ISSN (e): 2250 – 3005 || Volume, 07 || Issue, 09|| September – 2017 ||
International Journal of Computational Engineering Research (IJCER)

KEBCA – A Keyword Extraction Based Clustering Algorithm

George Louis Raja[1], F. Sagayaraj Francis[2], P. Sugumar[3]
[1] Research Scholar, Department of Computer Science and Applications, SCSVMV University, Enathur, Kanchipuram, India
[2] Professor, Department of Computer Science and Engineering, Pondicherry Engineering College, Pondicherry, India
[3] Assistant Professor, PG and Department of Computer Applications, Sacred Heart College, Tirupattur, India
Corresponding Author: George Louis Raja

ABSTRACT
In this paper, we propose a document clustering method that tends to arrive at clusters of text documents which are assumed to be related together in their keywords. The method first uncovers the keywords of significance from a text document by means of the Keygraph method. The revealed keywords from the documents are represented in a graph, with the keywords as nodes, and the semantically related keywords are linked through edges. Then, they are clustered based on how strongly the keywords in two graphs (documents) are connected, more strongly they are connected, more is the probability that they can be belonging to the same cluster.
Keywords: *Clustering, Keyword, keyword Extraction, Keygraph, Semantic Clustering*

Date of Submission: 22-08-2017 Date of acceptance: 09-09-2017

I. INTRODUCTION

Keygraph is a keyword extraction algorithm proposed by H. Sayyadi, L. Raschid [1]. KeyGraph is based on grouping of terms in a graph, the nodes of the graph will represent the co-occurrence among them. These groups are used to identify the concepts which depicts the context of the documents. Each groups present in the graph are ranked according to their frequency and the relationship of each term in a group, and the term with more relationship among other terms in a group is termed as the keyterm. This method is applied repeatedly to identify the keywords present in a text document.

The major Phases of Keygraph is specified as below

1. Extracting Foundations: In this phase the rudimentary and premise concepts are grouped as clusters, they are obtained by the estimation of co-occurrence of relationships.

2. Extracting columns: The relationships between groups obtained in step I and the terms are arrived at in this phase.

3. Extracting roofs: The terms that best describe the documents are identified based on the strength of their relationship among other terms in a cluster.

II. KEYWORD EXTRACTION

The process of keyword extraction is identifying the most appropriate terms that best describe the content of a text document. This process is a preliminary stage in text data mining, where in a documents context can be well derived based on the extracted keywords. Its application can also be extended in Natural Language Processing and Information Retrieval. The process is also termed as Key Phrase Extraction, Key Term Extraction or Key segment Extraction. The process of keyword extraction tends to gather the terms that best describe the document, however another variation to keyword extraction called as keyword assignment leads to gather keywords from a predefined vocabulary.

III. KEYWORD EXTRACTION METHODS

Keyword Extraction methods are broadly classified into supervised and unsupervised methods. The Existing methods for automatic keyword extraction can be classified as: 1) Statistical Oriented Methods and 2) Machine Learning Methods (3) Linguistic Approaches.

(1) Statistical Oriented Methods : These methods will filter out the terms in a document, based on the number of occurrences of terms available in a document, the terms that are present frequently are identified to be the key terms. But this approach does not suit the purpose of keyword extraction, since it does not consider

the semantically important terms which may occur less frequently, but can best describe the context of the document.

(2) Machine Learning Methods: These methods are supervised counter parts of key word extraction methods. These methods apply the concept of Artificial Intelligence, Neural Network and Fuzzy Logic to derive the key terms of a document.

(3) Linguistic Approaches: These methods exploit the linguistic features of a document by considering the Lexical Alignment, the Syntactic and Semantic Structure of the document to identify the keywords in a document.

In a nutshell, the supervised key word extraction method require a pre-designed model with a pre-specified set of keywords. These keywords are normally extracted by human interference which is injected into the vocabulary. These keywords are later used to induce the learning process of the system. This method is not best of the breed since the manual intervention is extremely difficult and inconsistent and time consuming. These training requires domain expertise which makes things more complex.

The unsupervised methods, are affected by inaccuracy and inconsistency. The output produced by the same method in different iterations may not be the same leading to the above mentioned problems.

The proposed clustering method uses Keygraph method which was described in section I, which falls under the unsupervised clause.

