Communications of the IIMA

Volume 4 | Issue 4 Article 4

2004

Semi-Automatic Query Expansion Approach to Web- Based Information Retrieval

ChaoYang Zhang
University of Southern Mississippi

Kuo Lane Chen
University of Southern Mississippi

Huei Lee
Eastern Michigan University

Hong Lan
Louisiana Tech University

QiJun Chen
University of Vermont

See next page for additional authors

Recommended Citation

Zhang, ChaoYang; Chen, Kuo Lane; Lee, Huei; Lan, Hong; Chen, QiJun; and He, JiangYan (2004) "Semi-Automatic Query Expansion Approach to Web- Based Information Retrieval," *Communications of the IIMA*: Vol. 4: Iss. 4, Article 4.
Available at: http://scholarworks.lib.csusb.edu/ciima/vol4/iss4/4

Semi-Automatic Query Expansion Approach to Web- Based Information Retrieval

Authors
ChaoYang Zhang, Kuo Lane Chen, Huei Lee, Hong Lan, QiJun Chen, and JiangYan He

Semi-Automatic Query Expansion Approach to Web-Based Information Retrieval

ChaoYang Zhang

Department of Computer Science and Statistics, University of Southern Mississippi, Hattiesburg, MS 39406
(601)266-5510, Fax: (601)266-6452, chaoyang.zhang@usm.edu

Kuo Lane Chen

School of Accountancy and Information Systems, University of Southern Mississippi, Hattiesburg, MS 39406
(601)266-5954, Fax: (601)266-4642, chenku60@yahoo.com

Huei Lee

Department of Computer Information Systems, Eastern Michigan University, Ypsilanti, Michigan 48197
(734) 487-4044, Fax: (734)487-1941, huei.lee@emich.edu

Hong Lan

Department of Computer Science, Louisiana Tech University, Ruston, LA 430072
hla002@latech.edu

QiJun Chen

Department of Computer Science, University of Vermont, Burlington, VT 05405
qchen@cs.uvm.edu

JiangYan He

Department of Computer Science, University of Vermont, Burlington, VT 05405
jhe@cs.uvm.edu

ABSTRACT

The query used for Web searching is usually short and may not be able to reflect the intrinsic semantics of the user information need. The purpose of the paper is to take into account user information feedback, and to develop a semi-automatic query expansion approach to improve the effectiveness of Web searching. A search engine has been developed using the vector information retrieval model to validate the semi-automatic query expansion approach. The experiments show that this approach may improve the effectiveness of web searching.

INTRODUCTION

Unlike data retrieval from database which aims at searching all objects that satisfy clearly defined conditions such as those in a regular expression or in a relational algebra expression, Web searchers emphasize on retrieving all Web pages satisfying the user information need from a large collection of Web pages that are not always well-structured and may be semantically ambiguous. A carelessly chosen query may not be able to find the valuable information. The Web pages returned by the Web search engine may contain the same words as the query but they are not relevant to the user information need. The searcher may not exactly understand the meaning of searching using a set of words and the user-specified words may not reflect the intrinsic semantics of text, which makes query formulation and Web searching frustrating sometimes.

Many current search engines often provide advanced query operators in the user interface. Advanced query operators may be helpful for effective searching. However, the new research has reported that generally the query operators provide little or no benefit, and moreover, they are counter productive in some cases (Eastman & Jansen, 2003). Only 10% of Web searchers utilize advanced query operators in their Web searching. Most Web searchers have problems with Boolean logic and only use simple and short query for Web searching. The average query submitted is only two (or three) words long. In addition, the size of the Web increases dramatically and search engines can search a large collection of Web pages, e.g. Google can search 4.28×10^9 Web pages. Without detailed knowledge of collection make-up and of retrieval environment, most users find it difficult to formulate queries which are well designed for Web searching. This difficulty motivates us to develop techniques to expand the query automatically or semi-automatically so that it can better reflect the user information need and hence improve the effectiveness of Web searching.

