IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Extracting most significant data about the user queries from the search engine by K-means++ algorithm

[1] Dr.A. Gomathi
Department of Computer Science and Engineering
Vivekanandha College of Technology for Women
Elayampalayam, Tiruchengode, Tamil Nadu, India

[2]Dr.D. Sathiya
Department of Computer Science and Engineering
Vivekanandha College of Technology for Women
Elayampalayam, Tiruchengode, Tamil Nadu, India

Abstract

Today every process requires very important as well as updated data about that process or work. In order to acquire those data, employees from various fields searching through different search engines. Only very few times search engines are helping to get the users expected data but many times it can provides only the approximate data about the user expectation. To avoid this state we used the algorithm named K-means++ over the search engine documents to extract the most relevant information, because this algorithm uses the special mathematical method to find the out the successive cluster center of each cluster documents and only the first center is random selection from the data unlike K-means algorithm, which randomly selects all the cluster centers. Segmentation fusion is applied to provide the most resultant list, which is accepting as input of each cluster's documents gradually. Performance of the K-means++ algorithm is compared and evaluated with the measures like purity and F-measure, which shows the novel algorithm has providing better relevant result while comparing with other traditional algorithms. Finally, this algorithm always works with large data set, so the researchers can utilize this algorithm to their innovative ideas comes to reality with high precision and recall.

Keywords: K-means++ algorithm, Clustering, Segmentation fusion, Performance measures.

I. INTRODUCTION

Whenever searching the information through the search engine, lot of documents are listed but while considering the relevance to the user expectation is very poor. There is a need of innovative technique that should retrieve the data from the search engine with more relevant or suitable to the user request. Though many advanced search engines are available today, only few queries are getting the relevant results to the user expectation, because of this reason user indented to post their queries with different combination of its related terms again and again until get their actual expectation.

To conquer this problem along with k-means algorithm the following concepts such as fuzzy technique, query expansion and semantic similarity [1] between the documents are applied. In semantic similarity, ontology is developed to refer the useful meaning to the retrieved documents. Information can also be retrieved based on the sentence level [2] like sentence can be used as a unit to retrieve the relevant outcomes based on the user query. A multidimensional search [3] provides the relevant rank list of document based on the dimensions such as structure, metadata and keyword. The relevant list was generated by combining the score of each dimension. Fuzzy constraints over the dimensions are helped to increase the ranking accuracy.

To avoid the various uncertainties, fuzzy c-means and vector space model [4] was tested; because, the document uncertainty leads to irrelevant resultant documents when retrieving the documents from the search engine. In order to avoid these states, we proposed an extended version of traditional method

called K-means++ algorithm with segmentation fusion over a search engine. This model exhibits an efficient clustering and fuses the documents from the clusters, which are excessively nearer over the user query.

This research work is organized as follows: Section II describes the related work; Section III depicts the proposed work includes K-means++ algorithm, Segmentation fusion and Performance measures; Section IV concludes the research paper.

II. RELATED WORK

There are many techniques are available to do the information retrieval. The proposed method introduces the efficient way of retrieval to solve the problems in the existing techniques. K-means++ algorithm and fusion technique make the relevant retrieval process easier over the search engines. For clustering purpose k-means algorithm [1] was used to retrieve the documents with help of ontology, which performs the extraction based on the semantic similarity instead of matching keyword. A method called Query-Document-Dependent Fusion (QDDF) gains the possible fusion category for every search results based on user queries as well as documents [5]. Here, QDDF rearranges the former multimodal fusion method and formulated as dual-phase and regression-based fusion technique in order to increase the effectiveness and efficiency of the query response. Fuzzy Granular System [6], which uses the two dimensions like similarity and granularity to increase the efficiency of the domain-based information retrieval with help of automated fuzzy ontology scheme. While extracting information from the search engine, fuzzy relations [7] are used to index as well as ranking along with the help of crawler, ranking algorithms, user request with its position and indexer. Segmentation fusion (SegFuse) [8] is used to combine the documents from different engines based on its probabilistic value of the particular document and this method comes under probabilistic fusion category. Here, SegFuse used as a merger to merge most relevant document form different clusters. Number of documents are allowed to SegFuse are gradually increased in order to pick the suitable documents from different clusters. This work surely provides the accurate resultant list from the search engine than existing systems.

