International Journal of Soft Computing and Engineering (IJSCE)
ISSN: 2231-2307, Volume-2, Issue-3, July 2012

An Approach for Extracting the Keyword Using Frequency and Distance of the Word Calculations

Ashwini Madane, Devendra Thakore

Abstract— ***A significant word used in indexing or cataloguing is regarded as a Keyword. Keywords provide a concise and precise high-level summarization of a document. They therefore constitute an important feature for document retrieval, classification, topic search and other tasks even if full text search is available. 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. Extracting keywords manually is an extremely difficult and time consuming process, therefore it is almost impossible to extract keywords manually even for the articles published in a single conference. Therefore there is a need for automated process that extracts keywords from documents.

This paper concentrates on the extracting the keywords by understanding the linguistic, non-linguistic and various other approaches but applying the simple statistics approach.

Keywords: Keyword Extraction methods, Keyword Frequency Count, Stemming, Tokenization.

I. INTRODUCTION

The keyword is the word which makes the reader to understand what the document or the content is all about. The main criterion for a word to be a keyword is the relevance of that word with respect to that document. It should reveal the meaning or the idea about the document.

The keywords can be extracted using various algorithms as well as manually. The keyword extraction manually consists of many errors as well as the process of extraction is too lengthy and time consuming. So it is preferred that the keyword extraction is made automated.

The task of automatic keyword extraction is to identify a set of words, representative for a document.

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

Manuscript received on July, 2012.

Ms Ashiwini.Madane BE. MTech(pursing) Bharati Vidyappeth University, College of Engineering, Pune – 43.

Prof Devendra Thakore BE. MTec .PHd (Pursing) Bharati Vidyappeth University, College of Engineering, Pune – 43.

III LINGUISTIC APPROACH

Linguistic approaches use the linguistic features of the words, sentences and document. Methods which pay attention to linguistic features such as part-of-speech, syntactic structure and semantic qualities tend to add value, functioning sometimes as filters for bad keywords. During automatic keyword extraction from multiple party dialogue episodes, the advantages of using the lexical resources are compared to a pure statistical method and relative frequency ratio. [1] Terms are vetted as keywords based on three features: document frequency (TF), collection frequency (IDF), relative position of its first occurrence in a document and the term's part of speech tag.

IV MACHINE LEARNING APPROACH

The Key phrase Extraction Algorithm (KEA) [26] uses the machine learning techniques. The process is separated into two phases: term-weighting and keyword extraction. First, a set of feature vectors is generated from different encyclopaedia domains. The same procedure is then performed on a corpus of newspaper articles. The encyclopaedia vectors are compared with the article vectors using a similarity calculation so as to separate the latter into different domains, after which they are sorted, producing the final set of feature vectors. In the second phrase, keyword extraction, a segment is analysed such that the most relevant domain is selected for it using the pre-existing feature vectors. [1] Phoneme recognition software is employed to do the analysis, looking for the best fit between a segment's vectors and that of one of the encyclopaedia domains. When the best fitting domain is chosen, its keywords are then assigned to the radio news segment and naive Bayes formula for domain-based extraction of technical key phrases.

V MIXED APPROACHES

Other approaches about keyword extraction mainly combine the methods mentioned above or use some heuristic knowledge in the task of keyword extraction, such as the

position, length, layout feature of the words, html tags around of the words,

VI SELF APPROACH

The self-approach embeds both the machine learning approach and the simple statistical approach. The process of keyword extraction goes like this. The steps are explained here.

Step 1: Clean the document by pre-processing it. By pre-processing we find those characters and those stop words which do not have the priority of becoming the keywords. The tokens are found by using the special characters and white spaces, dots, new line and are removed. Even if the propositions repeat themselves throughout the document do not consider them as a keyword. Eliminate the numbers and the non-alphanumeric alphabets. This process can be done using the machine learning algorithms.

Step 2: 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.

Step 3: Here the keyword extraction depends on the two criterions or attribute, frequency and the position. The density of the word plays important role only after pre-processing the document and removing pre-positions.

The number of repetitions of words are calculated and divided by the total number of words in the document by which the density of the word is found. The metric is the density of the word found by the formula

$$\text{Word density} = \frac{\text{Frequency of the word}}{\text{total number of words in the document}}$$

Now find the document density with respect to term. It is found by using the formula

Document density w.r.t.word =

$$1 + \frac{\text{Total number of Document}}{\text{number of documents that contain the term/word}}$$

Now the keyword metric can be found by using the formula

Frequency =

$$word\ density * \log(Document\ density\ w.r.t.word)$$

Step 4: The position of the word according to the paragraph that it exists in can be another identifying feature of the keywords since it is also expected them to be in the beginning and end of the paragraph. Furthermore, the position of the word in the sentence may be an identifying feature. For instance, while more important terms are found in the beginning or end of the sentence in English. Here calculate the words distance from the stating of the first word of the staring paragraph and the ending paragraph.

Step 5: based on the above two metrics Rank the words with respect to the frequency of the words and display the list of the ranked keywords. These words are the keywords that stored in for later references.

Based on the above algorithm, we have used Reuters-21578 dataset. We have taken just a single paragraph of the document and have found a list of keywords.

RESULTS

These are some set of keywords which are extracted from the Reuters-21578 “reut02-013”.

mln dlrs,
basis points,
month
mln shares
Corp
mln vs
mln dlrs contract
mln dlrs technology
options exchange
chartered bank
Net 47.5 mln
real estate
pct yearly turnover
pct loan-to-price ratio
bank board
sale
pct preferential return
mln note
pct rise
pct increase
mln revs
IBM announcement
april 14 reuter
company
personal computers
Pan Am
available seat miles
Shr 5.56 dlrs
Compaq
New York
March traffic
Beneficial Corp
March 21-31 period
Federal Home Loan Bank Board

VII 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] P. D. Turney, Learning Algorithms for Keyphrase Extraction, Information Retrieval, 1999
[2] E. Frank, G. W. Paynter, I. H. Witten, C. Gutwin, and C. G. Nevill-Manning. Domain-specific keyphrase extraction. In IJCAI, pages 668--673, 1999.
[3] R. O. Duda, P. E. Hart, D. G. Stork, Pattern Classification, page 20, Wiley-Interscience, 2000.
[4] G. Salton and M. J McGill. Introduction to Modern Information Retrieval. McGraw-Hill, 1983.
[5] Thomas Bayes. An essay towards solving a problem in the doctrine of chances. Philosophical Transactions of the Royal Society (London). 53:370-418, 1763
[6] P. Domingos and M. Pazzani. On the optimality of the simple bayesian classifier under zero-one loss. 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.