Several techniques for automatic query expansion have been proposed, such as automatic local analysis and automatic global analysis. Automatic global analysis techniques, based on a global similarity thesaurus, are expensive since the collections of Web pages are so large and ever-changing. In a local analysis strategy, the documents retrieved from a given query are used to determine terms for automatic query expansion. The underlying assumption is that the top m ranked answers are relevant to the user information need. The assumption is questionable in the

Web-based information retrieval because a short query can retrieve some Web pages in the top ranked list which contain the keywords in the query but not relevant to the user information need. To refine the automatic query expansion technique, we have developed a semi-automatic query expansion technique to interactively take into account the user relevance feedback during the Web searching process and to use it for reformulating the query. The expanded query updates the ranked list and improves the effectiveness of Web searching. A search engine based on the vector model has been developed to validate the semi-automatic query expansion approach.

In the next section, we provide the details of the semi-automatic query expansion approach. Following that, we briefly describe the implementation and development environment. The final section discusses the results and its implications for future research.

SEMI-AUTOMATIC QUERY EXPANSION

When searching online text collection, the searcher inputs a query and the search engine returns a ranked list of Web pages. The first query is an initial attempt to retrieve the valuable information. The searcher may examine the retrieved Web pages to determine if they satisfy the information need. This examination process may provide useful relevance feedback for reformulating the initial query. With the relevance feedback in the first Web searching attempt, it is expected that the expanded query can better reflect the user information need and is able to improve searching effectiveness.

There are several ways to calculate the modified queries (Carpineto, et. al., 2001; Wen, Nie, & Zhang, 2002). One good starting point is the standard Rochio method and its variants (Baeza-Yates, 1999), as shown in Eq (1).

$$\vec{q}_m = \alpha\vec{q} + \frac{\beta}{|D_r|}\sum_{\forall \vec{d}_j \in D_r} \vec{d}_j - \frac{\gamma}{|D_i|}\sum_{\forall \vec{d}_j \in D_i} \vec{d}_j \qquad (1)$$

where

$\vec{q}$: original query;

$\vec{q}_m$: reformulated query;

D_r: set of relevant documents retrieved, as judged by the user;

D_i: set of irrelevant documents retrieved;

$\vec{d}_j$: vector of weights of index terms in document j;

α, β, γ: tuning constants.

In Eq. (1), the first term is the original query, the second term adds new words selected from the relevant Web pages, and the third term subtracts words obtained from irrelevant Web pages. We can set tuning constants in Eq. (1), e.g. $\alpha = \beta = \gamma = 1$. If only a positive feedback strategy is used, the constant γ is set 0. In classic automatic query expansion techniques, the top *m* pages are assumed to be relevant to user information need and the others are irrelevant. However, the top *m* pages may contain irrelevant pages and these irrelevant pages are used as positive feedback in the automatic query expansion technique, which may affect the searching effectiveness. This observation suggests us to differentiate relevant and irrelevant Web pages in the top ranked list retrieved from the previous search and to develop a semi-automatic query expansion approach which takes into account user's opinion interactively to reformulate the initial query.

The semi-automatic query expansion involves in two steps: (1) determining the relevance of some retrieved pages, and (2) expanding the original query with new terms and reweighting the terms in the expanded query. For convenience, the following notations are used: *C* is the entire collection of Web pages, C_u is the set of Web pages that are not retrieved by the search engine and *R* is the set of the Web pages retrieved from the initial query. Thus, we have $C = C_u \cup R$. The searcher only examines some Web pages in *R* and determines whether they are relevant to the user information need or not, ignoring the rest Web pages. *R* consists three parts and is expressed as $R = R_r \cup R_i \cup R_u$, where R_r, R_i and R_u are sets of relevant pages, irrelevant pages and unexamined pages, respectively. The entire collection consists of R_r, R_i, R_u and C_u, i.e. $C = R_r \cup R_i \cup R_u \cup C_u$, as shown in Figure 1.

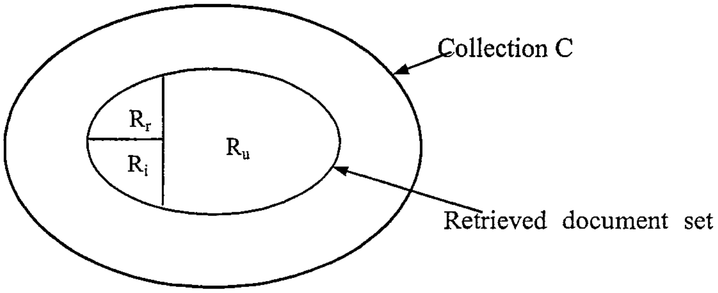


Figure 1: The entire collection and the retrieved document set

In semi-automatic query expansion approach, we make the following assumptions:

1. A simple query with a few keywords retrieves a larger set of ranked Web pages, and the examined Web pages contain relevant Web pages R_r and/or irrelevant Web pages R_i, judged by the searcher.
2. Only those pages examined by the searcher are used for query expansion. R_r is used for positive relevance feedback and R_i for negative relevance feedback. The pages that have

not been retrieved and those that have been retrieved but have not been examined are not taken into account for query expansion.

