SIKS Advanced Course

Probabilistic Methods for Entity Resolution and Entity Ranking

20-21 April 2009, Conference Center Woudschoten, Zeist

Monday (April 20, 2009)

10.00 - 10.30 **An introduction to the course**
Maurice van Keulen and Djoerd Hiemstra (University of Twente)

10.30 - 11.00 Coffee break

11.00 - 12.30 **Probabilistic Databases**
Maurice van Keulen (University of Twente)

12.30 - 13.30 lunch

13.30 - 15.00 **Probabilistic data integration approach to entity resolution**
Ander de Keijzer (University of Twente)

15.00 - 15.30 Coffee break

15.30 - 17.00 **Evaluating Entity Ranking in Wikipedia**
Arjen de Vries (CWI and TU Delft)

17.00 - 18.00 **Exercise session**

Tuesday (April 21, 2009)

9.00 - 10.30 **Basic Retrieval models for entity ranking**
Djoerd Hiemstra (University of Twente)

10.30 - 11.00 Coffee break

11.00 - 12.30 **Retrieving Entities**
Maarten de Rijke (University of Amsterdam)

12.30 - 13.30 Lunch

13.30 - 15.00 **Entities in chains**
Antal van de Bosch (University of Tilburg)

15.00 - 15.30 Coffee

15.30 - 16.30 Exercise session and closing

Course Description

The course addresses database, search, and natural language processing problems that need to relate structured data and unstructured text to real world entities such as persons, organizations, geographic locations, etc. In many applications it is not enough to mark a database field or a text string like "Vandenberg" as a person name, but the system should also relate it to a real world entity, for instance to the famous Enschede rock guitarist "Adrian Vandenberg", or possibly to his band "Vandenberg". Entities, such as Adrian Vandenberg, might occur with different text strings in databases and texts: Other sources might refer to him as "Adje van den Berg", as "Adje, or simply as "he". Entity extraction and entity resolution is needed in applications that do not have full

control of the data, such as applications working on natural language text, Web 2.0 applications that support user-generated data, and applications that need to integrate structured data from several sources. The SIKS course will address methods to reason about real world entities in applications from several angles, covering topics as:

- named entity extraction,
- co-reference resolution,
- statistical language modeling,
- entity retrieval and entity ranking,
- expert search,
- data integration, and
- probabilistic databases

The course will have two exercise sessions at the end of each day. After the course, the students are able to describe problems of entity resolution and entity ranking in several applications, and to apply advanced modeling techniques in these areas, for instance: EM-training of language models, probabilistic random walks, and possible world semantics.

Abstracts

Probabilistic data integration approach to entity resolution

by Ander de Keijzer (University of Twente)

The main problem with entity resolution is caused by semantics hidden in the data. Semantical problems can only be accurately solved by humans and, hence, entity resolution is a time consuming task. The only reason that a human is really needed, is that when decisions about equality are made, unlikely combinations are discarded, if we could keep alternatives and associate confidence scores with them, the process can be automated, since no data is lost. At a later stage in time, e.g. query time, a user can provide feedback and eliminate incorrect data. We will discuss the integration problem in detail and provide a framework for data integration using uncertain data.

Probabilistic Databases

by Maurice van Keulen (University of Twente)

Data exchange between embedded systems and other small or large computing devices increases. Since data in different data sources may refer to the same real world objects, data cannot simply be merged. Furthermore, in many situations, conflicts in data about the same real world objects need to be resolved without interference from a user. In this report, we report on an attempt to make a RDBMS probabilistic, i.e., data in a relation represents all possible views on the real world, in order to achieve unattended data integration. We define a probabilistic relational data model and review standard SQL query primitives in the light of probabilistic data. It appears that thinking in terms of 'possible worlds' is powerful in determining the proper semantics of these query primitives.

Basic retrieval methods for entity ranking

by Djoerd Hiemstra (University of Twente)

In this talk, I will give an introduction into four basic probabilistic models for information

retrieval: the binary independence probabilistic model, the language modeling approach, Google's pagerank, and probabilistic latent semantic indexing. I will then present four cases in which these models are applied to results of information extraction to retrieve for instance persons or concepts. Models will be explained in tutorial style. Goal of the lecture is to make the students aware of the consequences of modeling assumptions. After the course, students should be able to choose a model of information retrieval that is adequate in new applications of information retrieval. *Expert search* and *concept-based video search* are used as examples of such new retrieval applications.

Entity Ranking in Wikipedia

by Arjen de Vries (TU Delft and CWI)

Information retrieval evaluation assesses how well systems identify information objects relevant to the user's information need. Traditional evaluations have used the following working definition of relevance: If you were writing a report on the subject of the topic and would use the information contained in the document in the report, then the document is relevant. Here, a document is judged relevant if any piece of it is relevant (regardless of how small that piece is in relation to the rest of the document). Many realistic user tasks seem however better characterized by a different notion of relevance. Often, users search for specific entities instead of just any type of documents. Example information needs include "Countries where one can pay with the euro" or "Impressionist art musea in The Netherlands". To evaluate retrieval systems handling these typed information needs, the Initiative for Evaluation of XML Retrieval (INEX) has started to build a test collection for entity retrieval in Wikipedia, where the entities are assumed to correspond to Wikipedia entries. The talk discusses the consequences of modifying the definition of relevance on retrieval system evaluation.

Retrieving Entities

by Maarten de Rijke (University of Amsterdam)

Now that document retrieval has become somewhat of a commodity, the information retrieval community is increasingly considering tasks that revolve around entities rather than documents. Examples include product search, finding answers or locations, and profiling people or organizations. In this talk I will review some recent work on entity retrieval at the University of Amsterdam. Important building blocks for this work include named entity normalization and association finding. And prominent applications that will be discussed include expert finding and online media analysis. The talk is based on joint work with Sisay Fissaha Adafre, Leif Azzopardi, Krisztian Balog, Maarten Marx, Valentin Jijkoun, Mahboob Khalid, and Wouter Weerkamp.

Entities in Chains

by Antal van den Bosch (Tilburg centre for Creative Computing, Tilburg University)

In larger information systems that require sentence-level text analytics, named entity recognition provides information on the "who" and "what" in the text. Named entity recognition has often been taken as a generic task with about four to eight types (persons, organizations, locations, ...), but in restricted domains, domain-specific entities are more typically the goal of analysis. As the state of the art is pushed forwards, attention has been shifting to two more complicated tasks: (1) linking entities and linguistic anaphors (such as pronouns) up into co-reference chains that run through an entire discourse or text, and (2) detecting entities with a high paraphrase rate, i.e. referred to with many different expressions, such as events and dates. I provide an overview of the current developments, and highlight the biggest challenges.