III. PROPOSED WORK

This proposed work try to give more relevant result from the search engine in very different way, because it uses efficient clustering algorithm as well as fusion technique to do the same. By limiting the least relevant documents indentifying the most suitable document related to the user query.

A. Structure of the proposed system

The architecture of the proposed work is shown in figure 1. In this system user can post their query through the user interface.

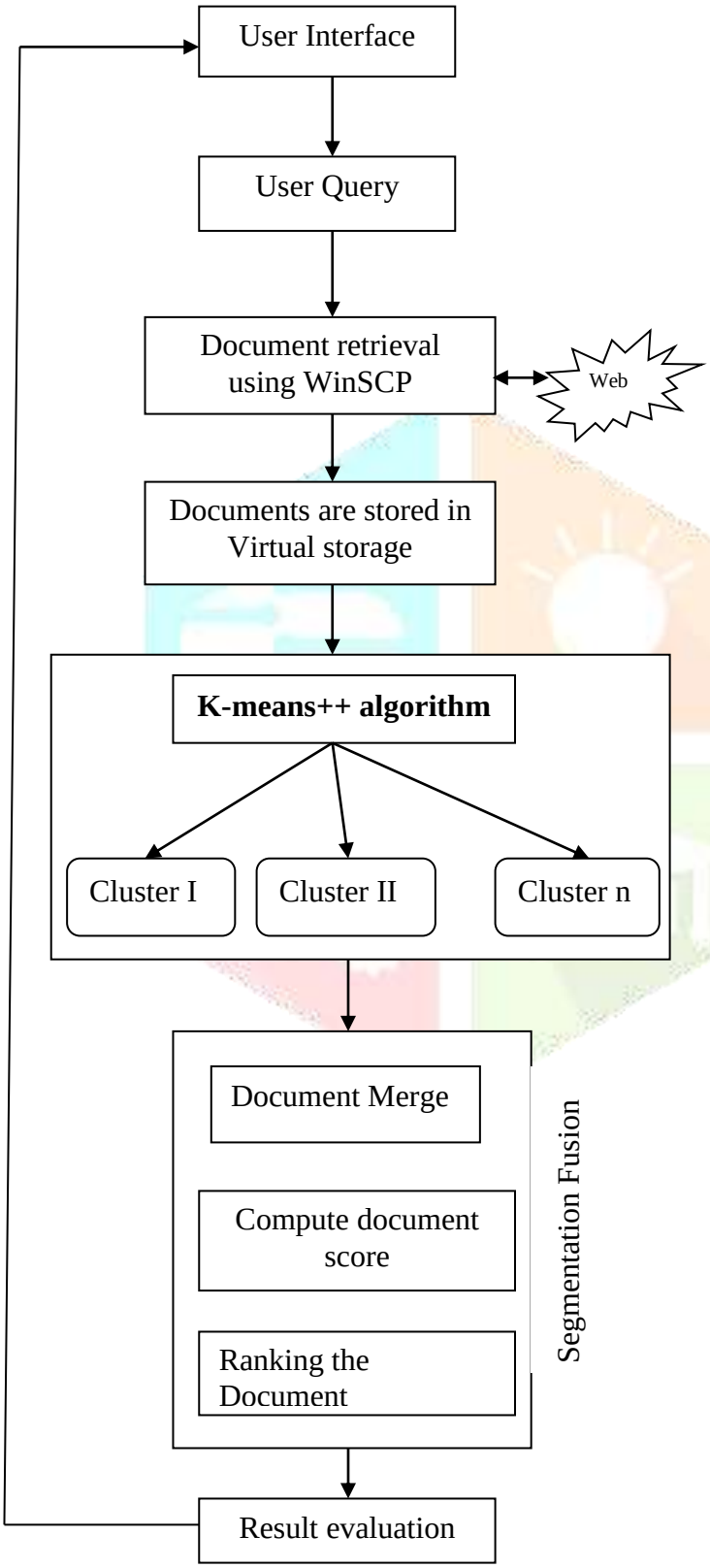

Figure 1. Architecture of proposed system

As usual search engine retrieve the documents and store them virtually in the local system. To extract the documents from search engine WinSCP tool [11] is used. It is open source free software, which can transfer the files from remote computer system to local computer and vice versa. Then documents are classified by K-means++ algorithm by selecting the first cluster center randomly then second onwards automatically will select based on the mathematical formula. So, the remaining clusters do not require selecting the centroid randomly unlike K-means algorithm. Because of this feature for large data set also K-means++ algorithm works very well. Segmentation fusion technique merge the clustered document based on its document priority. Then the performance measure such as purity and Mean Average Precision will prove the efficiency of the resultant list of documents after applying this proposed method.