The first assumption is intuitive and reasonable from the observation of Web searching practice. The second assumption differentiates relevant and irrelevant pages in the retrieved page set, which excludes those pages whose relevance are uncertain, and hence refines the query expansion. Based on the above assumptions, semi-automatic query expansion can be described by the following equation:

$$\vec{q}_m = \alpha\vec{q} + \frac{\beta}{|R_r|}\sum_{\forall\vec{d}_j \in R_r} \vec{d}_j - \frac{\gamma}{|R_i|}\sum_{\forall\vec{d}_j \in R_i} \vec{d}_j \qquad (2)$$

Eq. (2) has the similar format as Eq. (1) but conveys different meaning. The semi-automatic query expansion approach differs from classic automatic query expansion techniques for the vector model in that (1) it takes into account user relevance feedback; (2) it distinguishes relevant and irrelevant documents in all documents examined, where automatic query expansion technique assumes that the all top ranked documents are relevant; (3) it only uses those examined documents for query expansion, while automatic technique uses all documents retrieved in the initial search; and (4) semi-automatic query expansion is faster than automatic technique, since it only processes documents in the subsets R_r and R_n, instead of entire answer set R. The second search with the expanded query may be performed either on entire collection or only on the answer set returned from the initial query. The latter may accelerate the searching process and save CPU time.

IMPLEMENTATION AND RESULTS

To validate the semi-automatic query expansion approach, a search engine has been developed using vector information retrieval model. A full description of implementation of Web search engine is beyond the scope of this paper. Here, we briefly introduce the models and develop environment.

The development environment and models:

Platform: Sun Solaris 5.8

Web Server: Java Web Server 2.0 and Tomcat 4.1.8

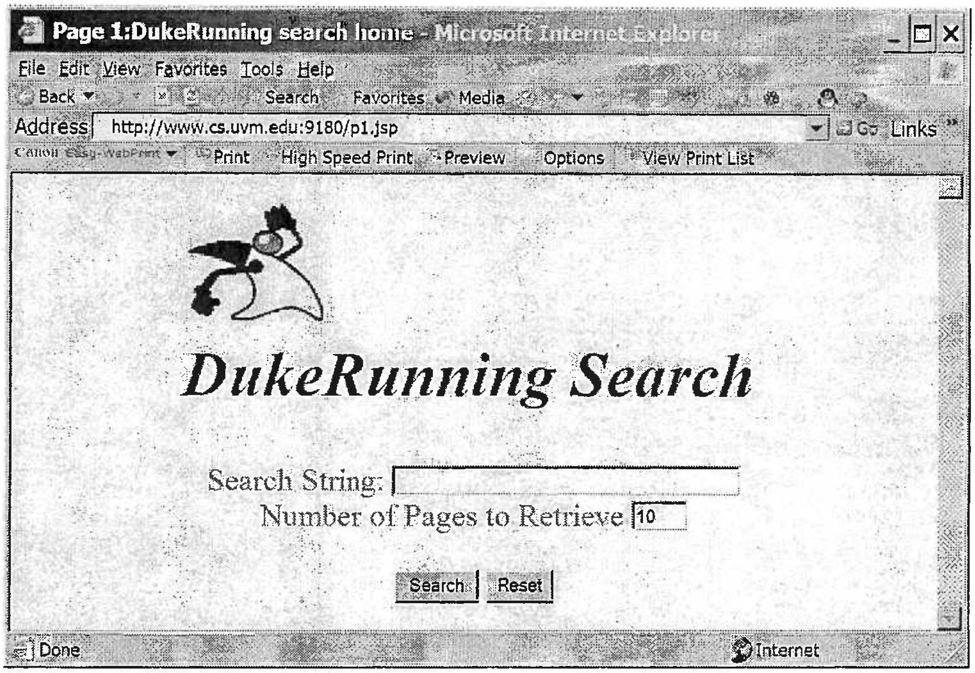
Programming language: Java, Java Servlet/JSP

Database: Oracle 9.0.1 running on Sun Solaris 5.8

IR models/techniques: vector model, Portal's algorithm

Query Expansion: automatic and semi-automatic query expansion approaches.

