Characterizing Web Resources for Improved Search

Position Paper

Luis Gravano
Computer Science Department
Columbia University
http://www.cs.columbia.edu/˜gravano
gravano@cs.columbia.edu

Web resources are extremely diverse, not only along every conceivable topical and non-topical dimension, but also in terms of the access interface that they present to users. Current search engines ignore crucial non-topical dimensions of web resources that could be used to improve the quality of query results.

As an important initial step to exploit such dimensions for web search, we have focused on geographical relevance. Web sites containing information on restaurants or apartment rentals, for instance, are relevant primarily to web users in geographical proximity to these locations. In contrast, an on-line newspaper may be relevant to users across the United States. We have studied how to mine the web and automatically estimate the geographical scope of web resources by using web hyperlinks and the actual content of web pages. For example, we can map every web page to a location based on where its hosting site resides. Then, we can consider the location of all the pages that point to, say, the Stanford Daily home page. By examining the distribution of these pointers, we can conclude that the Stanford Daily is of interest mainly to residents of the Stanford area, while The Wall Street Journal is of nation-wide interest. Similar conclusions can be drawn for other resources by analyzing the geographical locations that are mentioned in their pages.

We have developed and evaluated algorithms for computing geographical scopes, described in [2], and implemented a geographically-aware search engine for on-line newspapers, accessible at `http://www.cs.columbia.edu/˜gravano/-GeoSearch`. Figure 1 shows a screenshot of the search engine with the results for query "startups business" for a user with zip code 94043, which corresponds to Mountain View, California. The first article is from The Nando Times, a national on-line newspaper. Our system has determined that this newspaper's geographical scope is the whole United States, hence the coloring of the map next to the corresponding article. The second article returned is from the San Jose Mercury News, a newspaper based in San Jose, California, whose technology reports have followers across the country. Our search engine has classified this newspaper as having a national geographical scope as well. The last article returned originated in a newspaper whose geographical scope consists of the entire state of California, which is marked with a solid color on the map, plus a few cities scattered across the country, indicated by placing a dot in their corresponding states [2].

More generally, given the web's heterogeneity, it becomes crucial to extract and exploit attributes of web documents that do not directly relate to the subject being discussed and are still important for satisfying access needs. Examples of such attributes include:

- A document's popularity, "authoritativeness," and general reputation (e.g., [1, 3, 4, 5]).
- A document's structure and type (e.g., a newspaper article vs. a reference article vs. an informal discussion of the same subject).
- A document's linguistic complexity, sublanguage, and intended audience (e.g., medical information for patients vs. journal articles for physicians on the same disease or treatment).
- A document's geographical and temporal scope (e.g., news articles in a local vs. national newspaper).

A promising and potentially influential research direction to improve the quality of web search engines is to continue to investigate how non-topical information, including but not limited to the types given above, can be efficiently extracted from documents and how such information affects document relevance to specific users' queries and document quality.

Acknowledgments
This position paper describes past and on-going work performed jointly with Junyan Ding, Vasilis Hatzivassiloglou, Narayanan Shivakumar, and others.

REFERENCES

1. Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. In *Proceedings of the Seventh International World Wide Web Conference (WWW7)*, April 1998.
2. Junyan Ding, Luis Gravano, and Narayanan Shivakumar. Computing geographical scopes of web resources. In *Proceedings of the Twenty-sixth International Conference on Very Large Databases (VLDB'00)*, September 2000.

Figure 1: Search results from our geographically-aware search engine.

3. Monika Henzinger. Link analysis in web information retrieval. *IEEE Data Engineering Bulletin, Special Issue on Next-Generation Web Search*, 23(3), September 2000.

4. Jon Kleinberg. Authoritative sources in a hyperlinked environment. In *Proceedings of the Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 668–677, January 1998.

5. Davood Rafiei and Alberto Mendelzon. What is this page known for? Computing Web page reputations. In *Proceedings of the Ninth International World-Wide Web Conference (WWW9)*, May 2000.