An Expertise Recommender using Web Mining

Anupam Joshi, Purnima Chandrasekaran, Michelle Shu Yang, Ramya Ramakrishnan
Department of Computer Science and Electrical Engineering
University of Maryland, Baltimore County
[joshi, pchand1, syang3, rrmaka1]@cs.umbc.edu

Introduction

In this report we explore techniques to mine web pages of scientists to extract information regarding their expertise, build expertise chains and referral webs, and semi automatically combine this information with directory information services to create a recommender system [9] that permits query by expertise. We experimented with existing techniques that have been reported in research literature in recent past (including our own prior work), and adapted them as needed. We developed software tools to capture and use this information.

Rationale and Approach

The problem described in the RFP is common to many large organizations. Specifically, how can one gather information about the expertise of persons and build a system that would allow queries against it. *Present methods are essentially static.* Individuals describe their capabilities using some "keywords", and these are matched in response to a query. For a large organization, the exercise of gathering this information is fairly massive and time consuming. Moreover, once entered, the keywords remain unchanged until explicitly altered. More importantly, keywords are often too restrictive to capture the expertise of the individual, and people end up mentioning their primary expertise. Narrative descriptions of research in web pages or titles/abstracts of papers often provide a far richer description of a person's overall expertise, both primary and others. We present a method *where we used text-mining techniques to semi-automatically extract expertise-*related information from individual web pages. This not only saves on human effort, but also allows for automatic updates via web spiders that re-mine a page when it changes beyond a certain degree. Later in this section, we detail some text mining approaches.

Studies [7] have shown that in most organizations, the informal network of colleagues is one of the most important sources of disseminating expertise related information. In other words, when looking for a person with a particular expertise, people tend to ask friends and colleagues, who in turn might refer them to other colleagues and so on, thus building a referral chain [4]. Katz et al. [4] have shown how such "social networks" (in our context, people an individual knows augmented with information about their expertise) can be built by analyzing web pages for author lists or mention of particular names, or email headers for sender and recipient fields. Effectively, this is looking for structured data patterns in the web pages. We postulate that some of the newer research in text mining can be used to extract similar information from the unstructured information (research descriptions etc) in the web pages as well. Moreover, given "organization description" type pages (say within individual NASA units), this type of information can also be directly accessed. Given that analyzing email headers leads to potential privacy issues, we avoided using them in building the referral chains.

Related to the expertise chains is the notion of collaborative filtering. In collaborative filtering, people rate "information", and these ratings are used by the system to "recommend" appropriate information to other users. For example, our W^{3}IQ system [5, 2, 1] uses ratings provided by users to present appropriate URLs in response to queries by other users who share a similar interest. In the context of the current work, collaborative filtering can be used to share information about expertise across users, and build the referral chains. The system can be enhanced to allow NASA employees to describe and rate (in a secure and private manner) the expertise of other employees they know. As such, a person searching for a particular expertise would be able to benefit from the knowledge of a co-worker about that expertise. However, like in collaborative filtering, the open question here is the formation of appropriate groups - given divergent ratings and opinions about someone's expertise, whom do you believe? Consider for example that Jill, who runs the information visualization department, is searching for someone with expertise in hurricanes. The system determines that many of her colleagues have referral chains pointing to people with this expertise. Which ones should the system present to her in response to her query? The obvious answer is to respond with the names that are pointed to by the people whom she agrees with (or thinks like) most often. We have shown [1] a system that can bootstrap itself from no (or merely declarative) knowledge about similar interest groups to over time learn the groupings between individuals based on the similarity of choices they make. We have also shown [3, 8] how people's information access behavior can be mined to form these groups. We originally planned to explore this avenue in this research effort, but decided to concentrate on other non-intrusive techniques first.

As the preceding discussion would make clear, being able to mine information from free text descriptions in employee web pages is an important component of our system. There are classical text mining type approaches that we have explored in this report. These include mechanisms to identify keywords from a document corpus, such as Term Frequency Inverse Document Frequency (TFIDF). However, by intelligently using domain constraints, we could improve the performance of such methods by exploring the semi-structured nature of the employee web pages and their publications mentioned therein.

Note that in the preceding, we have essentially talked about automating the discovery of expertise information from existing unstructured data sources such as web pages. However, this task could be done more efficiently if the creator of the web page could annotate it to provide expertise-related information. We originally proposed, and then abandoned the idea because of lack of interest from NASA colleagues, to create an XML Scheme (similar to RDF) which would allow such annotations, and a tool to permit them to be easily added to web pages.