A Web searching example is used to analyze the effectiveness of the approach proposed in the paper. The 5000 web pages from the root http://www.uvm.edu were collected by the web spider. Each of the parsed web pages is preprocessed, and all data, such as URLs, index terms and the corresponding frequencies, are stored in the database. The interface of the search engine is shown in Figure 2 in which the user can enter a query and set the number of the web pages to be displayed.



Page 1:DukeRunning search home - Microsoft Internet Explorer

Address http://www.cs.uvm.edu:9180/p1.jsp

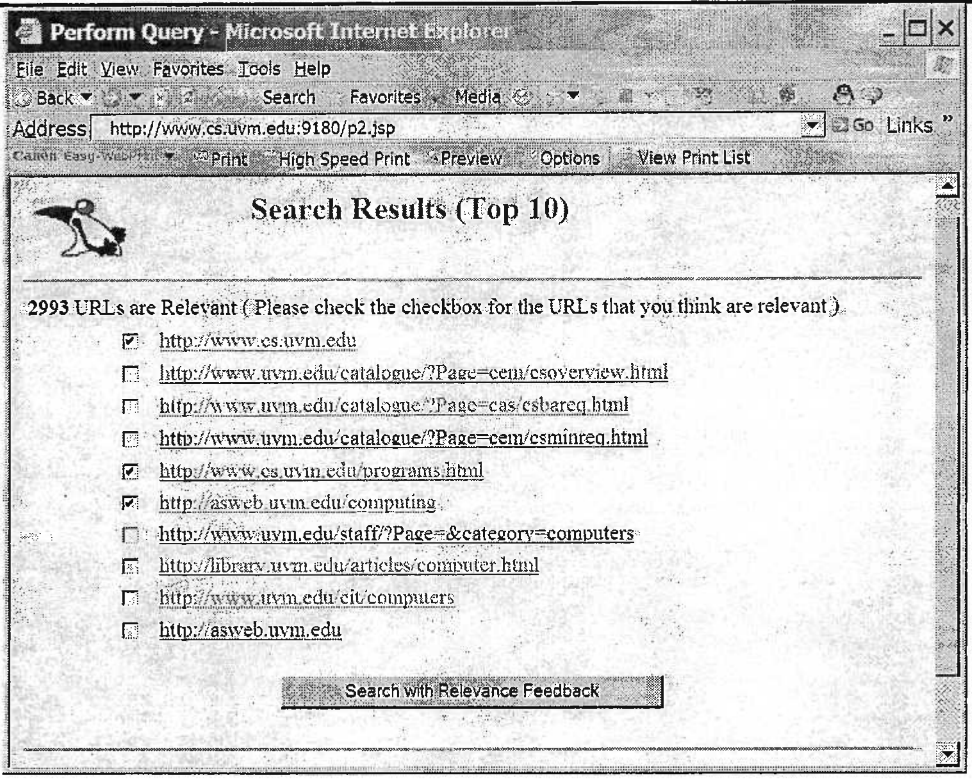
DukeRunning Search

Search String:

Number of Pages to Retrieve 10

Figure 2: Search engine interface

In the experiment, the initial query is "computer science information". A total of 2993 documents are returned. The top 10 retrieved pages are displayed in Figure 3. There is a checkbox for each web page.


Perform Query - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://www.cs.uvm.edu:9180/p2.jsp
Go Links
Print High Speed Print Preview Options View Print List
Search with Relevance Feedback

