Singaporean Journal Scientific Research (SJSR)
ISSN: 2231 - 0061Vol.4, No.1pp. 246 - 249

available at: : http://www.iaaet.org/sjsr

Kea-Mean Clustering Approach for Text Mining

[1]**Shobha Sanjay Raskar and** [2]**D.M. Thakore**
[1,2]Bharati Vidyapeeth University , BVUCOE, Department of computer Science and Engineering, Dhankawadi , Pune, India
[1]shobha.raskar@gmail.com

ABSTRACT

"Text mining"refers to the application of data mining techniques to automated discovery of valuable or interesting information from unstructured text. Increasing numbers of textual data has led to the task of mining useful or interesting frequent itemsets from very large text database. Text mining refers to the application of data mining techniques to automated discovery of valuable or interesting information from unstructured text.A Keyphrase extraction method relies on word frequency and position instead of document inherent semantic information. The machine learning approach first builds a prediction model using training documents with known keyphrases, and then uses the model to find keyphrases in new documents. KEA used for training and testing purpose. Keyphrases can help users get afeel for the content of a collection; provide sensible entry points into it.keyphrases are useful, not many documents have keyphrasesassigned to them, and manually assigning keyphrases to existingdocuments is costly. Therefore, there is a need for automatickeyphrase extraction.Keyphrase extraction is method of automatically extracting important phrases from a text. Kea mean clustering algorithm improves K mean algorithm by combining it with the kea keyphrase extraction algorithm.This provides efficient way to extract test documents from massive quantity of resources. This algorithm develop faster algorithm for clustering.

Keywords: keyphrases, Training and Testing, Text mining, Feature calculation

1. INTRODUCTION

The access to a large quantity of textual documents turns out to be effectual because of the growth of the digital libraries, web, technical documentation, medical data and more. These textual data comprise of resources which can be utilized in a better way. Text mining is major research field due to the need of acquiring knowledge from the large number of available text documents, particularly on the web.[1]. Both text mining and data mining are part of information mining and identical in some perspective. Text mining can be described as a knowledge intensive process in which a user communicates with a collection of documents. In order to mine large document collections, it is require pre-processing the text documents and saving the data in the data structure, which is suitable for processing it further than a plain text file. Information Extractionis defined as the mapping of natural language texts like text database, WWW pages, electronic mail etc. into predefined structured representation, or templates which, when filled, represent an extract of key information from the original text[2].

A.Text mining process

Text mining is multidisciplinary field, which includes text analysis, information retrieval, clustering, information extraction, categorization, visualization, machine learning, data mining and database technologyTextmining process includes,

- **Text preprocessing**: Syntactic/semantic text analysis i.e. finding the corresponding position for each word.
- **Feature Generation**: Text document is represented by words or features it contain and their occurrences. Two main approach of document representation, bag of words and vector space.
- **Feature selection**: Simple counting
- **Text/data mining**: classification and clustering

Text preprocessing includes tokenization, word stemming and the application of a stop words removal technique. Many of the most frequently used words in English are worthless in IR and in text miningthese words are called stop words.Tokenization is referred as the procedure of splitting the text into words or terms.

i. Stemming: There are a number of implicit options available to both schemes with regards to stemming-

a. Lovins Stemmer;

b. Porter Stemmer

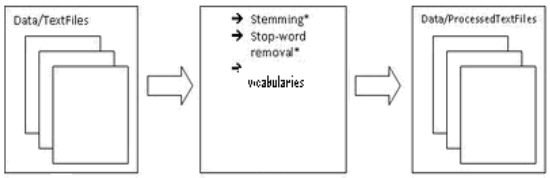

Figure 1 : Text mining Flow

c. Partial Porter Stemmer: Porter Stemmer has 5 multi-part stages which allow the "miner" to be more or be less conservative in their stemming process. The first stage Porter stemmer is a popular methodology, handling basic plurals e.g. horses becomes horse, processes becomes process, men does not become man; and

ii. Stop words:This is process of Removal of stopwords.

ii. Vocabularies: The "miner" is able to dicate a dictionary, thesaurus or list of terms when undertaking controlled indexing. Also, in terms of implementation, these can be in either text form or resourse description format ("rdf").

2. CLASSIFICATION

Classificationtries to put specific documents into groups known in advance. The same basic means can be used as in clustering. Statistical classification includes precision and recall. Table gives difference between precision and recall.

Precision	Recall
1) It is measure of extractness.	1) It is measure of completeness.
2) Precision is equal to number of relevant document retrieved by search divided by total number of documents retrieved by that search	2) Recall is equal to number of relevant document retrieved by search divided by total number of documents.
3) It measure how precise search is .	3) It measure how complete search is .
4) Higher precision means less unwanted documents.	4) Higher recall means less missing documents.

