Chapter 81
Revealing Groups of Semantically Close Textual Documents by Clustering: Problems and Possibilities

František Dařena
Mendel University in Brno, Czech Republic

Jan Žižka
Mendel University in Brno, Czech Republic

ABSTRACT

The chapter introduces clustering as a family of algorithms that can be successfully used to organize text documents into groups without prior knowledge of these groups. The chapter also demonstrates using unsupervised clustering to group large amount of unlabeled textual data (customer reviews written informally in five natural languages) so it can be used later for further analysis. The attention is paid to the process of selecting clustering algorithms, their parameters, methods of data preprocessing, and to the methods of evaluating the results by a human expert with an assistance of computers, too. The feasibility has been demonstrated by a number of experiments with external evaluation using known labels and expert validation with an assistance of a computer. It has been found that it is possible to apply the same procedures, including clustering, cluster validation, and detection of topics and significant words for different natural languages with satisfactory results.

INTRODUCTION

People and companies have many opportunities to express their opinions related to a wide variety of topics. The media used for such communication include personal web pages and blogs, social networks, discussion boards, e-mail, instant messages, and others. Various subjects can benefit from a high availability of information, which also demands bigger involvement, knowledge, information processing and decision making skills. Due to huge volumes of data that is often freely available for many different

DOI: 10.4018/978-1-5225-1759-7.ch081

subjects there is a need for approaches that enable to use the data for decision making. Since most of the data is available in an unstructured textual form, disciplines focusing on this type of data have gained on their significance during the last few years (Miner at al., 2012).

Because of inadequate time and effort that would be needed in order to reveal the knowledge hidden in the data, the processing cannot be often done manually by humans. Instead, the application of computer based automated methods is a more desirable choice. This is enabled by the availability of increased computational speed and memory sizes of ordinary computers as well as by the development of new algorithms that are able to address various needs and problems. Instead of a traditional methodology employing human operators for reading the documents, statistical analysis, and data mining techniques based on the non-linguistic structure of the documents (Dini & Mazzini, 2010), intelligent computer-based analysis called text mining might arrive at new and unforeseen results.

Text mining is a branch of computer science that uses techniques from data mining, information retrieval, machine learning, statistics, natural language processing, and knowledge management (Berry & Kogan, 2010). The greatest potential of text mining applications is in the areas where large quantities of textual data are generated and collected. These areas include, besides others, categorization of newspaper articles or web pages, e-mail filtering, organization of a library, customer complaints (or feedback) handling, marketing focus group programs, competitive intelligence, market prediction, extraction of topic trends in text streams, discovering semantic relations between events, or customer satisfaction analysis (Cao et al., 2014; Koteswara Rao & Dey, 2011; Miner at al., 2012; Nassirtoussi, 2014; Weiss et al., 2010). Text mining involves tasks such as text categorization, term extraction, single- or multi-document document summarization, clustering, association rules mining, or sentiment analysis (Feldman & Sanger, 2007).

At the end of the last century, machine learning gained on its popularity and became a dominant approach to text mining (Sebastiani, 2002). *Machine learning* is a discipline that focuses on modification or adaptation of computer behavior based on the past experience (the data in this case) so the behavior gets better in the future. Such an adaptation depends on whether there is the right behavior specified. If there is, it means that there is a set of examples with correct answers (actions) provided. In this case we talk about *supervised learning*. During the learning process a computer tries to generalize the knowledge to be able to react correctly to all, even previously unseen inputs. When the correct responses are not provided, a computer tries to find some patterns based on similarities between the inputs. This approach is known as *unsupervised learning* (Marsland, 2009). The common goal of both approaches is to achieve accuracy comparable to that achieved by human experts.

Supervised learning which is the most common type of learning problem (Dittrich, 1995) is also popular in text mining (Sebastiani, 2002). Therefore many text mining tasks require that the data items to be processed have assigned labels that categorize the data. Then the classifier, which is a function that generalizes the knowledge about how to assign correct labels to the data, uses the labels in the process of learning. In some cases, especially when the learned model suffers from high variance (it is too much fit to the exemplar data), having more labeled data is a possible direction of further improvements (Cawley & Talbot, 2010).

Unfortunately, unavailability of the labeled data is often a major problem. As new data constantly occurs, it is nearly impossible to have the labels assigned to the data in a reasonable time and in reasonable amounts. The labeling process itself is also very demanding. Reading just tens or hundreds of documents and assigning the labels correctly requires effort of many people for many hours. Even when the people (annotators) are experts in the given field, resulting quality of the labeled data collection is not obliged