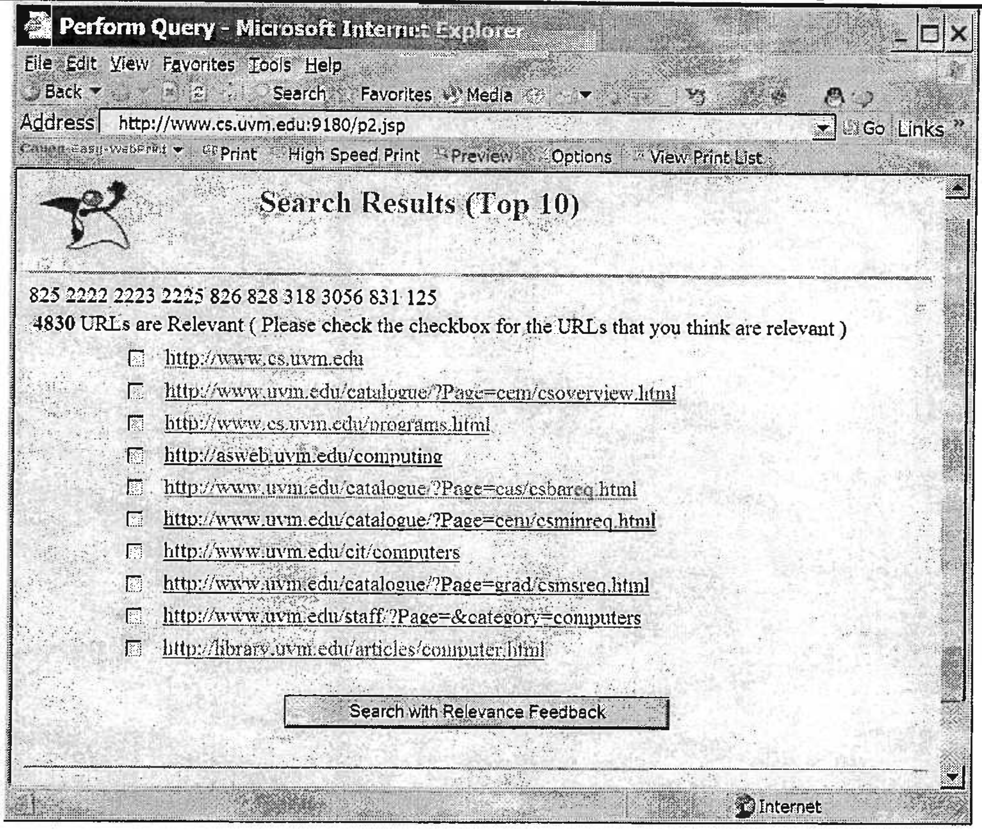
Search Results (Top 10)

2993 URLs are Relevant (Please check the checkbox for the URLs that you think are relevant)

☑ http://www.cs.uvm.edu
☐ http://www.uvm.edu/catalogue/?Page=cem/csoverview.html
☐ http://www.uvm.edu/catalogue/?Page=cas/csbareq.html
☐ http://www.uvm.edu/catalogue/?Page=cem/csminreq.html
☑ http://www.cs.uvm.edu/programs.html
☑ http://asweb.uvm.edu/computing
☐ http://www.uvm.edu/staff/?Page=&category=computers
☐ http://library.uvm.edu/articles/computer.html
☐ http://www.uvm.edu/cit/computers
☐ http://asweb.uvm.edu

Figure 3: Retrieved web pages with the initial query

The searcher examines some of the retrieved web pages and determines whether they are relevant to the information need or not. If the web page is relevant, which is judged by the searcher, the corresponding check box must be checked. In Figure 3, seven URLs have been examined and three of them checked. The relevance feedback has been taken in account in the new Web search. The query is automatically expanded and the new query consists of more keywords. A total of 4830 web pages are returned in the next search with the relevance feedback and semi-query expansion. The top 10 retrieved web pages are given in Figure 4. It is noted that the three checked Web links in Figure 3 have equal or higher ranks in Figure 4. The semi-automatic query expansion approach updates the ranks of the web pages so that the Web pages satisfying the information need may have higher ranks in the new Web search. For example, the relevant Web pages with ranks 9 and 10 in Figure 3 that satisfy the user information need have higher ranks (ranks 7 and 4, respectively) in new web search, as shown in Figure 4. The above preliminary results show that the semi-automatic query expansion may improve the effectiveness of web-based information retrieval by taking into account the relevance feedback to expand the query and hence update the ranks of the relevant Web pages.


