect and organize bookmarks under certain patterns that follow power law distributions.

M. Kitsuregawa et al. (Eds.): WISE 2005, LNCS 3806, pp. 611–612, 2005.

The proposed framework has been materialized into a hybrid peer-to-peer system, which we call *Searchius* (`http://searchius.ceid.upatras.gr`), that produces search results by strictly collecting and analyzing bookmark collections and their structures. Conceptually, Searchius can be positioned somewhere between search engines and web catalogs.

Searchius can be easily expanded and updated in an ad-hoc manner through asynchronous connections initiated by end-users. It can overcome shortcomings of algorithms based on link analysis (e.g., web islands), where information unreferenced by other sites is not being indexed. Moreover, Searchius is not capital intensive, since it concentrates on a small portion of the data that typical search engines collect and analyze. However, the collected data is of the highest interest for users. To order pages by importance, Searchius uses an aggregation function based on the preference to pages by different users, thus avoiding the expensive iterative procedure of PageRank. This allows for efficient implementations of several personalization algorithms. Finally, the way people organize their bookmarks can be used to segment the URL space to relative sub-spaces. This property can be exploited to provide efficient solutions to additional applications, including the construction of web catalogs and finding related URLs. Note that Searchius does not collect any personal data from end-users.

We have also conducted an extensive experimental analysis of the characteristics of bookmark collections and a comparative experimental study with Google to measure the quality of the search results. Our experiments with a collection of bookmark sets from 36.483 users showed that: (i) people collect and organize bookmarks in ways which follow power law distributions; (ii) there are diminishing returns on the rates that (distinct) URLs and search keywords are discovered; (iii) the number of bookmarks that a given URL has accumulated is linear to the number of users; (iv) important URLs are discovered early on the database construction phase; and (v) there is a consistently significant overlap between the results of Searchius and those produced by Google. More details on the experimental results as well as on Searchius can be found in [5].

References

1. S.Brin, R.Motwani, L.Page, and T.Winograd. What can you do with a web in your pocket. In Bulletin of the IEEE Computer Society Technical Committee on Data Engineering, 1998.
2. R.Cuha, R.McCool, and E.Miller. Semantic Search. In Proc. *World Wide Web Conference*, 2003.
3. T.Haveliwala. Topic Sensitive Page Rank. In Proc. *World Wide Web Conference*, 2002.
4. G.Jeh and J.Widom. Scaling Personalized Web Search. In Proc. *World Wide Web Conference*, 2003.
5. A. Papagelis and C. Zaroliagis. Searching the Web through User Information Spaces. CTI Tech. Report TR 2005/09/01, September 2005.
6. M. Richardons and P. Domingos. The Intelligent Surfer: Probabilistic Combination of Link and Content Information in Page Rank. Volume 14. MIT PRess, Cambridge, MA, 2002.