

2008-09 Upcoming DHS Brown Bag Talks

MIAS Multi-Modal Information Access and Synthesis Center

The University of Illinois at Urbana-Champaign,

Title: Making Sense of Unstructured Information

Speaker: Dan Roth, MIAS Center Director and Professor of Computer Science, UIUC

Abstract:

Recent studies have shown that over 85% of the information organizations deal with is *unstructured* – the vast majority of which is text in different forms. A multitude of techniques has to be used in order to enable intelligent access to this information and to support transforming it to forms that allow sensible use of the information.

The fundamental issue that all these techniques have to address is that of semantics – there is a need to move toward understanding the text at an appropriate level, beyond the word level, in order to support access, knowledge extraction and synthesis.

We will discuss some of our research in these directions, addressing several dimensions of text *understanding* that can facilitate access to and extraction of knowledge from unstructured text, transforming it to forms that are useful to different users in different settings, and integrating it along multiple dimensions and with existing institutional resources.

Title: Semantic Abstraction and Integration across Text Documents and Data Bases

Speaker: Dan Roth, MIAS Center Director and Professor of Computer Science, UIUC

Abstract:

Intelligent access to information requires semantic integration of structured databases with unstructured textual resources.

We will discuss some of our research towards identifying mentions of entities in text, identifying relations that may hold between entities and concepts in texts and determining whether different mentions of real-world entities, within and across documents, actually represent the same concept. We will also discuss using these capabilities to search unstructured text and to integrate information extracted from unstructured data with existing institutional knowledge bases.

Title: Discovering and analyzing topic patterns in text collections

Speaker: ChengXiang Zhai, Associate Professor of Computer Science, UIUC

Abstract:

With the explosive growth of online information, we have an urgent need for powerful text mining tools to help us digest and exploit the huge amount of information. A common task occurring in many different applications is to extract topics from text collections (e.g., major topics covered in blog articles about "Hurricane Katrina") and analyze their variations over various kinds of context such as time, location, or sources (e.g., how the coverage of subtopics differs in different places or at different time). In this talk, I will present a suite of general methods for discovering topics and analyzing their variations in text collections.

These methods are all based on probabilistic modeling of topics in text. They can be used either in an unsupervised manner to automatically learn topics and variations or in a semi-supervised manner to incorporate any user preferences. All the methods are general and can be applied to many different mining tasks, such as temporal text mining, spatiotemporal text mining, author-topic analysis, cross-collection comparative analysis, and sentiment summarization. Sample experiment results from these applications will be presented.

Title: Entity Search: Finding Stuff on the Web, Directly and Holistically

Speaker: Kevin Chang, Associate Professor of Computer Science, UIUC

Abstract:

What have you been searching lately?

With so much data on the Web, we often search for various "stuff" (e.g., a phone number, a paper, a name, a date). However, current search engines only *indirectly* take us to pages. With the scale of the Web, the stuff that we are looking for usually appears in many pages, but current search engines find pages *individually*. In support of developing more direct and holistic search mechanisms, the Web Indexing and Search for Dynamic Mining (WISDM) project at the University of Illinois is building a new search system for finding our target "entities." For such entity search, I will motivate its query semantics, develop the search mechanism, and demonstrate the current prototype in several real-world application scenarios.

Title: Real-Time Anomaly Mining in Massive Data Streams

Speaker: Jiawei Han, Professor of Computer Science, UIUC

Abstract: Traditional data mining systems and methods assume that data resides on disks or in main memory, and a data mining process can scan the data sets multiple times to uncover interesting patterns. However, real-time systems and other dynamic environments often generate tremendous (potentially infinite) volumes of stream data in fast speed. Many applications require real time mining of unusual patterns in data streams, including finding unusual network or telecommunication traffic, real-time pattern mining in video surveillance, and detecting suspicious on-line transactions or terrorist activities. Recently there have been substantial growing research interests in developing methods for querying and mining stream data. In this talk, I will first present an overview of recent developments on stream data mining and outline what are the major challenging research problems for mining dynamics of data streams in multi-dimensional space. In particular, I will be addressing the following issues in detail: (1) multi-dimensional on-line analysis methods for discovering unusual patterns in stream data; (2) dynamic stream classification methods; and (3) mining clusters in stream data.

Title: Exploring the Power of Links in Information Network Mining

Speaker: Jiawei Han, Professor of Computer Science, UIUC

Abstract: Information network analysis has been studied extensively in web search with algorithms like PageRank and HITS invented that explore page links for discovery of authoritative web pages and hubs. We show that the power of links can be systematically explored in the mining of information networks for tasks like link-based classification, clustering, information integration, and veracity analysis. Some recent results of our research that explore the crucial information hidden in links will be introduced, including (1) multi-relational classification, (2) user-guided clustering, (3) link-based clustering, (4) object distinction analysis, and (5) veracity analysis. We will discuss how to apply these techniques for DHS applications as well. The technical material is based on one thesis of my students that has received ACM SIGKDD 2008 Dissertation Award.