International Journal of Multimedia and Ubiquitous Engineering
Vol.12, No.3 (2017), pp.57-64
http://dx.doi.org/10.14257/ijmue.2017.12.3.06

Relevance Mining of Multi-Keyword Search to Originate Initial Cluster Centers for K-Means Algorithm

Zhaoyang Qu[1], Wei Ding[1], Tiehua Zhou[1*], Nan Qu[2], Ling Wang[1] and Hongbo Lv[3]

[1]Department of Computer Science and Technology, School of Information Engineering, Northeast Dianli University, Jilin, China
qzywww@nedu.edu.cn,wxj3429@126.com,
thzhou55@163.com,smile2867ling@163.com
[2]Jiangsu Provincial Company Branch Overhaul, Jiangsu, China
649993290@qq.com
[3]State Grid Jilin Electric Power Co., Ltd, Changchun, China
zhth0213@163.com

Abstract

Multi-keyword search take the words as the input query to find the information resources where these keywords occur and look for ways to connect these words using information on referential integrity constraints. Text clustering is an effective way to analyze textual document in retrieval system. K-means algorithm is most popular and simple tool widely used in data clustering analysis. In this paper, a novel method is proposed to address the initial cluster centers problem in k-means algorithm based on relevance mining of multi-keyword search. In proposed work, the initial cluster centers have obtained using frequent pattern for each keyword and after that k-means algorithm is applied to gain optimal cluster centers in database. The performance of the proposed algorithm is tested on the 20-Newsgroups collection. The experiment results show the proposed method is better way to represent the clusters.

***Keywords**: Multi-keyword search, k-means clustering, frequent pattern*

1. Introduction

The World Wide Web is a hypertext collection of electronic documents, which are becoming a more and more important and hot research area in recent decades. Text retrieval systems process query based on keyword over relational databases, aim at allowing users to efficiently retrieve relevant texts by matching query keywords and text description over large textual collections. Retrieval of the relevant natural language textual document is more challenging tasks in order to discover similar texts together, which enable to quickly query large textual collections, and make it possible to easily present to the relevant documents in them. Unfortunately, there are certain difficulties to gather those data in a right way.

Data clustering is an effective way for data analysis that is associated to data mining which analyzes query specific textual document in retrieval system [1-2]. Data clustering has been considered for use in a number of different areas of text mining and information retrieval [3], such as organizing large document collection, finding similar documents, recommendation system, duplicate content detection, search optimization. Clustering is a technique that is used to distinctly differentiate data points such that similar points are assigned to same cluster while points with different properties are assigned to different clusters. If clusters are analyzed properly, they can give meaningful information to the

[*] Corresponding author

Copyright ⓒ 2017 SERSC

users. Grouping texts into meaningful categories is a challenging problem, which can reveal relevance between keyword and document. This problem faced in clustering is the identification of clusters in given data.

A commonly and widely used approach for clustering is based on k-means in which the data is partitioned into k number of clusters. In this method, clusters are predefined which is highly dependent on the initial identification of elements representing the clusters well. The goal is to use k-means clustering algorithm for query dependent clusters in textual document, and help end users to identify query from textual documents in which they are interested. For each point, k-means work like this; if the number of point is less than the number of cluster then we assign each point as the centroid of the cluster. Each centroid will have a cluster number. If the number of point is bigger than the number of cluster, for each point, we calculate the distance to all centroid and get the minimum distance. This point is said belong to the cluster that has minimum distance from this point. Since we are not sure about the location of the centroid, we need to adjust the centroid location based on the current updated point. Then we assign all the points to this new centroid. This process is repeated until no point is moving to another cluster anymore.

Several works in clustering have focused on improving the clustering process such that the clusters are not dependent on the initial identification of cluster representation. Generally, relevance mining is the process of interacting with keywords and performs clustering of meaningful documents based on descriptions that are inherent in the documents themselves. In this paper, a modified k-means algorithm for text retrieval is developed where frequent pattern is used to generate the initial number of clusters and the cluster centers, because the high quality result of k-means algorithm is dependent on the initial cluster centers. The good cluster is directly proportional to its centers. In our research, clusters will be generated according to the multi-keyword in the user's query and related texts will be found. We apply k-means clustering algorithm to group cohesive words for each keyword according to frequent pattern due to the user play in active role in the retrieval process and the system should employ any information that can be provided by the interacting user. This requires an intimate interplay between the search process and the user.

