Content Based Key-Word Recommender

Mona Amarnani
Student, Computer Science and Engg. Shri Ramdeobaba College of Engineering and Management (SRCOEM), Nagpur, India

Dr. C. S. Warnekar
Former Principal,Cummins College of Engineering,Karve nagar,Pune, India,

Abstract: The author uses the keywords to summarize the document. Keywords represent a significant aspect of informative documents like research papers. Thus, they span the semantic conceptual contents of the corresponding text. They serve as reference points in understanding the gist of a text document. It is often required to identify keywords from a document to summarize it. Automatic Keyword Recommender helps in identification of Key Words & is a current need. Recommender systems are a subclass of Information filtering systems, typically producing a list of suggestions of actual or similar items predominately required by the user. Various candidate items are compared before final recommendation, through two basic approaches for recommendation, collaborative or content-based filtering of the corresponding document. Here we attempt to develop a content based key-word recommender system. The system incorporates a token based keyword recommender system and a Group of token based keyword recommender system. The token based keyword recommender system is based on the criterion used by authors. Such criteria includes selection based on words appearing in main titles of corresponding research paper or frequently used word, etc. The group of token based keyword recommender system is based on the frequently used group of context based semantically relevant tokens. The keywords recommended by the system so developed are then compared with the original keywords provided by the author in their standard paper format. A reverse engineering of this process could be applied to improvise the efficiency of search engine.

Keywords: Keyword, Recommender, Extraction, Tokenization, Stop words, Stemming, POS tagging, Word Frequency.

I. INTRODUCTION

Keyword is a significant aspect of informative documents like research papers. They indicate the topic area and the methodology of the research paper. Ideally, they span the semantic conceptual contents of the corresponding text. The author uses the keywords to summarize the document in order to reflect the specificity of the paper. They serve as reference points in understanding the gist of a text document. Automatic Keyword Recommender helps in identification of Key Words & is a current need.

Recommender systems are a subclass of Information filtering systems, typically producing a list of suggestions of actual or similar items predominately required by the user. Various candidate items are compared before final recommendation, through two basic approaches for recommendation, collaborative or content-based filtering of the corresponding document. In collaborative filtering the system maintains a database of many items and involves machine learning by finding other similar items whose contents strongly correlate with the current item. It weights all items with respect to similarity with the active item. A key advantage of the collaborative filtering approach is that it does not rely on machine analyzable content and therefore it is capable of accurately recommending complex items without requiring an understanding of the item itself. Content based filtering system recommendations are based on information on the content of items rather than on other correlated items.

The present system extends the scope of recommendation system for key-word extraction from plain text document. It extracts words of great significance in the body of the document using criteria based on the logistics given in the sequel. The system implements content based filtering, thus the other co-related documents are not required and machine learning is not involved in the system. This is more advantageous as it recommends to documents specified by set of unique keywords. The system is also able to recommend new and un-familiar keywords. The system could also be helpful to summarize huge business documents for Management Information System. It can be extended to guide automatic text summarization. In the burgeoning cyber world activity, various search engines are employed to retrieve the desired information from voluminous text data in response to query. Hence the relationship between query-string and text contents is important in order to populate the search results with meaningful links. It is suggested that a reverse engineering process of this research may be extended to improvise the efficiency of web search engine.

II. CONTENT BASED KEY-WORD RECOMMENDER

As seen from the block diagram of the proposed system, the soft copy of a plain text document forms input to the system. The key-word extraction uses certain logistics which are explained in the sequel.

Fig 1. Key-word Recommender System

The system is developed in two user selectable modules -

1. Token based Keyword Recommender System
2. Group of tokens based Keyword Recommender System

In a **Token based Keyword Recommender System**, content based filtering is done in order to identify and extract tokens from the body of the document. It implements simple Statistical approach for key-word extraction eliminating the need for training data. Frequently occurring individual tokens in the entire body of the input text document are recommended as key-words by the system. Before extracting the individual tokens some pre-processing is done on the body of the paper to remove less significant words.

In a **Group of Token based Keyword Recommender System** uses linguistics approach. Here content based filtering is done in order to extract a group of tokens from the body of the paper. The system is simple and uses the linguistic features of the words and their relation with sentences and documents. The linguistic approach includes the standard features like lexical analysis, syntactic analysis, etc. Thus the system recommends key-words that are semantically relevant group of tokens. The system groups minimum two to four tokens to extract the key-words for recommendation. Parts of speech(POS) tagging is done before key-word extraction as a part of pre-processing thereby semantically relevant group of tokens identified by their frequent occurrence are recommended as key-words.

