IMPLEMENTATION OF KEA-KEYPHRASE EXTRACTION ALGORITHM BY USING BISECTING K-MEANS CLUSTERING TECHNIQUE FOR LARGE AND DYNAMIC DATA SET

M. Arshad, Asst. Professor, IT-Dept., SIRT College, Bhopal-462021, m_arshoo@yahoo.com

Abstract

In most traditional techniques of document clustering, the number of total clusters is not known in advance and the cluster that contains the target information cannot be determined since the semantic nature is not associated with the cluster. To solve this problem, this work proposes a new clustering algorithm based on the Kea[1] key phrase extraction algorithm which returns several key phrases from the source documents by using some machine learning techniques. In this work, documents are grouped into several clusters like Bisecting K-means, but the number of clusters is automatically determined by the algorithm with some heuristics using the extracted key phrases. By this it is easy to extract test documents from massive quantities of resources.

Keywords – KEA, K-means, Bisecting K-means, Clusters, Key Phrase Extraction, etc.

Introduction

Current research in the area of text mining tackles problems of text representation, classification, clustering, information extraction and modeling of hidden patterns. Text classification is an important functionality of text mining. Text classification based on content is a useful and challenging work, in face of dynamically growing huge amount of texts in various fields on the Internet and the developing technology in natural language processing and machine learning. As we know that "Current research in the area of text mining tackles problems of text representation, classification, clustering, information extraction or the search for and modeling of hidden patterns" As we know text clustering is one of the important text mining functionalities and now it becomes a natural activity in every organization.

In this paper we present the following task:

1. Extracting key phrases from given document by the automatic key phrase extraction.
2. Applying clustering algorithm on these extracting key phrases and generating clustered documents.

KEA (Automatic Key-phrase Extraction)

Keyphrases provide semantic metadata that summarize and characterize documents. Kea identifies candidate keyphrases using lexical methods, calculates feature values for each candidate, and uses a machine learning algorithm to predict which candidates are good keyphrases. The machine learning scheme first builds a prediction model using training documents with known keyphrases, and then uses the model to find keyphrases in new documents. The system is simple, robust, and publicly available.

Kea's extraction algorithm has two stages:

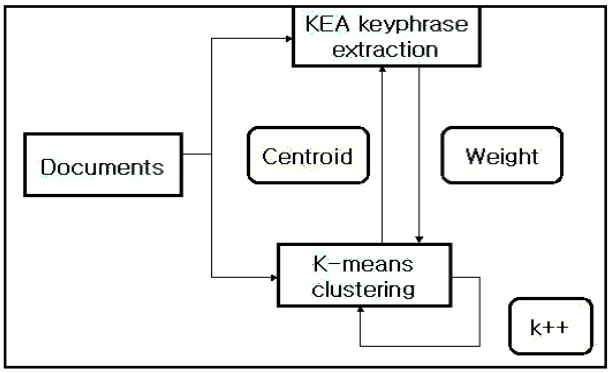
1. Training: Create a model for identifying keyphrases, using training documents where the author's keyphrases are known.
2. Extraction: Choose keyphrases from a new document, using the above model.

Bisecting K-means Clustering

In this section we discuss general issues related to the K-means clustering algorithm and introduce the bisecting K-means algorithm. There are many ways to enhance the basic K-means algorithm. But to keep things simple, we chose a very simple and efficient implementation of the K-means algorithm. For instance, we select our initial Centroids by randomly choosing *K* documents. However, we did choose to update Centroids incrementally, i.e., as each point is assigned to a cluster, rather than at the end of an assignment pass as is indicated in the K-means algorithm. The reason is that we noticed that incremental updates were more effective, i.e., produced results with better overall similarity and lower entropy.

For that we will use a Bisecting K-means algorithm as our primary clustering algorithm. This algorithm starts with a single cluster of all the documents and works in the following way:

1. Pick a cluster to split.
2. Find 2 sub-clusters using the basic K-means algorithm. (Bisecting step)
3. Repeat step 2, the bisecting step, for ITER times and take the split that produces the clustering with the highest overall similarity.
4. Repeat steps 1, 2 and 3 until the desired number of clusters is reached.



Finally, note that bisecting K-means has a time complexity which is linear in the number of documents. If the number of clusters is large and if refinement is not used, then Bisecting K-means is even more efficient than the regular K-means algorithm (In this case, there is no need to compare every point to every cluster centroid, since to bisect a cluster we just consider the points in the cluster and their distances to two centroids).

Kea-means Clustering

The Kea-means clustering algorithm is our proposed new clustering method that improves the K-means algorithm by combining it with the Kea keyphrase extraction algorithm. The Kea-means clustering tries to solve the main drawback of K-means that the number of total clusters is pre-specified in advance. In Kea-means algorithm, documents are clustered into several groups like K-means, but the number of clusters is determined automatically by the algorithm heuristically by using the extracted keyphrases.

The main idea of Kea-means is the following: Initially, the number of clusters *k* is set to 1. As in K-means, *k* clusters are formed by generating *k* centroids, and the similarities between centroids and documents are measured, and a document is assigned to the cluster that has the nearest centroid. Then, the Kea algorithm is applied to each document that is nearest to the corresponding centroid to extracted keyphrases. These keyphases are used to assign weights to other phrases. Now, the distance between the weighted documents and centroids are measured, and if the measured values do not reach to the threshold value, the value of *k* is increased by 1. This process is repeated until the measured distance exceeds the threshold value. At this point, the number of clusters *k* is determined, and the K-means algorithm is now used for actual clustering.

