International Journal of Electrical and Computer Engineering (IJECE)
Vol. 5, No. 4, August 2015, pp. 782~787
ISSN: 2088-8708

Scientific Documents Clustering Based on Text Summarization

Pedram Vahdani Amoli, Omid Sojoodi Sh.
Faculty of Electrical, Computer and IT Engineering, Qazvin Islamic Azad University
Nokhbegan Blvd, Qazvin, Iran

Article Info

Article history:

Received Jan 27, 2015
Revised Apr 25, 2015
Accepted May 18, 2015

Keyword:

Scoring
Summarization
Text Clustering
Text Mining

ABSTRACT

In this paper a novel method is proposed for scientific document clustering. The proposed method is a summarization-based hybrid algorithm which comprises a preprocessing phase. In the summarization phase unimportant words which are not frequently used in the document are removed. This process reduces the amount of data for the clustering purpose. In this proposed method after the preprocessing phase, Term Frequency/Inverse Document Frequency (TFIDF) is calculated for all words in the document and BM25 in caluculated for words in sentences and summed over the document to score each word in document level. In next phase, Text summarization is performed based on BM25 scores. After that document clustering is done according to the scores of calculated TFIDF. The hybrid progress of the proposed scheme, from preprocessing phase to cluster labeling, gains a rapid and efficient clustering method which is evaluated by 400 English texts extracted from scientific articles of 11 different topics. The proposed method is compared with CSSA, SMTC and Max-Capture methods. The results demonstrate the proficiency of the proposed scheme in terms of computation time, and comparative efficiency using F-measure criterion.

Corresponding Author:

P. V. Amoli,
Faculty of Electrical, Computer and IT Engineering,
Qazvin Islamic Azad University,
Nokhbegan Blvd, Qazvin, Iran.
Email: pedramvahdani@gmail.com

1. INTRODUCTION

Information retrieval is an area concerned with searching for documents to make some analysis and the critical part of its approachs is to represent the content of a document and one of these apparchs is clustering. Normally, documents are represented as bag of words which means they occur independently. Many researchers have tried to group words into topic in order to represent the importance of relations between words with in a group. [1]

Document clustering or text clustering is one of the main terms in text mining. It refers to the process of grouping documents with similar contents or topics into clusters to improve both availability and reliability of documents [2-3]. The text documents are grouped together as clusters on the basis of their similarities and into different groups on the basis of dissimilarities between them, this concept forms the foundation of text document clustering [2-4]. Data clustering is a difficult task in computing and that is because of its hard computationally [5].

Recently there have been some researches on text clustering based on sentence which highlights the importance of measuring words in sentence level. It is because of higher measurement result that a word gains in sentence level compare to document level.

A. Skabar, Kh. Abdalgader [6] introduce an approach base on sentence level which is also fuzzy and operate on input data by using graph representation in EM framework.

K. Jeyalakshmi, R. Deepa, M. Manjula [7] present a novel hierarchical fuzzy relational clustering algorithm that operates on relational input data. Thier algorithm uses a graph representation of the data, and operates in a Fuzzification Degree framework in which the graph centrality of an object in the graph is interpreted as likelihood.

E.R.M. Seno and M.D.G.V. Nunes [8] they propose an evaluation framework based on an incremental and unsupervised clustering method which is combined with statistical similarity metrics to measure the semantic distance between sentences.

S. Sharma and V. Gupta [9] has presented a text clustering approach that uses Karaka for processing Punjabi text and process only top twenty term of each document for clustering.

There are three kinds of problems in document clustering. The first one is how to define similarity of two documents. The second problem is how to decide appropriate number of document clusters in a text collection and the third one is how to cluster documents precisely corresponding to natural clusters. Therefore the researchers have tried to satisfy these issues by proposing their algorithm. In order to achieve a good clustering result variety of methods has been proposed by using techniques such as dimension reduction, pattern analysis, semantic processing and etc. By the way, beside accuracy another problem of text clustering is timing which also important for measuring and efficiency.