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 of P (j).
- And a combination of the above.

III. TOKEN BASED KEY-WORD RECOMMENDER SYSTEM

For a meaningful application of the aforesaid logistics, following pre-processing of the text document P (j) is helpful.

- Tokenization – Retaining nonempty sequence of characters, excluding spaces and punctuations from P (j).
- Stop Words - Removing function words and connectives like the, a, an, is, for, and, which, etc.
- Stemming - Reducing inflected (like ed, s, es, ing, etc. or sometimes derived) words to their root form.
- Frequency counting of the text-words of P (j) in order to select predominant words.

The token based recommender system is developed by a suitable amalgamation of the above pre-processing and logistics steps. Keywords thus extracted from P (j) by the developed system are subsequently compared with author's list of key words, with a view to judge the effectiveness of the system.

This entire process is illustrated through the block diagram.

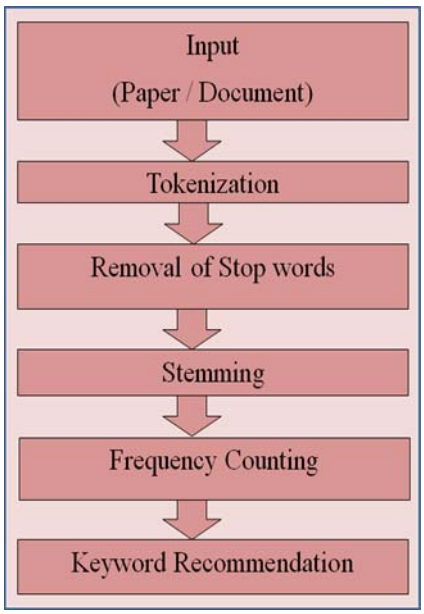

Fig 2. Block Diagram of Token based Key-word Recommender System

3.1 DEVELOPMENT STEPS OF TOKEN BASED KEY-WORD RECOMMENDER SYSTEM

A. Input to the system is a plain text document P (j)

Since the system focuses more on research papers, its softcopy may be used as input. The keywords of author are stored using –

$$K(i, j) = f(P(j))$$

where,

P (j) → The list of any published document in a standard format.
K (i, j) → The list of keywords or index terms
where,
'i' represents the current keyword of jth document.
f → The linking function.

B. Tokenization

Tokenization is the process of breaking up a document into sentences. Sentences into words or phrases. These phrases or meaningful elements are called tokens. Tokens are usually separated by whitespace characters or space or line breaks or punctuation characters. The stream of characters in P (j) needs to be broken up into distinct meaningful units (or tokens) before any further processing beyond the character level. If the text is perfectly punctuated, a simple program can produce tokens by removing punctuations. For example NLTK tokenizer package, including Simple tokenizer, White Space tokenizer and Regular Expression tokenizer, has been used in the present system.

C. Removing Stop Words

Stop words in a document have less significance in searching meaningful key-words. There is not one definite list of stop words which all tools use. These are short function words *like* and, a, an, the, is, etc. These words are filtered out prior to key-word extraction from the document in order to improve the efficiency of the system.

D. Stemming

Stemming is the process for reducing inflected or derived words to their root form. The morphological root of the word may not be probably similar to the stem. Such base or root words called as stem help in locating typical

keywords. Stemming algorithms include strong stemming algorithm or a weak stemming algorithm. A strong stemming algorithm removes the inflectional suffixes (s, ing, es, ed) as well as derivational suffixes (able, ability. A weak stemming algorithm removes only the inflectional suffixes (s, ed). Key-word recommender system uses a strong stemming algorithm. OpenNLP English stemmers are interfaced and used by the system.

After stemming is done, minimised version of P (j) called PShort(j) or PS (j) is produced, which is further used for frequency counting.

3.2. Case Study and Results of Token based Key-word Recommender System

The above automatic extraction process is applied to a set of 30 input research papers P (j). The key-words thus recommended by the system are then compared with the author's list of keywords or index terms.

