Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 123 (2018) 271–277

Procedia Computer Science

www.elsevier.com/locate/procedia

8th Annual International Conference on Biologically Inspired Cognitive Architectures, BICA 2017

Text clustering as graph community detection

Elizaveta K. Mikhina, Vsevolod I. Trifalenkov

National Research Nuclear University "MEPhI" (Moscow Engineering Physics Institute)
Kashirskoe highway 31, 115409, Moscow, Russian Federation
eli-mikhina@yandex.ru

Abstract
This article suggests a method of text clustering that does not depend on any user-set parameters. Text documents and connections between them are represented as graph nodes and edges and graph community detection method is thus applied to the text clustering problem. The method was tested against news articles collections and proved effective – manual and automatic clustering of text documents in collections were same or really close.

Keywords: text clustering, non-parameter clustering, graph community detection, modularity

1 Introduction

At present, processing and analyzing of information often becomes impossible without the use of computer facilities. However, despite the rapidly growing power of computing resources, many tasks are quite difficult to be fully computerized.

There are areas of activity closely related to the analysis of text documents. An example of such documents may news articles that are closely related, but not identical to one another. To computerize the task of analyzing such data, it is necessary to solve the clustering problem, meaning to group data units according to the measure of their similarity.

The existing text clustering algorithms have a significant drawback, namely, the result of clustering depends strongly on a set of parameters. Moreover, in most cases, user can not know in advance how these parameters will affect the final result of the algorithm application. This reduces the computerization potential of such solutions, and also requires that user thoroughly understand both the problem being solved and the mathematical features of the algorithm.

10.1016/j.procs.2018.01.042

2 Text Clustering

Text clustering involves dividing the source collection of texts (documents) into disjoint subsets (clusters) so that each cluster should consist of similar objects, and the objects of different clusters differed significantly among themselves (Manning, 1999).

The initial data for text clustering consists of unstructured texts; texts used as sample collections in experiments described in this article are news articles in Russian.

The task of cluster analysis can be formulated as follows: based on data contained in the matrix of features, split a set of objects (texts) into a number of clusters (subsets) so that each object belongs to one and only one subset. (Nizametdinov, 2012) In this case, close objects belong to the same cluster, and distinct objects belong to different ones.

The concept of proximity and distinction is formally stated by setting a distance function or similarity function.

Based on the matrix of features using the similarity function, a matrix of text mutual similarity is calculated, in which the search is made for maximum values that do not lie on the diagonal of the matrix. The row number and column number of the maximum element correspond to the numbers of the closest documents.

3 Proposed clustering method

The main problem of most clustering algorithms lies in the fact that their result depends heavily on some set of parameters (Kiselev, 2005), and, in most cases, the user can not know in advance how these parameters will affect the final result of the clustering. The key feature of the method proposed by the authors of this article is that it does not depend on any user-set parameters. The decision to assign documents to one cluster is made on the basis of the mutual similarity of the documents with respect to other documents in the collection.

In addition, the proposed method differently estimates the minimum value of the similarity function for different clusters. This allows to select both dense clusters with high minimum value of the similarity function between documents within cluster and sparse clusters with comparatively low value.

The text of the document is considered as a set of words without considering their relative positions. The high-dimensional feature space is the entire set of words used in the document collection, and the non-zero values of the document vector correspond to the words and expressions that occur in it.

To improve the quality of the feature space, the preliminary filtering of the features is performed:

- Stop-words exclusion (built-in stop-words list MSSQL (Microsoft Developer Netbook, 2015));
- Minimal morphological processing (reduction of words to the initial form);
- Evaluation of the importance of the word in the document (based on TF-IDF (Ramos,2003));
- Evaluation of significant bigrams (based on approach described in (Mikhina, 2016)).

As the similarity function authors used the cosine of the angle between a pair of vectors in the multidimensional space corresponding to a pair of documents in the collection.

$$sim^{ij} = \cos(D^i, D^j) = \frac{(D^i \mathrm{x} D^j)}{|D^i| * |D^j|}, \tag{1}$$

where $(D^i \mathrm{x}\, D^j)$ is the scalar product of vectors D^i and D^j, while $|D^i| * |D^j|$ is the product of the lengths of the vectors D^i and D^j.

In the proposed method, text clustering is considered as a particular case of graph community detection (in Russian-language literature the term graph approximation is often attributed to the problem (Il'ev, 2016)). For further clustering, the document collection is represented as a weighted graph

$$G \coloneqq (V, E), \tag{2}$$