ISSN : 2278 – 1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 1, Issue 5, July 2012

Keyword Extraction through Applying Rules of Association and Threshold Values

Mr. J.Naveenkumar

BharatiVidyapeeth Deemed University, College Of Engineering, Pune

pro_naveen@hotmail.com

ABSTRACT— ***Keywords are a set of significant words in an article that gives high-level description of its contents to readers. They provide a concise and precise high-level summarization of a document. Keywords are useful tools as they give the shortest summary of the document. A keyword is identified by finding the relevance of the word with or without prior vocabulary of the document or the web page. Generating the word statistics as well associating the words and categorizing it is the main focus which is pinned down in this paper.***

***Keywords*— Keywords, Keyword extraction, threshold, association rules.**

I. Introduction

Keywords are a set of significant words in an article that gives high-level description of its contents to readers. Identifying keywords from a large amount of on-line news data is very useful in that it can produce a short summary of news articles. As on-line text documents rapidly increase in size with the growth of WWW, keyword extraction has become a basis of several text mining applications such as search engine, text categorization, summarization, and topic detection. Manual keyword extraction is an extremely difficult and time consuming task; in fact, it is almost impossible to extract keywords manually in case of news articles published in a single day due to their volume. For a rapid use of keywords, we need to establish an automated process that extracts keywords from news articles.

II. Related work

Extracting keywords from a text is closely related to ranking words in the text by their relevance for the text. To first approximation, the best keywords are the most relevant words in the text. Determining the right weight structure for words in a text is a central area of research since the late 1960's ([1]). In 1972 Spark Jones (reprinted as [2]) proposed a weighting for specificity of a term based on 1 + log (#documents=#term occurrences). This term weighting, which has become known as TF.IDF, is subsequently refined in [3], studied in the light of latent semantic analysis by [4], given a detailed statistical analysis by [5], and a probabilistic interpretation by [6]. An information theoretic explanation of TF.IDF is given. [7]

III. Keys for extraction

In order to decide whether to label a word as a key, the words in the document must be distinguished by using features and the properties of keywords have to be identified. The first possible feature that comes into mind is the frequency, which is the number of times a keyword appears in the text. It is obvious that the more important phrases will be more used in a text. Usually prepositions such as "the, that, this, etc." appear much more than any other words (even those which are keys) even though they have no value as a keyword. Therefore, we have to extend the concept of a keyword. If a word appears much more frequently in a document than with respect to other documents, this can be another distinguishing feature of that word on deciding whether it is a keyword or not.

Combining these two properties, we obtain the metric TF x IDF (standing for Term Frequency x Inverse Document Frequency) score, which is the standard metric used in Information extraction [2], and for a word W in document D, is defined as.

TF x IDF (P,D) = P (word in D is W) x [- log P (W in a document)].

IV. APPROACH

A key word extraction is relevant technique for a number of text mining related tasks including document retrieval, webpage retrieval, document clustering and summarization. Keywords provide a concise and precise high-level summarization of a document. A keyword is identified by finding the relevance of the word with or without prior vocabulary of the document or the web page. Key word extractor plays important role in fetching the relevant information from repositories. The user enters the phrases or the word from which the keyword is made. This keyword is then searched in the repository and according to various parameters the page containing the information is fetched and given to the user.

The key word extractor is a logical entity which takes the input from the user, Based on the input the keyword is searched on various document collected by the web crawler in repository. This approach will be useful for ranking the pages based on the context as well keywords. A file from Repository is taken and the words or phrases entered by the user is taken and considered it as keyword. The keywords are then matched with the file and frequencies of the words are counted. There are various parameters which should be taken into account like word frequency, word weighting, and word relations.

For association rules, support(s) and confidence(c). The rule WiWj has support s in the collection of documents D if s percentage of documents in D contain WiWj . The support is calculated by the following formula:

$$\text{Support (WiWj)}= \frac{\text{Support count of WiWj}}{\text{Total number of documents D}}$$

The rule Wi x Wj holds in the collection of documents D with confidence c if among those documents that containWi, c % of them contain Wj also. The confidence is calculated by the following formula:

$$\text{Confidence (Wi/Wj)}= \frac{\text{Support(WiWj)}}{\text{SupportWi}}$$

Such sets are called the frequent keyword sets and use the identified frequent keyword sets to generate the rules that satisfy a user specified minimum confidence (called minimum configuration). The frequent keywords generation requires more effort and the rule generation is straightforward.

Every article written or published lays emphasis on certain words within it. Such words are distinguished from other words in the style of their usage i.e. in the form they are being used. They can either be in italics or bold or underlined or have a different font color as compared to other words in the paragraph. Such words are what we refer to as keywords.

Our approach for keyword extraction starts with reading a bag of text i.e. document. The document is pre-processed and the noisy data are removed. The structure of the document is analysed for overall extraction of the keywords. Then the structured document is fed to algorithm for extracting the keyword.

The document is read line by line in order to find out the words count. The words counts are the frequencies which are recorded line by line and the summation of counts of each word is used and the Max (Word count) of the word is the keyword.