We have outlined several recent approaches from literature that can serve as the basis for creating the expertise recommender system. When implementing the system, we studied the tradeoffs involved in a centralized expert directory service versus a distributed / hierarchical directory service. Issues studied involve scalability, complexity of the system, ease of maintenance, efficiency in query execution and linkage sought with the existing X.500 infrastructure etc. We extensively consulted with NASA colleagues when designing the system in light of these tradeoffs.

System Implementation

As explained in the previous section, the existing methods to query a person's expertise use static data from large databases that are updated quite infrequently. Moreover this method is very restrictive as an individual describes his expertise in terms of a few keywords that are matched to return a response to a query.

Our approach to the problem is to mine narrative text, web pages, departmental reports, progress reports, etc., which are rich sources of expertise related information. This method also allows for automatically updating expertise information via web spiders that re-mine pages that have changed beyond a certain degree.

We have developed a tool that mines web pages of scientists at NASA and extracts information regarding their expertise, builds expertise chains and referral webs, and links this information with the existing X.500 directory services to create a recommender system that permits query by expertise.

We started with Webcrawl, a public domain tool that takes a given URL, downloads the page, and follows all the URLs referenced in that page. It repeats this process on all the downloaded pages, recursively, until no more pages are left to download or no more links to follow. Initially, we limited the search space to all the pages in the gsfc.nasa.gov domain. But now the system has been enhanced to handle employee web pages that reside outside the domain being crawled but are linked to pages in the NASA domain. This is done by allowing the crawler to visit all external links upto 3 levels from the domain being crawled. The crawling of external links is limited to 3 levels so as to avoid the problem of "infinite crawling" wherein the web crawler endlessly crawls the web returning lots of extraneous data. The document corpus created by crawling the following NASA domains that are accessible on the Internet have been used as the source for extracting expertise information.

- Space Science Directorate, Code 600
  `http://space.gsfc.nasa.gov`
- Space Science Data Operations office, Code 630
  `http://ssdoo.gsfc.nasa.gov`
- Laboratory for High Energy Astrophysics, Code 660
  `http://lheawww.gsfc.nasa.gov/docs/lhea/LHEAstaff/Staff.html`
- Laboratory for Astronomy and Solar Physics, Code 680
  `http://lasp-nts1.gsfc.nasa.gov/`
- Goddard Space Flight Center - Astrochemistry Branch
  `http://www-691.gsfc.nasa.gov/`
- Laboratory for Extraterrestrial Physics, Code 690
  `http://lep694.gsfc.nasa.gov/code690.html`

The next step is to identify key words in the archived content. We used the Term Frequency Inverse Document Frequency (TFIDF), which is a classical text mining approach. In this approach, first a set of stop words is removed from the archive document. Stop words are those words that occur very commonly and can be straightaway rejected as non-key words. Example stop words include "a", "and", "the", "that", "where", etc. The remaining words in each document are 'stemmed' by removing root word extensions, so multiple related words map on to their word root. This further reduces the number of candidate key words. TFIDF then determines the occurrence of the words in various documents and prioritizes the words, so that a word occurring in least number of documents with highest frequency are picked over other words. The first 'N' number of words is returned as the keywords, where 'N' is user selectable.

However, we found that in the web accessible employee resume, which is a major source of the employee's areas of interest/expertise, the expertise terms occurred just a few times. Thus, both TF and DF for the expertise information are low and hence, not selected as keywords by the above algorithm. This scheme is likely to work better for descriptive texts like annual reports, progress reports, etc. across many groups in GSFC because then the TF increases due to the multiple occurrence of potential keywords in the said texts while the DF remains the same. These pages are more likely to satisfy the underlying principles of TFIDF.

As a workaround, we studied the format of the web resumes and saw that they are not unstructured text narratives as assumed by TFIDF. We exploited the semi-structured nature of the resumes by looking for particular ways the expertise is described (in research interests and publication titles).

The document corpus returned by the webcrawler is analyzed to determine employee resumes. We noticed that the first string within the BODY tag of the resumes was the full name of the employee and hence used this heuristic. Then, keyword extraction in these resumes is restricted to particular sections like the "Research Interests" and the "Selected Publications" section. The HTML tags delimiting the text following the heading "Research Interests/Research Area/Expertise" is determined and extracted for further processing. This text is specified either as a narrative or as a list of interests. If there is a "Selected Publications" section, then again the text within the tags delimiting the section is extracted. Stop words are eliminated from the extracted text and the expertise of the employee returned is updated into a mSQL database that has been provided with a Web based query interface.

