Research Paper Engineering

Key Phrase Extraction Algorithm

Akash Arora	M Tech Computer Science Bhagwant University, Ajmer
Er. Abhishek Choudhary	M Tech Computer Science Bhagwant University, Ajmer

KEYWORDS

Introduction

Many journals ask their authors to provide a list of ***keywords*** for their articles. We call these ***key phrases***, rather than keywords, because they are often phrases of two or more words, rather than single words. We define a ***key phrase list*** as a short list of phrases (typically five to fifteen noun phrases) that capture the main topics discussed in a given document. This paper is concerned with the automatic extraction of Key phrases from text.

Key phrases are meant to serve multiple goals. For example, (1) when they are printed on the first page of a journal article, the goal is summarization. They enable the reader to quickly determine whether the given article is in the reader's fields of interest. (2) When they are printed in the cumulative index for a journal, the goal is indexing. They enable the reader to quickly find a relevant article when the reader has a specific need. (3) When a search engine form has a field labeled ***keywords***, the goal is to enable the reader to make the search more precise. A search for documents that match a given query term in the ***keyword*** field will yield a smaller, higher quality list of hits than a search for the same term in the full text of the documents. Key phrases can serve these diverse goals and others, because the goals share the requirement for a short list of phrases that captures the main topic of the document.

We define ***automatic key phrase extraction*** as the automatic selection of important, topical phrases from within the body of a document. Automatic key phrase extraction is a special case of the more general task of ***automatic key phrase generation***, in which the generated phrases do not necessarily appear in the body of the given document. In our document collections, an average of about 75% of the author's key phrases appear somewhere in the body of the corresponding document. Thus, an ideal key phrase extraction algorithm could (in principle) generate phrases that match up to 75% of the author's key phrases.

The most closely related work involves the problem of automatic index generation (Fagan, 1987; Salton, 1988; Ginsberg, 1993; Nakagawa, 1997; Leung and Kan, 1997). One difference between key phrase extraction and index generation is that, although Key phrases may be used in an index, Key phrases have other applications, beyond indexing. Another difference between a key phrase list and an index is length. Because a key phrase list is relatively short, it must contain only the most important, topical phrases for a given document. Because an index is relatively long, it can contain many less important, less topical phrases. Also, a key phrase list can be read and judged in seconds, but an index might never be read in its entirety. Automatic key phrase extraction is thus a more demanding task than automatic index generation.

Key phrase Extraction Algorithm

Text mining is powerful tool to find useful and needed information from huge data set. For context based text mining, Key phrases are used. Key phrases provide brief summary about the contents of documents. In document clustering, number of total cluster is not known in advance. In K-means, if pre specified number of clusters modified, the precision of each result is also modified. Therefore Kea ,is algorithm for automatically extracting Key phrases from text is used. In this kea algorithm, number of clusters is automatically determined by using extracted Key phrases. Kea means clustering algorithm provide easy and efficient way to extract test document from large quantity of resources. Key phrase play important role in text indexing, summarization and categorization. Key phrases are selected manually. Assigning Key phrases manually is tedious process that requires knowledge of subject. Therefore automatic extraction techniques are most useful.

M. Arshadet al.[1]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 deter-mined 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[2] key phrase extraction algorithm which returns several key phrases from the source documents by using some machine learning techniques. In this 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.

The Primary objective of this paper 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.

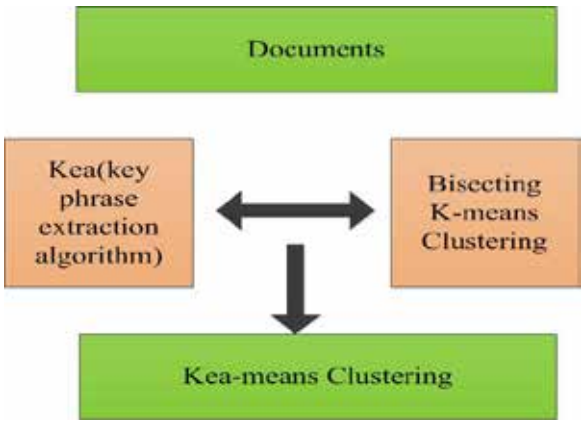

Figure 3.1 KEA-Bisection k-means clustering system architecture

Anoop Kumar Jainet.al[3] The objective of paper is to make search engine results easy to make document cluster. Document clustering algorithms attempt to group similar documents together. In this paper the proposed method is a phrase based clustering scheme which based on application of Suffix Tree Document Clustering (STDC) model.