In this paper for the aim of achieving effective clustering a new approach has been proposed. As effectiveness and accuracy of text clustering depend on the pureness of clusters. Therefore, selecting of an item that is mostly related to a specific cluster is the most critical issue. Hence, measuring of documents has been made based on their words thus proper approach for validating useful words is the most important issue. For this purpose the new approach that is presented in this paper use a new way to select useful words by eliminating non-useful words and measuring each words based on their sentences. By that way timing has been monitored to be counted as important issue because it affects efficiency.

2. RESEARCH METHOD

In this section the proposed hybrid method is described. The aim is to attain a fast and efficient text clustering method. The proposed algorithm consists of four main phases as a) preprocessing phase, b) word weighting and scoring phase, c) summarization phase and d) clustering phase.

By the way each text has been read from saved txt file which holds the content of extracted website article.

2.1. Preprocessing Phase

In the preprocessing phase, frequently used words which are commonly existed through over the text are removed. Hence the amount of data for clustering is reduced. After that, stemming is performed by Porter algorithm [10]. This action helps the algorithm to find different aspects of the same stems to search the frequent items more efficiently.

2.2. Weighting and Scoring Phase

In this phase, Term Frequency/Inverse Document Frequency (TFIDF) is calculated for each word in the document and okapi BM25 is calculated in the sentence level and summed in the document. The BM25 for each word is considered as the weight factor for the next step in summarization.

2.3. Summarization

The objective of summarization is to remove non-important words from processing further. Therefore, clustering will be fast and it would be more efficient with the amount of data reduced. BM25 formulae is used for text summarization as

$$Score(S,R) = \log\frac{N - n(R) + 0.5}{n(R) + 0.5} * \frac{f(R,S) * (k+1)}{f(R,S) + k * \left(1 - b + b * \frac{|S|}{avg|S|}\right)} \quad (1)$$

In which $Score(S,R)$ is the score value of word R in the sentence S. Term $f(R,S)$ is the frequency of the word R in the sentence S while $|S|$ and $avg|S|$ are sentence length and averaged length over the text. Parameter N is the number of sentences in the text and $n(R)$ is the number of documents comprising the word R. Constants b and k are found optimally as $b = 0.75$ and $k = 2$ as it will be explained in details in the experiments. It should be noted that if the BM25 value is less than one for a word, the word will be eleminated as a non-important one. Hence, a set of important frequent words are created for the document.

2.4. Clustering

The clustering process is performed by considering the TFIDF values of frequent words. The algorithm can be summarized as follow

a) If the first document is analyzed put it in one cluster, otherwise go to step (b).
b) Evaluate the document with every cluster and sum the weights of the similar words between this document and other documents in each cluster. If no cluster remains, go to step (c).
c) Find the largest weight. If it was not zero, assign the document to the cluster and otherwise go to step (d).
d) Assign a new cluster to this document.

Finally, the clusters are labeled as

a) Find for each cluster the word with largest weight.
b) Among the weights in the cluster find the largest one.
c) Label the cluster with the word with largest weight.

3. RESULTS AND ANALYSIS

To evaluate the efficiency of the proposed method, 400 English texts in 4 groups of experiments were used. For each group 10, 30, 50, 70, 90 and 100 samples were chosen. All texts have been extracted from scientific websites articles of different topics (see figure 1).

