A Group of Token Based Approach for Key-Word Recommendation

[1] Mona Amarnani, [2] C. S. Warnekar

[1] M. Tech Scholar, Computer Science and Engg. Shri Ramdeobaba College of Engineering and Management (SRCOEM), Nagpur, India

[2] Former Principal,Cummins College of Engineering,Karve nagar, Pune, & Sr. Professor JIT, Nagpur, India

Abstract

In today's world, the accelerated pace and steady growth of text document generation raises the questions about access and discoverability of data. Lot of applications is being derived for information retrieval and natural language processing. Keywords are important aspect of any plain text document. Careful selection of keywords helps the researchers to gauge the contents and summarize it. Automatic keyword extraction is a process in which the key-words are systematically extracted from a text document. Here we attempt to present a content based system for automatic key-word extraction and recommendation. The recommendation includes statistical and linguistic approaches for key-word extraction. This will help to achieve the goal of automatic extraction of key-words and command the velocity of document generation to provide solutions to problems such as access and discoverability.

Keywords: *Keyword, Extraction, Word Frequency, Summarize.*

1. Introduction

Keywords are important aspect of any plain text document. They serve as a dense summary for a document. The aim of automatic keyword extraction and recommendation is to find a small set of terms that describes a specific document. This will be helpful in document collection and information retrieval. Applications of Keyword Extraction [10] include Domain-Based Extraction of Technical Key phrases, Spoken Language Processing with Term Weighting, Spoken Text Keyword Extraction with Lexical Resources, Keyword Extraction with Thesauri and Content Analysis and Linguistic Features as Error Correction in Keyword Extraction. Methods of Automatic Keyword Extraction can be categorized into four categories [7]

1. Statistical Approach – based on statistical information of the words and do not need training.
2. Linguistics Approach - based on linguistic features of the words.
3. Machine Learning Approaches – a supervised learning approach which applies the model to find keywords from new documents.
4. Other Approaches – based on the combination of above three approaches.

Here we present an approach of key-word recommendation which includes statistical and linguistic approaches for extracting key-words. This method can be used extract tokens or set of tokens which may be helpful in making the document searchable hence enhancing its accessibility and discoverability.

2. Related Work

Since keyword is the smallest unit which express meaning of complete document, many applications can take benefit of it such as automatic indexing, manuscript summarization, information recovery, classification, clustering, filtering, cataloguing, topic detection, web searches, tracking, report generation, information visualization, etc is discussed by David B. Bracewell and Fuji REN[1]. Methods such as machine learning, by A. Hulth[5], have been used for Finding potential terms from a single document. By adding linguistic knowledge to the representation, experiments on automatic extraction of keywords from abstracts using a supervised machine learning algorithm are discussed.

Eliminating the need of machine learning for extraction of key-words, Barker and Cornacchia (2000)[3] discuss an algorithm by means of POS patterns where the number of words and the frequency of a noun phrase and the frequency of the head noun is used to determine what terms are keywords. Boguraev and Kennedy [2] extract technical terms based on the noun phrase patterns suggested by Justeson and Katz [12]. Daille et al. [13] applied statistical filters and extracted noun phrases.

3. Proposed Approach

In the study it is seen that term frequency is the best filter candidate of the scores investigated. To extract potential terms, problem while using POS patterns lies in how to

restore the relevant terms and restrict the number of terms. This paper attempts to develop a system which uses the combination of statistical and linguistic approaches.
Plain text document is given as input to the system. The system performs automatic key-word extraction using certain logistics or criteria.

3.1 Logistics of Key-Word Recommender System

The present system uses the following logic to select the set of keyword from the body of the paper –

- Frequency of a word appearing in body of the jth plain text document P (j).
- Words appearing in title and sub titles (including titles of the diagrams) of P (j).
- Words appearing in the abstract of P (j).
- And a combination of the above.

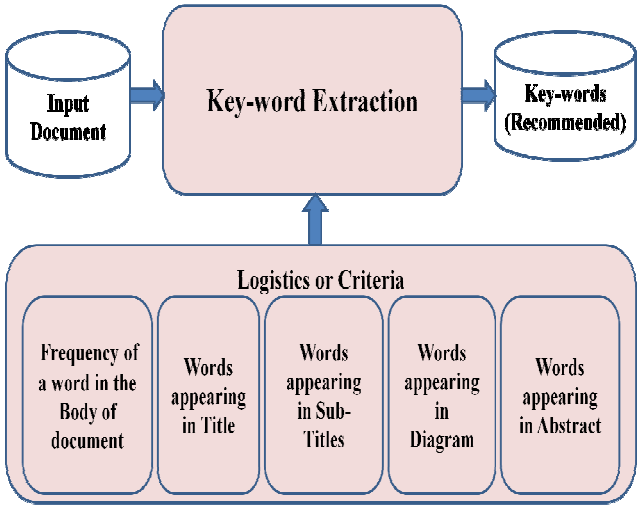

Fig. 1 Key-word Recommender System.

3.2 Proposed algorithm for Key-word Extraction

Step 1. Input document
Step 2. Tokenization
Step 2. Removal of Stop words
Step 3. Parts of speech(POS) tagging
Step 4. Grouping
Step 5. Frequency Counting

Step1. Input document P (j) - A research papers may be used as input. The keywords of author are stored using K (i, j) = f (P (j)). where, P (j) indicates the list of any published document in a standard format. K (i, j) indicates the list of keywords or index terms and f is the linking function.

