ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 11, November 2014

Keyphrase Extraction using supervise learning

Shobha S. Raskar[1], Salma H. Pathan [2]

Modern Education Society COE, MESCOE, Department of Computer Science and Engineering,

Wadia College Campus, Pune India[1,2]

Abstract: Text mining is knowledge intensive process in which a user communicates with a collection of documents. Text categorization is a kind of "supervised" learning where the categories are known beforehand and determined in advance for each training document. Manually extraction of keywords is slow, expensive and tedious. Therefore automatic keyword extraction is necessary. Keyphrases help the users to get idea about the content of document. Kea-means clustering used for extracting test document from large quantity of text data. In Kea-means algorithm, documents are clustered into several groups like K-means, but the number of clusters is determined automatically by using the extracted keyphrases. Set of training documents and machine learning is used to determine phrases are keyphrase or not.

Keywords: Text mining, Keyphrases, clustering, supervise learning.

I. INTRODUCTION

Number of digital document is grouping exponentially, caused by ever-grouping use of computer. The need for efficient search, indexing, categorization over those documents becomes ever more important. Text categorization is a kind of "supervised" learning where the categories are known beforehand and determined in advance for each training document. Keyphrases describe the content of document, thus enabling the user to decide whether a particular document is relevant for user's information need. A keyphrase is defined as meaningful and significant expression consisting of one or more words in documents [1].

Keyword extraction can be seen as supervised learning. In machine learning, first set of training document is provided to system, each of which has human chosen keywords. Then training model applied to find keyword from new test document. Text mining process includes, Text preprocessing, feature generation, Feature selection classification and clustering.

A. Clustering :

Clustering method can be used in order to find groups of documents with similar content. Each cluster consists of a number of documents. The quality of clustering is considered better if the contents of the documents within one cluster are more similar and between the clusters more dissimilar. K-mean algorithm use for clustering.

II SUPERVISE LEARNING

Machine Learning (ML) is an area of artificial intelligence concerned with the development of techniques which allow computers to "learn" by the analysis of data sets. The focus of most machine learning methods is on symbolic data.

Kea mean algorithm has training and extraction phase. The training data is used to form a model that takes these features and predicts whether or not a candidate phrase will actually appear as a key-phrase, this information is known for the training documents. Then the model is applied to extract likely key-phrases from new documents. The most well-known and widely used learning algorithm to estimate the values of the weights is the Back Propagation (BP) algorithm. Generally, BP algorithm includes the following six steps:

1. Present a training sample to the neural network.
2. Compare the network's output to the desired output from that sample. Calculate the error in each output neuron.
3. For each neuron, calculate what the output should have been, and a scaling factor, how much lower or higher the output must be adjusted to match the desired output. This is the local error. Adjust the weights of each neuron to lower the local error.
4. Repeat the steps above on the neurons at the previous level.

III FEATURE CALCULATIONS

One of the simple weighting is TF-IDF. This weight is numerical statistic which reflects how important word is to document in collection of corpus. This value increases proportionally to number of times word appears in document Two features are calculated for each candidate phrase and used in training and extraction. They are: *TF-IDF*, a measure of a phrase's frequency in a document and *first occurrence*, which is the distance into the document of the phrase's first appearance. Algorithm have been proposed to classify candidate phrase into either keyphrasse or not most important feature for classifying candidate phrases are frequency and location of phrase in the document. TF-IDF (Term Frequency, Inverse Document Frequency) is a basic technique to compute the relevancy of a document with respect to a particular term." The relevancy of a document to a term can be calculated from the percentage of that term shows up in the document. The count of the term in that document divide by the total number of terms in it, which called as the "***term frequency***"

On the other hand, if this is a very common term which appears in many other documents, then its relevancy should be reduced. The count of documents having this term divided by total number of documents, which called as the "***document frequency".*** The overall relevancy of a document with respect to a term can be computed using both the term frequency and document frequency.

