BEYOND INFORMATION RETRIEVAL: A SURVEY

[1]J.ASHOK, [2]KAMALAKAR RAMINENI, [3]E.G.RAJAN

[1]Professor and Head , Department of Information Technology, GCET,Hyderabad,India.
[2]Lecturer, Satavahana University, Karimnagar.India.
[3]President, Pentagram Research Center, Hyderabad,India

ABSTRACT

The World-Wide Web is developing very fast. Currently, finding useful information on the Web is a time consuming process. The search is still potentially combinatorial explosive, so we put a resource limitation on search activity. This limit is expressed as a maximum number of accesses to non-local Web nodes per minute. Current information retrieval tools mostly use keyword search, which is unsatisfactory option because of its low precision and recall.

Keywords: *IP, Boolean searching, Westlaw is Natural, FREESTYLE, TARGET, PHOAKS, Autonomy.*

1. INTRODUCTION

The days of the traditional abstracting and indexing services are waning, as abstracts and bibliographic data become a commodity. However, there are tremendous opportunities for those organizations willing to look beyond the status quo to the new possibilities enabled by the latest wave of advanced technologies. Features like automatic extraction of key concepts or names, collaborative filtering to help with trends analysis, visualization techniques, and more, can take information past the retrieval stage and into the management arena.

2. A MOMENT IN HISTORY

For the past quarter century, the focus of information has been on finding it. Information retrieval has always been important, but Information Providers (IPs) have been able to get away with not giving answers, rather merely providing a means to the answer. Here's a list of proof - you go find out which one gives you the answer. But the advent of computerized databases was so much better than searching through the equivalent print products, or the primary literature directly, that its usefulness was not questioned until something better came along. Remember the 1960's and 70's? The rise of the mainframe systems like Lockheed's Dialog, SDC's Orbit,BRS, ESA-IRS, Lexis, and others. At the time, the only way to provide computerized access to databases was with large, expensive mainframe systems and owned software. Users could dial in at 300 baud and databases were primarily bibliographic proofs and abstracts due to the costs of storing and accessing the data. Databases became available online as their producers switched to computerized typesetting methods for easier printing of the journals. Enter the 1980's and the personal computer. Modem speeds increased to 9600 baud and the number of databases provided through the traditional online services grown rapidly into the thousands. However, the search language was still boolean and required a trained operator to retrieve relevant results. Search services were sold and consolidated, but profits remained high. Despite talk about branching out to an end-user market, the pricing strategies and complicated search language restrained the market to professional searchers.

In all this time, very little was done to enhance the searching of the databases. Rather, the services concentrated on loading more and more files.The early 1990's brought a few attempts at enhancing the search languages to encompass more than Boolean operators. Westlaw was first, with it's WIN (Westlaw is Natural) natural language searching. Users were instructed to type in their queries as they would ask a colleague, and the system would interpret the important words in the

phrase. Of course, the user could twist the search by changing the word weightings, adding or deleting terms, etc. Lexis quickly followed with its FREESTYLE features, which enabled the user to search without any knowledge of boolean operators. They also added the WHERE and WHY features to enable the user to analyze the results. Dialog also entered the relevancy ranking with its TARGET command.Then the world turned upside down with the advent of the World Wide Web in the mid 90's. All of a sudden online became a household word, and everyone was searching. Storage and retrieval costs dropped dramatically, enabling a high-end PC to surpass the capabilities of the early mainframes at a fraction of the cost. As computers sprouted on every desktop, and networks became the norm, the little known search and retrieval world suddenly came into it's own. Suddenly everyone had computerized information of some sort and needed a way to store and then find it.

Throughout this time there had been a number of companies developing and selling enhanced search and retrieval software. Most notably, PLS started in 1986 and was used by many companies, including Congressional Quarterly and Dow Jones to enhance the search capabilities of their respective databases - but not on the traditional online services. America Online uses PLS software, and became so dependent upon them that they bought the company earlier this year. Other companies, provided similar sophisticated search and retrieval software. But the focus was always on finding the information, not on providing the answer to a question.

