WEB SEARCH, FILTERING, AND TEXT MINING: TECHNOLOGY FOR A NEW ERA OF INFORMATION ACCESS

BRUCE CROFT

NSF Center for Intelligent Information Retrieval, Computer Science Department, University of Massachusetts, Amherst, MA 01003-4610 USA
E-mail: croft@cs.umass.edu

Received: 30th May 2000 / Published 11th May 2001

ABSTRACT

Much of the information in science, engineering and business has been recorded in the form of text. Traditionally, this information would appear in journals or company reports, but increasingly it can be found online in the World-Wide Web. Tools to support information access and discovery on the Internet are proliferating at an astonishing rate. Some of this development reflects real progress but there are also many exaggerated claims. The focus of this presentation will be to review the important technologies for text-based information access on the Web and to describe the progress that is being made by researchers in these areas.

INTRODUCTION

Ten years ago, the primary technologies being used to construct large information systems were database systems, information retrieval systems, and information filtering systems. Database systems were used to handle large volumes of structured data and to provide guarantees of reliability and consistency despite systems failures and high volumes of update transactions. Information retrieval systems were used to search large databases of text, such as scientific abstracts, legal materials, or newspaper stories. Information filtering or "clipping" services provided periodic updates in the form of text stories, mostly in the business domain, based on user profiles.

In the relatively short period since, there have been many developments that have affected how information technology is talked about and used. The most important of these have been the growth of the Internet and the availability of cheap hardware. The technologies for the large information systems discussed today include the Internet (and intranets and extranets), Web search, portals, agents, collaborative filtering, XML and metadata, and data mining.

There are many opinions about the current technology for information systems, including that everything is different, everything is the same, and everything is a mess. What people generally do agree on is that there is much more data on-line, much of that data is unstructured (i.e. text, image, video), and that the data is much more distributed than in the past. This statement is usually applied to the Web in general, but it also applies, with some reservations, to scientific information.

This paper provides a brief description of some of the new technologies and reviews their current status and future research directions.

SEARCH ENGINES

One of the major tools for information access is the search engine. Most search engines use information

Published in "Chemical Data Analysis in the Large: The Challenge of the Automation Age", Martin G. Hicks (Ed.), Proceedings of the Beilstein-Institut Workshop, May 22nd – 26th, 2000, Bozen, Italy

http://www.beilstein-institut.de/bozen2000/proceedings/croft/croft.pdf

retrieval techniques to rank Web pages in presumed order of relevance based on a simple query. Compared to the bibliographic information retrieval systems of the 70's and 80's, the new search engines must deal with information that is much more heterogeneous, "messy", more varied in quality, and vastly more distributed or "linked".

In the current Web environment, queries tend to be short (1-2 words) and the potential database is very large and growing rapidly. Estimates of the size of the Web range from 500 million to a billion pages, with many of these pages being portals to other databases (the "hidden Web").

In response to this huge expansion of potential information sources, today's Web search engines have emphasized speed and coverage, with less importance attached to effectiveness. With the growing number of complaints about "information overload", however, this is beginning to change. Similarly, most Web search engines use a centralized architecture where "Web crawlers" gather Web pages and a single, very large index is created. An approach like this has inherent scalability problems.

There has been a growing awareness that effective information retrieval is a hard problem. Indeed, in a recent Turing Award lecture, it was identified as a software "grand challenge". To address this challenge, researchers in information retrieval and related areas of computer science are proposing new retrieval models and techniques to support distributed architectures, summarization, question answering, cross-lingual retrieval, better interfaces, and multimodal search.

Retrieval models provide the underlying framework for a search engine. In other words, they are the basis for the algorithms that score and rank the Web pages. Recent developments in this area include ranking algorithms based on link structure (e.g. www.google.com) and language modeling. The algorithms based on link structure analyze link patterns to identify sites that are highly linked. This is similar to the citation analysis techniques developed in the 1970s for scientific articles. Probabilistic techniques based on language modeling are the basis of effective algorithms for a variety of language tasks, such as speech recognition and machine translation, and are beginning to demonstrate effectiveness improvements in large-scale experiments. This work is also being used in the development of cross-lingual techniques, where queries are given in one language and the results are found across a variety of other languages.

There has also been considerably more work recently that is applying natural language processing techniques to the problem of information retrieval. Much of this work is being done under the title of "question answering". The goal of this type of information access is to produce a concise answer to well-formulated queries. In the case of simple queries such as "What is the boiling point of water?" both the answer and the task are well-defined. For other questions such as "What is the best drug for treating high blood pressure?" the answer is much less well-defined and will probably require combining data from a variety of sources. Techniques for distributed retrieval and summarization will be part of the solution.

Researchers in the area of distributed search are developing techniques for identifying relevant information sources, describing their contents, and combining results from multiple searches. Summarization researchers are looking at ways of generating a variety of different types of summary for single documents and groups of documents. The summary types include lists of keywords, extracted sentences, and generated text. Visualization techniques and techniques for automatically generating taxonomies are also important.

