ISBN 978-981-11-0008-6
Proceedings of 2016 6th International Workshop on Computer Science and Engineering (WCSE 2016)
Tokyo, Japan, 17-19 June, 2016, pp. 706-709
doi: 10.18178/wcse.2016.06.125

Analysis of Automatic Clustering of Textual Information Using by Mountain Clustering Method

Kyaw Zaw Ye, Fedorov A. R., Shiryaev A. P. [+], Gagarina L. G. and Yanakova E. S.

Russian Federation, National Research University of Electronic Technology

Abstract. Methods of searching and categorizing information become very important in view of the growing volumes of unstructured information in the Internet. One of such methods is cluster analysis, which is presented by a variety of algorithms. This article is devoted to the research and development of heuristic methodic to perform automatic clustering of text information effectively. The proposed solution uses a subtractive clustering algorithm to determine the number of clusters and performs splitting of articles into clusters using an algorithm of the k-means family. A computational experiment by a collection of articles from the web resource Wikipedia.ru demonstrates that this method using the algorithms k-means and k-medoids is suitable for automatic clustering of textual information, but it requires further improvements.

Keywords: clustering, cluster analysis, k-means, k-medoids, mountain clustering.

1. Introduction

Today, the Internet is practically unlimited. In order to find the useful information, it is necessary to filter out from "information garbage", and from knowledge, only indirectly related to the subject of the information request. Algorithmic and social filters are set between a person and valuable information for him, as well as a combination of the first and the second one. Algorithmic filters are the technology of computer data processing. Filters of the social nature are the Internet community. Filters do not remove “unnecessary”, and only move “necessary” to the fore. The information that has not passed through these filters does not disappear – it remains in the open access. Modern filters can prompt the user where to find information, but they can hide it, considering it as uninteresting or unnecessary (filter bubble). Therefore, so much frankly incorrect or simply ridiculous information come in response to requests for information on any topic.

How to allocate from all the variety of information that one that is necessary for us? One of the possible solutions of this problem is to use cluster analysis to categorize and search information. Automatic clustering of the data has no unique solution. There are many different algorithms, but there is no best clustering quality criterion unambiguously.

Among non-hierarchical clustering algorithms there are popular algorithms of k-means family [1]. The algorithms of this family are fairly simple and effectively used in processing of large volumes of data. They are used for preliminary splitting into groups of a large set of data after which specifying clustering can be carried out. But for algorithms of this family it is required to set a quantity of clusters that should be produced by the split. Development of integrated algorithm will allow to get rid of this problem.

The purpose of the research is improving the efficiency of algorithms of automatic clustering of textual information due to use integrated clustering algorithm in terms of large amounts of data.

2. Theory & Methodology

[+] Corresponding author. Tel.: +8-916-907-82-74
E-mail address: alex-sh2@yandex.ru

Let's consider the algorithms k-means and k-medoids. K-means minimizes the total square deviation of points of clusters from the centres of these clusters. K-medoids is similar to k-means algorithm, but it differs in that k-medoids seeks the centres of clusters at each iteration as medoids of points, not as a mean of points. That is, the cluster center must be one of its points.

The number of clusters is one of the input parameters of these algorithms. The choice of optimum value of this parameter can be a very difficult task often solved only experimentally. The union of k-medoids or k-means with algorithm of mountain clustering allows solving the problem of setting the number of clusters for these algorithms.

The following tasks have been solved in this research:

1. The scheme of the analysis process of articles is constructed and analysed.
2. The comparative analysis of the algorithm efficiency using clustering of k-means [1,2] and k-medoids [5] on the example of a Russian-language collection of articles received from the Wikipedia.ru web resource is carried out.
3. The integrated algorithm of clustering uniting mountain clustering [1,2,6] and one of algorithms of splitting is developed: k-means or k-medoids.
4. Advantages of the integrated algorithm of clustering are proved and systematized [2,3,4].
5. A significant selection of content and its "cleaning" using Stemming (clipping from the word endings and suffixes) and lemmatization are carried out (bringing the word to its normal dictionary form).
6. By means of the measure of Okapi Bm-25 [7] distribution of scales is constructed, and keywords for each article are allocated.

Quality criteria of clustering represent degree of compliance of obtained splitting to the ideal decision. In this research quality of clustering was estimated by the following formal criteria: Ratkowsky index [8], Maulik-Bandoypadhyay index [9,10], Score function [9,10].