3. TECHNOLOGIES TO USE

3.1 BEYOND SEARCHING

Now let's switch and look at the World Wide Web and the information that is provided with these new search engines. YAHOO! was the first service to help users find information on the web, but it is not a search engine. It is a classification system utilizing many people to correctly categorize the hundreds of thousands of web sites. It is still one of the most popular sites, perhaps emphasizing the importance of editorial enhancements. (An interesting note from the search engine conference was the consensus from the search engine companies like Northern Light, Infoseek, Verity, etc. that YAHOO had won, and they are all starting to focus on classifying and categorizing data, even though automatically rather than investing in the human effort that YAHOO employs.) A number of search engines became available on the Web, most springing from University research projects, such as Lycos from Carnegie Mellon, Inktomi from Berkeley, etc. In the early days (1996-1997) the search engines competed on the number of pages indexed and the speed of retrieval. Now, however, most users realize that it is completely irrelevant if the search engine retrieves 100,000 or 1 million records. Either way it is too many to effectively use. However, the search engines are also utilizing a variety of advanced technology features to help the users get the information they want, even if it is not what they asked for. Excite (Company) employs concept searching, automatically expanding the search to look for similar ideas and synonyms and alternatives of the search words. This fuzzy searching sometimes gives you very strange results, but more often than not retrieves relevant pages that would have been missed in a standard boolean search. Northern Light goes beyond the web and combines data from its special collections - full-text data licensed from IAC, Soft Line and others. And, its unique offline categorization allows users to separate and browse different themes or concepts within their search results. This is almost always better than either reverse chronological order or straight relevancy ranking. It is relevancy ranking within concepts.

3.2 COLLABORATIVE FILTERING

Collaborative filtering has been effectively used by a number of sites to make recommendations to a user based on what other users have done. For instance, at CDNOW, you can use the Album Advisor to suggest additional albums in which you may be interested, based on your designation of your three favorite artists. Or, at the Amazon.com online bookstore, you may receive a suggestion for other books to purchase as you scan any given title. Studies have shown that users prefer personal recommendations over any other kind of information. A truly unique site in this area is PHOAKS, People Helping One Another Know Stuff, which uses collaborative filtering to analyze people's recommendations in newsgroups. Essentially you can search on any topic, and the most frequently mentioned sites will rise to the top of the results list. The PHOAKS folks are careful to warn you, that frequent mention is not necessarily a good thing. Of course, the searching function on all these sites is simply taken for granted.

3.3 DATA EXTRACTION

Once you have found information that may be relevant to your query, extracting the key pieces

of data to answer your question is the next important step. Or, extracting key pieces of data from a source to enhance the search and retrieval mechanism may also be important. A lot of the technology in this area grew out of the intelligence community. IsoQuest, for instance, started with extracting personal and company names from all kinds of data for the US intelligence community, and is now a recognized leader in developing data extraction, text analysis, and indexing tools based on natural language processing technology. Their clients include such companies as the Thomson Corporation, IBM, Disclosure, InfoSeek and News Edge. The technology is primarily used to extract company or individual names and then match them with a centralized database of names. This enables one to easily find news articles or other information about a company or person, no matter what name variant is used. The key to success here is to build an adequate dictionary to help the software identify the object you want to extract. Currently, the key applications relate to business information, but with development of a proper dictionary, the technology could certainly be applied to any number of subject areas and objects. Similar technology is employed by the data mining and knowledge management systems marketed for intranet applications.

3.4 DATA VISUALIZATION

No matter how compelling your data, a picture is still worth a thousand words. The latest technologies let you graphically display trends and patterns, perhaps providing a different insight into the data. You can see examples of this in a couple of the web search engines. Alta Vista, through its refine option, lets you view a graphical representation of the concepts that are co-appearing with your query words, and use it to enhance your search. The simplest form of visualization mapping is illustrated in the online yellow pages like Big Yellow and Big Book. Find a company and click a button to have the software display a map showing the location of your selected company. Many of the principles of Geographic Information Systems are incorporated into the data visualization technologies.

SemioMap has been a pioneer in this field, creating a search and retrieval software that displays results in a three dimensional concept map. The software extracts key phrases and concepts, and presents the results in a fashion that lets you move through related concepts, drill down to different levels of detail, or jump to the documents referenced. This provides a way to view and work with the interrelated concepts in even large volumes of information. Similarly, Manning & Napier's MAPIT software provides a way to extract and display trends in the patent literature. The system matches words and concepts contained in patent applications and supporting documents, and returns the analysis to the user as a series of visualized comparisons.

3.5 AGENT TECHNOLOGY

Repetitive tasks can easily be replicated by software agents. Examples include the shopping agents that comb the internet comparing prices of specific products, like books and CDs. Acses compares prices and shipping charges for books at over 25 online stores. Other agents, like those from AgentSoft, follow scripts to search and retrieve information from the web on your designated topics. Similar software is used to monitor stock prices, or to "filter" news stories for customized news sites.

4. PATTERN RECOGNITION

Some of the advanced agent technology is based on pattern recognition. Essentially the software uses small pieces of less accurate information that combined together give increasing precision. It operates by calculating the probability of seeing ***x*** if we see ***y***, and then what is the probability of seeing ***z*** if both ***x*** and ***y*** are present, and so on. The small pieces of information are accumulated by analyzing user feedback, so the agent learns as you use it. For instance, the Autonomy software is widely used by a variety of customers, from law enforcement agencies looking to spots trends or patterns of behavior, to companies wanting to automatically categorize and sort data. In addition, publishers like Macmillan use it to help profile and customize data for their website users.

