CIKM'10 Tutorials

Tutorial 1: Harvesting Knowledge from Web Data and Text

Tuesday, October 26, 2010 ***(Room: Algonquin)***

Tutorial Organizers:
Hady W. Lauw, Ralf Schenkel, Fabian Suchanek, Martin Theobald, and Gerhard Weikum

Time: ***9:00 – 12:30***

Abstract:

The Web bears the potential of being the world's greatest encyclopedic source, but we are far from fully exploiting this potential. Valuable scientific and cultural content is interspersed with a huge amount of noisy, lowquality, unstructured text and media. The proliferation of knowledge-sharing communities like Wikipedia and the advances in automated information extraction from Web pages give rise to an unprecedented opportunity: Can we systematically harvest facts from the Web and compile them into a comprehensive machine-readable knowledge base? Such a knowledge base would contain not only the world's entities, but also their semantic properties, and their relationships with each other. Imagine a "Structured Wikipedia" that has the same scale and richness as Wikipedia itself, but offers a precise and concise representation of knowledge, e.g., in the RDF format. This would enable expressive and highly precise querying, e.g., in the SPARQL language (or appropriate extensions), with additional capabilities for informative ranking of query results. The benefits from solving the above challenge would be enormous. Potential applications include 1) a formalizedmachine-readable encyclopedia that can be queriedwith high precision like a semantic database; 2) a key asset for disambiguating entities by supporting fast and accurate mappings of textual phrases onto named entities in the knowledge base; 3) an enabler for entity-relationship-oriented semantic search on the Web, for detecting entities and relations in Web pages and reasoning about them in expressive (probabilistic) logics; 4) a backbone for natural-language question answering that would aid in dealing with entities and their relationships in answering who/where/when/ etc. questions; 5) a key asset for machine translation (e.g., English to German) and interpretation of spoken dialogs, where world knowledge provides essential context for disambiguation; 6) a catalyst for acquisition of further knowledge and largely automatedmaintenance and growth of the knowledge base. While these application areas cover a broad, partly AI-flavored ground, the most notable one from a database perspective is semantic search: finally bringing DB methodology to Web search! For example, users (or tools on behalf of users) would be able to formulate queries about succulents that grow both in Africa and America, politicians who are also scientists or are married to singers, or flu medication that can be taken by people with high blood pressure. The search engine would return precise and concise answers: lists of entities or entity pairs (depending on the question structure), for example, Angela Merkel, Benjamin Franklin, etc., or Nicolas Sarkozy for the questions about scientists. This would be a quantum leap over today's search where answers are embedded if not buried in lots of result pages, and the human users would have to read them to extract entities and connect them to other entities. In this sense, the envisioned large-scale knowledge harvesting fromWeb sources may also be viewed as machine reading.

Tutorial 2: Online Advertising Business Models, Technologies and Issues

Tuesday, October 26, 2010 ***(Room: York)***

Tutorial Organizers:
Deepak K Agarwal, James G. Shanahan

Time: ***9:00 – 12:30***

Abstract:

Over the past 15 years online advertising, a $65 billion industry worldwide in 2008, has been pivotal to the success of the world wide web. This success has arisen largely from the transformation of the advertising industry from a low-tech, human intensive, "Mad Men" way of doing work (that were common place for much of the 20th century and the early days of online advertising) to highly optimized, mathematical, computer-centric processes (some of which have been adapted from Wall Street) that form the backbone of many current online advertising systems. This half-day tutorial (3-4 hours) focuses on helping researchers and developers attending CIKM acquire new skills as well providing an account of the latest developments in this diverse and relatively new scientific discipline of online advertising. In doing so it will provide a clear and detailed overview of the technologies and business models that are transforming the field of (online) advertising along the following themes: business models; market size and scope; forward and spot markets; machine learning and statistical technologies; economic models; new directions such as social advertising and behavioral targeting; challenges and open issues that face online advertising. Participants will learn about: • the primary business models that make online advertising • the online advertising markets, their revenues and sizes • forward and spot markets • linear programming, quadratic programming, Markowitz model • auction models, game theory and how it can be used to analyze online auctions • metrics and evaluation practices (online evaluation through designed experiments like AB tests, Factorial designs; offline evaluation through logged data) • Modern collaborative filtering methods • Bayesian models, online learning and multi-armed bandit schemes. • learning to rank • predict ad quality/click-thru-rates • automatically targeting ads both in sponsored search and in contextual advertising • understand the specific issues that face online advertising such as privacy, deception and fraud (in particular, click fraud, the spam of online advertising) and current approaches • how advertising is being leveraged in Web 2.0 applications such as social networks, and video/photo-sharing • new directions: behavioral targeting, social advertising, data exchanges • the open research issues that face the field of online advertising.

Tutorial 3: Real World Text Mining

Tuesday, October 26, 2010 ***(Room: Algonquin)***

Tutorial Organizers:
Ronen Feldman and Lyle Ungar

Time: ***14:00 – 17:30***

Abstract:

The proliferation of documents available on the Web and on corporate intranets is driving a new wave of text mining research and application. Earlier research addressed extraction of information from relatively small collections of well-structured documents such as newswire or scientific publications. Text mining from the other corpora such as the web requires new techniques drawn from data mining, machine learning, NLP and IR. Text mining requires preprocessing document collections (text categorization, information extraction, term extraction), storage of the intermediate representations, analysis of these intermediate representations (distribution analysis, clustering, trend analysis, association rules, etc.), and visualization of the results. In this tutorial we will present the algorithms and methods used to build text mining systems. The tutorial will cover the state of the art in this rapidly growing area of research, including recent advances in unsupervised methods for extracting facts from text and methods used for web-scale mining. We will also present several real world applications of text mining. Special emphasis will be given to lessons learned from years of experience in developing real world text mining systems, including how to handle informal texts such as blogs and user reviews, infer sentiment towards various objects and how to build scalable systems.

Tutorial 4: Mobility Data: Modeling, Management, and Understanding

Tuesday, October 26, 2010 ***(Room: York)***

Tutorial Organizers:
Stefano Spaccapietra, Esteban Zimányi, Chiara Renso

Time: ***14:00 – 17:30***

Abstract:

Applications using mobility data are blooming in every economic area. Many research efforts have been devoted to developing concepts, models, theories and tools to properly apprehend mobility data and make it manageable for the benefit of these applications. This proposal is authored by experts that have significantly contributed to the field in complementary ways. The tutorial aims at discussing different facets of mobility data, from spatio-temporal data modeling to data aggregation and warehousing and to data analysis, and from basic definitions to state-of-the-art concepts and techniques. One peculiarity of the tutorial is to use a consistent framework that facilitates understanding of all different facets. The learning objective is to enable attendants to understand the many issues at hand and identify the research perspectives they may be willing to investigate to contribute to the domain.