The steps to follow the proposed system as follows:

- User input the query through user interface.
- Information extracted as documents from the search engine.
- Applying the K-means++ algorithm, which do the following process for each clusters.
  - Select the first cluster center randomly.
  - Use mathematical formula to select the centriod for remaining cluster.
  - Clustering the documents based on its center.
- Merge the documents using segmentation fusion technique.
- Evaluate the resultant documents.
- Relevant documents displayed in the user interface.

B. *Document clustering algorithm*

The proposed system uses the very efficient clustering algorithm named K-means++ algorithm to cluster the extracted documents from the search engine. It is a probabilistic means of K-means algorithm and it was introduced by David Arthur and Sergei Vassilvitskii at 2007 [9]. This research work proves that the K-means++ algorithm provides the suitable beginning for any cluster centers than traditional algorithms. The K-means++ algorithm has very few modifications with traditional K-means algorithm.

The following steps to be followed in the proposed algorithm:

- Select the cluster center for the first cluster as c1.
- Select the next cluster ci, by picking c2 from x' ϵ X with help of the probability of $d(x)^2 / \sum_{x \in X} d(x)^2$, here d(x) is the data point with closer distance with cluster centroid.

- Second step should be repeated up to the k^{th} cluster; here k represents the total number of clusters.
- Repeat the steps from two to four until no more cluster can be formed with K-means algorithm.

This algorithm guarantees that estimated proportion is O (log k) [9]; here k depicts the total number of clusters used. Already many authors were applied the K-means for information retrieval, but K-means ++ algorithm outperforms the former one.

C. Segmentation fusion technique

Segmentation fusion (SegFuse) is a probabilistic fusion method [8], which can do the better merging process over the documents retrieved from the search engines. In the proposed system SegFuse is used to combine the documents from various clusters in order to provide the better resultant document list about the user queries. Normally, SegFuse allows the documents from different search engines, and then separate the documents into various sizes of chunks [8]. These are allowed to merge based on the probabilistic score of the each document; here probability value can be calculated based on the relevance of document to the user query.

The following steps to be processed to merge the documents from clusters:

- Input the different cluster documents.
- SegFuse, partition the each cluster documents with various sizes, which may rise linearly.
- Compute the probability values (final score) for each document from all the clusters.
- Merges (rank) the documents from different clusters based on the final score of that particular document.

The following equations are used to find the final score of the each document in the cluster.

$$D^c = S_d - D^c min / D^c max - D^c min \quad (1)$$

$$\text{SegFuse} = \sum_{c}^{N} P(\text{rel}|d_k^c) x D^c + P(\text{rel}|d_k^c) \quad (2)$$

Here, from the equation (1) S_d represents the ranking score of the document in the particular cluster. $D^c max$ and $D^c min$ depict the maximum and minimum score of the document in the cluster. D^c represents the normalized score of the retrieved relevant documents.

From the equation (2), c and k stand for number cluster and number of document in each cluster respectively. N the number of clusters covering the document d.

D. Performance Evaluation

In order to check the quality of clustering popularly used measure named F-measure is used. This is an external measure method and covers the both recall and precision of information retrieval [9]. Each cluster has its own documents, recall and precision of each may vary. The F-measure can be calculated with the following formula.

$$F = \sum_i \frac{Di}{D} maxF(i,j) \quad (3)$$

From the equation (3) i and j represents the cluster i and cluster j. Here, D represents the number of documents includes both clusters i and j. Threshold of the F-measure is from 0 to 1 and the higher value this measure shows greater quality of the cluster.

IV. EXPERIMENTS AND RESULT

By testing the merged documents we prove K-means++ performs better than the traditional systems. Here, evaluation metric called Mean Average Precision (MAP) [10] is used to evaluate the documents score along with F-measure. For different sample documents MAP values are compared as in table 1.

Table 2 shows the experimental results with three categories of samples. Initial sample is taken as 5 to 10 documents, corresponding F-measure for both K-means++ and K-means algorithm's metrics are 0.75 and 0.69 respectively. This shows the gradual improvement over the proposed method. At the same time MAP produces the metrics 0.478 and 0.459 for the above algorithms, here also proved the same. While considering the third case (150 to 200 documents) F-measure and MAP score 0.53 & 0.50 and 0.402 & 0.397 respectively.

