

Link Analysis For Collaborative Knowledge Building

Harris Wu, Michael D. Gordon, Kurt DeMaagd, Nathan Bos

University of Michigan

701 Tappan Street

Ann Arbor, MI 48197, USA

01-734-647-7667

{harriswu,mdgordon,demaagdk,serp}@umich.edu

ABSTRACT

We present an ongoing research project utilizing navigation and hyperlink data to aid collaborative knowledge building. We allow collaborators to personally organize documents and other research resources and make references to them. We combine their personal organizations and references to develop a unified, hierarchical categorization of these resources. We analyze collaborators' navigations to identify prominent research activities as well as the key documents related to these activities. We examine prominence over time to identify research trends.

Categories and Subject Descriptors

H.5.4 [Information Interfaces and Presentation]: Hypertext/Hypermedia - navigation.

General Terms

Algorithms, Management, Design, Human Factors.

Keywords

Navigation analysis, link analysis, knowledge management.

1. INTRODUCTION

Much hyperlink research has been done to improve the performance of a search engine or usability of a web site. Broader hyperlink research has been identified as necessary for knowledge management issues [5]. In our research, we develop an approach to support researchers working on an emerging research topic. By analyzing how researchers link documents and navigate among them, we gain the ability to organize a research corpus by concept and improve researchers' access to others' complementary resources. In the following we first describe the problems that our research website approach addresses. We present our data collection and analysis techniques. We contrast our hierarchical hyperlink analysis and Markov-chain based navigation analysis with previous research. We conclude the paper with future work.

2. RESEARCH WEBSITE

A group of researchers has been building up a body of research around how information and communication technology relates to social responsibility, an important but relatively undefined research area. The group is working on related but distinct aspects of this question using an online collaboration tool

developed using Lotus Notes (Domino) to organize their working documents and electronic source materials including uploaded documents and URLs. Researchers with different perspectives have been free to contribute to existing directories or create new directories. As a result, hundreds of resources have been stored in about 30 public directories and with no clear boundaries or hierarchy among them. It has also been difficult to identify major areas of research activities within the online research community. Since the research area is a burgeoning, dynamic area, the focus areas of research has shifted over time. These issues had started to severely hamper the collaborative research, as was made evident by researchers' complaints. Our project thus was aimed at answering these complaints. Our aim was to allow each researcher to have a personal view on the overall collection; to consolidate these personal views into one more representative of the collection's overall structure; and to identify current foci and trends within the collection over time.

We constructed a website (elab-linux3.bus.umich.edu) for these researchers as an experiment using link analysis to address these issues. In the new website, working documents, electronic research resources (word and pdf documents and URLs), and researchers' homepages are all called *nodes*. Researchers can create, modify, and comment on nodes, and organize them into "buckets" (supernodes) that link to a list of other nodes. Researchers can also organize nodes by attaching the nodes as citations to their working documents. Researchers also vote on nodes' usefulness and bookmark nodes so they can be accessed conveniently. All nodes in the website are dynamically served by server-side scripts. The server-side scripts capture detailed information about researcher's navigation sessions, store users' node actions and transform them to hyperlinks when presenting the nodes. For example, citations result in hyperlinks in the working documents, and bookmarking results in hyperlinks on researchers' homepages. Navigations are captured as quadruples (session with user ID, from node, to node, timestamp) in a database.

We apply hierarchical cluster analysis to these researcher-made associations (represented by hyperlinks) in order to obtain a hierarchical clustering of all resources within the website. The associations used for our analysis are either from researchers' organization of nodes into buckets or from the citations they attach to their working documents. We place the associations into a matrix with documents as rows and buckets or citing documents as columns. The resulting matrix is similar to a document-keyword matrix and can be analyzed using cluster analysis [6]. We perform hierarchical cluster analysis on the matrix using SPSS®. The resulting hierarchy of document clusters functions as

a unified, overall categorization of documents which overlooks the idiosyncrasies of individual researchers. It allows researchers to zoom in and out using the hierarchy. While we expect the hierarchy to be particularly useful for newcomers to the website, it is also useful for existing researchers to compare their own personal perspective with the unified view. The hierarchy also makes it easier for existing researchers to identify resources that they have overlooked, or resources in unfamiliar subtopics.