IV. KEBCA – KEYWORD EXTRACTION BASED CLUSTERING ALGORITHM

This section describes the proposed clustering algorithms which first uncovers the key terms in a document using the keygraph method. These keyterms and documents are represented in a graph, which is later inspected and used to cluster the text documents into groups.

The Algorithm for the above mentioned process is given below

ALGORITHM KEBCA (Dataset A)

INPUT: Dataset A Containing the documents to be clustered

OUTPUT: Clusters upon the documents of the Dataset A

Step 1: Extract keywords from the documents using KeyGraph algorithm.

for each Document D in Dataset A

Extract the Keywords from the document

End for

Step 2: Constructing a Document Keygraph

Create an adjacency matrix that represents the documents as columns and keywords as rows

Find the degree of relationship among the keywords using the matrix

Step 3: Clustering

Cluster the documents according to the relationship level

End KEBCA

Example

Step 1: Extract keywords from the documents using KeyGraph algorithm

Consider the following Documents A and B and after the application of the keygraph algorithm, this step extracts the keywords from the given documents A and B as follows.

A Cluster is therefore a collection of objects which are similar between them and are dissimilar to the objects belonging to other clusters

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group called a cluster are more similar in some sense or another to each other than to those in other groups clusters

The list of keywords extracted by the keygraph algorithm from Document A are {belonging, dissimilar, collection, cluster, objects, similar}. The set of keywords for Document B being {called, task, set, cluster, analysis, objects, group, sense, similar}.

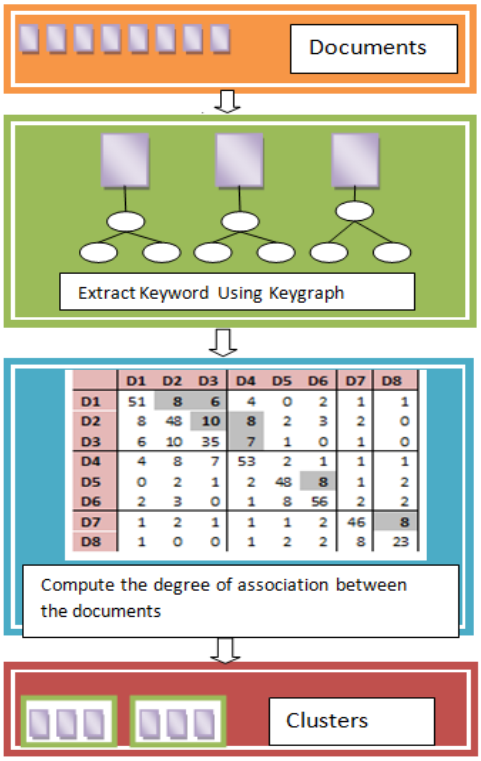
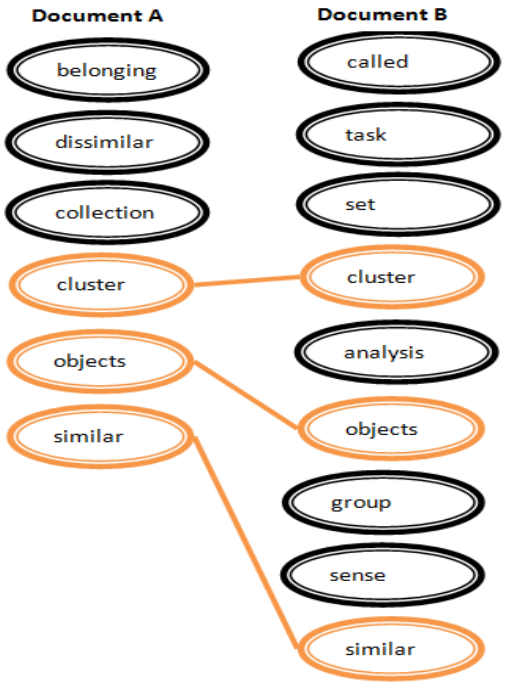


Figure 1- Phases of KEBCA

Step 2: Constructing a Document Keygraph

After, the keywords of the Text document are identified, the keywords are represented as nodes. The edges represent the terms that are related from Document A and B. in this case there are three such edges from the terms of A to B, namely {cluster, object , similar} . hence the degree of relatedness among the document A and B would be 3.



Degree of association among Document A and B is 3

Step 3: Clustering

After the degree of relatedness among the documents are computed, the documents and the keywords are represented in a matrix. The documents are represented as rows and the terms as columns. We have taken a set of 8 documents for sample, and arrived at the degree of relatedness as below.

