The Technology of Phrase Browsing Applications

Workshop held in conjunction with the First ACM-IEEE Joint Conference on Digital Libraries

June 28, 2001
Roanoke, Virginia, USA

Co-chairs: Nina Wacholder, Rutgers University and Craig Nevill-Manning, Google

Introduction

Phrase browsing applications provide information seekers with access to text content via structured lists of index terms. The index terms, which may be identified by a variety of techniques, are phrases that have been automatically extracted from full text documents. Browsing applications support interactive navigation of index terms and provide direct access to the original documents via the index terms. Terms are presented to users in ways that allow them to either 'drill down' from a shorter, more general term to longer, more specific ones or to navigate from one term to other related ones via graphical interfaces.
A major advantage of these systems is that they provide users with index terms instead of requiring users to devise the terms themselves, an especially difficult task for an information seeker looking for information about an unfamiliar domain. Because the terms are extracted from documents, phrase-browsing systems can be easily integrated with full-text search and are complementary to standard information retrieval systems. Browsing systems are distinct from organizational systems based on ontologies that do not correspond directly to collection content.

Issues related to the identification of terms and the development of browsing applications, sometimes called phrase browsing, have been discussed in the digital library, information retrieval, and natural language processing communities (for example, Liddy and Myaeng 1993; Nevill-Manning et al. 1997; Anick and Tipirneni 1999; Wacholder et al. 2000). The usability of electronic indexes has also been investigated, for example, by Milstead 1994 and by Hert et al. 2000.The motivation for the workshop was to bring researchers working in this area together in order to address the question of where the technology of phrase browsing currently stands and what are the most important areas for research to speed of development of practically useful phrase browsing applications.

Technology of Phrase Browsing Application was held on June 28, 2001 in conjunction with the first ACM-IEEE Joint Conference on Digital Libraries (JCDL '01). It was attended by 14 people. The workshop brought together researchers with two distinct foci:

1. Development of efficient techniques and methods for navigating and browsing phrases that provide access points to full text documents. In general, these methods work with any index terms, regardless of the technique by which these terms are identified.
2. Development and analysis of techniques for effectively extracting index terms from full text, for systematically determining which phrases are most useful, and for organizing and classifying terms in ways that that take into account human searching behavior and knowledge of language.

Participants agreed that the opportunity to see how advanced the state of the art is in each of these areas was one of the major benefits of the workshop. Probably the single most important conclusion that participants reached was on the importance of bring together research on methods for browsing and hierarchically organizing phrases with research on identification of index terms in text. There was also general consensus that phrase-browsing applications have exciting potential for improving information access in digital libraries.

Papers

The workshop included five papers. The workshop papers and some of the presentations are available at http://www.scils.rutgers.edu/~nina/phrasebrowsing/workshop062801.

Gordon Paynter and Ian Witten's paper, "A combined phrase and thesaurus browser for large document collections", describes an interface that combines a hierarchy of terms extracted automatically from the documents with a manually constructed thesaurus. The collection of documents comes from the web site of the Food and Agriculture Organization (FAO) of the United Nations, and AGROVOC, a thesaurus developed by the FAO. Linking terms extracted from the documents with those in manually constructed thesaurus allows the user to move between terms that are used in the text and in the thesaurus in order to expand or narrow down their search. For example, the thesaurus lists *robinia* as a synonym for *locust*. If the user starts by looking for the technical term *robinia*, the system suggests that the information seeker should search not just for phrases that contain *robinia*, but also *locust.* This interface is an example of how an automatically constructed list of phrases can be merged in a straightforward way with a hand-built thesaurus, thereby combining the advantages of document coverage provided by automatic identification of terms with the advantages of a set of terms systematically organized for access by people.

Gordon Paynter, Craig Nevill-Manning and Ian Witten's paper, "Phrase hierarchy inference", describes two algorithms for identifying a hierarchy of overlapping phrases from a sequence of discrete symbols. The first approach builds a suffix tree or suffix array and performs all of its processing in memory. The second approach makes multiple passes over disk files to count phrases and build expansion lists of terms. The first approach is conceptually simple but memory intensive; it requires primary memory equal to several times the size of the input text. The second approach is more practical because it requires slightly less memory than the size of the input text. Both techniques produce a hierarchy of phrases chosen purely on the basis of frequency rather than on linguistic grounds.

James Cooper's paper, "The technology of lexical navigation" demonstrates a convenient technique for moving between related documents and terms without having to formulate an exact query to retrieve these related entities. This system is built with data extracted from documents by IBM's Textract mining system, which identifies proper names and technical terms. Textract also automatically identifies term relations; they include named relations such as *similar-to* and unnamed relations based on computation of mutual information. The browsing interface is a graphical display of terms extracted from the document and linked by the named and unnamed relations. The relations are stored in a database and returned as an array of Java objects. The advance reported on in this paper is the use of the SOAP (Simple Object Access Protocol) serializer and deserializer to construct the XML data stream and the reconstruct the object on the client side. This efficient method allows users to find related information in disparate documents.

