A Survey of Approaches on Mining the Structure from Unstructured Data

Frederik Hogenboom
fhogenboom@ese.eur.nl

Flavius Frasincar
frasincar@ese.eur.nl

Uzay Kaymak
u.kaymak@ieee.org

Econometric Institute
Erasmus University Rotterdam
PO Box 1738, NL-3000 DR
Rotterdam, the Netherlands

Nowadays, it is increasingly important to be able to handle large amounts of data more efficiently, as anyone could generate or collect a lot of information about almost anything at any given time. However, distinguishing between relevant and non-relevant information quickly and responding to newly obtained data of interest adequately, still remains a cumbersome task. Therefore, a lot of research aiming to alleviate and support the increasing need of information has been conducted during the last decades. At first, research was primarily focused on information retrieval, but currently the focus has shifted to information extraction or data mining.

A common problem with the data mining approaches is the handling of unstructured data, often being described using natural (human-understandable) language. One can consider natural languages to be human languages that have evolved naturally in a community, such as for instance English or Dutch. In order to retrieve or process large amounts of data efficiently, it is desired to make use of machines. However, one major problem needs to be overcome in order to make automated structure mining more accurate. As natural languages are by definition not machine-understandable, there is a need for a way to convert natural languages to a more formal representation by using Natural Language Processing (NLP).

Throughout the years, a lot of NLP systems have been created, and new NLP systems still emerge. Although these systems have similarities, they can also differ greatly between one another. Different systems employ various techniques, but are also built for different purposes and may vary in their focus. One could distinguish between three main approaches to assist in an NLP application. First of all, there are statistics-based approaches, which mainly utilize statistics and mathematical models based on probability theory. Then, there are pattern-based approaches, which use linguistic patterns to extract data from texts. Finally, there are hybrid approaches, which combine methods from the former two approaches.

Statistical approaches are commonly used for natural language processing applications. These methods are data-driven and rely solely on (automated) quantitative methods to discover relations, i.e., these approaches use (large) text corpora to develop generalized models that approximate linguistic phenomena. Statistical approaches are not restricted to basic statistical reasoning that is based on probability theory, but encompass all quantitative approaches to automated language processing, such as probabilistic modeling, information theory, and linear algebra.

In contrast to statistics-based approaches, pattern-based approaches are based on linguistic or lexicographic knowledge, as well as existing human knowledge regarding the contents of text that is to be processed. This knowledge is mined from corpora by means of patterns. One could distinguish between several patterns, i.e., lexico-syntactic and lexico-semantic patterns. The former patterns are a combination of lexical representations and syntactical information, whereas the latter patterns are a combination of lexical representations and semantic information.

Although theoretically the distinction between statistical and pattern-based approaches is crisp, in practice it appears to be difficult to stay within the boundaries of a single approach. For instance, applying solely pattern-based algorithms successfully is hard, because these approaches often need for instance bootstrapping or some initial clustering, which can be done by means of statistics. Also, researchers find it difficult to have a purely statistical approach without using some present (lexical) knowledge. Often, an arbitrary approach to NLP can be considered as mainly statistical or pattern-based, but there is also an increasing number of researchers that equally combine data-driven and knowledge-driven approaches.

In his talk, Frederik Hogenboom, PhD student at the Erasmus University Rotterdam, will present the main approaches for text mining and identify the main challenges of this research field.