Text Mining at Detail Level Using Conceptual Graphs

Manuel Montes-y-Gómez, Alexander Gelbukh, and Aurelio López-López

1 Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE), Mexico
{mmontesg,allopez}@inaoep.mx
2 Centro de Investigación en Computación (CIC-IPN), Mexico
gelbukh@cic.ipn.mx

Abstract. Text mining is defined as knowledge discovery in large text collections. It detects interesting patterns such as clusters, associations, deviations, similarities, and differences in sets of texts. Current text mining methods use simplistic representations of text contents, such as keyword vectors, which imply serious limitations on the kind and meaningfulness of possible discoveries. We show how to do some typical mining tasks using conceptual graphs as formal but meaningful representation of texts. Our methods involve qualitative and quantitative comparison of conceptual graphs, conceptual clustering, building a conceptual hierarchy, and application of data mining techniques to this hierarchy in order to detect interesting associations and deviations. Our experiments show that, despite widespread misbelief, detailed meaningful mining with conceptual graphs is computationally affordable.

Keywords: text mining, conceptual graphs, conceptual clustering, association discovery, and deviation detection.

1 Introduction

Text mining is an emerging research area that can be roughly characterized as knowledge discovery in large text collections, thus combining knowledge discovery and text processing methods. It is concerned mainly with the discovery of interesting patterns such as clusters, associations, deviations, similarities, and differences (Feldman, 1999; Mladenić, 2000; Ciravegna et al., 2001).

Current methods of text mining tend to use simplistic shallow representations of text, e.g., keyword sets or keyword frequency vectors. On one hand, such representations are easy to obtain from texts and easy to analyze. On the other hand, however, they usually restrict the knowledge discovery results to be thematic relations between different texts (such as frequent co-occurring topics in the set of texts).
To obtain more useful and meaningful results, richer text representations (i.e. representations that allow not only expressing the topic but also how it is treated) are necessary. In this paper, we describe a method for text mining that uses conceptual graphs (Sowa, 1999) for representing text contents. We show how the conceptual graph representation allows discovering in a set of texts meaningful and detailed patterns, i.e., those distinguishing not only entities (topics) but also actions, attributes and their relations.

Basically, our method considers three traditional data mining descriptive tasks: clustering generation, association discovery, and deviation detection. These classical data mining tasks are well known in the literature (see also Fayyad et al., 1996; Agrawal et al., 1996; Arning et al., 1996; Feldman and Hirsh, 1996; Lent et al., 1997; Mannila, 1997), though the existing methods allow discoveries only at thematic level.

The difference between the thematic (traditional) and detailed (our) levels of analysis is the following. Thematic description treats the text as a set of tokens (keywords or fixed multiword expressions such as differential equations), which are atomic in the sense that they can be either equal or completely different. This limits the type of discoveries to the statistics of co-occurrence and intersections of the sets of such atomic tokens (Feldman et al., 1998). Conceptual graph representation, however, permits us to describe the text as a set of phrases or sentences, having their own internal structure (being not atomic), which can thus partially differ and partially coincide. We will show how such detail level representation allows us to symbolically describe and numerically measure the similarity within a set of (different) expressions, which in turn allows for much richer analysis than simple statistics of keyword repetitions.

The paper is organized as follows. Section 2 describes previous work on text mining and discusses main limitations of current approaches. Section 3 presents our method for doing text mining using conceptual graphs as underlying text representation. Section 4 shows some experimental results that illustrate our method. Finally, section 5 concludes our discussion.

2 Text Mining

The problem of analysis of large amounts of information has been solved to a good degree for information that has a fixed structure such as well-structured databases. Sophisticated methods for uncovering interesting patterns hidden in this kind of large data sets are known under the generic name of data mining (see Han and Kamber, 2001; detailed discussion of data mining is outside of the scope of this paper). However, this problem remains unsolved for weakly structured information such as unrestricted natural language texts.

Text mining has emerged as a new area of knowledge discovery and text processing that attempts to fill the gap in mining methods (Feldman, 1999; Mladenić, 2000; Ciravegna et al., 2001). It can be defined as data mining applied to textual data, i.e., as the discovery of new facts and world knowledge from large collections of texts that (unlike in natural language understanding) do not explicitly contain the knowledge to be discovered (Hearst, 1999). Naturally, the goals of text mining are similar to those of data mining; for instance, it also attempts to discover clusters, trends, associations,