3. Experiments

The collection of articles (1 264 articles) from the Wikipedia.ru web resource was used as the initial data for the efficiency analysis of clustering algorithms. According to the obtained weights there were allocated about 10 keywords for each article. In operation with fewer keywords there was deterioration of the quality decomposition.

There were obtained 426 clusters as a result of applying the mountain clustering algorithm for the collection of 1264 articles. The principal component analysis (PCA) was used to reduce the dimensionality of data [11] in rendering. The proximity of two articles suggests a positive correlation, and diametrically opposed to the location - the negative.

To compare performance of two algorithms the number of clusters obtained by using mountain clustering algorithm is used as the initial data for the algorithm k-means. Thus, the collection of articles is divided into 426 clusters. Initial cluster centres are set randomly. The clusters of one element were not being observed, but they are presented in the partition. We carry out an expert evaluation of conformity of the articles to obtained groups. The cluster example is presented in Table 1.

Table 1: Cluster obtained by mountain clustering algorithm and k-means

Cluster members	Keywords
Greco-Persian Wars	pers, themistocles, xerxes, greco-persian, athenian, greek, persian, herodotus, cimon, spartan
Ancient Egyptian religion	osiris, set, egypt, god, isis, mythology, egyptian, sarcophagus, cult, deity
Charles IV, Holy Roman Emperor	wenceslaus, luxembourg, charles, czech, margrave, prague, holy, brandenburg, moravia, emperor

All cluster members are historical subjects; we can conclude that this distribution is adequate. But there is a total collection of historical articles and other topics which were not included in the cluster.

After analysing of all clusters, we conclude that we have a fairly large partition error. Statistical measure of the weights of keywords and k-means algorithm are contributed to this error. K-means is sensitive to noise and determines only spherical clusters.

The number of clusters obtained by mountain clustering algorithm was used for the k-medoids algorithm. Clustering with k-medoids has given the partition that in many cases differs from obtained with k-means partition. The initial values of centers were also selected randomly. The cluster example is presented in Table 2. We carry out an assessment of the partition results with the k-medoids and compare the results with the k-means.

Table 2: Cluster obtained by mountain clustering algorithm and k-medoids

Cluster members	Keywords
Russian Empire	title, empire, rank, imperial, reform, princely, duma, estate, table, family
Russia	federation, million, russia, federal, crimea, population, billion, rsfsr, thousand, rus
Russians	slavonic, people, nation, costume, ethnic, marker, anthropological, old russian, ethnonym, slavic

This cluster is identical with that obtained using the k-means algorithm. Graph of this cluster with the "Russians" centroid is highlighted in Fig. 1.

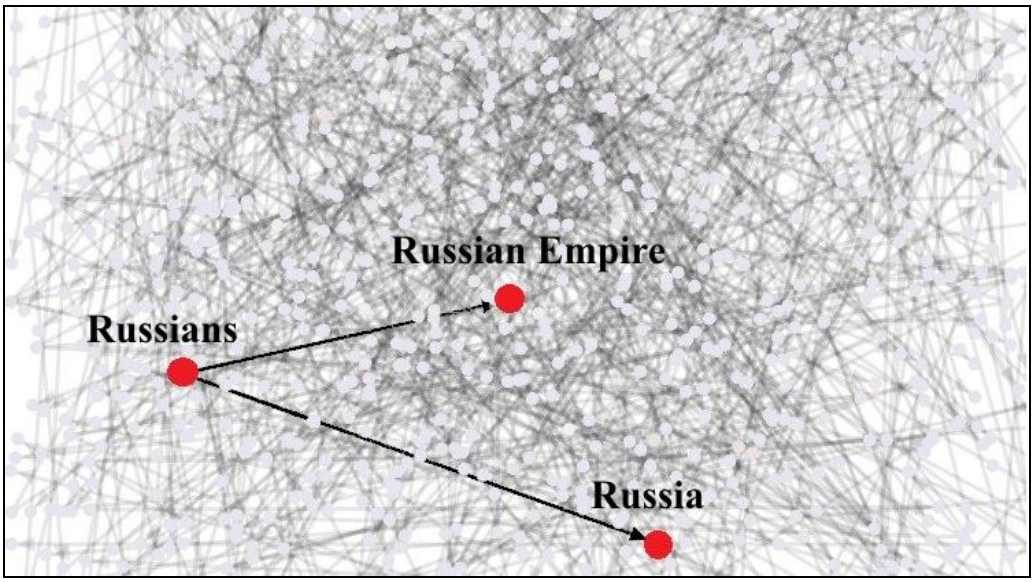

Fig. 1: Cluster "Russian" obtained by mountain clustering algorithm and k-medoids on the resulting graph of articles.