5. CLASSIFICATION AND CLUSTERING

Many of the technologies listed above do various sorts of clustering and classification. The important feature is to identify the key concepts within a document and then pull all the information on those topics together and display it in a way that the user understands where to go. Northern Light created a classification system with more than 20,000 terms in a loosely hierarchical arrangement. All web pages or documents from the special collection are mapped to this classification scheme offline, and category tags are attached to each

record. However, the folders that display after a search are created dynamically, based on these tags.

6. VIRTUAL COMMUNITIES

In an age of technology, personalization is important. Answers do not come from the literature alone. They come from discourse with one's peers, from grey literature, from international websites, from personal databases, and more. Information clubs, like BioMedNet or the Engineering Information Village, provide excellent examples of serving a variety of resources to their members. In every subject area, these online clubs have the potential of eliminating the traditional professional societies. They provide most of the services that members want, in a fashion more suited to the modern professional. And yet, you see few of the professional societies in the forefront of these kinds of member services.

7. CONCLUSION

Information use and management is more important than ever. The data that the traditional services are providing are still important, they are just not the end product. Just about anyone with a PC can produce a bibliographic database nowadays, and fewer and fewer organizations are willing to pay for expensive human indexing to find journal articles. The Information Providers need to explore the amazing array of technologies available and identify those that will help them provide solutions to their customer base.

Unfortunately for some of the traditional players, their expertise is obsolete. We no longer need huge data storage centers and proprietary software. We need distributed systems and linkages. We can't tolerate months of delay to read an abstract, and then a longer wait to retrieve the full text of an article, which then must be analyzed for the answer to our question. We need software that will quickly distill the information and display it in the context that the user prefers. As we have heard a thousand times, content is still king. But it is content in context - distributed in the right way to the right audience at the right time for the right price. Those who own content should be hustling after the technology wizards. Partner with the young companies creating the enabling technologies, or buy them before Microsoft does. Don't reinvent the wheel. To be successful in the new information age the marriage of content and technology should be managed by those who know content, and merely enabled by the technology. Many new products fail because the technology is leading the design. The bottom line is that people pay for answers, and don't really care where or how they get it, as long as they get it in a timely, accurate fashion.

REFERENCES

[1] Berners-Lee et al.(1994), Learning Information Retrieval Agents:" Experiments with Automated Web Browsing". AAAI Spring Symposeum on Information Gathering.

[2] "Integrating external information sources to guide Worldwide Web Information Retrieval". Technical Report CS96-474.

[3] Google Search Engine http://google.stanford.edu

[4] N. Schmidt-Maenz and M. Koch(2005), Patterns in search queries in: *Data Analysis and Decision Support*, D. Baier, R. Decker and L. Schmidt-Thieme, eds, Springer, Heidelberg, 2005, pp. 122–129.

[5] Search Engine Watch http://www.searchenginewatch.com

AUTHOR PROFILES:

Prof J.Ashok is currently working as Professor and Head of Information Technology at Geethanjali College of Engg. & Technology, Hyderabad, A.P, INDIA. He has received his B.E. Degree from Electronics and Communication Engineering from Osmania University and M.E. with specialization in Computer Technology from SRTMU, Nanded,INDIA. His main research interest includes neural networks, data retrieval process and Artificial Intelligence. He has been involved in the organization of a number of conferences and workshops. He has been published around 20 papers in national and International journals and conferences. He is currently doing his Ph.D from Anna University.

Kamalakar Ramineni is working as Lecturer at Satavahana University, Karimnagar, A.P, INDIA. He has received M.Sc Degree in Computer Science and currently pursuing Ph.D at Kakatiya University. His main research interest includes data retrieval process by using search engines.

Dr.E.G.Rajan is an Electronics Engineer and a Professor of signal Processing having about 30 years of experience in teaching, research and administration. He has a number of publications to his credit. He is a Professional member of ACM and he is an editor of the Journal of AMCE, Barcelona, Spain. He received his Ph.D degree in electrical Engineering from Indian Institute of Technology(IIT),Kanpur, U.P., M.E. degree in Applied Electronics from Madras University. His Contribution to the state of the art of Electronic welfare has been well recognized in the Government and industrial sectors. He received distinguished Scientist and man of the Millennium award from who is who Bibliographical records,Cambridge,2000. He authored many books like, symbolic Computing, Signal and Image Processing, Electronic Order of Battle Records of Military Radars, Computers and Information Technology. Two of his Fifteen Ph. D scholars are involved in the design of digital Circuits using Oganic molecules and some of them are working in the Novel area of Genomic Signal Processing. He has brought out more than 25 original concepts.