The Referral chains mentioned earlier in the section has been built from the co-authorship of publications in the "Selected Publications" section of the web resumes. We saw that the names of co-authors are specified following the same conventions as any technical publications. They are specified as the last name followed by a comma and preceded or succeeded by two/three single letter initials each followed by a period. We look for these patterns in the "Selected Publications" section and extract the data. The extracted data are the co-authors who have published with the employee whose resume is being parsed. This data is updated into the mSQL database to build the referral chain. This data is also used for stop word elimination from the text in the Publications.

Certain directorates did not have any resumes that could be mined. Hence, we used documents containing lists of publications of the branch. For example, branch publications were used to mine expertise information of scientists in the Earth & Space Data Computing Division (ESDCD)-Code 930. A *Perl* program parses these documents, extracts the list of co-authors and the keywords that form the title of each of the publication. These keywords are added as the areas of expertise for the first author of the publication. The rest of the authors are added as his/her collaborators. When a user of the Expertise Recommender queries the system with one of these key words, it will return the first author as the principal researcher and the rest of the co-authors as the referrals.

We have also built a distributed version of the expertise extractor. In this mode, the Webcrawl and associated expertise extractor software are installed in a set of workstations. They are invoked from shell scripts running on a workstation that acts as a master. We used the Secure Shell, called *ssh*, to invoke these programs remotely in a secure way to crawl different web domains to speed up information extraction. Since the backend of the Expertise extractor is a centralized database, it can be populated with the information extracted from the crawled web pages at individual workstations independently of every other workstation. This obviates the need for any synchronization across the multiple distributed processes, and also results in linear performance improvements with the number of participating workstations/processes. Thus the scripts running on individual workstations simultaneously crawl multiple domains, create the document corpus locally, extract expertise information, build referral chains and update the central database.

