Effective Approaches For Extraction Of Keywords

Jasmeen Kaur [1], Vishal Gupta [2]

**[1] ME Research Scholar Computer Science & Engineering, UIET,
Panjab University Chandigarh, (UT)-160014**

**[2] Assistant Professor Computer Science & Engineering, UIET,
Panjab University Chandigarh,(UT)-160014**

Abstract

Keywords are index terms that contain most important information. Automatic keyword extraction is the task to identify a small set of words , keyphrases or keywords from a document that can describe the meaning of document. Keyword extraction is considered as core technology of all automatic processing for text materials. In this paper, a Survey of Keyword Extraction techniques have been presented that can be applied to extract effective keywords that uniquely identify a document.

Keywords:*Keyword Extraction, Approaches, Keywords and Document .*

1. Introduction

Keywords play a crucial role in extracting the correct information as per user requirements. Everyday thousands of books , papers are published which makes it very difficult to go through all the text material ,instead there is a need of good information extraction or summarization methods which provide the actual contents of a given document. As such effective keywords are a necessity. Since keyword is the smallest unit which express meaning of entire document , many applications can take advantage of it such as automatic indexing , text summarization , information retrieval , classification , clustering , filtering , cataloging , topic detection and tracking , information visualization , report generation , web searches , etc.[1]

Existing methods about Automatic Keyword Extraction can be divided into four categories:-

- Simple Statistics Approach :

These methods are simple and do not need the training data. The statistical information of the words can be used to identify the keywords in the document. Cohen uses N-Gram statistical information to automatically index the document. N-Gram is language and domain independent. Other statistical methods include word frequency, TF*IDF, word co-occurrence, etc[7].

- Linguistics Approach :

These approaches use the linguistic features of the words mainly sentences and documents. The linguistic approach includes the lexical analysis, syntactic analysis discourse analysis and so on.

- Machine Learning Approaches :

Keyword Extraction can be seen as supervised learning, Machine Learning approach employs the extracted keywords from training documents to learn a model and applies the model to find keywords from new documents. This approach includes Naïve Bayes, Support Vector Machine, etc.

- Other approaches :

Other approaches about keyword extraction mainly combines the methods mentioned above or use some heuristic knowledge in the task of keyword extraction, such as the position, length, layout feature of words, html tags around of the words, etc.

Various extraction methods discussed are for single document but these can further applied to multiple documents as per their suitability [12]. Keywords are extracted by identifying the noun phrase as noun phrase comprises very crucial information about the text document. The keywords are choosed based on their linguistic features [2] and informative features [6] such as highlighted words. Query-focused and the words in abstracts or titles can be the part of candidate keywords. Methods such as co-occurrence [8],[14] and machine learning [15],[16] have been used for extracting keywords from a single document. Topics are detected using keyword clustering. In addition extracting and clustering related keywords based on history of query frequency [11] is also one of the methodology adopted. In the following sections various approaches for selecting effective keywords are elaborated.

2. Identify Noun Phrase As Keyword

Nouns contain bulk of information and this keyword extraction algorithm requires a morphological analyzer and rules or grammar for finding simple noun phrases. Since noun phrases are extracted and become candidate keywords. The noun phrases are scored and clustered and then clusters are scored. The shortest noun phrase from the highest scoring clusters are then used as keywords.
The keyword extraction algorithm[1] overview:

a) Morphological Analysis
- Word Segmentation
- Part of Speech Tagging
- Stemming

b) NounPhrase Extraction and Scoring
- Noun Phrase Extracted
- Stopwords Removed
- Noun Phrase Scored using UnigramFrequency:

$$UF(NP)= {}^{|NP|}\Sigma_{i=0}UnigramFrequency(Wi)$$

$$Score(NP)=\frac{UF(NP)*NPF(NP)}{|NP|}$$

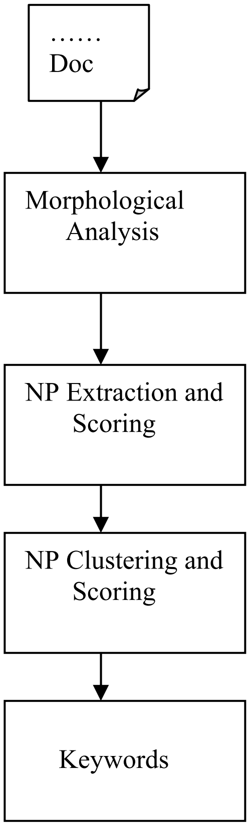

Fig .1. Keyword Extraction Algorithm

c) NounPhrase Clustering and Scoring
- Clusters are formed having a common word
- Clusters are scored

$$Score(Cluster)= \frac{{}^{|cluster|}\Sigma_{i=0}Score(NP_i)}{|cluster\ |}$$

d) Choosing Keywords
- Shortest length noun phrase choosed as keyword.

