Information & Knowledge Management

Practical Knowledge Representation for Efficient Information Access

Bill Woods, Sun Microsystems

Finding information and organizing information so that it can be found are two key problems for any knowledge-management system. This talk describes an experiment combining the respective strengths of humans and computers in a knowledge-based system to help people find information in text. Unlike many previous attempts, this system demonstrates a substantial improvement in search effectiveness by using linguistic and world knowledge and exploiting sophisticated knowledge-representation techniques. The system uses taxonomic subsumption technology on a large scale to organize and access domain-independent linguistic and world knowledge. It integrates syntactic, semantic, and morphological relationships to help solve some of the paraphrase problems that occur when there are terminology differences between what you ask for and what you need to find.

The MERL SpokenQuery System

Bhiksha Raj, Peter Wolf, MERL

In information retrieval systems one usually types the keyterms into a query engine, such as Google, which then returns all documents pertinent to the query. There are, however, several situations where it is inconvenient or impossible to type in queries, such as when the device used for information retrieval is too small for a keyboard, e.g., PDAs or cellphones, or when hands-free operation is required, e.g., while driving a car. In these cases it is much more convenient for the user to be able to speak the query, rather than to type it. The SpokenQuery system is an enabling technology for such a spoken interface for information retrieval systems, or more generally for database access.

The conventional approach to this task would be to use a speech recognizer to convert the spoken utterance to a text transcription, which would then be passed on to an information retrieval engine. The IR engine would be unaware that the query was spoken and not typed. There are several problems with this approach. 1) Speech recognition engines make mistakes, which can result in poor performance. 2) Speech recognition engines may not have the specialized words that characterize many documents in their vocabularies. 3) Text-based IR systems do not have an indexing mechanism that can cope with errors in the query.

The MERL SpokenQuery system architecture tackles all three problems. First, it uses the search space of the recognizer, as represented by the recognition lattice, instead of the recognition output, to perform retrieval. Second, it passes information from the index back to the recognizer in order to enable the recognizer to better identify important keyterms in the spoken query. Finally, the document index is based on an SVD-based representation that permits comparison of vector representations of the recognizers search space against the index. In this talk we will describe all of these components, and discuss their merits and limitations.

Question Answering: Is More Always Better?

Susan Dumais, Microsoft Research

This talk describes a question-answering system designed to capitalize on the tremendous amount of data now available online. Most question-answering systems use a wide variety of linguistic resources. We focus instead on the redundancy available in large corpora as an important resource. We use this redundancy to simplify the query rewrites that we need to use, and to support answer mining from returned snippets. Experimental results show that question-answering accuracy can be greatly improved by analyzing more and more matching passages. Simple passage ranking and N-gram-extraction techniques work well in our system, making it efficient to use with many backend retrieval engines.

Turning the Web Into a Knowledge Base: Information Extraction with Finite-State Models

Andrew McCallum, U Mass, Amherst, formerly of WhizBang Labs

The Web is the world's largest knowledge base. However, its data is in a form intended for human reading, not manipulation, mining and reasoning by computers. Today's search engines help people find web pages. Tomorrow's search engines will also help people find "things" (like people, jobs, companies, products), facts and their relations.

Information extraction is the process of filling fields in a database by automatically extracting sub-sequences of human-readable text. Finite-state machines are the dominant model for information extraction both in research and industry. In this talk I give several examples of information extraction tasks performed at WhizBang Labs, and then describe two new finite-state models designed to take special advantage of the multifaceted nature of text on the web. Maximum entropy Markov models (MEMMs) are discriminative sequence models that allow each observation to be represented as a collection of arbitrary overlapping features (such as word identity, capitalization, part-of-speech and formatting, plus agglomerative features of the entire sequence and features from the past and future). Conditional random fields (CRFs) are a generalization of MEMMs that solve a fundamental limitation of MEMMs and all other discriminative Markov models based on directed graphical models. I introduce both models, skim over their parameter estimations algorithms, and present experimental results on real-world tasks.

(Joint work with Fernando Pereira, John Lafferty, Dayne Freitag, and many others at WhizBang Labs.)

