Extending Retrieval with Stepping Stones and Pathways

IIS-0307867

Principal Investigator

Edward A. Fox
Dept. of Computer Science
Virginia Polytechnic Institute and State University (Virginia Tech)
M/C 0106, 660 McBryde Hall
Blacksburg, Virginia 24061
Phone: (540) 231-5113
Email: fox@vt.edu
http://fox.cs.vt.edu

Keywords

Information Retrieval
Document Classification
Literature-based Discovery
Data Fusion

Project Summary

Millions of users of information retrieval systems each day seek useful items, but many are frustrated with the results received. Often, better results can be found when a query is split into parts that cover different but connected aspects of the information need. Sometimes a small set of documents is needed as an answer instead of a single document.

This project researches an alternative interpretation of user queries and presentation of the results. Instead of returning a ranked list of documents, the result of a query is a connected network of chains of evidence. Each chain is made of a sequence of additional concepts (**stepping stones**). Each concept in the sequence is logically connected to the next and previous one, and the chains provide a rationale (a **pathway**) for the connection between the two original concepts. To increase the user's understanding of the chain, it is desirable that the stepping stones be justified by concrete documents, along with the connections (relationships) among those documents.

This approach has the potential to improve retrieval results whenever there is a mismatch between a user's understanding of the collection and the actual collection content. A probabilistic retrieval scheme is employed with: (1) a framework based on belief networks which combine multiple sources of evidence; and (2) user feedback at document, cluster (group), and relationship (e.g., citation or hypertext link) levels.

Query results in this interpretation are networks of document groups representing topics, each group relating and connected to other groups in the network that partially answer the user's information need. New and more effective representations and techniques help users visualize these results. The user is part of the retrieval process, and can manipulate the network of content. The system can provide deeper support of the user's need, in a way that goes well beyond traditional relevance feedback.

Project Impact

This project will involve scores of students (including those in minority and under-represented groups) in information retrieval studies, leading to several world-accessible electronic theses and dissertations, and improved search methods. The same technique used to harvest the test collection can be used to harvest documents from a group of seed documents, where the seeds documents are selected by the teacher from important research papers in important topics for the course syllabus. Stepping stones and pathways can suggest intermediate topics and documents connecting the topics provided by the teacher though the seed papers. Citation analysis and metadata can provide highly related papers, both as alternatives and complements to the ones provided by the teacher. Because of this research, scientists and the public in general should be able to more easily search for answers to complex questions. Software and collections developed will be available to other researchers. Results will be shared at workshops and conferences.

Goals, Objectives and Targeted activities.

The first year of this project will concentrate on creating and enhancing an effective retrieval system for small/medium collections where our hypothesis may apply. It will study the application of our query interpretation to both test collections and live, publicly available ones. The study of these collections will lead to a better understanding of the characteristics of the queries that can be answered by our technique, regarding parameters like minimum length, number of topics, and generality of terms. The retrieval model will be refined to include other sources of information and the effect of source dependencies on retrieval. It will also study users regarding the effectiveness of the method as compared to traditional retrieval systems, when allowing a first level of user feedback to modify the resulting network. We will disseminate the results of year 1 at IR and digital library conferences, like ACM SIGIR and JCDL/ECDL/ICADL.

The second year will concentrate on studying our query interpretation using big collections, including the tradeoffs between interactive retrieval and the construction of high-quality networks. Building on the result of user studies in year 1 the user interface will be redesigned to implement full user feedback and improve the user's understanding of the collection, and will evaluate the best representation in terms of stepping stones / link labeling. The results of the project will be publicly available as a software package, as papers describing the user studies and scalability approaches, and as a worldwide-accessible electronic dissertation completed by graduate research assistant Fernando Das-Neves.

Area Background

There area of information retrieval includes many approaches to combining evidence to improve the matching between queries and documents. They try to capture the query "concept" and document semantics, in different ways, in each method of interpretation of documents (mostly unstructured). However, using document structure and connections can improve retrieval. Structural information (links, citations, etc.) is today on par with content, and cannot be ignored. Google, the most successful search engine today, utilizes linking as the most important mechanism to decide the ranking of a relevant document, and ResearchIndex [Lawrence99] uses reference counting to rank matching documents. How to combine and take advantage of richer structure and multiple sources of information also has been explored in IR under the name of data fusion.

Building connections between apparently independent topics has been studied in the past under the name of "literature-based discovery". These studies proved that it is possible to build meaningful connections among seemingly unrelated concepts or document sets. The main difference between these studies and ours is that they depend on the extensive use of a pre-existing classification system and an accurate classification of documents. Other approaches that relied more on free-form text needed the involvement of user experts who know a great deal about the connections sought after. Furthermore, chains of relationships previously studied have been limited to one intermediate step, with very specific types of relations, like for example, illness→ (symptom, effect) ←treatment.

Area References

Dasigi, V. *Information Fusion Experiments for Text Classification*. In Proceedings of the 1998 IEEE Information Technology Conference, pp. 23-26, 1998.

Dean, J., Henzinger, M. R. *Finding related pages in the World Wide Web*. In Proceedings of the Eighth International World Wide Web Conference, 1999.

Fox, E., and Shaw, J. *Combination Of Multiple Searches*. In Proceedings of the Second Text Retrieval Conference *(*TREC-2), pp. 243-252, 1993.

Gordon, M., Dumais, S. *Using Latent Semantic Indexing for Literature Based Discovery*. Journal of the American Society for Information Science and Technology, Vol. 49, Issue 8, pp. 674-685, 1998.

Haines, D., Croft, B. *Relevance feedback and inference networks*. In Proceedings of SIGIR'93, pp. 2-11, 1993.

Kleinberg, J. *Authoritative Sources in a Hyperlinked Environment.* In Proceedings of ACM-SIAM Symposium on Discrete Algorithms, pp. 668-677, 1998.

Kindsay, R., Gordon, M. *Literature-Based Discovery by Lexical Statistics*. Journal of the American Society for Information Science and Technology, Vol. 50, Issue 7, pp. 574-587, 1999.

Modha, D., Spangler, W. S. *Clustering hypertext with applications to Web Searching*. In Proceedings of Hypertext 2000, pp. 143-152. 2000.

Silva, I., Ribeiro-Neto, B., Calado, P., Moura, E., Ziviani, N. *Link-based and content-based evidential information in a belief network model*. In Proceedings of SIGIR 2000, pp. 96-103, 2000.

Swanson, D. *Two Medical Literatures that are logically but not Bibliographically connected*. Journal of the American Society for Information Science and Technology, Vol. 38, Issue 4, pp. 228-233, 1987.

Swanson, D. *Complementary Structures in Disjoint Scientific Literatures*. In Proceedings of SIGIR'91, pp. 280-289, ACM Press, 1991.

Swanson, R., Smalheiser, N, Bookstein, A. *Information Discovery from Complementary Literatures: Categorizing Viruses as Potential Weapons*. Journal of the American Society for Information Science and Technology, Vol. 52, Issue 10, pp. 797-812, 2001.

Turtle, H., Croft, W. *Inference networks for document retrieval*. In Proceedings of SIGIR'90, pp. 1-24, 1990.

Webber, M. Vos, R, Klein, M., and de Jong-van den Berg. *Using Concepts in literature-based discovery: simulating Swanson's Raynaud-fish-oil and migraine-magnesium discoveries*. Journal of the American Society for Information Science and Technology, Vol. 52, Issue 7, pp. 548-557, 2001.

Project Website

http://fox.cs.vt.edu/SSP/