Steve Jones' paper, "Using keyphrases to support flexible reading of on-line documents", describes Phrasier, a system designed to help users make sense of documents that they read online. Phrasier provides four views: 1) *the topic overview* consists of a list of terms that represent overall document content; 2) the *topic location view* shows users, by means of a black bar, where in the document a selected phrase frequently occurs; 3) the *document content view* allows the user to control the degree of visual differentiation between the keyphrases and document content; and 4) the *summary view* displays sentences that have been identified as important in the document.

Nina Wacholder's paper, "The Intell-Index System: Using NLP techniques to organize a dynamic text browser" describes Intell-Index, a phrase browsing system designed for testing the effectiveness of natural language processing techniques for automatic identification of index terms. Simplex noun phrases (noun phrases heads and their content bearing pre-modifiers, e.g. *digital library*) are used in Intell-Index because 1) they can be relatively easily identified using shallow linguistic knowledge, as compared to complex noun phrases (e.g., *digital library in the humanities*) and 2) they include single word noun phrases (e.g., *metadata*), which terms that consist of repeated word sequences do not. Decisions about which index terms to include in a phrase browsing system have important implications for the number, specificity and comprehensibility of index terms identified in full-text documents.

Panel discussion

Judith Klavans' talk, "Browsing and phrases: lessons from the trenches" focused on the distinction between content ("what you see") and form ("how you see it"). In terms of content, metadata and definitions of technical terms are examples of types of information that can usefully enhance phrase browsers, which Klavans illustrated with examples from research in which she and her students are involved (e.g., Klavans and Muresan 2001, Klavans and Whitman 2001). In terms of form, she demonstrated a user interface for browsing definitions and information extracted from these definitions developed at the Columbia University Digital Research Center <http://www.cs.columbia.edu/digigov/>. She also reiterated the need for further study of information seeker's use of data and content and for development of evaluation metrics for phrase browsing applications.

Elizabeth Liddy's talk, "Phrasing technologies, applications and challenges", described an approach to information access that relies on identification of expressions and of relationships among those expressions that can be identified by natural language processing. Liddy reported on a number of applications that have used terms identified by this approach over the past ten years, including document representation and indexing, query representation and expansion, automatic summarization, analysis of transcripts of focus groups; and metadata generation. One example was the DR-LINK system (Liddy and Myaeng 1993), a phrase browsing application that used linguistically motivated units such as noun phrases and proper names to help users select documents that meet a specified information need. Liddy also identified a number of challenges for the development of phrase applications such as dirty data, phrase boundary detection, selection of subsets of useful phrases, and evaluation of the contribution of phrases to larger tasks.

Craig Nevill-Manning suggested that it is important to bring together the two disparate emphases of research in phrase browsing: natural language processing and algorithmic efficiency. The combination will produce simultaneously more plausible phrases and browsing structures, and more practical implementations. The resulting techniques will allow ubiquitous use of phrase browsing systems, which was a clear imperative from the workshop.

Two main issues were raised during the period of discussion that concluded the workshop: whether phrase browsing systems have been adequately evaluated for usability, and how to resolve the tension between efficiency and quality of phrases.

The question of whether phrases have yet been scientifically shown to be useful for information access arose in the discussion. Participants cited several studies that have shown that they are, including Gutwin *et al*. (1999), Jones and Paynter (2001), Jones, S. (1999), Peñas *et al.* (2001), Wacholder *et al*. (2001). However, there was general agreement that a lot more work is needed in this area and that the question remains open of whether providing information access via phrases actually speeds up or otherwise improve the results of the search process.

The second main question that the workshop raised is whether computationally efficient schemes produce plausible phrases, and conversely whether techniques based on natural language processing are fast enough for multi-gigabyte corpora such as Medline and the world-wide web. In answer to this question, two workshop participants presented informal numbers for their systems.

- Jim Cooper of the IBM TJ Watson Research Center reported on the processing speed of Textract (mentioned above), the IBM system for identifying proper names, technical terms and relationships among terms. It took Textract about 32 minutes to process 107,000 documents (472 Mb) for proper names, technical terms, and named relationships. This translates to roughly 885 MB of text an hour, on a 600 MHz Windows 2000 machine with 1 GB of memory. A preprocessing step, written in Java and not yet optimized, converts these 107,000 files into 53 files of about 9 MB each; this preprocessing took an additional 30 minutes. Processing of the unnamed relations for the 107,000 documents currently takes about four hours; work on speeding up this process will take place this year.

- Anselmo Peñas of Universidad Nacional de Educación a Distancia (UNED), Spain, reported that their part-of-speech tagging system tags over 215,000 documents (1020Mb) of Spanish text in about two

hours. The original tagger, called MACO, was developed at the Technical University of Catalonia (UPC) in conjunction with the University of Barcelona (UB). Some heuristics for Spanish were added to the tagger in order to process a large collection.