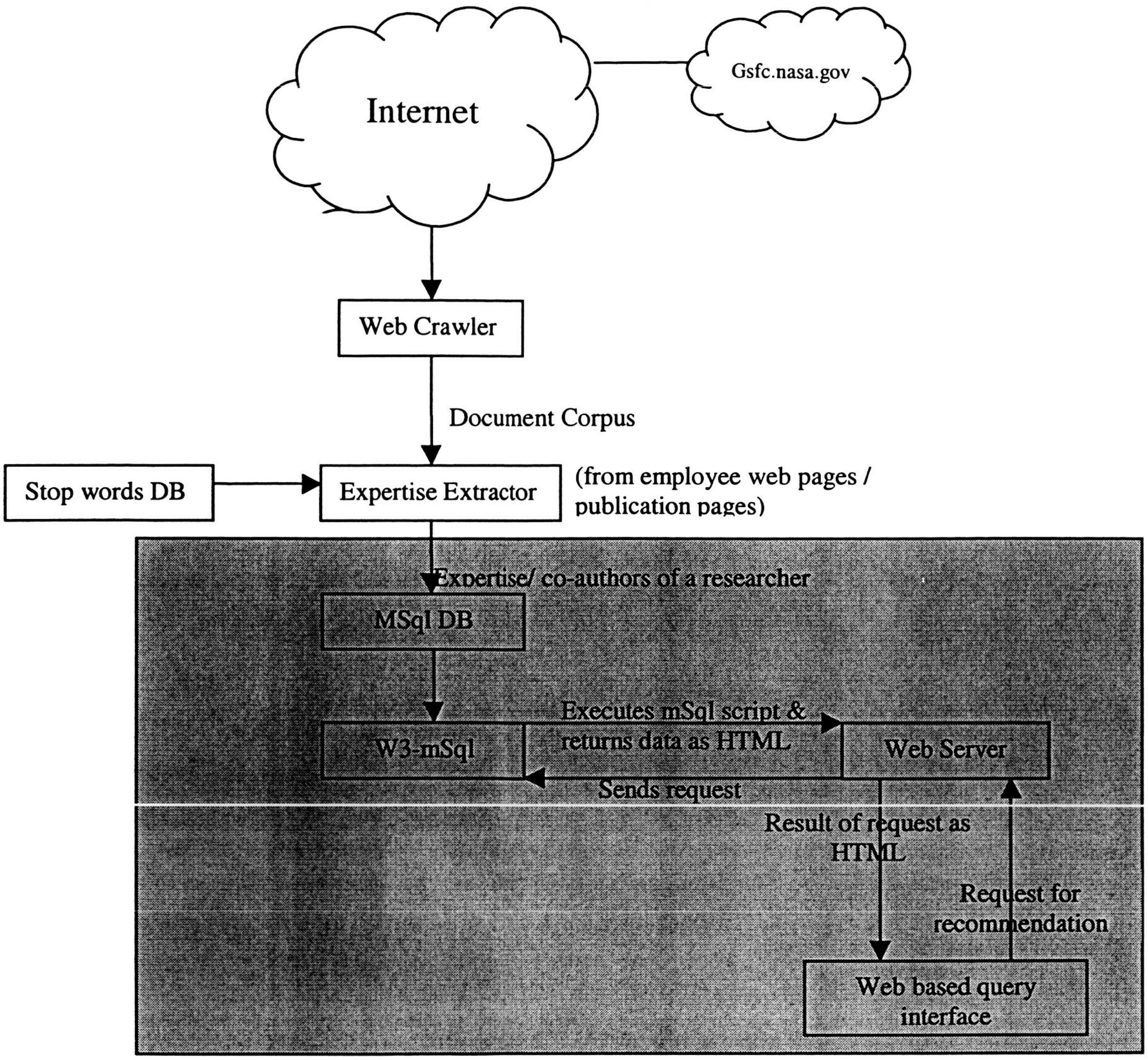

Internet
Gsfc.nasa.gov
Web Crawler
Document Corpus
Stop words DB
Expertise Extractor
(from employee web pages / publication pages)
Expertise/ co-authors of a researcher
MSql DB
W3-mSql
Executes mSql script & returns data as HTML
Sends request
Web Server
Result of request as HTML
Request for recommendation
Web based query interface

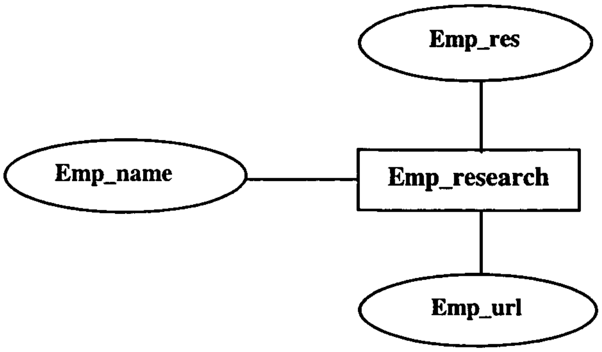
Figure 1. Architecture of the system

A web-based query interface to this central database has been developed with links to NASA's X.500 directory services. This has been developed using W3-mSQL v 2.0, which is the WWW interface package to the mSQL database. Lite (a programming language provided by W3-mSQL) code to query the database has been embedded into HTML documents. The HTML document is processed through the W3-mSQL binary that generates HTML code on the fly for the embedded Lite code. This web-based interface can be used to query by name or it can be used as an Expertise Recommender by querying on expertise area. When queried on expertise area, the system returns a list of employees who have the queried expertise, links to their X.500 data and a referral list for each employee sorted so as to give priority to co-authors who share the expertise being queried.

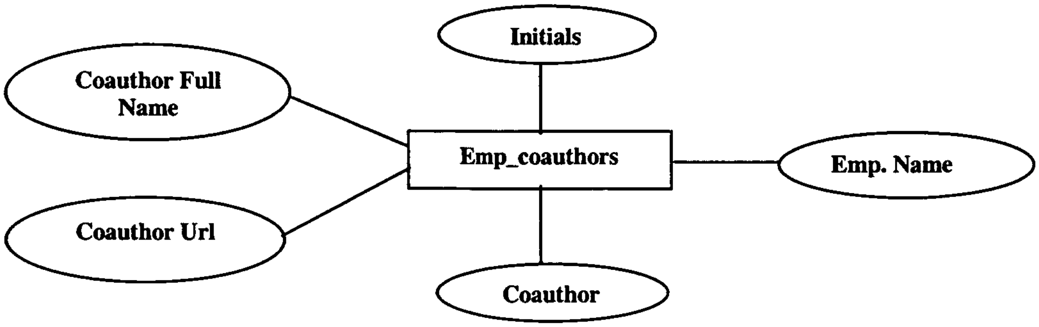
Database Schema

The expertise recommender database has two tables – emp_research & emp_coauthors as shown below.

emp_research relation contains the employee name, expertise and a link to their web page if it exists.



emp_coauthors relation contains the employee name, coauthorship, coauthor full name and a link to their web page if it exists.



