Abstract

Vast amount of data in text form contributed to the development of tools that, among others, automatically group texts thematically. As a result, finding the information contained in them becomes easier. To categorize articles by subject, one should pay attention to: a way in which the data is represented and an algorithm used to detect clusters. Data representation should keep information about the topic. Thus, usually, the following assumptions are made: firstly, order of words does not matter, because only their distribution is relevant, and secondly, inflection of words is not important, but their meaning is. One can therefore characterize each text as a vector of frequencies of words occurring in it. However, such an approach is highly inefficient due to the high dimensionality of data (equal to the total number of unique words found in all the texts) and their sparsity. A common practice is to group words and then represent texts using vectors of frequencies of groups of words. Therefore, the component of processing text data that is often used, is to reduce words to their root. Such an approach, however, does not include spelling errors, typos or deficiencies of diatric marks, which are often found in text data. Then, one can determine the measure of distance on strings (strings of any length over a finite set called the alphabet), assigning words to predetermined groups (defined by basic forms). The aim of this study is to investigate the influence of the selection of these distances on the quality of automatic thematic categorization of texts based on articles from Polish Wikipedia. In the first step, words are grouped using the selected distances. Furthermore, articles are represented as frequencies of particular groups of words that appeared in the text. On the basis of the obtained data, we cluster the articles. Evaluation of the results is carried out on the basis of the groups with real categories assigned to each of the texts of Wikipedia. Note that the use of such a data set is associated with several major challenges. Firstly, Polish language is grammatically very complex and contains a large number of words. Next, set of texts from Polish Wikipedia is relatively large, which raises the need for adequate data management and process optimization due to limited computing and storage resources. Lastly, some data analysis methods cannot be effectively used on large data sets.

Keywords: metrics over space of strings, strings, thematic categorization, clustering