	D1	D2	D3	D4	D5	D6	D7	D8
D1	51	8	6	4	0	2	1	1
D2	8	48	10	8	2	3	2	0
D3	6	10	35	7	1	0	1	0
D4	4	8	7	53	2	1	1	1
D5	0	2	1	2	48	8	1	2
D6	2	3	0	1	8	56	2	2
D7	1	2	1	1	1	2	46	8
D8	1	0	0	1	2	2	8	23

The closely related documents which have similar degree of relatedness would form clusters. Hence from the example above the clusters formed are {D,D2,D3,D4} {D5,D6} and {D7,D8}.

V. EXPERIMENTAL RESULTS

The experiment conducted using a set of ten random documents belonging to three categories, resulted in the following clusters based on the K-Means, Vector Space Model and NoGoDi Methods.

Method	Cluster 1	Cluster 2	Cluster 3
K-means	**A,B,D,H,I**	**C,J**	**E,F,G**
Vector-space model	**A,B,C,H,I,J**	**D,G**	**E,F**
NoGoDi	**A,B,C,D**	**H,I,J**	**E,F,G**
KEBCA	**A,B,C,D**	**H,I,J**	**E,F,G**

For the clusters generated, the following table summarizes the clustering quality of the methods.

Method	Precision %	Recall %	F-Measure
K-means	**65**	**23**	**34**
Vector-space model	**40**	**17**	**24**
NoGoDi	**80**	**50**	**62**
KEBCA	**80**	**70**	**74**

VI. CONCLUSION

We have tried to approach the text clustering problem with keyword extraction technique. Since, the context of a text document can be well described by its keyterms. We have utilized the keygraph method to deduce the keywords from text documents. After then based on the strength of association among the documents based on these keywords, we have tried to cluster down the documents. The trial experiments revel that this method performs better than the existing methods for the selected set of documents. Compare this method with atleast 3 existing text document clustering method and prove that this method works better.

REFERENCES

[1] H. Sayyadi, L. Raschid. "A Graph Analytical Approach for Topic Detection", ACM Transactions on Internet Technology (TOIT), 2013

[2] Snehalata M. Lad. Keyword Extraction from Conversation Text Document and Recommending Document using Fuzzy Logic Based Weight Matrix Method. International Journal of Advanced Research in Computer Science, Vol.7, No.4, August 2016, 34-38.

[3] Maryam Habibi and Andrei Popescu-Belis. Keyword Extraction and Clustering for Document Recommendation in Conversations. IEEE, 2013, 1-14.

[4] His-Cheng Chang, Chiun-Chieh Hsu. Using Topic Keyword Clusters for Automatic Document Clustering. Proceedings of the Third International Conference on Information Technology and Applications, IEEE, 2005.

[5] Youngsam Kim, Munhyong Kiml, Andrew Cattle, Julia Otmakhova. Applying Graph-based Keyword Extraction to Document Retrieval. International Joint Conference on Natural language Processing, October 2013, 864-868.

[6] aryam Habibi and Andrei Popescu-Belis. Keyword Extraction and Clustering for Document Recommendation in Conversations. IEEE, 2015, 746-759.

[7] Mohammad Rezaei, Najlah Gali, Pasi Franti. CIRank: A Method for Keyword Extraction from web pages using Clustering and distribution of nouns. IEEE/ WIC /ACM International Conference on Web Intelligence and Intelligent Agent technology. 2015, 79-84.

[8] FangJiang, GuoheLi, XueYun, Xiang Yue. Semantic –based Keyword Extraction Method for Document. International Journal of u- and e-Service, Science and Technology. Vol.8, No.5, 2015, 37-46.

[9] Shete Nikita U., Bhor Jayesh B., Zaware Vandana B., Thube Aswini S. Survey of Document Recommendation based on Keyword Extraction and Clustering from textual Conversation. International Journal of Research in Advent Technolgoy, February 2017, 1-4.

International Journal of Computational Engineering Research (IJCER) is UGC approved Journal with Sl. No. 4627, Journal no. 47631.

George Louis Raja. "KEBCA – A Keyword Extraction Based Clustering Algorithm." International Journal of Computational Engineering Research (IJCER), vol. 7, no. 9, 2017, pp. 08–11.