Relevancy =TF* log(1/DF) ……Equation 1

Keywords are a fundamental part of information retrieval (IR) and as such they have been studied extensively. They are used for everything from, searching to describing a document. The most common selection/weighting schemes are based on collection statistics or using supervised machine learning algorithms. To extract the terms from a document, the following process is common,\

- Extract words by tokenize the input streams
- Make the words case-insensitive (e.g. transform to all lower case)
- Filter out stop words
- Stemming (e.g. transform cat, cats, kittens to cat)
- the word count of per word/doc combination
- the total number of words per doc
- the total number of docs per word. And finally compute the TF-IDF.

IV .KEYWORD EXTRACTIONS

Task of automatic keyword extraction is to identify a set of words representive for document. Keyword extraction identify small set of words, keyphrases , keywords which describe meaning of document. Keyword search enables efficient scanning of large document collections. Text categorization techniques can be applied to assign appropriate key-phrases to new documents. The training documents provide a predefined set of key-phrases from which all key-phrases for new documents are chosen. For each key-phrase the training data defines a set of documents that are associated with it. For each key-phrase, standard machine learning techniques are used to create a "classifier" from the training documents, using those associated with it as positive examples and the remainder as negative examples. Given a new document, it is processed by each key-phrase's classifier.

V. EXPERIMENT AND RESULT

To extract keyphrases from documents model has to build, which can be used for the purpose of extraction, and train system from some known facts using supervised learning. Hence the implementation of algorithm consists of two steps, Training and extraction. Training is a process for making the machine learn something from the environment by experience. The learning system adapts its parameters by some algorithms to generate the desired output patterns from a given input pattern.

A. Experiment : Keyword search

System is implemented using java. Input for the system is text files. Text files collected from UCI Dataset. After training and testing stage keyphrases extracted from the text documents. These training documents are text documents and must have the ending " .txt". Their key phrase documents are also text documents with the same name but with the ending .key".A model is created for identifying keyphrases, using training documents where the author's keyphrases are known. Keyphrases are choosen from a new document, using the model created during the training process. System is tested using five text files. During testing phase keyphrase are extracted from files.

No.	Text file	Keyphrases
1.	bostid_b12sae.txt	Salinity Salt tolerance Halophytes Brining Tolerance Saline water Western Australia Atriplex Leaf protein Soil salinity
2.	copyiirr_ii02re.txt	Agroforestry Biophysics Trees Forestry Intercropping Natural resources management Terrace cropping Erosion Philippines Farm planning
3.	foodfirst_ff08ne.txt	Bangladesh Landowners Food aid Famine Jute Merchants Landlessness Family planning Cooperative farming First aid
4.	gtz_g24ine.txt	Animal power Milling Mills Flours Draught animals Wheels Figs Tillage Energy sources Friction
5.	iirr_ii02re.txt	Agroforestry Biophysics Trees Forestry Intercropping Natural resources management Terrace cropping Erosion Philippines Farm planning

Table 1.1 : Result of keyphrases

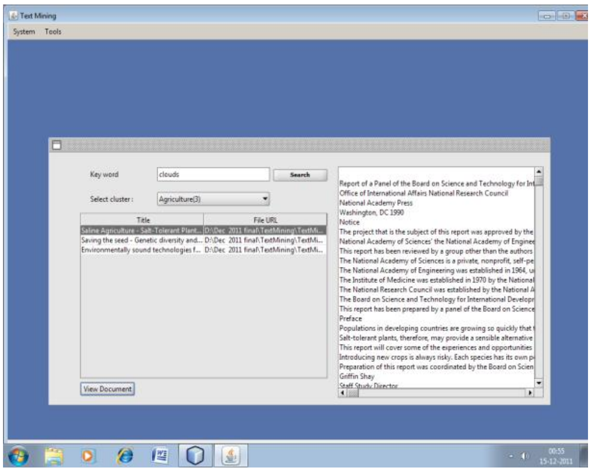
ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 11, November 2014

The result of extraction of keyphrases depends very much on the domain of the training set of documents, because these are the training instances that make the system learn.