In addition to the hierarchical document clustering just described, we apply Principal Clusters Analysis from our earlier work [7] to navigation data in order to identify the most prominent conceptual *navigation clusters* as well as the most important documents within these clusters. The navigation data is collected in a navigation frequency matrix where “from nodes” and “to nodes” are rows and columns, respectively, and the frequencies of navigation between node pairs are the entries in the matrix. Principal Clusters Analysis is an SVD (singular value decomposition) based technique. As a data reduction technique it tells with what percentage the most prominent clusters have represented users’ navigation activities. The resulting clusters represent focus areas of research activities, or topics. Principle Clusters Analysis does short-run equilibrium analysis for Markov chains, which is appropriate for Web navigations, which have a limited number of steps. We treat navigations as Markov chains with documents as states and normalized navigation frequencies as transition probabilities between the states. The equilibrium analysis produces key starting and ending points of navigations, or hubs and authorities [3]. Analysis results are represented by a number of “clouds” representing navigation clusters, with links of various strengths between them indicating the likelihood of inter-cluster navigations. Each cloud contains the links to the key hubs and authorities within that navigation cluster. Dynamic navigation data allow us to analyze navigations for different time periods, thus researchers not only have a high-level view of recent research activities, but also see past research activities to observe the trend of the research. Studying individuals’ navigations also allow us to identify “who is doing what”, which can suggest possible collaboration opportunities.

3. RELATED WORK

In [4], the goal of navigation analysis is index page synthesis: the automatic creation of navigational pages that consist of a set of links on a particular topic. A similarity matrix is created from co-occurrence frequencies between documents in the same user session. Entries below a threshold value are removed from the matrix, then maximal cliques or connected components found from this sparse matrix are clusters. This analysis tends to create small closely coupled document clusters with no particular order of importance. All the weak links, known to be critical in networks [2], are thrown out by the threshold-based data reduction.

In [8], the goal of navigation analysis is to predict and assist users’ navigation on a website. A hierarchical clustering algorithm finds three kinds of conceptual clusters based on similarities in in-links, similarities in out-links, and similarities in both. Given a user’s current page and a set of visited pages and clusters as history, a Markov model is used to predict the most likely next

steps of the user. Analysis in [8] produces three conflicting cluster hierarchies.

Both [4] and [8] leave some pages unclustered. These analyses are limited in helping online collaborators, who typically want a high-level view of important research activities, some starting points and key resources, and an inclusive categorization.

Previous research including [4] and [8] have used Web access logs to reconstruct navigations. Each log record typically contains the timestamp, the URL and originating IP address. Hits from an IP address within a time frame are assumed to come from a single user session for a certain task. There are several issues with using Web logs. User requests may have been channeled through the same IP address. Users may have multiple sessions each for a different task. It is difficult to justify what time frame constitutes a user session. Not all navigations are captured due to browser caching. In our research we use embedded coding in server-side scripts to capture the whole navigation sequence of a user session, which addresses all the aforementioned issues with web logs.

4. CONCLUSION AND FUTURE WORK

Online communities have shown to be promising for knowledge co-creation [1]. Our website gives researchers a personal perspective (personal organization of resources), a unified perspective (synthesized hierarchy of resources) and a dynamic perspective (prominent research activities and trends) of research. Initial feedback from user interviews is very positive. We plan to evaluate our analysis results using researchers’ votes and bookmarks, along with online questionnaires. We also plan to capture navigations from a group of users using a proxy server to tunnel their HTTP requests. Using this approach we can capture users’ navigations to any websites on the Internet. The proxy approach is best for a collocated team but should also be viable for distributed collaborators.

5. REFERENCE:

- [1] Gordon, M., Fan, W., Rafaeli, S., Wu, H., Farag, N. The architecture of commKnowledge. Intl. Journal of Electronic Business, 1(1), pp. 69-82.
- [2] Granovetter, M. (1973). The strength of weak ties. American Journal of Sociology, 78(6), 1360-1380.
- [3] Kleinberg, J. M., (1999). Authoritative sources in a hyperlinked environment. Journal of ACM, 46: 604-632.
- [4] Perkowitz, M., Etzioni, O. (2000). Towards adaptive Web sites. Artificial Intelligence 118: 245-275.
- [5] Ricardo, F. J. (2001). Hypertext and Knowledge Management. In Proc. of ACM Hypertext’01, pp. 217-226.
- [6] Van Rijsbergen, C.J. Information Retrieval. Butterworths, London, 1979.
- [7] Wu, H., Gordon, M., Demaagd, K., Fan, W. (2002). Principal Clusters Analysis. In Proc. of 18th Intl. Conf. on Information Systems, Barcelona, Spain, Dec 2002.
- [8] Zhu, J., Hong, J., and Hughes, J. (2001). Using Markov Models for Web Site Link Prediction. In Proc. of ACM Hypertext’02, pp. 131-13.