Perform Query - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://www.cs.uvm.edu:9180/p2.jsp
Go Links
Print High Speed Print Preview Options View Print List
Search with Relevance Feedback
Internet

Search Results (Top 10)

825 2222 2223 2225 826 828 318 3056 831 125
4830 URLs are Relevant (Please check the checkbox for the URLs that you think are relevant)

- □ http://www.cs.uvm.edu
- □ http://www.uvm.edu/catalogue/?Page=cem/csoverview.html
- □ http://www.cs.uvm.edu/programs.html
- □ http://asweb.uvm.edu/computing
- □ http://www.uvm.edu/catalogue/?Page=cas/csbareq.html
- □ http://www.uvm.edu/catalogue/?Page=cem/csminreq.html
- □ http://www.uvm.edu/cit/computers
- □ http://www.uvm.edu/catalogue/?Page=grad/csmsreq.html
- □ http://www.uvm.edu/staff/?Page=&category=computers
- □ http://library.uvm.edu/articles/computer.html

Figure 4: Retrieved web pages with the semi-automatic query expansion

The experiments with other queries also give similar results: the relevance feedback can update the ranks of the relevant web pages and the returned top ranked list may contain more relevant Web pages.

DISCUSSION AND CONCLUSION

The semi-automatic query expansion approach has been proposed and used for query reformulation, and a workable search engine based on the vector IR model has been implemented to validate the effectiveness of the semi-automatic query expansion approach. 5000 Web pages were collected by the spider from a given root URL. The performance of the search engine depends on the underlying system, Web server and database. The indexing data structure can improve the searching performance.

Given an initial query, the Web pages in the answer set are examined. The relevance of each Web page in the answer set is specified by checking the corresponding checkbox. The relevance feedback is taken into account in semi-automatic query expansion. The expanded query is used for re-ranking the answer set retrieved from the initial query or for searching the entire collection again. The experiments show that, if the next Web searching with expanded query is applied for the answer set retrieved from the initial query, the semi-automatic query expansion can update the ranks of the returned web pages, and hence improve the effectiveness of Web searching. If the next search is applied for the entire collection, some new Web pages that cannot be retrieved in the first search are added to the top ranked list of the answer set and they are relevant to the user information need. Searching collection C again can improve the coverage and ranking of Web-based information retrieval but it needs more CPU time.

No benchmark is available for measuring the performance of the Web-based information retrieval because the collection of the web pages in real world is huge and it is impossible to objectively determine the relevant set and the actual ranks of the web pages for various queries. The experiments in the paper are conducted on a small collection. The ongoing research is to validate the semi-automatic query expansion approach on standard larger collections, such as TREC collection, and to compare it with other's work using a more objective criterion. The links structure is also an important aspect to be considered in the future work to improve the coverage, relevance and ranking in Web searching (Zeng & Bloniaz, 2004).

This exploratory study has shown the promise of the semi-automatic query expansion approach proposed in the paper in improving the performance of the Web-based information retrieval. However, it has several limitations of the semi-automatic query expansion approach include 1) the overhead to select those web pages used as relevance feedback, 2) the additional execution time to expand the query, and 3) lack of an objective way to determine the relevance of each Web page. It is noted that only a few web pages in the answer set are selected and used for feedback. These web pages have already been preprocessed and index terms have been stored in the database, and hence, the execution time of query expansion is not significant.

REFERENCES

Baeza-Yates, R. & Ribeiro-Neto, B. (1999). "Modern Information Retrieval," ACM press.

Carpineto, C., Mori, R. D., Romano, G. & BiGi, B. (January 2001). "An Information-Theoretic Approach to Automatic Query Expansion," *ACM Transactions on Information System*, 19(1), 1-27.

Eastman, C. M. & Jansen, B. J. (October 2003). "Coverage, Relevance, and Ranking: The Impact of Query Operators on Web.Search Results," *ACM Transactions on Information System*, 21(4), 383-411.

Gauch, S., Wang, J., & Rachakonda, S. M. (July, 1999). "A Corpus Analysis Approach for Automatic Query Expansion and Its Extension to Multiple Databases," *ACM Transactions on Information System*, 17(3), 250-269.

Wen, J., Nie, J. & Zhang, H. (January, 2002). "Query Clustering Using User Logs," *ACM Transactions on Information System*, 20(1), 59-81.

Zeng, J. & Bloniaz, P. A. (April, 2004). "From Keywords to Links: an Automatic Approach," *Proceedings of the International Conference on Information Technology: Coding and Computing*, Las Vegas.