Text files are input for system. Using training phase system is learned and using testing phase kephrases are extracted from input test files. Enter keyword which is to be searched, and then clusters are formed. After selecting cluster, list of documents are shown, which are under that cluster as shown in figure 1.1.

Figure 1.1 GUI for cluster formation



Title	File URL
Saline Agriculture - Salt-Tolerant Plant...	D:\Dec 2011 final\TextMining\TextMi...
Saving the seed - Genetic diversity and...	D:\Dec 2011 final\TextMining\TextMi...
Environmentally sound technologies f...	D:\Dec 2011 final\TextMining\TextMi...

Report of a Panel of the Board on Science and Technology for Int
Office of International Affairs National Research Council
National Academy Press
Washington, DC 1990
Notice
The project that is the subject of this report was approved by the
National Academy of Sciences' the National Academy of Enginee
This report has been reviewed by a group other than the authors
The National Academy of Sciences is a private, nonprofit, self-pe
The National Academy of Engineering was established in 1964, u
The Institute of Medicine was established in 1970 by the National
The National Research Council was established by the National A
The Board on Science and Technology for International Develop
This report has been prepared by a panel of the Board on Science
Preface
Populations in developing countries are growing so quickly that
Salt-tolerant plants, therefore, may provide a sensible alternative
This report will cover some of the experiences and opportunities
Introducing new crops is always risky. Each species has its own p
Preparation of this report was coordinated by the Board on Scien
Griffin Shay
Staff Study Director

Table 1.2 : Result of cluster formation

	Keyword	Clusters formed	No. Of documents per clusters
1.	Clouds	Agriculture Developing Countries Technologies Energy Environment Information Kit Plant Small	3 3 3 2 2 2 3 2
2.	Plants	Agriculture Developing Countries Technologies Energy Environment Information Kit Vegetable oil Small	3 3 3 2 3 2 2 2
3.	banana	Agriculture Developing Countries Technologies Energy Environment Information Kit Vegetable oil Women	3 3 3 2 2 3 2 2

CONCLUSION

Keyphrase and K-mean clustering algorithm is important for obtaining the appropriate cluster context and low quality clustering results will decrease extraction performance. Kea mean algorithm provides efficient way to extract test documents from large quantity of resources. Kea keyphrase algorithm that extracts several keyphrases from source documents by using some machine learning techniques. Supervised learning requires a trainer, who supplies the input-output training instances. The learning system adapts its parameters by some algorithms to generate the desired output patterns from a given input pattern. In supervise learning, set of training document is provided to system, each of which has human chosen keywords, which increases accuracy.

REFERENCES

1. Xiaojun Wan and Jianguo Xiao, "CollabRank: Towards a Collaborative Approach to Single document Keyphrase Extraction", Proceeding of 22nd Intenational Conference on Computational Linguistics , Aug 2008 , PP 969-976.
2. Xiaojun Wan and Jianguo Xiao, "CollabRank: Towards a Collaborative Approach to Single document Keyphrase Extraction", Proceeding of 22nd International Conference on Computational Linguistics , Aug 2008 , PP 969-976.
3. Kamal Sarkar, Mita N, Suranjan G, " A New Approach to Keyphrase Extraction Using Neural Networks", IJCSI, Vol 7, Issue 2. No.1 , March 2010, ISSN 1694-0784.
4. D. Sanchez, M. J. M, I. Balanco, " Text Knowledge Mining : An Alternative to Text Data Mining ", 2008 IEEE International Conference on Data Mining Workshops.
5. M. F. Eltibi, W. Asho ,"Initializing K-Means Clustering Algorithm using Statistical Information", International Journal of Computer Applications (0975 – 8887) Volume 29– No.7, September 2011