These speeds are fast enough for processing sizable collections but not for processing the entire web.

Challenges

In addition to the questions discussed above, several open issues were raised:

- How to identify the most important phrases for inclusion in a phrase browsing system from the list of candidate terms extracted from a document;
- How to present complex information seekers in ways that is helpful but not distracting;
- What criteria are most suited for evaluating phrase browsing technology;
- How can the list of terms presented to the user be flexibly adapted to user's level of domain expertise and to corpus characteristics?

Conclusion

The workshop brought together for the first time a small community of researchers working on phrase browsing systems. It crystallized several open research questions the area, and exposed the most significant tradeoff: efficiency versus quality. In setting the research agenda in this area, and demonstrating the state of the art, the workshop was extremely productive, and our discussions will hopefully result in novel published research and practical fielded systems in the next few years.

On-line phrase browsing systems

Workshop participants reported on several phrase-browsing systems that can be accessed via the web.

Phrasier:	http://www.cs.waikato.ac.nz/~stevej/Research/Phrasier/ *[some problems with this URL - ed.]*
Phind/Greenstone:	http://www.nzdl.org/phind/
Phind/BioIR:	http://www.bioir.org/phrase.html
Website Term Browser:	http://rayuela.lsi.uned.es/wtb
Intell-Index:	http://www.cs.columbia.edu/~nina/IntellIndex

References

Anick, P. and Suresh T. (2000) "The paraphrase search assistant: terminological feedback for iterative information seeking", *Proceedings of COLING 2000.*

Gutwin, C., Paynter, G.W., Witten, I.H., Nevill-Manning, C.G. & Frank, E. (1999). "Improving Browsing in Digital Libraries with Keyphrase Indexes," *Decision Support Systems* 27(1-2): 81-104.

Hert, Carol A., Elin K. Jacob and Patrick Dawson (2000) "A usability assessment of online indexing structures in the networked environment", *Journal of the American Society for Information Science* 51(11), 971-988.

Jones, S. (1999). "Phrasier: an interactive system for linking and browsing within document collections using keyphrases". *Interact 99: Seventh IFIP Conference On Human-Computer Interaction*, Edinburgh Conference Centre, Riccarton, Edinburgh, Scotland 30th August - 3rd September 1999, pp 483-490.

Jones, S. and Paynter, G.W. (2001) "Human evaluation of Kea, an automatic keyphrasing system", *Proceedings of First ACM/IEEE-CS Joint Conference on Digital Libraries*, pp.148-156, June 24-27, 2001, Roanoke, VA.

Liddy, E.D. & Myaeng, S. H. (1993). "DR-LINK's linguistic-conceptual approach to document detection", *Proceedings of First Text Retrieval Conference (TREC-1).* NIST.

Klavans, J.L. and Muresan, S. "Evaluation of DEFINDER: A system to mine definitions from consumer – oriented medical text", *Proceedings of First ACM/IEEE-CS Joint Conference on Digital Libraries*, pp.201-202, June 24-27, 2001, Roanoke, VA.

Klavans, J.L. and B. Whitman (2001) "Extracting taxonomic relationships from on-line definitional sources using LEXING", *Proceedings of First ACM/IEEE-CS Joint Conference on Digital Libraries*, pp.257-258, June 24-27, 2001, Roanoke, VA.

Milstead, Jessica L. (1994) "Needs for research in indexing", *Journal of the American Society for Information Science*.

Nevill-Manning, Craig G., Ian H. Witten and Gordon W. Paynter (1997) "Browsing in digital libraries: a phrase based approach", Proceedings *of the DL97*, Association of Computing Machinery Digital Libraries Conference, 230-236.

Peñas, A., Gonzalo, J. and Verdejo, F., (2001) "Cross-Language information access through phrase browsing: Applications of natural language to information systems", *Proceedings of 6th International NLDB Workshop 2001*, Madrid, Lecture Notes in Informatics (LNI), Series of the German Informatics society (GI-Edition). Volume P-3, pp. 121-130, ISBN-3-88579-332-6, ISSN-1617-5468.

Wacholder, N., Evans, D.K. and Klavans, J.L. (2001) "Automatic identification and organization of index terms for interactive browsing", *Proceedings of First ACM/IEEE-CS Joint Conference on Digital Libraries*, pp.126-134, June 24-27, 2001, Roanoke, VA.

Wacholder, Nina, David Kirk Evans, Judith L. Klavans (2000) "Evaluation of automatically identified index terms for browsing electronic documents", *Proceedings of the Applied Natural Language Processing and North American Chapter of the Association for Computational Linguistics (ANLP-NAACL) 2000*. Seattle, Washington, pp. 302-307.