Table 1. Comparison with Existing System

System/ Metrics	Precision	Recall
K-means	0.0224	0.875120
Fuzzy c-means	0.0360	0.880065
Proposed system	0.0391	0.882065

Table 2. Experiment results of various set of documents.

Sl. no	No of Doc.	K-means++		K-means	
		F-measure	MAP	F-measure	MAP
1	05-10	0.75	0.478	0.69	0.459
2	50-100	0.61	0.433	0.58	0.426
3	150-200	0.53	0.402	0.50	0.397

Here, we found the degradation over the measures, because the relevance of the documents is reduced while increasing the document count. Finally, K-means++ algorithm proves that the information retrieval over the search engine must more relevant about the user query than the tradition system such as K-means algorithm and C-means algorithm, etc. The proposed method surely works for large number of documents also and Young researchers kindly utilize this algorithm for their part of their research.

V. CONCLUSION AND FUTURE WORK

The proposed system uses the K-means++ algorithm in order to improve the information retrieval from the search engine by selecting the random clutser center for the initial cluster and the remaining clusters centers are automatically selected using the mathematical fomula. This way proposed system is differ from K-means algorithm. The experiments proved that the K-means++ algorithm with segmentation fusion provides accurate performance increases than the traditional algorithms with the metrics F-measure and MAP. The K-means++ system can be used to any type of data classification without any hardness of clustering process.

Further process can be extended with any soft computing technique to extract the multimedia, spatial and Geographical, etc. informations with more accuracy and efficiency.

REFERENCES

[1] Kajal Joshi and Rashmi Chauhan, "Ontology based Fuzzy Classification of Web Documents for Semantic Information Retrieval", IEEE, 2013.

[2] Jiali Zuo, Mingwen Wang, Jianyi Wano and Wenbing Luo, "Information Retrieval Model Combining Sentence Level Retrieval", International Conference on Asian Language Processing, IEEE, 2013.

[3] Wei Wang, Christopher Peery, Ame ´lie Marian, and Thu D. Nguyen, "Efficient Multidimensional Fuzzy Search for Personal Information Management Systems", IEEE Transactions on Knowledge and Data Engineering, Vol. 24, No. 9, September 2012.

[4] Chandrani Ray Chowdhury and Prachet Bhuyan, "Information Retrieval using Fuzzy c-means Clustering and Modified Vector Space Model", IEEE, 2010.

[5] Zhonghua Li, Bingjun Zhang, Yi Yu, Jialie Shen, and Ye Wang, "Query-Document-Dependent Fusion: A Case Study of Multimodal Music Retrieval", IEEE Transactions on Multimedia, Vol. 15, No. 8, December 2013.

[6] Raymond Y.K. Lau and L. Song, "Mining Fuzzy Ontology for Fuzzy Granular IR Systems", WCCI, IEEE World Congress on Computational Intelligence June, 10-15, 2012.

[7] Joseph B. Lee and Eunjin Kim," Fuzzy Web Information Retrieval System with Fuzzy Thesaurus using Fuzzy Relational BK-Products", IEEE World Congress on Computational Intelligence, 10-15, 2012.

[8] Milad Shokouhi, "Segmentation of Search Engine Results for Effective Data-Fusion", School of Computer Science and Information Technology RMIT University, 2007.

[9] Bashar Aubaidan, Masnizah Mohd and Mohammed Albared, "Comparative Study of K-Means and K-means++ Clustering Algorithms on Crime Domain", Journal of Computer Science 10 (7): 1197-1206, 2014 ISSN: 1549-3636 © 2014 Science Publications doi:10. 3844 / jcssp. 2014. 1197.1206 Published Online 10 (7), 2014.

[10] Lillis. D, Toolan. F, Collier. R, and Dunnion. J, " ProbFuse: a probabilistic approach to data fusion", In Proc. ACM SIGIR conf., pages 139–146, Seattle, Washington, 2006.

[11] http://www.techspot.com/downloads/1487-winscp.html.

[12] Supriya.Y and Gayathri.K, "Study of Query Optimization in Cloud", International Journal of Scientific Research in Science and Technology",ISSN: 2395-602X, Vol.6, Issue 2, pages 92-94, 2019.