The rest of this paper is organized as follows. In the next section, we provide some technical background and discuss related work. Section 3 describes the proposed method in more detail. Experimental result is evaluated in Section 4 and conclusion of the paper is summarized in Section 5.

2. Related Work

Data clustering is a powerful tool for data analysis which can be used to discover the similarity or dissimilarity between groups of data points such that points in one group are more similar than other groups. Hence, large numbers of algorithms have been widely studied by broad literature for data clustering task, such as informatics, multimedia, marketing, meteorology, geology, medical, *etc.* [4-7]. In the clustering domain, k-means is most popular and simple tool [8-9], due to it is fast and sensitive, and easy to implement.

There are also many other clustering objectives related to k-means which are commonly studied. In real life, these algorithms are also some problems, for example, k value cannot be determined, the initial cluster centers are randomly selected and the presence of noise points, lack of big data processing capabilities, *etc.* And it is quite difficult to choose the number of clusters present in final result. In order to improve the k-means algorithm, the researchers propose a variety of solutions. [10] proposed k-means++, a random sampling based approximation algorithm for Euclidean k-means which achieves a factor of O(logk). [11] showed the k-means problem is NP-hard for points in the plane.

Copyright ⓒ 2017 SERSC

[12] proposed by local density of data points to estimate the initial cluster center. These approaches improve the performance of the k-means algorithm to a certain extent.

Currently, research about improved k-means algorithm is mainly concentrated in the number of clusters determined, selecting of the cluster center and improving of the clustering and other criteria. [13] proposed a medoid-based k-partitions approach called Clustering Around Weighted Prototypes (CAWP), which works with a similarity matrix. In CAWP, each cluster is characterized by multiple objects with different representative weights. [14] proposed ranked-k-means, which is essentially less sensitive to the initialization of the centers. They presented the ranked set sampling design and explain how to reformulate the k-means technique under the ranked sample to estimate the expected centers as well as the clustering of the observed data. [15] used multiple utility functions to build a theoretic framework for k-means based consensus clustering. They transformed consensus clustering to k-means based clustering. The authors have also handled incomplete basic partitions. Their suggested algorithm performs very well on multiple datasets and was comparable to the state of art algorithms in this field.

The most relevant to our discussion are the k-median and the k-center objectives. The k-median problem is typically studied in the setting where the centers are one of the data points. The k-median problem has been a testbed of developing new techniques in approximation algorithms, and has constantly seen improvements even until very recently [16-18]. In the k-center problem the objective is to pick k center points such that the maximum distance of any data point to the closest center point is minimized. [19] focused on comparing MWK-means with other existed methods to refine and initialize centers of the clusters in k-means.

Frequent pattern is used to represent the relevance mining between keywords if they either correspond to similar document or if they are semantically related. Here two texts are considered to be related if they share common keywords, and they will lie in the same group. Hence the documents, which are highly associated with the same keywords, are clustered together [20].

3. Proposed Method

In this section an initialization method is proposed for k-means algorithm based on relevance mining of frequent pattern. The proposed method is used to obtain most favorable cluster points for generating the initial cluster centers rather than random or user specified cluster centers.

3.1. Preliminaries

As described in [21], the basic idea of k-means algorithm is as follows: First, randomly selected k items, each item as a cluster center. For each remaining item according to its distance from the center of each cluster, assign it to the nearest cluster. And then cluster all items into k groups. Next, calculate the average of each cluster and set them as the new clustering center. This process will to be done recursively until the objective function converges. Here the objective function is the sum of the distance of each item to its center. Specifically, the process of k-means algorithm shows as follows:

Step 1: Randomly selected k items from database, each item as an initial clustering center $C_1, C_2, \ldots, C_k$.

Step 2: The rest items are assigned to the nearest cluster, if $d_{ij}(p_i, C_j) < d_{im}(p_i, C_m)$, then $pi \in Cj$.

Step 3: The average of all the items in each cluster is calculated as new center of this cluster, *i.e.*, $C_i' = \frac{1}{n}\sum_{p \in C_i} p$.

Step 4: Repeat steps 2 and 3 until the objective function converges.

