



**International School of Information Management
University of Mysore,**

ISiM Special Lecture by

Dr. Mark T. Maybury

The MITRE Corporation
202 Burlington Road
Bedford, MA 01730, USA

maybury@mitre.org

Tel: (781) 271-7230 Fax: (781) 271-2780

<http://itc.mitre.org>

January 24, 2009 : 3.30PM

Title of the talk

**Next Generation Information Access:
State of the Art Tools and Methods**

Detailed Report

MOTIVATION

Are you ever frustrated not being able to find what you want to on the web? When you find what you are interested in, would you like to be able to read the material in half the time with no information loss? Would you like to be able to automatically extract information from massive document collections? Would you like to be able to read foreign language sites? Would you like the ability to simply ask a question and get an answer? Would you like to benefit from your past efforts to find things and/or to benefit from the efforts others have made to find things? Curious about how Google, Amazon, or Flickr work? If you want answers to these questions and want to increase your own search ability or if you are just interested in the cutting edge technology for search, this course is for you.

OVERVIEW

Today most users need to manually perform data searches, extract information from retrieved data, summarize and interpret the results, and form conclusions based on the results. In contrast, intelligent information access tools can facilitate these activities for the engineer, analyst or manager throughout the process, decreasing task time and increasing comprehensiveness and accuracy of search if tools are appropriately chosen and applied. The purpose of this course is to provide an overview and demonstrations of

the most important intelligent information access technologies: information retrieval, summarization, information extraction, text clustering, question answering and collaborative filtering/social bookmarking.

BENEFITS OF ATTENDING

Participants who attend the course will:

- Learn the current state of the art in information retrieval, summarization, information extraction, text clustering, question answering, and collaborative filtering/social bookmarking.
- Understand the applicability of tools for specific discovery and analysis tasks
- See live demonstrations of several of the technologies if internet connectivity is available

While the course will not make specific tool recommendations, it will provide participants with a list of internet or commercially available tools in each of these categories.

Keywords

Information access, information retrieval, document summarization, information extraction, text clustering, and question answering, collaborative filtering, social bookmarking.

Target Audience

This course is intended for engineers, analysts, tool builders and program managers who want to get a solid understanding of the range of capabilities available to intelligently access information. The course will be useful to researchers and practitioners interested in using or designing intelligent information access systems.

Prerequisites

There is no prerequisite knowledge required, although general knowledge of information technology will enhance the value of this course for participants.

OUTLINE

The one day course (8 am to 5 pm) will describe, demonstrate, and explain the state of the art in the following technologies:

- Information retrieval
- Summarization
- Information extraction
- Text clustering
- Question answering
- Machine translation
- Collaborative filtering and social bookmarking

INFORMATION RETRIEVAL

The state of the art in information retrieval will be summarized to show how ranked lists of documents can be obtained from queries. Information retrieval methods such as indexing and query transformation will be discussed along with how one evaluates the results. Today, systems can return documents across languages relevant to a particular subject with around 80% precision but low recall (or vice versa). Advances in automated query expansion and relevancy feedback have achieved near human retrieval performance. We summarize results from NIST's annual Text Retrieval Conference (TREC). We describe how search engines exploit statistics (e.g., Term-frequency-inverse document-frequency (TFIDF) and co-occurrence, structure (e.g., Google's use of link analysis) and format to enhance retrieval. Emerging systems exist that provide content based retrieval of speech, imagery, and video. We contrast search engines, directories, metacrawlers, and content providers. We summarize how the role of the semantic web in enhancing retrieval.

SUMMARIZATION

Summarization is the technology process that distills the most important information from a source (or sources), and produces an abridged version of the information as either an abstract or an extract. With newspaper text, analyst can summarize documents to 20% of their source size without information loss, saving themselves 50% of task time (Mani and Maybury, 1999). We outline summarization

evaluations conferences such as SUMMAC, the Japanese Text Summarization Challenge, and the Document Understanding Conference summarization evaluation (<http://duc.nist.gov>).

INFORMATION EXTRACTION

Delving deeper, information extraction is used to identify semantic elements within a body of text, for example, entities such as people places or things, properties such as characteristics of entities, or relationships. Current systems are able to extract named entities in news with over 90% accuracy and relations among entities at 70-80% accuracy.

TEXT CLUSTERING

Text clustering is the process of detecting topics within a document collection, creating a taxonomy of these topics, assigning documents to the topics, and then labeling these topic clusters so they can more easily be used by various tools. A number of commercial tools are available on line (e.g., vivismo.com) and they may include visualization of search results such as link node diagrams (e.g., www.kartoo.com).