One of the key aspects of improving the effectiveness of Web search involves getting better descriptions of the user's information need. A short one or two word query is generally not descriptive of the actual information need and is not helpful to the search engine. Techniques such as automatic query expansion and machine learning through relevance feedback have been developed to address this problem. The growing ubiquity of wireless devices is also leading to a new interest in voice interfaces, which bring a variety of new challenges and opportunities to the designer of Web search engines, including dealing with longer queries.

There has also been considerable research on multimedia and multimodal retrieval. Multimedia retrieval involves algorithms for representing and comparing image and video data. A number of promising techniques have been developed, but large scale experimentation has not been done except for some specialized tasks such as face retrieval. Multimodal retrieval involves frameworks for combining evidence from multiple sources, such as image and text, into overall estimates of relevance for complex objects.

XML/METADATA

XML is a new standard developed for Web page markup or, more generally, for describing the structure of data that is more loosely formatted than a standard database schema instance. It is related to the older SGML and HTML markup standards. There has been a considerable amount of publicity about XML and an increasing number of compatible tools are becoming available. The XML standards activity has also expanded to include the definition of ontologies for the description of document content in addition to the structure. MPEG7 has related aims for video data.

There is no doubt that these efforts on standardizing format and content through XML and metadata will have a large impact on future information systems. There is, however, less reason to believe that this approach will solve the access problem. Manual indexing using controlled vocabularies is one of the oldest methods of text representation that has been used in information retrieval systems. There is abundant evidence that this approach does not scale and, in general, is of limited effectiveness when used as the only representation. The most effective applications of this type of indexing are in limited domains with a substantial investment in ontology development, such as medicine (MESH) and chemistry (Chemical Abstracts). Developing XML-based ontologies in such domains would be useful if done in conjunction with content-based searching and categorization. Categorization is a technique for automatically assigning labels (or controlled vocabulary terms) to new documents. A considerable amount of research has been done in this area and, with enough training data, good results can be achieved. As more XML data described using ontologies and metadata becomes available, categorization techniques will become viable.

INFORMATION FILTERING

Information filtering has been around for some time in the form of "current awareness" systems. A number of Web tools provide this functionality (often under the "agent" label). Most of the applications of this technology are in the business and news domains. Many of these systems use simple Boolean matching techniques, although there has been much research and a number of new companies applying machine learning techniques to this problem. Effective filtering is, however, as difficult as effective search, and the problems involved with proactively sending too much data that is not relevant to the users have resulted in varying levels of acceptance. Many of the

techniques being developed to improve search, however, will also result in more effective filtering so we can expect to see more applications involving this technique in the future.

Collaborative filtering is a complementary technique based on matching user preferences that has become popular in e-commerce applications. It remains to be seen whether the combination of content-based and collaborative filtering will improve information access in scientific and engineering contexts.

TEXT DATA MINING

A considerable amount of research is being carried out under the heading of text data mining. This includes a variety of techniques such as information extraction, clustering, and discovery of associations or "rules". All of these techniques combine statistical methods with some level of linguistic analysis. In contrast to data mining using relational database systems, where a number of commercial packages are available, text data mining is still an open research issue. Evaluation of research in this area is also difficult, since many of the results are presented with examples instead of statistical data.

Information extraction techniques are designed to extract "facts" from text. In many cases, this means very simple facts such as names of companies, people, and monetary amounts, but in general this technique can be used to extract more complex information, such as filling a database according to a template or schema. Extraction is a key component of text data mining since it provides the objects for the statistical analysis. Much of the research in this area has been done with newspaper text, but results with scientific text are beginning to be reported. There has also been recent work focusing on information extraction based on the structure of Web pages.

Clustering is used to group related information. This technique has been well-studied in information retrieval but has recently been the subject of a number of new papers. Information extraction and clustering can be used with other techniques to discover interesting associations in text databases. The applications of this type of discovery have been mostly based on business information, but it may also be useful in scientific and engineering contexts.

"Literature-based discovery" is an interesting area of research that has been underway for some time and is an early example of text data mining. By analyzing the literatures of related fields for topics that are related but not connected by direct reference, Swanson and his colleagues at the University of Chicago have found a number of connections in the medical literature (specifically, Medline abstracts) that have been the subject of follow-up scientific investigations.

CONCLUSIONS

The Web is a huge, relatively unstructured and sometimes unreliable source of information. The development of XML and ontology standards for metadata will promote sharing and introduce a limited amount of structure to the Web, but they are not the whole solution to the information problem. Many new tools are being developed to exploit unstructured information and to make it more useful to specific user communities such as scientists. These tools can also be used for information access and discovery with scientific literature and databases. Techniques such as text data mining, however, will require considerably more research and experimentation before their effectiveness can be established.

Papers describing research in a number of these areas and extensive references to other papers can be found in [1].

REFERENCES AND NOTES

[1] Croft, W.B. (editor), *Advances in Information Retrieval*, Kluwer Academic Publishers, Boston, **2000**.