3. Term Frequency-Inverse Document Frequency

TF-IDF weight evaluates the importance of a word to a document in a collection. Importance increases proportionally to number of times a word appears in document but is offset by frequency of word in corpus [3].
Term T_i in particular document D_j
Term frequency is,

$$tf_{ij}=(n_{i,j})/ \Sigma_k n_{k,j}$$

where $n_{i,j}$ is number of occurences of considered term(t_i) in document d_j
and denominator is the sum of number of occurrences of all terms in $d_{j.}$

Inverse Document Frequency[20] is a measure of general importance of term obtained by dividing number of all documents by number of documents containing the term and then taking logarithm of the quotient

$$Idf_i=\frac{log\ |D|}{|\{d:t_i€d\}|}$$

Where, |D|is total number of documents in corpus and the denominator is number of documents where t_i appears.
Hence

$$(tf\text{-}idf)_{i,j}=tf_{i,j}\times idf_i$$

But the limitation of this method is that it does not work for single document since there are no other documents to compare keywords to algorithms, so it will choose keywords based on term frequency.

4.Selection Based On Informative Features

Words are found in various forms of writing in documents which provides additional information about the importance of words[6]. There are various informative features such as,
- Words emphasized by application of bold, italic or underlined fonts,
- Words typed or written in upper case,
- The size of the font applied,
- Normalized Sentence Length , which is the ratio of number of words occurring in sentence over

number of words occurring in the longest sentence of the document,

- Cue-phrases are sentences beginning with summary phrase(in conclusion or in particular) and transition phrase like however, but, yet, nevertheless.

5. Query Focused Keyword Extraction

According to this method, keywords correlate to the query sentence and denote the main content of document. It calculates query related feature and then obtains importance of word [9]. The whole system worked as follows,

- Query
- Sentence-Pruning
- Query-Related Feature
- Keywords are selected

The relevant degree of words w1 and w2 is calculated by taking window of length K words. All words in window are said to co-occur with first word with strengths inversely proportional to distance between them.

If n(w1,k,w2) is the number of w1 and w2 co-occur in the window, where k denotes real distance between w1 and w2 when they are co-occurred.

The relevant degree R(w1,w2) is calculated by,

$R(w1,w2)= {}^{K}\Sigma_{k=0}w(k)*n(w1,k,w2)$

Then query-related feature of word w_i is ,

$F1(w_i)={}^{qwt-1}\ \Sigma_{\ j=0}\ R(w_i,w_j)$

6. Position Weight Algorithm

Words in different positions carry different entropy as if same word appears in introduction and conclusion paragraphs, the word carries more information. This Position Weight method record the importance of a word position. It uses three important elements,

- Paragraph weight
- Sentence weight
- Word weight

If the paragraph(p_j) is main title or subtitle, leading or concluding paragraph, it carries more weight than common paragraph.

First and concluding sentences (s_k) are more important than the example sentences which are weighed nearly zero.

Likewise words (w_r) that are capitalized plus some digits are heavily weighed than other common words[2].

The total weight of the term t in document is the sum of weights of all positions it appears. If term t appears m times in document d, its PW is

$PW(t,d)={}^{m}\Sigma_{i=1pw}(t_i)$

Where pw of a term in a specific position as

$Pw(t_i) = pw(t_i,p_j).pw(t_i,s_k).pw(t_i,w_r)$

For preprocessing, text chunking and elimination of stop wordsthat are included in the Fox stop list[17] have been carried out and leaving special words having transmissible or negative meaning like ‘however’, ‘nevertheless’ and etc. Next is to stem the words using Krovetz algorithm[18] based on WordNet Dictionary[19]. Last is to calculate the PW on the algorithm described.

7. Keyword Extraction Using Conditional Random Field(CRF) Model

Conditional random Field (CRF) model works on document specific features. CRF [7] is a state of art sequence labeling method and utilize most of the features of documents sufficiently and effectively for efficient keyword extraction. At the same time, keyword extraction can be considered as string labeling. Here, keyword extraction based on CRF has been discussed. Using CRF model in keyword extraction has not been investigated previously. The results show that CRF model outperforms other machine learning methods such as support vector machines, multiple linear regression model, etc. in the task of keyword extraction.