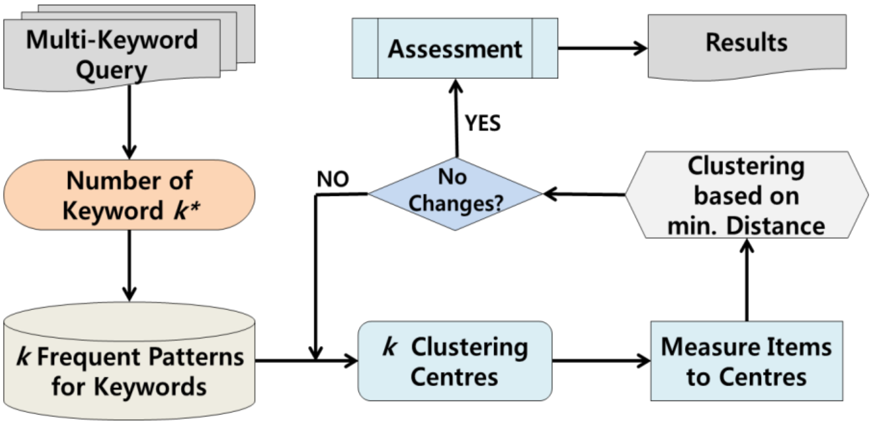


Figure 1. Flow Chart of Proposed Method

3.2. Scenario

In text retrieval system, a document needs to be compared to every other one in the database to determine the closest match for retrieval or other operations. Text clustering is defined as an organization method where the interested documents, which have some similarity along keywords, are kept close or the documents that differ from each other, are kept further apart. In this paper, the clustering problem can be defined by number of keywords and number of partitions in the database. The procedure of proposed method is illustrated in Figure 1. For a given multi-keyword query, the mean intensities of k-means method are calculated iteratively for each frequent pattern and then the document is partitioned by relevance words into the cluster according to the closest words.

3.3. Frequent Pattern for K-Means Clustering

K-means is based on the idea that a center point can represent a cluster. Therefore, the goal of our proposed method is to partition the meaningful and relevant words into k groups in which frequent pattern is used for given keywords, which are frequently encountered in documents. The cluster of words is necessary for two reasons: (1) for faster retrieval, and (2) for determining the value of document uniqueness. The value of document uniqueness depends on the quantity of words in a cluster because the larger the quantity in a cluster corresponds to a smaller value of object uniqueness. Hence, document uniqueness is related to the result of text retrieval.

Consider an example of documents database D, which is including a large number of words $D=\{p_1, p_2, \ldots \}$. For a given multi-keyword query $Q=\{q_1, q_2, \ldots, q_{k^*}\}$, the working of proposed method is given below:

First, for any two keywords q_i and $q_{i'}$, mining the frequent pattern based on a given threshold value, *i.e.* $d(q_i + q_{i'}), p_j)\leq\mu$, to generate the range of the initial centers. The number of cluster $k=[k^*(k^*-1)]/2$.

Then, obtain the initial cluster centers using the following function:

$$C_i = \frac{1}{n}((q_i + q_{i'}) + \sum_{j=1}^{m} p_j) \tag{1}$$

where n=m+2.

Next, calculate the distance d as measure of each point to assign cluster center Ci using following function:

$$d_{\mathrm{min.}} = \left(\left\|C_i - p_j\right\|^2\right)^{\frac{1}{2}} \tag{2}$$

If the distance between the point and cluster is less than the minimum distance, assign this point to its nearest cluster. As a result of this assignment, the cluster representation, or center, may change. The centroid is recalculated as an average of properties of all points in the cluster. In addition, the distance of the affected cluster from every other cluster, as well as the minimum distance between any two clusters, and the two clusters that are closest to each other is recalculated.

Finally, the iterative computation of centers for each cluster is updated until no point move from one cluster to another, and then, return the final cluster centers. The algorithm of proposed method is summarized in Figure 2.

To get the desirable results, the proposed method is performed to cluster similar words in the same group. The uniqueness of document is obtained through the frequent words that belong to the cluster compared to the total number of documents in the database.

```
Input: Multi-keywords query $Q=(q_1, q_2, ..., q_{k^*})$,
        words in database $D=(p_1, p_2, ... )$
Output: $k$ clusters
For each two keywords $q_i$ and $q_{i'}$
    mine frequent pattern according to
    $d(q_i, q_i) \leq \mu$ ;
    For k frequent patterns, k=[k*(k*-1)]/2
        Calculate distance d in every frequent pattern to assign cluster center
        $C_i = ((q_i + q_{i'}) + (p_1 + \cdots + p_m))/n$
        For each point $p_j$,
            If $d(C_i, p_j) \leq d$
                assign $p_j$ to cluster $C_i$
            Endif
     Repeat
Endfor
Return k clusters
```

Figure 2. Algorithm of Frequent Pattern for K-Means Clustering

4. Experiments

Table 1. The Performance of Accuracy

No. of keywords	Methods	Accuracy
5	k-means MWK-means proposed method	0.2781 0.3627 0.402
10	k-means MWK-means proposed method	0.3827 0.4911 0.557
20	k-means MWK-means proposed method	0.5763 0.6852 0.7388

The experiments of proposed method are evaluated on 20-Newsgroups [22]. This dataset is a collection of 20,000 newsgroup documents, partitioned evenly across 20 different newsgroups. The documents in each category of our experimental datasets are chosen at random from the corresponding group. All the performance compared with standard k-means algorithm and MWK-means algorithm.