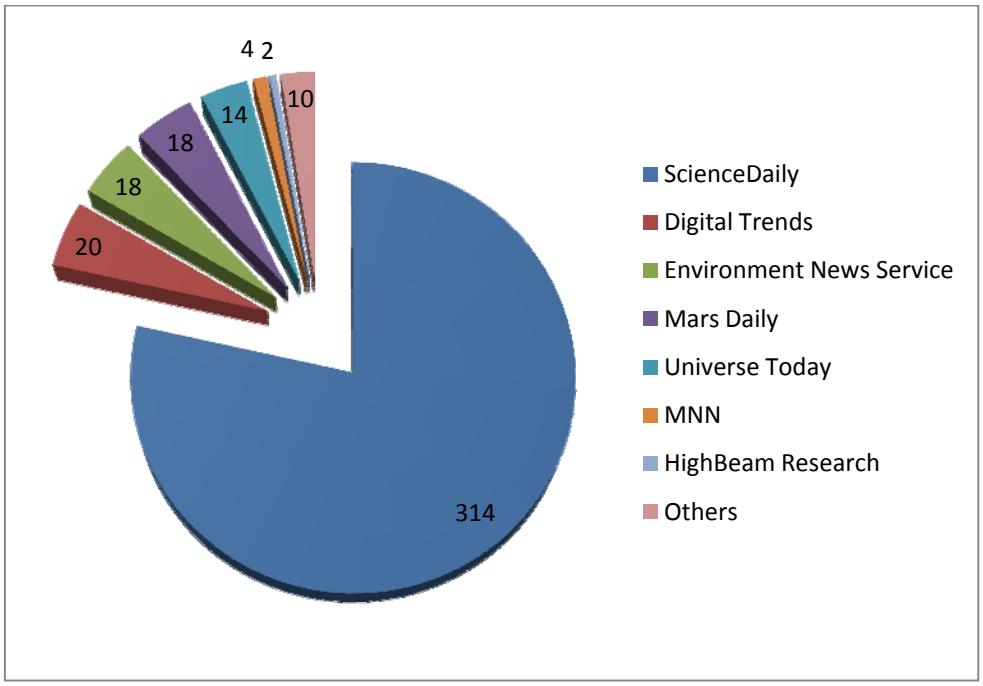

Figure 1. Used databases article for the experiments

The experiments contain two parts. In the first part, two constants in the BM25 formula are investigated to find the optimum values. After the method is optimized, it will be compared with CSSA [9], SMTC [11] and Max-Capture [12] methods in the second part of the experiments.

In the BM25 formulae there are two parameters k and b. To investigate the impact of parameter k on the efficiency of the proposed method it is set to 1, 2, 3 and 4. Tables 1 & 2 show the results of F-measure and computation times of the method for different values of parameter k.

Table 1. F-measure for 4 different values of parameter k

F-measure	k=2 b=0.75	k=1 b=0.75	k=3 b=0.75	k=4 b=0.75
DS1	0.50	0.51	0.51	0.50
DS2	0.51	0.48	0.51	0.50
DS3	0.62	0.59	0.66	0.62
DS4	0.51	0.46	0.50	0.49
Average	0.53	0.51	0.54	0.52

Table 2. Run time (second) of the proposed method for 4 different values of parameter k

Time(Sec)	k=2 b=0.75	k=1 b=0.75	k=3 b=0.75	k=4 b=0.75
DS1	34.33	35.50	35.16	35.16
DS2	27.16	27.16	27.00	27.00
DS3	33.33	34.83	34.00	34.16
DS4	26.16	28.50	27.83	27.83
Average	30.25	31.50	31.00	31.04

The results given in Tables 1 and Table 2 show that the best value for k is 2. For assessment of the parameter b, 4 different values 0.25, 0.75, 1 and 1.25 are set. Tables 3 & 4 show the results of F-measure values and computation time for different values of parameter b.

Table 3. F-measure for 4 different values of parameter b

F-measure	b=0.75 k=2	b=1 k=2	b=1.25 k=2	b=0.25 k=2
DS1	0.5	0.51	0.53	0.46
DS2	0.51	0.52	0.46	0.47
DS3	0.62	0.63	0.65	0.68
DS4	0.51	0.5	0.46	0.5
Average	0.535	0.54	0.525	0.5275

Table 4. Run time (second) of the proposed method for 4 different values of parameter b

Time(Sec)	b=0.75 k=2	b=1 k=2	b=1.25 k=2	b=0.25 k=2
DS1	34.33	35.83	36.00	35.16
DS2	27.16	27.50	27.83	27.16
DS3	33.33	35.00	35.16	34.16
DS4	26.16	28.50	28.83	27.83
Average	30.25	31.70	31.95	31.08

According to the results given in Tables 3 & 4, the optimum value for parameter b is 0.75. In Table 3, F-measure value for $b = 0.75$ is better than the value for $b = 1.25$ and $b = 0.25$. There is only a slight weakness compared to $b = 1$. However considering the results of both Tables, it can be inferred that the optimum value for parameter b is $b = 0.75$ comparatively.

In the second part of the experiment, the proposed method is compared with the CSSA [9], SMTC [11], and Max-Capture methods [12]. Two criteria are used for the assessment as computation time and F-measure. The result of computation time for the comparing methods is given in figure 2. As it can be seen from the results, the proposed method outperforms the other methods in term of computation time.