Phase presents a way for finding information from a collection of indexed documents by automatically extracting association rules from them. Association rules have already been used in TM [7, 10, 11, 15, 16, 17, 18, 19]. Below we define and describe the association rules in the context of TM. Given a set of keywords { } A = w1, w2... wn and a collection of indexed documents D = {d1,d2,..., dm }, where each document IDIS a set of keywords such that di x A. Let Wi be a set of keywords. A document i d is said to contain Wi if and only if Wi x di . An association rule is an implication of the form Wi ⇒Wj where WiA ,WjA and Wi ∩Wj =φ .There are two important basic measures.Frequent keywords generation requires more effort and the rule generation is straightforward.

Every article written or published lays emphasis on certain words within it. Such words are distinguished from other words in the style of their usage i.e. in the form they are being used. They can either be in italics or bold or underlined or have a different font colour as compared to other words in the paragraph. Such words are what we refer to as keywords.

The task of keyword extractor begins with the scanning of file which in an article can be a paragraph. First, the document text is split into an array of words by the specified word delimiters. This array is then again broken into sequence of continuous words at phrases end and stop words. Stop words are words such as the, them, their, etc. When the extractor counters stop words, they areimmediately removed from the sequence of further processing. Fig 1 shows the process of keyword extraction through the self process which is explained as below.

The remaining sequence of words is further processed. The first task is to match every individual word in the sequence with all the remaining words in every remaining paragraph in the article. If a match is found in the comparison, the extractor ought to increase the frequency of the respective word in the frequency graph of the phrasal words created by the extractor in the memory. This frequency is termed as the Keyword Score.

The additional task done by the extractor is to reflect the time consumption done during the extraction. The last but one step in the process is to refer to the corpus, i.e. the data-dictionary in order to determine the meaning of the extracted word. The final step performed by the keyword extractor is to display the context-based word extracted along with their respective Keyword Scores.

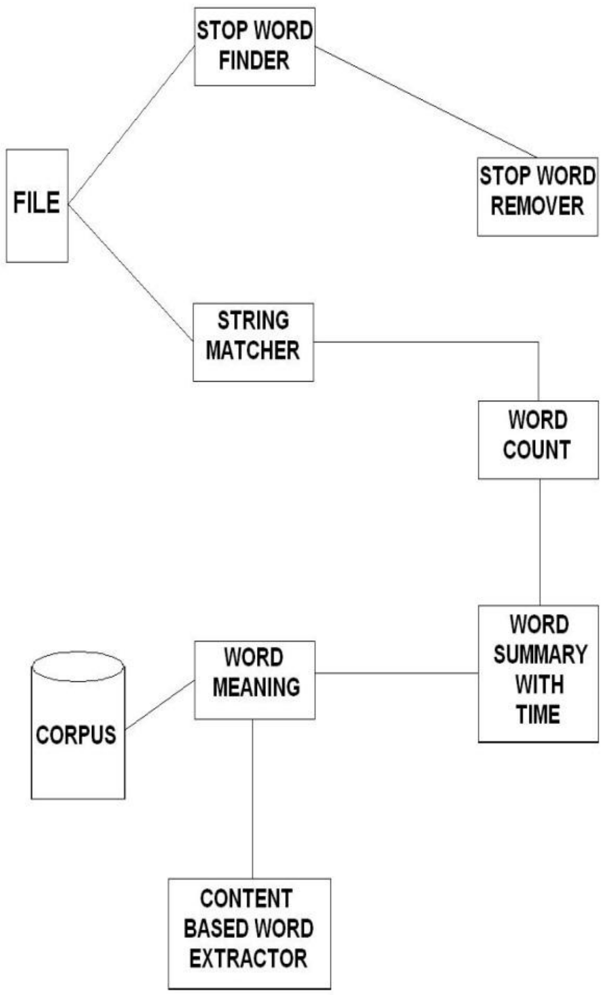

Fig1: Keyword Extraction Self Process

V. CONCLUSION

Keyword Extraction is necessary for many purposes. They are used in many areas varying from search engines to text categorization. There are proposed methods for automatic keyword extraction from documents. Our method uses self-approach, which uses the frequency count of the word and the distance of the word to the beginning of the text, paragraph and the sentence to identify keywords in the text.

REFERENCES

[1] Turney P. D., Learning Algorithms for Keyphrase Extraction, Information Retrieval, 1999.

[2] Frank E., Paynter G. W., Witten I. H., Gutwin C., and Nevill-Manning C. G.. Domain-specific. keyphrase extraction. In IJCAI, pages 668--673, 1999.

[3] Duda R. O., Hart P. E., Stork D. G., Pattern Classification, page 20, Wiley-Inter science, 2000.

[4] Salton G. and McGill M. J. Introduction to Modern Information Retrieval. McGraw-Hill, 1983.

[5] Bayes Thomas. An essay towards solving a problem in the doctrine of chances. Philosophical Transactions of the Royal Society (London). 53:370-418, 1763

[6] Domingos P. and Pazzani M.. On the optimality of the simple bayesian classifier under zero-oneLoss. Machine Learning, 29(2/3):103-130, 1997.

[7] http://www.uni-weimar.de/medien/webis/research/events/tir-10/proceedings/wartena10-keyword-extraction-using-word-co-occurrence.pdf.

[8] A Text Mining Technique Using Association Rules Extraction, HanyMahgoub, DietmarRösner, Nabil Ismail and FawzyTorkey, International Journal of Information and Mathematical Sciences 4:1 20

Biography

J.Naveenkumar – Completed my M.Tech Computer. Projects completed in the field of software quality assurance. I have a keen interest in the field of Information retrieval and indexing. Researching in the field of keyword extraction and context based web search.