In addition to the random assignment of initial centroids we have considered the use of obtained centroids by the mountain clustering algorithm as the initial points for the k-means and k-medoids algorithms. Splitting articles into clusters for k-medoids and k-means has worsened from an expert point of view, but indicators for Maulik-Bandoypadhyay index and Score function criteria have notable improved. Ratkowsky index remained practically unchanged. Thus, the random assignment of initial centers is preferred.

According to the quality of the criteria k-means has the best indicators, but according to expert estimates k-medoids algorithm shows the best results.

4. Conclusion

In this paper, we have carried out research and developed a heuristic method that allows performing automatic clustering of text information. The proposed solution uses the mountain clustering algorithm to determine the number of clusters and performs splitting of articles into clusters using the algorithm of k-means.

A comparative analysis of k-means and k-medoids algorithms was carried out for using in the method. According to the quality criteria, k-means clustering has the best structure of clusters than k-medoids, but from an expert point of view in k-means a greater number of partitioning errors was identified.

The inaccuracy of this technique is due to the following factors:

1. For stemming we used the Porter stemming algorithm [12] that has noticeable inaccuracy in monosyllabic and two-syllable words. Monosyllabic words were included in the research.
2. The quality of lemmatization depends on the dictionary. We used modification of the dictionary of Zaliznyak for the lemmatization.
3. K-means and k-medoids are algorithms of clear clustering and they are sensitive to noise and outliers.

The computational experiment has shown that in general this methodology using k-means and k-medoids algorithms is suitable for automatic clustering of textual information, but it requires further improvements.

5. References

[1] Khaled Hammouda, Fakhreddine Karray, A Comparative Study of Data Clustering Techniques, University of Waterloo, Ontario, Canada, Volume 13, Issues 2-3, November 1997, pp. 149-159.

[2] Marta Marrón Romera, Miguel Angel Sotelo Vázquez, and Juan Carlos García. Comparing Improved Versions of 'K-Means' and 'Subtractive' Clustering in a Tracking Application. *Proc. of the 11th international conference on Computer aided systems theory 2007*. 2007, pp. 717–724.

[3] K Yang Qing, Liu Ye, Zhang Dongxu and Liu Chang. Improved k-means algorithm to quickly locate optimum initial clustering number. *Proc. 30th Chinese Control Conference*. China, Yantai: IEEE. 2011, pp. 3319 – 3322.

[4] Younjeong Lee, Ki Yong Lee, Jaeyeol Rheem. Speaker identification based on subtractive clustering algorithm with estimation number of clusters. *Proc. 8th International Conference*. Karlovy Vary, Czech Republic. 2005, pp 249-256.

[5] Xin Jin, Jiawei Han. Encyclopedia of Machine Learning. Springer US. 2010, pp. 564-565.

[6] Priyono, A., Ridwan, M., Alias, A.J., Rahmat, R.O.K., Hassan, A., Ali. Generation of fuzzy rules with subtractive clustering. Journal Teknologi, Volume 43. 2007, pp.143–153.

[7] Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze. An Introduction to Information Retrieval. Cambridge University Press. 2009, p. 233.

[8] Bernard Desgraupes. *Clustering Indices*. University Paris Ouest Lab Modal'X . 2013.

[9] Sandro Saitta, Benny Raphael, and Ian F.C. Smith. A Bounded Index for Cluster Validity. *Proc. 5th International Conference. Leipzig*, Germany. 2007, pp. 174-187.

[10] Sivogolovko E.V. Methods of evaluating crisp clustering. *Proc. Computer tools in education*. 2011, **4**(96): p. 14.

[11] Pomerantsev A. Principal component analysis (PCA), 2008 [Online]. Available: http://rcs.chemometrics.ru/Tutorials/pca.htm. [Accessed: 20-Jan-2016].

[12] Karen Sparck Jones and Peter Willet. *Readings in Information Retrieval*. Morgan Kaufmann Publishers Inc,.1997.