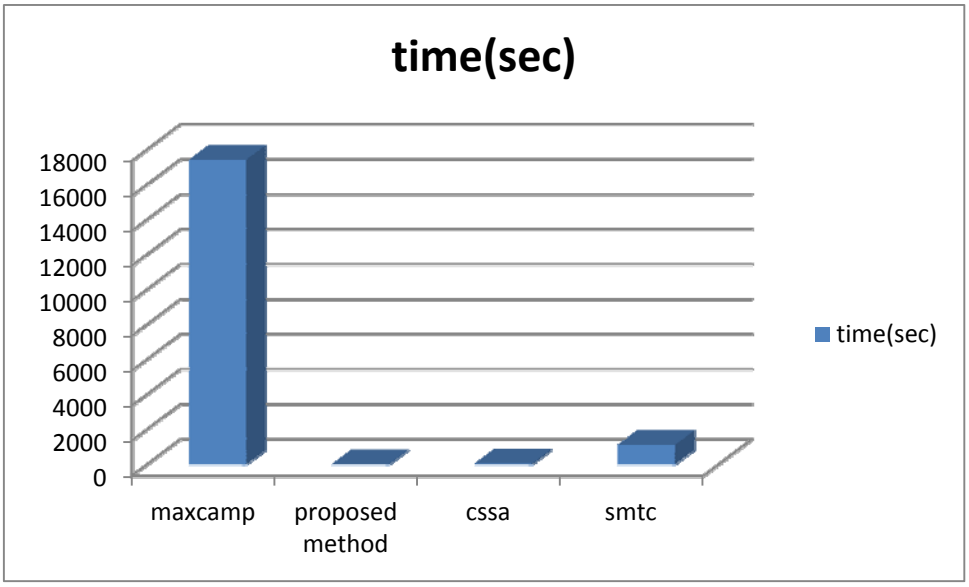

Figure 2. Processing time for CSSA [9], SMTC [11], Max-Capture [12], and proposed method

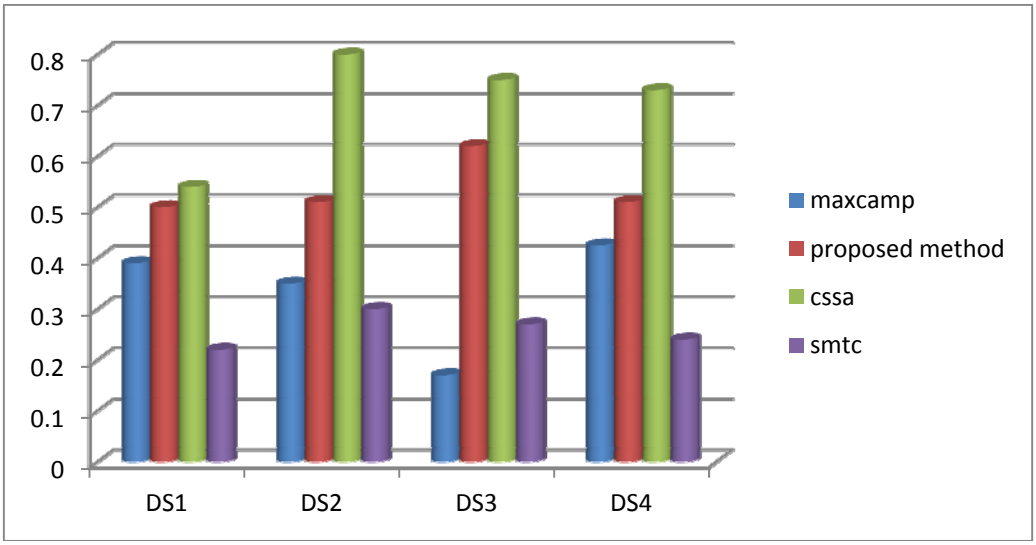

Figure 3. Comparison of the proposed method with CSSA [9], SMTC [11] and Max-Capture [12] methods

The comparison of F-measure results for these four methods has been shown in figure 3. This result shows that the CSSA method yield the better results compared to other methods. Although in the first dataset the proposed method had almost similar result with CSSA but in the other three dataset it could not repeat it again. Over all the proposed method attains a better performance compared to SMTC and Max-Capture methods by the F-measure. Both the methods SMTC and Max-capture use all the words in the document and this feature affects the clustering performance in terms of processing time and accuracy. The difference in the accuracy of the proposed method and CSSA method is caused by the differences in the concept of scoring methods for clustering.