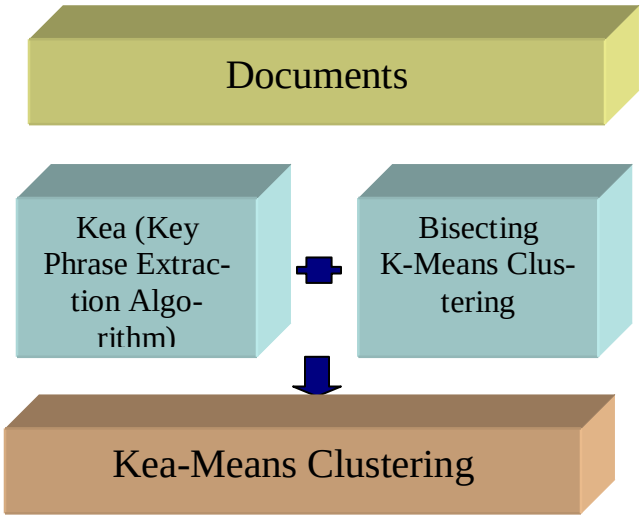
The main characteristics of the Kea-means clustering algorithm can be summarized by comparing with previous clustering algorithms in which the task of clustering is done without knowing the semantic nature of each cluster, the Kea-means clustering recognizes this semantic nature of clusters by using the keyphrases extracted from the documents in the cluster.

K-means must specify the number of clusters *k* in advance by the user, which results in the change of clustering results as the value of *k* changes.

Kea K-means clustering system architecture

Kea-means solves this problem by automatically determining this number. While the K-means algorithm uses the cosine measure or the Euclidean distance measure to calculate feature values, our Kea-means clustering algorithm uses these two measures simultaneously. Hence, the similarity of two documents is computed by the following expression:

$$Sin(d1; d2) = cosine(d1; d2) / Euclidean(d1; d2)$$

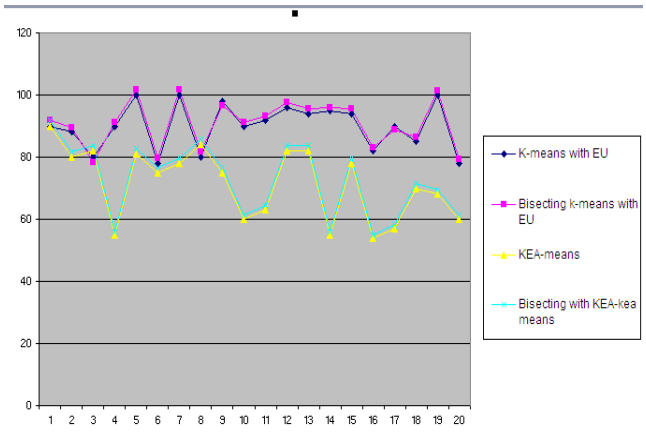


KEA-Bisection k-means clustering system architecture

Conclusion

The Primary objective of this thesis is to propose a new clustering algorithm based on the Kea Key phrase algorithm that we use here to extract several Key phrases from source Text documents by using machine learning techniques. The Kea bisecting K-means clustering algorithm gives easy and efficient way to extract text documents from large amount of Text documents.

The Kea Key phrase extraction algorithm is automatic extracting key phrase from text , Our results shows that kea can an average match between one and two of the given key phrase chosen . By this we can consider this to be good performance. The consistently good quality of the clustering that it produces, bisecting K-means is an excellent algorithm for clustering a large number of documents.



Experimental results for comparison between K-means and proposed Kea Bisecting k-means algorithm. In this experiment we use the kea's data set. Kea-means is also learn by these data set, in this data set, I collect 100 text documents and then applying my proposed (Kea bisecting K-means) algorithm to generate clusters.

References

[1] Ian H.Witten Gorden W.Paynter Carl Gutwin and Craig G, "KEA: Practical Autamatic Keyphrase Extraction," IEEE Magazine, August 2007.

[2] Michael Steinbach, George Karypis and Vipin Kumar "A Comparison of Document Clustering Techniques". 2008

[3] Juhyun Han, Taehwan Kim and Joongmin Choi " Web Document Clustering By Using Automatic Keyphrase Extraction'. IEEE Magazine 2007

[4] Frank, E., Paynter, G.W., Witten, I.H., Gutwin, C. and Nevill-Manning, C.G. (1999) "Domain-specific keyphrase extraction." Submitted to IJCAI.

[5] C. Clifton, R. Cooley, J. Rennie. TopCat: Data mining for topic identification in a text corpus. IEEE Trans. Knowledge and Data Engineering, 16(8):949- 964, August 2004.

[6] A. Likas, N. Vlassis, J. Verbeek. The global k-means clustering algorithm. Pattern Recognition, 36(2):451- 461, February 2003.

[7] I. Witten, G. Paynter, E. Frank, C. Gutwin, C. Nevill- Manning. Kea: Practical automatic keyphrase extraction. Proc. 4th ^{AC}M Conference on Digital Libraries, 254-255, August 1999.