Table 1 presents the results in terms of the accuracy. For the performance of accuracy, which is the ratio of true results T_r among the total number of database defined as follows:

$$Accuracy = \sum_{i=1}^{k} \frac{T_r}{n} \tag{3}$$

Based on commonly used performance measures in text retrieval, recall is the fraction of relevant documents that are successfully retrieved. Figure 3 presents the experiment results of recall based on different number of clusters. There is a slow growth from the number of cluster 15 to 21 due to most of relevance documents have been retrieved when k=15.

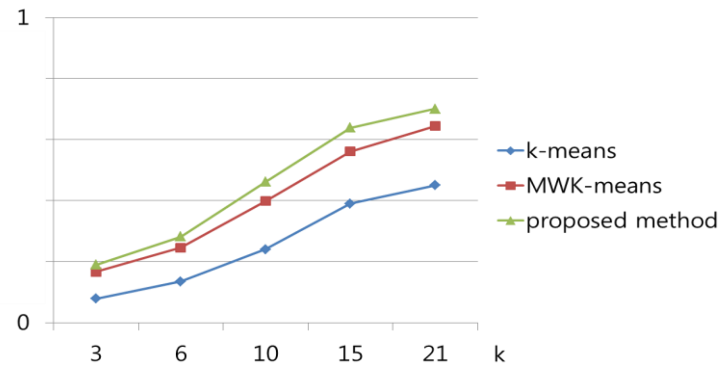


Figure 3. The Performance of Recall

5. Conclusion

In text retrieval system, the content of a document can be expressed in terms of different words. K-means clustering algorithm is applied to group the textual database into various clusters. In this paper, we focus on the initialization problem, which is formulated by: how to initialize initial cluster centers for k-means algorithm? The proposed method initializes multi-keyword for k-means algorithm based on relevance mining of frequent pattern. The experiment results show the proposed method is very helpful for finding the specific text in search query. In the future, we plan to extend our research to create the system for different applications and evaluate the effectiveness of the proposed method.

Acknowledgments

This work was supported by the National Science Foundation project of China (No. 51277023), by the Jilin province science and technology development plan project (21060623004TC), by the Project-sponsored by SRF for ROCS, SEM, and by the Education Department Foundation of Jilin Province (No.201698).

Copyright ⓒ 2017 SERSC

References

[1] R. Varadarajan, V. Hristidis, “A System for Query-Specific Document Summarization”, Proceedings of the 15th ACM International conference on Information and knowledge management, Arlington, USA, (**2006**) November 05-11.

[2] C. R. Chowdary, P. S. Kumar, “An Incremental Summary Generation System”, Proceedings of the 14th International Conference on Management of Data, Mumbai, India, (**2008**) December 17-19.

[3] S. Muppidi, M.R.Murty, “Document Clustering with MapReduce using Hadoop Framework”, Journal of Recent and Innovation Trends in Computing and Communication, vol. 3, Issue 1, (**2015**), pp. 409-413.

[4] G.Gan, C.Ma and J.Wu, “Data Clustering: Theory, Algorithms, and Applications”. SIAM Publishers, NZ, (**2007**).

[5] Z.Y.Qu, S.L.Hou, P.Y.Zhang, “The Realization of the substation 3D visualization Training Platform”. Journal of Northeast Dianli University,Vol. 1, (**2014**), pp. 75-79.

[6] Z.Y.Qu, X.D.Fan, H.T.Yu, N.Qu, “Smart Grid Text Knowledge Acquisition Model Based on Ontology”. Journal of Northeast Dianli University, Vol 34, Issue 05, (**2014**), 60-68.

[7] X.L.Guo, T.T.Yang, Y.C.Zhang, “Gesture Recognition Based on Kinect Depth Information”. Journal of Northeast Dianli University, Vol 36, Issue 02, (**2014**), pp. 90-94.

[8] J. Macqueen, “Some Methods of Classification and Analysis of Multivariate Observations”. Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability, California, USA, (**1967**) December 27 – January 7.