Table 1 : .Comparison of Precision and Recall

Following equation gives relation between Recall and Precision

F measure = 2PR/(P+R) ----------equation2.1

where P: Precision

R= Recall

Consider following data related with documents,

Table 2

	Relevant	Not Relevant
Retrieved	a	b
Not Retrieved	c	d

P= (number of relevant document retrieved)/

(Total number of document retrieved)

P= a/(a+b)----------equation 2.2

R= (no. of relevant document retrieved)/ (no. of all relevant document in Database)

R= a/(a+c)----------equation 2.3

Text clustering is defined as efficient way of sorting several documents, summarize and arrange text documents.

3. ABOUT KEYPHRASE

A list of keyphrases associated with a document may serve asindicative summary or document metadata, which helpsreaders in searching relevant information.Automatic keyphrase extraction is the identification of themost important phrases within the body of a document bycomputers rather than human beings. It normally involves theuse of statistical information. Keyphrase. When authorsassign keyphrases without a controlled vocabulary list,typically 70-90% of their keyphrases appear somewhere intheir documents. Keyphrases are similar to keywords, except that the document is summarized by a set of phrasesrather than words.Keyphrase extraction is a classification task: a documentcan be seen as a set of phrases, and a keyphrase extractionalgorithm should correctly classify a phrase as a keyphrase ora non-keyphrase. Machine learning techniques can be use for this task.If they are provided with a set of training datacomposed of both keyphrase examples and non-keyphraseexamples. The data are used to train the algorithm todistinguish keyphrases from non-keyphrases.

Keyphrases assigned particular weight and frequency is calculated in document i.e. how many times that keyphrase appear in document. Document keyphrases have been successfully used in the IR and NLP task: document indexing, document classification, document clustering and document summarization.

Keyphrases are manually assigned by authors, but most of articles do not keyphrases, at that time it is beneficial to extract keyphrases automatically.keyphrases give syntactical information about the documents. For example two documents about same topic “Health” can share a few common phrases e.g. “Food”, “Diet” and they can provide additional knowledge for each other to better evaluate

and extract keyphrases from each other. User would better understand a topic expressed in a document if the user reads more documents about the same topics.

There are fundamentally two different approaches for automatic extracting keyphrases:

i) **Keyphrase assignment**: It seeks to select the phrases from a controlled vocabulary that best describe a document. The training data associates a set of documents with each phrase in the vocabulary, and builds a classifier for each phrase. A new document is processed by each classifier, and assigned the keyphrase of any model that classifies it positively [4]. The only keyphrases that can be assigned are ones that have already been seen in the training data.

ii) **Keyphrase extraction**: This approach does not use a controlled vocabulary, but instead chooses keyphrases from the text itself. It employs lexical and information retrieval techniques to extract phrases from the document text that are likely to characterize it [3]. Keyphrase extraction consists of two steps: candidate phrase identification and keyphrase selection. In this approach, the training data is used to tune the parameters of the extraction algorithm.Both use machine learning methods, and need for training purposes a set of documents with keyphrases already attached. Keyword extraction classified as supervised or unsupervised.. Unsupervised methods usually assign candidate phrases considering various features. Phrase end with adjective are not allowed, and only phrase end with nouns are collected as the candidate phrase of the documents.

4. TRAINING AND TESTING

To extract keyphrases from documents first model is build which can be used for the purpose of extraction, and for this purpose system has to train from some known facts using supervised learning.Learning is an inherent characteristic of the human beings. By virtue of this,people, while executing similar tasks, acquire the ability to improve their performance.When this learning is done by a machine, it is usually referred to as 'machine learning'.Machine learning can be broadly classified into three categories:Supervised Learning: Supervised learning requires a trainer, who supplies the input-outputtraining instances. The learning system adapts its parameters by some algorithms togenerate the desired output patterns from a given input pattern.For the training dataset, tomark the candidate as an author assigned keyphrase or not an author assigned keyphrase,which is used for the learning process. The classifier learns two sets of numeric weightsfrom the discretized feature values.

There are two phases to Kea: learning a model of appropriate keyphrases, and use of the model to extract keyphrases from documents. To learn a model, Kea requires a set of training documents, for which there is a set of keyphrases (these might be provided by authors, or created by hand). Two attributes of phrases are used in building a model: the distance into a document where a phrase first occurs; and the TFxIDF value of the phrase (a measure of how frequently it occurs within a given document compared to other documents).