Model-Based Retrieval of Multimodal Information and Biosurveillance

Chung-Sheng Li & John R. Smith, IBM Watson

Most existing information retrieval applications are based on similarity retrieval of templates or examples, such as similarity retrieval of text and image documents. In such retrievals, the query usually consists of a number of keywords or phrase (for text), or features of an image or a segment of image. Each of the documents (text or image) in the database or digital library is usually represented as one or more vector(s) in a multi-dimensional feature space. The query processing of such similarity retrieval usually involves identifying in the feature space those vectors that have the smallest Euclidean distance to the vector that corresponds to the query target. This similarity retrieval paradigm, however, is not entirely suitable for many scientific and business decision support applications, which are mostly based on models.

The main challenge of applying models to large archives is scalability. Although most of the applications require the retrieval of only a very small subset of the results that maximize or minimize the model, almost all existing methods require applying the model sequentially over the entire region of the data. In this talk, I describe the SPIRE project which uses model-based information retrieval framework to address this challenge. In particular, the focus will be on models for extracting and searching complex geology structures (reflectors, horizons, faults, river delta lobe), locating high risk regions that might be vulnerable to Hantavirus pulmonary disease, extracting boundaries of wetland, and the latest effort on detecting bioterrorism from nontraditional data sources.

Search the Speech, Browse the Video

Alex Cozzi, IBM Almaden

This talk describes search and browse technologies within CueVideo, a multimedia research project at IBM Almaden Research that consists of an automatic multimedia indexing system and a client-server video-retrieval system. Their approach to multimedia retrieval is "Search the speech, browse the video". The video and audio are considered as two parallel media streams of information that are related by a common time line. Thus, they take advantage of the two parallel streams, using the audio stream for search and the video stream for quick visual browsing in a complementary manner to provide the desired video search functionality. The video indexing automatically detects shot boundaries, generates a shots table, and extracts representative key-frames as JPEG files from each of the shots. Several browsable video summaries are generated for rapid browsing. The audio processing starts with speech recognition followed by text analysis and information retrieval. Several searchable speech indexes are created, including an inverted word index, a phonetic index and a phrase glossary index. On the search, this talk focuses on the word and phonetic speech indexing. On the browsing, this talk focuses on video summaries and visualizations that assist in rapid browsing.

