

Web Information Retrieval

CSCI335.01 Research Paper

Fall 2006

Assigned: Monday October 23, 2006

Proposal due: Monday October 30, 2006

Information update due: Monday November 6, 2006

Paper due: Monday November 20, 2006

Presentation: November 27, 29, December 1, 2006

Now that you have some basic knowledge and experiences with information retrieval and one of its most popular applications to-date – web search, it is time to read some more research papers in one or more particular areas of IR and web search to gain a further understanding of the subject. You are asked to select an area of interest and read 3 - 5 research papers about the subject, write a survey paper and present your findings to the class. The final paper should be anywhere between 5 to 15 pages, as long as the topic is clearly presented and necessary background information covered. You may choose a in-depth study of a particular topic. Or you may choose to do a survey in a broad subject area.

You are encouraged to work in group of 2 - 3 students. The maximum number of students in a group is 4. You may choose to work individually. Regardless of the size of the group, each person will participate in the final oral presentation. The presentation for the group can be expected about 15 to 25 minutes long.

You are given a week of time to determine a subject and the group. I have a list of suggested topics below. You may choose other related topics as well. By the proposal due date, you are asked to submit a brief proposal indicating the topic, group members, and a list of three to five potential reference papers. The list should include the title and the author(s) of each paper, where and when each paper is published. You are not expected to read them all by then. But you should have done some basic search and start to collect some reference papers.

When *information update* is due, you are asked to update the reference list. By then you should have read the papers you want to reference. You may find some of the earlier references are not quite suitable and you decide to drop them. You may find some new references. Regardless the case, you are asked to submit an updated list of references.

When the final version of the paper is due, the reference list may be further updated as necessary.

Please refer to the research paper guidelines listed in the course web page
<<http://www.eg.bucknell.edu/~csci335/index.html#paper-guide>>.

Here is a suggested list of topics in IR and web search and possible starting research papers. The papers quoted are not necessarily the most important paper in the field. It is just a relevant paper. You are supposed to follow this initial paper and find other papers in the field. (It is almost like a starting URL, you need to index other papers from this starting point. Use the *hub* and *authority* model to expand the paper collections.) The list is in no particular order of importance or interests. You are certainly

encouraged to find your own interest area and research papers.

1. **Web crawling:** How to gather data from the web efficiently and less-intrusively? “Efficient Crawling Through URL Ordering” by J. Cho, H. Garcia-Molina, and L. Page, in *Proceedings of 7th World International Wide Web Conferences*, Brisbane, Australia, 14-18 April 1998
2. **Topic-specific search engines:** How to build topic-specific search engines? “AGORA: A Search Engine for Software Components”, *IEEE Internet Computing* vol. 2, no. 6, Nov/Dec 1999
3. **Mining the usage log:** How to collect and analyze web access log data? “In Search of Reliable Usage Data on the WWW” by J. Pitkow, in *Proceedings of the 6th International World Wide Web Conferences*, Santa Clara, California, USA, 7-11 April 1997
4. **Internet agents:** How does an Internet agent work? “Intelligent Agents on the Internet: Fact, Fiction, and Forecast”, by Oren Etzioni and Daniel S. Weld, in *IEEE Expert*, August 1995, pp. 44-49
5. **Meta-search engines:** What are they? How do they work? “Architecture of a Metasearch Engine that Supports User Information Needs”, by Eric J. Glover, Steve Lawrence, William P. Birmingham and C. Lee Giles, in the *Proceedings of The ACM Conference on Information and Knowledge Management (CIKM) '99* pp 210-216, Kansas City, MO, USA, November 1999
6. **Recommendation system:** We can use the web to get recommendations for movies, books, and others. How do these types of systems work? “An Adaptive Web Page Recommendation Service” by Marko Balabanovic, in *Proceedings of the First International Conference on Autonomous Agents*, Marina del Rey, CA, USA, February 1997
7. **Web’s link structures:** How are web pages linked to each other? How can we make use of the link structures? “Mining the Web’s Link Structure”, by Soumen Chakrabarti and others, *IEEE Computer*, pp. 60-66, August 1999

Last but not the least, check other ACM and IEEE journals and conferences for papers and ideas. And don’t forget the WWW conference series at <http://www.eg.bucknell.edu/~csci335/2006-fall/wwwconf.html>.

Use your library and web search skills to search for other relevant information. The annual WWW conference is an important source. If you have some paper in mind, you may find other citation information from *citeseer* <http://citeseer.ist.psu.edu/>.

Another good computer science literature resource page is <http://www.vldb.org/dblp/db/>