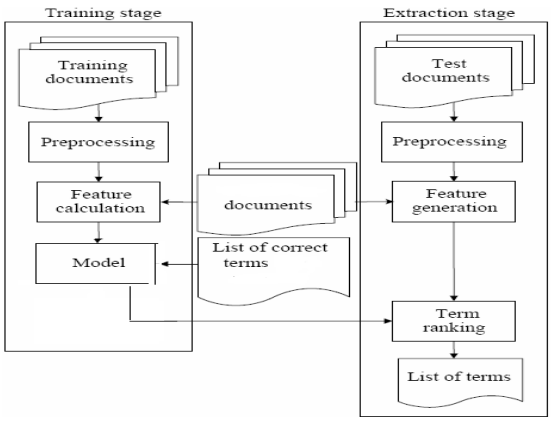

Figure 2 : Training and Extraction Stage

The above diagram illustrates a summary of the information retrieval and document preprocessing implementation of text mining. This kea-mean algorithm include two phases **Training and extraction**.

Training is a process for making the machine, learn something from the environment by experience.After the training of the system is over, trained model is ready to extract keyphrases from any test document.Select keyphrases from a new document, Kea determines candidate phrases and featurevalues, and then applies the model built during training. The model determines the overallprobability that each candidate is a keyphrase, and then a post-processing operation selects thebest set of keyphrases. For this the test document is passed through all thefirst three components to generate the feature vectors of each candidate.

Dataset is a text file with a number of tags identifying the individual parts of the each article. However, this structure is not separated by fixed part lengths - thus the need to apply logic to separate each article into individual text documents.

5. ABOUT K-MEAN AND FEATURE CALCULATION (TF-IDF)

It is partition based clustering algorithm. Consider algorithm select 40 documents as the initial cetroids and then iteratively assigns all documents to the closest cluster , and again compute the centroid of each

cluster, till the centroids do not change. The similarity between documents calculated by using cosine measure and Euclidean distance.

Given text Di= (ti1,Wi1; ti2,Wi2,……..)

Where tij= feature term

Wij= weight of tij in text

5.1 Feature calculation

Two features are calculated for each candidate phrase and used in training and extraction. Theyare: *TF -IDF*, a measure of a phrase's frequency in a document compared to its rarity in general use; and *first occurrence*, which is the distance into the document of the phrase's firstappearance.*TF-IDF*

This feature compares the frequency of a phrase's use in a particular document with thefrequency of that phrase in general use. General usage is represented by *document frequency*— the number of documents containing the phrase in some large corpusWeight act calculated according to TF-IDF formula, TF is the frequency of a term in the document. The more often a term occurs in the document, the more likely it is to be important for that document. The *standard TF* of a term *T* in a document *D* is calculated by,

Standard TF = no. of occurrences of *T* in *D*

…eqn 5.1

IDF is the rarity of a term across the collection. A term that occurs in only a few documents is often more valuable than a term that occurs in many documents. The *standard IDF* of aterm *T* is given by:

$$\text{Standard IDF}=\log\frac{\text{no. of documents in collection}}{\text{no. of documents occurs in}}$$

…..Eqn 5.2

TF×IDF is a common way of combining TF and IDF.The *distance* attribute is the position where a term first appears in the document. A term that occurs at the beginning of the document is often more valuable than a term that occurs at the end of that document. The *distance* of a term *T* in a document *D* is given by:

$$\text{Distance}=\frac{\text{no. of words before first appearance of T}}{\text{no. of words in D}}$$

…….Eqn 5.3

According to TF-IDF formula , calculate the weight of each key word in the corresponding text. TF X IDF measure of phrases frequently in document. TF is frequency of term ti in document dj.TF x IDF can select terms which occur frequently in parts of document but appear rarely in corpus. This feature compare frequency of phrases use in particular document with frequency of that phrase in general use. DF means document frequency is number of document containing the phrase in some large corpus.

6. CONCLUSION

Keyphrase and K-mean clustering algorithm is important for obtaining the appropriate cluster context and the low quality clustering results will decrease extraction performance.Traditional 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-means solves this problem by automatically determining this number.

Kea mean clustering algorithm improves K mean algorithm by combining it with the kea keyphrase extraction algorithm. This provides efficient way to extract test documents from massive quantity of resources. This algorithm develop faster algorithm for clustering.

REFERENCES

[1] Ah-hwee Tan, "Text mining : The state of the art and the challenges", In proceedings of the PAKDD Workshop on knowledge Discovery from Advanced Databases, pp. 65-70, 1999.

[2] Wilks Yoricks, "Information Extraction as a core technology ". International Summer Scholl , SCIE-97,1997

3] P. D. Turney, Learning algorithm for keyphrase extraction. Journal of Information Retrieval, 2(4), 303-36 2000.

4] S. T. Dumais, J. Platt, D. Heckerman and M. Sahami. "Inductive learningalgorithms and representations for text categorization." Proceedings of ACMCIKInternational Conference on Information and Knowledge Management, pp 148-155,1998

5] A New Approach to Keyphrase Extraction Using Neural Networks Kamal Sarkar, Mita Nasipuri and Suranjan GhoseComputer Science and Engineering Department, Jadavpur University,Kolkata-700 032, India IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 2, No 3, March 2010