Multimedia Indexing

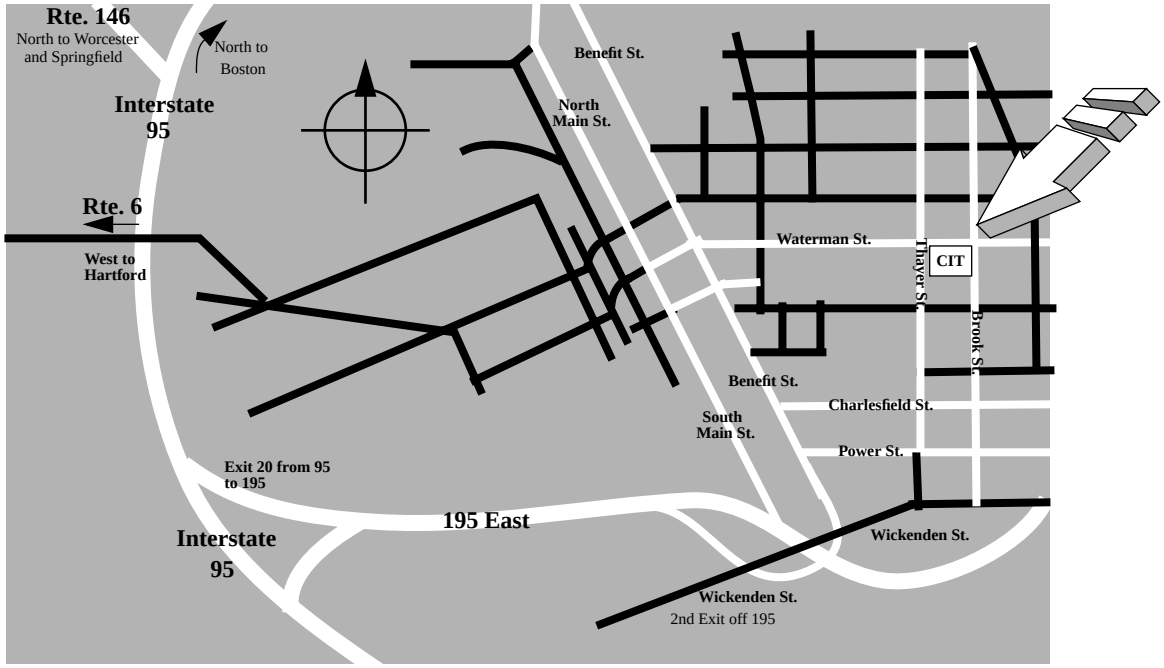
Pedro Moreno, Hewlett-Packard

During the last two years HP Cambridge Research Lab has explored indexing and search technologies and its application to several mulitimedia types such as audio, images and music. I'll describe some of the systems we have built with a special emphasis on our SpeechBot (http://www.speechbot.com) audio-indexing system. I'll give an overview of this audio search engine, its current limitations and some of the technologies we have explored to improve and extend its capabil-

continued →

SCHEDULE

8:30	B R E A K F A S T and Registration 4th floor, CIT Building
9:00	**Introduction and Overview** Eli Upfal, CS Chairman John Savage, IPP Director Emeritus Thomas Hofmann, Host
9:30	**Practical Knowledge Representation for Efficient Information Access** Bill Woods, Sun Microsystems
10:20	B R E A K
10:40	**The MERL SpokenQuery System** Bhiksha Raj, Peter Wolf, MERL
11:30	**Question Answering: Is More Always Better?** Susan Dumais, Microsoft Research
12:20	B U F F E T L U N C H and Posters
1:20	**Turning the Web Into a Knowledge Base: Information Extraction with Finite-State Models** Andrew McCallum, U Mass, Amherst
2:10	**Model-Based Retrieval of Multimodal Information and Biosurveillance** Chung-Sheng Li and John R. Smith IBM Watson
3:00	B R E A K
3:20	**Search the Speech, Browse the Video** Alex Cozzi, IBM Almaden
4:10	**Multimedia Indexing** Pedro Moreno, Hewlett-Packard
5:00	FINAL DISCUSSION and WRAP-UP
5:30	R E C E P T I O N (5th floor)



ities. Among others, I'll describe our experiements with Bayesian belief networks for topic segmentation, boosting for confidence scoring and particle (subword) based recognition for improved IR performance.

◆I◆I◆

This symposium is a benefit of membership in our **Industrial Partners Program**.
Member companies are: EMC, Invensys (Foxboro), Fidelity Investments, GTECH, Hewlett-Packard (Compaq), IBM, MERL, Microsoft and Sun. There is no charge.

EMAIL REGISTRATION

To: sjh@cs.brown.edu
By: Friday, November 8, 2002
Please include the following:
Name, title
Company, Department
Postal address
Phone/Fax

DIRECTIONS TO THE CIT BUILDING

- From I-95 N or S, take Exit 20 to I-195E.
- From I-195E take Exit 2, Wickenden St.
- Go LEFT on Wickenden, LEFT again at the 2nd light onto Brook St.
- The red-brick CIT Building (Center for Information Technology) is on the left at the intersection of Brook and Waterman (1st light).
- *Registration is on the 4th floor.*

PARKING

Because most of the visitor parking has been assigned to University employees, I'm afraid we're unable to provide parking. Street parking is usually available for early birds, but watch out for newly designated 2- and 3-hour zones, which used to be all-day spots. You might try the residential area NW of the CIT.

The 30th IPP Symposium

Department of Computer Science

BROWN UNIVERSITY

SYMPOSIUM on INFORMATION & KNOWLEDGE MANAGEMENT

Thursday November 14, 2002

Host: Professor Thomas Hofmann

INDUSTRIAL PARTNERS PROGRAM