[9] S.N.Sulaiman, N.A.M.Isa, “Adaptive Fuzzy-K-means Clustering Algorithm for Image Segmentation”. Journal of IEEE Transactions on Consumer Electronics, Vol. 56, Issue 4, (**2010**), pp. 2661-2668.

[10] D.Arthur, S.Vassilvitskii, “K-means++: The Advantages of Careful Seeding”. Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms, New Orleans, Louisiana, USA, (**2007**) January 7-9.

[11] M.Mahajan, P.Nimbhorkar, and K.Varadarajan, “The Planar K-Means Problem is NP-Hard”. Journal of Theoretical Computer Science, Vol. 442, (**2012**), pp. 13-21.

[12] L.Kaufman, P.J. Rousseeuw, “Finding Groups in Data: An Introduction to Cluster Analysis”. Wiley Series in Probability and Statistics, Wiley Publishers, USA, (**2008**).

[13] J.P.Mei, L.Chen, “Proximity-based K-Partitions Clustering with Ranking for Document Categorization and Analysis”. Journal of Expert Systems with Applications, Vol. 41, Issue 16, (**2014**), pp. 7095-7105.

[14] M. W.Ayech, D. Ziou, “Segmentation of Terahertz Imaging using K-Means Clustering based on Ranked Set Sampling”. Journal of Expert Systems with Applications, Vol. 42, Issue 6,(**2015**), pp. 2959-2974.

[15] J.Wu, H.Liu, H.Xiong, J.Cao, and J.Chen, “K-Means-based Consensus Clustering: A Unified View”. Journal of IEEE Transactions on Knowledge and Data Engineering, Vol. 27, Issue 1, (**2015**), pp. 155-169.

[16] K.Jain, M.Mahdian, and A.Saberi, “A New Greedy Approach for Facility Location Problems”. Proceedings of the 34th Annual ACM Symposium on Theory of Computing, Montreal, Canada, (**2002**) May 19-21.

[17] S. Li, O. Svensson, “Approximating K-Median via Pseudo-Approximation”. Proceedings of the 45th ACM Symposium on Theory of Computing, Palo Alto, USA, (**2013**) June 1-4.

[18] Z.Y.Qu, S.L.Hou, P.Y.Zhang, “The Realization of the substation 3D visualization Training Platform”. Journal of Northeast Dianli University,Vol. 1, (**2014**), pp. 75-79.

[19] R.C. Amorim, P.Komisarczuk, “On Initializations for The Minkowski Weighted K-Means”, Proceedings of the 11th International Conference on Advances in Intelligent Data Analysis, Helsinki, Finland, (**2012**) October 25-27.

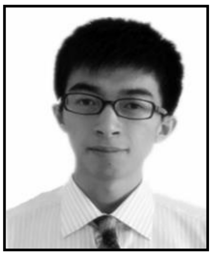
[20] N.A.Samat, M.A.A. Murad, M.T.Abdullah, and R.Atan, “Malay Documents Clustering Algorithm based on Singular Value Decomposition”. Journal of Theoretical and Applied Information Technology, Vol. 8, Issue 2, (**2009**), pp. 180-186.

[21] M.Muja, D.G. Lowe, “Fast Approximate Nearest Neighbors with Automatic Algorithm Configuration”. Proceedings of the International Conference on Computer Vision Theory and Applications, Lisboa, Portugal, (**2009**) February 5-8.

[22] http://qwone.com/~jason/20Newsgroups/

Authors

Zhao Yang Qu, is currently a professor in the Department of Computer Science and Technology, Northeast Dianli University. His research interests are mainly engaged in research work Smart Grid and Power of Information Technology, Virtual Reality, Network Technology.



Wei Ding, is currently a master student in the Department of Computer Science and Technology, Northeast Dianli University. His research interests are mainly in the areas of graph computing and smart grid.

Tie Hua Zhou, is currently an associate professor in the Department of Computer Science and Technology, Northeast Dianli University. He received the Ph.D degree in Computer Science from Chungbuk National University of Korea, in 2016. His research interests are mainly in the areas of multimedia image processing, data mining, spatial-temporal database and Energy Internet.

Ling Wang, is currently an associate professor in the Department of Computer Science and Technology, Northeast Dianli University in China. She received the Ph.D degree in Computer Science from Chungbuk National University of Korea, in 2013. Her research interests are mainly in the areas of data mining, graph computing, cloud computing, big data processing, Smart Grid and Energy Internet.

Copyright ⓒ 2017 SERSC