CRF model is a new probabilistic model for segmenting and labeling sequence data. CRF is an undirected graphical model that encodes a conditional probability distribution with a given set of features. In process of manual assignment keyword to a document, the content of document will be analyzed and comprehended firstly. Keywords which can express the meaning of document are then determined. Content analysis is the process that most of the units of a document such as the title, abstract, full-text, references and so on, be analyzed and comprehended. Sometimes, the entire document has to be read then summarize the content of document, and give the keyword finally. According to process of manual assignment keyword to a document, in this technique, the process is transferred to labeling task of text sequences. In other words, a word or a phrase can be annotated with a label by a large number of features of them. Therefore keyword extraction algorithm based on CRF has been devised to extract keywords. It uses CRF++[13] tool to extract keywords.

The different kinds of features used are:

- Local context features such as len, t, a, c, pos, etc.
- Global context features such as T, A, H, L, R, etc.

Process of CRF based keyword extraction:-

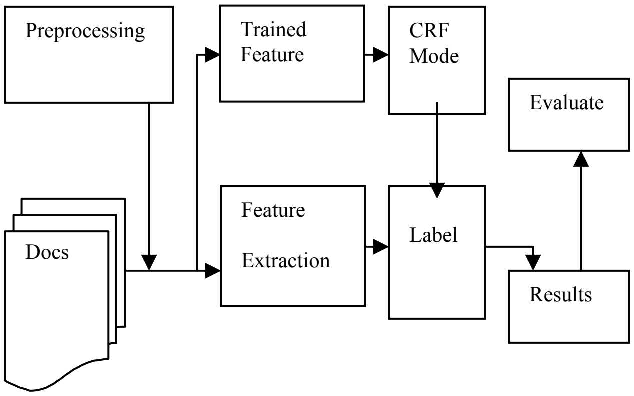

Fig.2. CRF based Keyword Extraction Process

7.1 Preprocessing and Features Extraction:-

The input is a document. Before CRF model training, transfer the document into the tagging sequences, i.e. a bag of words or phrases of a document. For a new document, the sentence segment has been conducted, and pos tagging Then the features mentioned above are automatically extracted. The output of the feature vectors, and each vector corresponds to a word or a phrase.

7.2 CRF Model Training:-

The input is a set of feature vectors by step above. A CRF model has been trained that can label the keyword type. In the CRF model, a word or phrase could be regarded as an example, the keyword has annotated by one kind of labels, such as 'KW_B', 'KW_I', 'KW_S', 'KW_N', 'KW_Y'. The tagged data are used to training the CRF model in advance. In the CRF++ the output is a CRF model file.

7.3 CRF Labeling and Keyword Extraction:-

The input is a document. The document is preprocessed and features are extracted. Then, keyword type has been predicted using CRF model. According to the keyword type, the keywords of the document are extracted.

7.4 Results Evaluation:-

Results of keyword extraction can be evaluated by comparing these results with the manual assignment results.

8. Conclusion

As with the noun phrase keyword extraction methodology, the only requirement is that the language have a morphological analyzer and rules for finding simple noun phrases. Since nouns contain bulk of the information, noun phrases are extracted and become candidate keywords. The noun phrases are scored and clustered and then the clusters are scored. The shortest noun phrase from the highest scoring clusters are then used as the keywords.

The Position Weight algorithm automatically extract keywords from a single document using linguistic features. The results show that the PW algorithm has a great potential for extracting keywords, as it generates a better result than other existing approaches.

Using TF-IDF Variants, there are six different values for every word and filtering can be done by using cross-domain comparison i.e. meaningless words have been removed. Furthermore, TTF(Table Term Frequency)[3] has been applied to more precise extraction of keywords.

CRF is a state of art sequence labeling method and utilize most of the features of documents sufficiently and effectively for efficient keyword extraction. At the same time, keyword extraction can be considered as string labeling. Here, keyword extraction based on CRF has been discussed. Using CRF model in keyword extraction has not been investigated previously. The results show that CRF model outperforms other machine learning methods such as support vector machines, multiple linear regression model, etc. in the task of keyword extraction.

9. REFERENCES

[1] David B. Bracewell and Fuji REN, " Multilingual Single Document Keyword Extraction For Information Retrieval", Proceedings of NLP-KE, 2005,pp. 517-522.

[2] Xinghua u and Bin Wu, " Automatic Keyword Extraction Using Linguistics Features ", Sixth IEEE International Conference on Data Mining(ICDMW'06), 2006.