Step2. Tokenization - Implements the retention of non empty set of characters, exclusive of spaces and punctuations.

Step3. Stop words Removal – less significant words, such as then, and, which, such, etc, are removed.

Step4. Parts of speech(POS) tagging - based on both its meaning as well its context(semantics), i.e. relationship with adjacent and related words in a sentence.

Step5. Grouping - Each word is grouped after POS tagging with its predecessor and successor word. Minimum two and maximum four words can be grouped. After grouping, each word occurrence with its associated POS tag is counted and compared. For each word, the groups which appear with the most frequent POS tag are considered. Rest of the groups for that particular word are discarded. This gives semantic weight age to groups before for recommendation of key-words.

Step6. Stemming - finds base form of each word.

Step7. Frequency Counting - PShort(j) or PS(j) is produced as a minimised version of P(j). First twenty frequent groups are recommended as key-words.

Step8. Key-word Recommendation - After applying the aforesaid key-word extraction algorithm, most frequent extracted sets of tokens are recommended as keywords.

This entire process is illustrated through the block diagram given below -

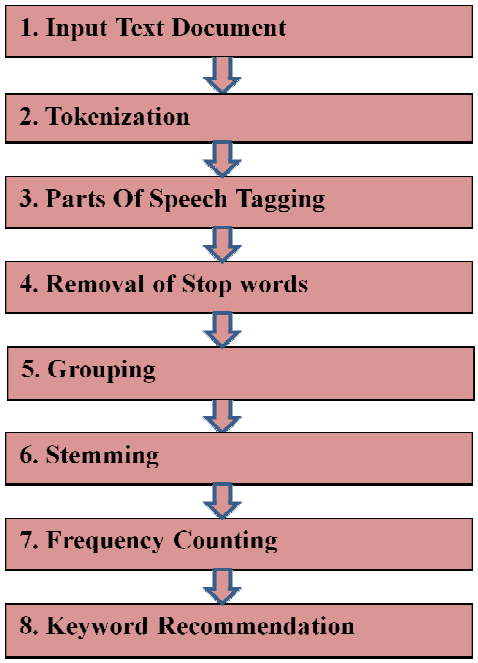

Fig. 2: Group of Token based Key-word Recommender System

Typically contextually relevant group of words may consist of single word or even entire natural language statement. Assuming an average value of eight words in

english like natural language statement, the step 5, was designed accordingly where the number of tokens can be specified by the user. This can easily be extended to other Indian language statements.

4. Conclusion and Future Scope

Machine learning key-word extraction approaches involve huge databases. Combination of the statistical and linguistic approaches suggested here improve the overall efficiency of key-word recommender systems, eliminating machine learning process. The new system also considers unpopular and unique words for recommendation as key-words. Business tender or legal documents are often voluminous and their reading and understanding their contents is quite time consuming. Gist of such big documents can be derived, using this system, in Management Information System (MIS). Reverse engineering of the present key-word recommender system could be used to enhance the overall efficiency of web search engine.

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. Branimir Boguraev and Christopher Kennedy. 1999. Applications of term identification technology: Domain description and content characterization. Natural Language Engineering, 5(1):17–44.
3. Ken Barker and Nadia Cornacchia. 2000. Using noun phrase heads to extract document key phrases. In Canadian Conference on AI.
4. Peter D. Turney. 2000. Learning algorithms for key phrase extraction. Information Retrieval, 2(4):303–336.
5. Hulth, " Improved Automatic Keyword Extraction Given More Linguistic Knowledge ", In Proceeedings of the Conference on Empirical Methods in Natural Language Processing,2003.
6. Marko Balabanovik And Yoav Shoham (Acm), March 1997/Vol. 40, No. 3. "Content-Based, Collaborative Recommendation".
7. Jasmeen Kaur and Vishal Gupta , "Effective Approaches For Extraction Of Keywords", IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 6, November 2010.
8. Branimir Boguraev and Christopher Kennedy. 1999. Applications of term identification technology: Domain description and content characterisation. Natural Language Engineering, 5(1):17–44.
9. Sugandha Dani And Dr. C. S. Warnekar, (Ijca), (2011). "Improvising Search Engine By Prioritizing Query String".
10. Michael J. Giarlo., "A Comparative Analysis of Keyword Extraction Techniques", Rutgers, The State University of New Jersey.
11. Imad A. Al-Sughaiyer, Ibrahim A. Al-Kharashi (Jasict) Vol.55, Issue3, January (2011). "Arabic Morphological Analysis Techniques: A Comprehensive Survey", Second International Workshop on Education Technology and Computer Science (ETCS), 2010, pages 673 – 675.
12. John S. Justeson and Slava M. Katz. 1995. Technical terminology: some linguistic properties and an algorithm for identification in text.
13. B´eatrice Daille, ´ Eric Gaussier, and Jean-Marc Lang´e. 1994. Towards automatic extraction of monolingual and bilingual terminology. In Proceedings of COLING-94, pages 515–521, Kyoto, Japan.