QUESTION ANSWERING

Question answering uses several of the previously discussed technologies. In question answering, questions are analyzed and augmented by the system, documents are retrieved using this augmented question, answers are extracted from these candidate documents, and a ranked set of possible answers is provided to the user. Using the best performing question answering system, an analyst can retrieve answers to simple factual questions from relevant documents at 75% accuracy (Maybury, 2004). On line question answering systems include AskJeeves (www.ask.com) and Language Computer Corp (www.languagecomputer.com).

MACHINE TRANSLATION

With over 6000 languages spoken in our global village, the demand for machine translation has fueled increasingly effective methods for automated machine translation from a source language (e.g., Chinese) to a target language (e.g., English). Methods have ranged from simple word to word translation to syntactic and deeper semantic understanding and transfer (possibly via an interlingua. In the past decade, large scale corpus-driven statistical models have yielded significant performance improvements. Today web-based, gist quality translation (e.g., Systran) is only a click away. Machine learning from parallel corpora and translation memories enable rapid development and customization. We will briefly summarize the state of the art and describe high quality commercial (e.g., Language Weaver) and free, gist quality tools on the web (e.g., Babelfish).

COLLABORATIVE FILTERING/ANNOTATION

Often wisdom arises from the collective. We will describe collaborative filtering systems like MovieLens or Amazon.com which enable discovery of relevant items based on users previous choices. Finally, we will explore systems such as Flickr (for photo sharing) and del.icio.us (for collaborative bookmarking) that enable discovery via collaborative annotation (tagging) and social bookmarking. We will discuss folksonomies

SUMMARY

Effectively exploited, intelligent information access systems promise many benefits. These include:

- More *strategic* management of intellectual resources – unlocking the full enterprise potential.
- More *efficient* knowledge discovery -- enabling more rapid knowledge discovery with less work.
- More *effective* knowledge application -- tailoring information access to individual needs.

AUDIO VISUAL REQUIREMENTS:

This course will require a projector for powerpoint. If live demonstrations are desired, an internet connection is needed.

INSTRUCTOR

As Executive Director of MITRE's Information Technology Division, Dr. Mark Maybury is responsible for the direction of MITRE advanced research and development for information and knowledge systems. Mark has organized international conferences, given tutorials, and published over fifty articles in the areas of language generation, multimedia presentation, text summarization, intelligent information retrieval and analysis. Mark is editor of *Intelligent Multimedia Interfaces* (AAAI/MIT Press, 1993), *Intelligent Multimedia Information Retrieval* (AAAI/MIT Press, 1997), *New Directions in Question Answering* (AAAI/MIT Press, 2004), co-editor of *Readings on Intelligent User Interfaces* (Morgan Kaufmann Press, 1998), *Ad-*

vances in Text Summarization (MIT Press, 1999) and *Advances in Knowledge Management: Classic and Contemporary Works* (MIT Press, 2001) and co-author of *Information Storage and Retrieval: Theory and Implementation, 2nd Edition* (Kluwer Academic, 2000) and co-editor of *Knowledge Management* (MIT Press 2000). Mark serves on the Board of the Object Management Group, is a former officer of ACM SIGART, and a member of the Steering Committee for ACM IUI. Mark received his B.A from College of the Holy Cross (1986), an MBA from RPI (1989), and his M.Phil. in Computer Speech and Language Processing (1987) and Ph.D. in Artificial Intelligence (1991) at Cambridge University, UK.

REFERENCES

1. Advanced Question Answering for Intelligence (AQUAINT) www.ic-arda.org/InfoExploit/aquaint.
2. Fukusima, T. and Okumura, M. 2001. "Text Summarization Challenge: Text summarization evaluation in Japan." Workshop on Automatic Summarization. Second Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL'2001). New Brunswick, New Jersey: Association for Computational Linguistics.
3. Google Text Mining References: http://directory.google.com/Top/Reference/Knowledge_Management/Knowledge_Discovery/Text_Mining/
4. Light, M. and Maybury, M. 2002. Personalized Multimedia Information Access: Ask Questions, Get Personalized Answers. *Communications of the ACM* 45(5): 54-59. (www.acm.org/cacm/0502/0502toc.html).
5. Mani, I. and Maybury, M., editors, 1999. *Advances in Automatic Text Summarization*. MIT Press
6. Maybury, M. T. (ed.) 1997. *Intelligent Multimedia Information Retrieval*. Menlo Park: AAAI/MIT Press. (<http://www.aaai.org:80/Press/Books/Maybury-2/>)
7. Maybury, M. T. editor. 2004. *New Directions in Question Answering*. AAAI/MIT Press.