4. CONCLUSION

In this paper a new efficient and very high speed approach for document clustering was proposed. The result of this clustering has shown that the new method of using text summarization for eliminating not-useful words has been effective in order to perform effective clustering. It also needs to be mention that this

method has gained incredibly low running time compare to all comparative methods. It has been believed this is a result of job reduction that has been implemented by summarization.

ACKNOWLEDGEMENTS

I would like to thanks my supervisor Dr.Sojoodi for his advice and comment that helped me through my project.

REFERENCES

[1] Ravi kumar V., K. Raghuveer, "Legal Documents Clustering and Summarization using Hierarchical Latent Dirichlet Allocation", IAES International Journal of Artificial Intelligence (IJ-AI), Vol. 2, No. 1, March 2013, pp. 27~35, ISSN: 2252-8938.
[2] Tian Weixin, Zhu Fuxi, "Text Document Clustering Based On The Modifying Relations", In *Proceedings of IEEE International Conference on Computer Science and Software Engineering*. 1 (12-14 Dec. 2008), 256-259.
[3] S. Murali Krishna and S. Durga Bhavani, "An Efficient Approach for Text Clustering Based on Frequent Itemsets", European Journal of Scientific Research ISSN 1450-216X, 42, 3 (2010), EuroJournals Publishing, Inc. 2010, 399-410.
[4] Le Wang, Li Tian, Yan Jia and Weihong Han, "A Hybrid Algorithm for Web Document Clustering Based on Frequent Term Sets and k-Means", In *Proceedings APWeb/WAIM 2007 International Workshops: DBMAN 2007, WebETrends 2007, PAIS 2007 and ASWAN 2007*.
[5] Parthajit Roy and J. K. Mandal, "A Novel Spectral Clustering based on Local Distribution", *International Journal of Electrical and Computer Engineering (IJECE)*, Vol. 5, No. 2, April 2015, pp. 361 – 370, ISSN: 2088-8708.
[6] Andrew Skabar, Khaled Abdalgader, "Clustering Sentence-Level Text Using a Novel Fuzzy Relational Clustering Algorithm", IEEE Transactions on Knowledge and Data Engineering, Vol. 25, No. 1, January 2013.
[7] K. Jeyalakshmi, R. Deepa, M. Manjula, "An Efficient Clustering Sentence-Level Text Using A Novel Hierarchical Fuzzy Relational Clustering Algorithm", International Journal of Advanced Research in Computer and Communication Engineering Vol. 3, Issue 2, February 2014.
[8] Ramiz M. Aliguliyev, "A new sentence similarity measure and sentence based extractive technique for automatic text summarization", Expert Systems with Applications 36 (2009) 7764–7772.
[9] Saurabh Sharma and Vishal Gupta, "Punjabi text clustering by sentence structure analysis", CS & IT-CSCP, pp. 237–244, 2012.
[10] M.F. Porter, "An algorithm for suffix stripping", 14(3): 130-137, 1980.
[11] Yuan L., "An Effective Chinese Short Message Texts Clustering Algorithm Based on The Ward's Method", 978-1-4577-0536-6/11/2011 IEEE, 1897-1899.
[12] Zhang W., T. Yoshida, X. Tang, Q. Wang, "Text clustering using frequent itemsets", 2010. Knowledge-Based Systems 23, pp. 379–388.

BIOGRAPHIES OF AUTHORS

P. V. Amoli obtained Bachelor Degree in Information Technology majoring Information system Engineering in 2008.
His interest topic includes Data mining.

Dr. Sojoodi received his B.Sc in Software Eng. from QIAU and M.Sc degree in AI from SRBIAU and PHD degree in AI from UPM.
He is with Islamic Azad Qazvin University as a lecturer. His research is in fields of Data mining. Further info on his homepage: http://qiau.ac.ir/sojodishijani.info