[3] Sungjick Lee, Han-joon Kim, " News Keyword Extraction For Topic Tracking ", Fourth International Conference on Networked Computing and Advanced Information Management, 2008 ,pp. 554-559.

[4] Meng Wang, Chao Xu, " An approach to Concept-Obtained Text Summarization ", Proceedings of IEEE, 2005,pp.1290-1293.

[5]Tingting He, " A Query-Directed Multi - Document Summarization System ", Sixth International Conference on Advanced Language Processing And Web Information

Technology,2007.

[6] Rasim M. Alguliev and Ramiz M. Aliguliyev,“ Effective Summarization Method of Text Documents ”, Proceedings of International Conference on Web Intelligence, IEEE, 2005.

[7] Chengzhi Zhang, “ Automatic Keyword Extraction From Documents Using Conditional Random Fields ”, Journal of Computational and Information Systems, 2008.

[8] Y. Matsuo and M. Ishizuka, “ Keyword Extraction from a Single Document using Word Co-occurrence Statistical Information ”, International journal on Artificial Intelligence Tools, vol.13, no.1, 2004,pp.157-169.

[9] Liang Ma, Tingting He,“ Query-focused Multi-document Summarization Using Keyword Extraction ”, International Conference on Computer Science and Software Engineering, IEEE,2008,pp. 20-23

[10]Christian Wartena, Rogier Brussee, “ Topic Detection By Clustering Keywords ” , 19th International Conference on Database and Expert Systems Application, 2008,pp. 54-58.

[11]Toru Onoda ,” Extracting and Clustering Related keywords based on History of Query Frequency ” , Second International Symposium on Universal Communication ,2008,pp.162-166.

[12] WWW wikipedia.org.

[13] CRF++: Yet Another CRFToolkit.http://Chasen.org/~taku/Software/CRF++.

[14] Y. Ohsawa, N E Benson, “ KeyGraph: Automatic indexing by cooccurence graph based on building construction metaphor ”, In Proceedings of The Advanced Digital Library Conference,1998,vol.12.

[15]A. Hulth, “ Improved Automatic Keyword Extraction Given More Linguistic Knowledge ”, In Proceeedings of the Conference on Empirical Methods in Natural Language Processing,2003.

[16]P Turney, (2000), “ Learning algorithms for Keyphrase Extraction ”, Information Retrieval, vol. 2, no. 4, pp. 303-336.

[17] C Fox, “ Lexical Analysis and Stoplists. Information Retrieval: Data Structures and Algorithms “, Prentice Hall, New Jersey, 1992, pp. 102-130..

[18] R Krovetz., “ Viewing morphology as an inference process ”, Proceedings of ACM-SIGIR93,1993,pp.191-203.

[19] G. Miller, “Wordnet: An on-line lexical database International Journal of Lexicography “,1990,vol. 3,no.4.

[20] Stephen Robertson, “ Understanding Inverse Document Frequency: on theoretical arguments for IDF ”, Journal of Documentation, Vol. 60, No. 5, 2004, pp 503-520.

[21] Zhang and H Xu, “ Keyword Extraction Using Support Vector Machines ”, In Proceedings of Seventh International Conference on Web Age Information Management China, 2006, pp.85-96.

AUTHORS’ INFORMATION

Author’s Biodata

Jasmeen kaur is pursuing ME in Computer Science and Engineering at Unversity Institute of Engineering and Tecnology , Panjab University, Chandigarh. Jasmeen Kaur did her B Tech in CSE from Bhai Gurdas Institute of Engineering and Technology Sangrur in 2007.She secured 80% marks in B Tech. She is carrying out her thesis work in the field of Natural Language Processing.

Second Author’s Biodata

Vishal Gupta is Lecturer in Computer Science & Engineering Department at University Institute of Engineering & Technology, Panjab university Chandigarh. He has done MTech. in computer science & engineering from Punjabi University Patiala in 2005. He was among university toppers. He secured 82% Marks in MTech. Vishal did his BTech. in CSE from Govt. Engineering College Ferozepur in 2003. He is also pursuing his PhD in Computer Sc & Engg. Vishal is devoting his research work in field of Natural Language processing. He has developed a number of research projects in field of NLP including synonyms detection, automatic question answering and text summarization etc. One of his research paper on Punjabi language text processing was awarded as best research paper by Dr. V. Raja Raman at an International Conference at Panipat. He is also a merit holder in 10th and 12th classes of Punjab School education board. in professional societies. The photograph is placed at the top left of the biography. Personal hobbies will be deleted from the biography.