Document clustering is one of the difficult and recent research fields in the search engine research. Most of the existing documents clustering techniques use a group of keywords from each document to cluster the documents. Document clustering arises from information retrieval domains, and "It finds grouping for a set of documents belonging to the same cluster are similar and documents belongs to the different cluster are dissimilar". The information retrieval plays an important role in data mining for extracting the relevant information for related to user request. Information retrieval finds the file contents and identifies their similarity. It measures the performance of the documents by using the precision and recall. In this paper we proposed a phrase based clustering scheme which based on application of Suffix Tree Document Clustering (STDC) model. The proposed algorithm is designed to use the STDC model for accurate equivalent representation of document and similarity measurement of the similar documents. This method of clustering reduces the grouping time and similarity accuracy as compared to other existing methods.

They proposed Suffix Tree Document Clustering (STDC) mechanism have four steps, these are as follows:

I. Document collection as input

II. Document Cleaning

III. Phrase Identification of Clusters

IV. Phrase Merging Clusters

Document clustering has initially been investigated in Information Retrieval mainly as a means of improving the performance of search engines by pre-clustering the entire corpus. The cluster hypo paper stated that similar documents will tend to be relevant to the same queries, thus the automatic detection of clusters of similar documents can improve recall by effectively broadening a search request. In our paper we are proposed a phrase based clustering scheme which based on application of Suffix Tree Document Clustering (STDC) model. Our proposed Suffix Tree Document Clustering (STDC) mechanism have four steps, these are Document collection as input, Document Cleaning, Phrase Identification of Clusters and Phrase Merging Clusters. We conclude that our proposed STDC technique is better perform based on document clustering than k-means clustering.

Khaled Hammoudaet. al[4], used the core phrase key phrase extraction algorithm. The core phrase key phrase extraction method is just the opposite of the traditional keyword based clustering. This method gives more accurate representation of clusters. The algorithm first constructs a list of candidate key phrases for every cluster. It scores each candidate key phrases on the basis of its features and then ranks the candidate key phrases by score. Finally it will select the top ranking key phrases for output.

In the survey of clustering data mining techniques, Pavel Berkhin [5]used the hierarchical clustering algorithm and linkage metric method. The cluster system is initialized as a set of singleton clusters in the hierarchical clustering. Then merges or splits the appropriate clusters iteratively until the stopping criterion is achieved. The similarity or dissimilarity of the cluster elements defines the appropriateness of a cluster to merge and split. From this we can understand that cluster can have similar points. To merge or split subsets of points rather than individual points, the distance between individual points has to be generalized to the distance between subsets. Such a derived proximity measure is called a linkage metric.

S. Zhu et al.[6] cluster Medline documents the semantic information of mesh thesaurus is applied by mapping documents into mesh concept vectors. To check the semantic similarity two steps are done. First, similarity between two MeSH main headings. Second, checks the similarity between two Mesh indexing sets. After the semantic similarity check, it is integrated with the content similarity and then spectral clustering is applied.

References

[1] M. Arshad(2012), IMPLEMENTATION OF KEA-KEYPHRASE EXTRAC-TION AL GORITHM BY USING BISECTING K-MEANS CLUSTERING TECHNIQUE FOR LARGE AND DYNAMIC DATA SET.
[2] Ian H. Witten Gorden W. Paynter Carl Gutwin and Craig G, "KEA: Practical Autamatic Keyphrase Extraction," IEEE Magazine, August 2007.
[3] Anoop Kumar Jain and Satyam Maheshwari(2013) "Phrase based Clustering Scheme of Suffix Tree Document Clustering Model",
[4] Khaled Hammouda and Mohamed Kamel, "Collaborative Document Cluster ing," 2007
[5] Pavel Berkhin, "A Survey of Clustering Data Mining Techniques,"1990
[6] S. Zhu, J. Zeng, and H. Mamitsuka, "Enhancing MEDLINE document cluster ing by incorporating mesh semantic similarity," ***Bioinformatics***, vol. 25, no. 15, pp. 1944–1951, Aug. 2009.