Future Work

The system can be enhanced by further research in the following areas:

Expertise Extraction

- Keywords extracted from employee resumes may be mapped to NASA "standard" terminology to build a hierarchical ontology so that finer levels of expertise areas from resumes map to broader fields of expertise. This would allow queries based on broader and more generally known terminology than those directly present in the resumes mined.
- Create seed fuzzy thesauri [6] that show the degree of similarity between various terms that describe similar expertise and thus solving the "synonym" problem.
- An updating mechanism based on underlying page changes, wherein only pages that have been modified since they were last mined may be used in updating the data source.
- Use more heuristics that will result in increased coverage of employee web pages / formats. The existing system can analyze and mine expertise information from more than 90% of the employee pages.
- Use word stemming so that different derivatives of a word map to their root word.
- Use of association rule type approaches to find key phrases as well as words that occur together. This would enable the capture of more complex expertise descriptions than simple keywords.

Referral Chains

- Explore other sources of building Referral Chains that are not intrusive.

- Techniques to integrate these Referral Chains into the querying process.
- We initially planned to take advantage of the collaborative filtering techniques to enhance the expertise recommender system and the results it returns in response to queries. However, we found that this requires an active participation by NASA research staff, and/or access to the NASA web servers' log files. However, as it was decided to perform this task as least intrusively as possible, we abandoned this enhancement.

Performance Tuning

- The response time of the queries on expertise can be improved by building precompiled views for the query and storing them in the database. This feature is not present in the current implementation, as views are not supported by the publicly available Mini SQL Version 2.0.11.

References

[1] A. Joshi, C. Punyapu, and P. Karnam. Personalization & Asynchronicity to support mobile web access. In *Proc. Workshop on Web Information and Data Management, ACM Conference on Information and Knowledge Management,* November 1998.

[2] A. Joshi, S. Weerawarana, and E. N. Houstis. Disconnected Browsing of Distributed Information. In *Proc. Seventh IEEE Intl. Workshop on Research Issues in Data Engineering,* pages 101-108. IEEE, April 1997.

[3] A. Joshi and R. Krishnapuram. Robust fuzzy clustering methods to support web mining. In S. Chaudhuri and U. Dayal, editors, *Proc. ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery,* June 1998.

[4] H. Kautz, B. Selman and M. Shah. Combining social and collaborative filtering. *Communications of ACM,* 40:63-65, 1995.

[5] R. Kavasseri, T. Keating, M. Wittman, A. Joshi, and S. Weerawarana. Web Intelligent Query – Disconnected Web Browsing using Cooperative Techniques. In *Proc. 1st. IFCIS Intl. Conf. On Cooperative Information Systems,* pages 167-174. IEEE, IEEE Press, 1996.

[6] G. Klir and B. Yuan. *Fuzzy Sets and Fuzzy Logic.* Prentice Hall, 1995.

[7] R. Kraut, J. Galegher, and C. Edigo. *Intellectual Teamwork: Social and Technological Bases for Cooperative Work.* Lawrence Erlbaum, Hillsdale, NJ, 1990.

[8] O. Nasraoui, R. Krishnapuram, and A. Joshi. Mining web access logs using a fuzzy relational clustering algorithm based on a robust estimator. In *Proc. WWW – 8th International World Wide Web Conference,* May 1999.

[9] P. Resnick and H. Varian. Recommender Systems. *Comm. ACM,* 40(3):56-58, March 1997.

NASA

Report Documentation Page

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.

4. Title and Subtitle
An Expertise Recommender using Web Mining

5. Report Date

6. Performing Organization Code

7. Author(s)
Anupam Joshi

8. Performing Organization Report No.

10. Work Unit No.

9. Performing Organization Name and Address
University of Maryland Baltimore County
302 Administrative Building, 1000 Hilltop Circle
Baltimore, Maryland 21250-5394

11. Contract or Grant No. NAS5-32337
USRA subcontract No. 5555-97-73

12. Sponsoring Agency Name and Address
National Aeronautics and Space Administration
Washington, DC 20546-0001
NASA Goddard Space Flight Center
Greenbelt, MD 20771

13. Type of Report and Period Covered Final
January 2000 - January 2001

14. Sponsoring Agency Code

15. Supplementary Notes
This work was performed under a subcontract issued by
Universities Space Research Association
10227 Wincopin Circle, Suite 212
Columbia, MD 21044
Task 97

16. Abstract

This report explored techniques to mine web pages of scientists to extract information regarding their expertise, build expertise chains and referral webs, and semi automatically combine this information with directory information services to create a recommender system that permits query by expertise. The approach included experimenting with existing techniques that have been reported in research literature in recent past , and adapted them as needed. In addition, software tools were developed to capture and use this information.

17. Key Words (Suggested by Author(s))
mining web pages

18. Distribution Statement
Unclassified--Unlimited

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price
Unclassified	Unclassified	1	