Invited Talk

Text Mining: Opportunities and Challenges*,*

Ronen Feldman , Instinct Software, Israel

The information age has made it easy to receive and store large amounts of data. The proliferation of documents available - on the Web, in corporate intranets, on news wires and elsewhere - is overwhelming. However, while the amount of data available to us is forever increasing, our ability to absorb and process this information has remained constant. Search engines only exacerbate the problem by making more and more documents available in a matter of a few keystrokes; so-called "push" technology makes the problem even worse by constantly reminding us that we are failing to follow critical news, events, and trends. We experience information overload, and miss important patterns even as they unfold before us.

Text Mining is a new and exciting area of research that tackles this problem through techniques borrowed from data mining, machine learning, information retrieval, natural-language understanding, case-based reasoning, statistics, and knowledge management to help people gain rapid insight into large quantities of semi-structured or unstructured text. Typically, it involves preprocessing a document collection (e.g., through text categorization or term extraction), storing , indexing and analyzing the intermediate representations (through distribution analysis, document clustering, trend analysis, and association rule discovery), and presenting the results graphically.

In my talk I will present some of the new challenges facing the text mining community, with particular focus on the representation of documents and the ability to provide better insights into document collections. I will also try to provide a consumer perspective into text mining while reviewing the business opportunities that currently exist in this area.

Invited Talk

ATTICS: A Toolkit for Text Classification and Text Mining

David Lewis, AT&T Research, USA

ATTICS is an extensible text classification system recently implemented in C++ at AT&T Labs and available for research purposes. In spirit it is a hybrid between text retrieval systems such as SMART and machine learning toolkits such as MLC++. The design, data model, and emphasis on online classifier application are unusual for either type of software. While ATTICS is primarily intended for online text classification, its flexible architecture and support for both textual and non-textual data make it applicable to many text mining tasks. I will discuss how ATTICS can be used for text mining, and when it is or isn't a better choice than other tools.