The following tables illustrates the comparison statistics of input research paper entitled – "KNOWLEDGE RETRIEVAL USING HYBRID SEMANTIC WEB SEARCH" -

TABLE 1 LIST GIVEN BY AUTHOR

Key-word given by Author	Number of times keyword		Frequency in body of paper
	Appears in main title	Appears in sub-title	
Knowledge	1	3	68
Retrieval	1	3	39
Semantic	1	3	80
Web	1	2	85
Search	1	1	37
Ontology	0	2	44
Clustering	0	2	1
User	0	2	32
Profiling	0	2	2

TABLE 2 LIST PRODUCED BY THE SYSTEM

Frequent words suggested by the key-word recommender	Frequency of word	Is it a key-word given by author?	Appears in Main Title?
Web	85	Yes	Yes
Semantic	80	Yes	Yes
Knowledge	68	Yes	Yes
Inform	46		
Use	45		Yes
Ontology	44	Yes	
Retrieve	39	Yes	Yes
Search	37	Yes	Yes
Data	32		
User	32		
Research	24		
Technique	20		
Compute	19		
Profile	19	Yes	
Relate	17		

Observations made from the above case study –

- 55.55% of author's key-words are present in main title.
- 46.6% of recommended keywords are present in author's list of keywords.
- 85.71% of recommended keywords appear in main title.

Observations made from the above case study –

- 55.55% of author's key-words are present in main title.
- 46.6% of recommended keywords are present in author's list of keywords.
- 85.71% of recommended keywords appear in main title.

3.3. Comparision of Key-words in Token based Key-word Recommender System

Keywords recommended by the system K (i, j) are then compared with the author's given list of keywords KAuthor (i, j) or KA (i, j) and the matching percentage is indicated. After repeating the above steps for a large number (30 at present) of cases, the efficacy of the key-word recommender system is established. The system recommends typical set of key words to the author of a research paper and may be used to highlight typical key-points of lengthy business deals or contracts to an entrepreneur.

Average observations from 30 research papers –

- 54.75% of the author's keywords are present in main title.
- 58.64% of words in main title are recommended keywords.
- 62.56% of words present in author's list of keywords are recommended keywords.

IV. GROUP OF TOKEN BASED KEY-WORD RECOMMENDER SYSTEM

For a meaningful application of the aforesaid logistics, following pre-processing of the text document P (j) is helpful.

- Tokenization – Nonempty sequence of characters, excluding spaces and punctuations from P (j).
- POS tagging – Parts of speech tagging which identifies words as nouns, verbs, adjectives, adverbs, etc.
- Stop Words - Removing function words and connectives like the, a, an, is, for, and, which, etc.
- Grouping – Context based grouping on the basis of POS tagging.
- Stemming - Reducing inflected (like ed, s, es, ing, etc. or sometimes derived) words to their root form.
- Frequency counting of the text-words of P (j) in order to select predominant words.

The group based key-word recommender system is developed after suitable pre-processing and logistics steps. Similar to token based, keywords thus extracted from P (j) by the developed system are subsequently compared with author's list of key words, with a view to judge the effectiveness of the system.

This entire process is illustrated through the block diagram.

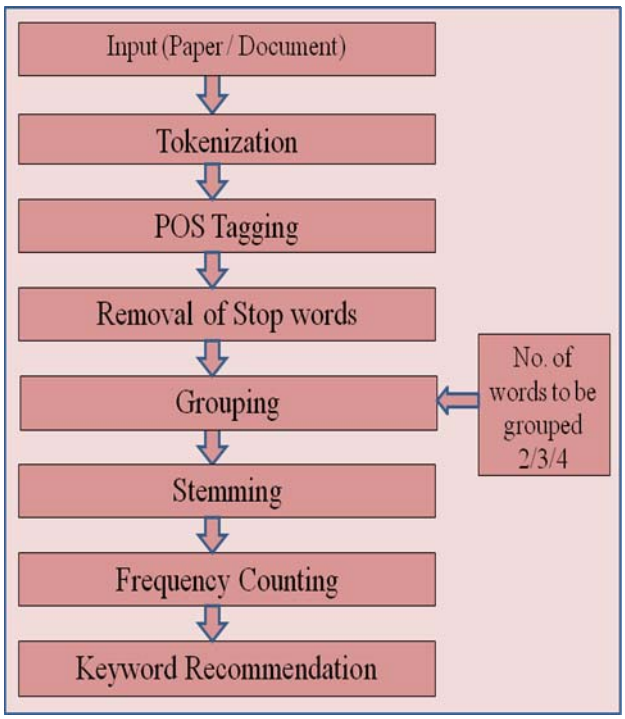

Fig 3. Block Diagram of Group of Token based key-word Recommender System

4.1. Development steps of Group of Token based Key-word Recommender System

A. *Input to the system is a plain text document P (j)*

Research paper softcopy may be used as input. The keywords of author are stored using –

$$K(i, j) = f(P(j))$$

where,

P (j) → The list of any published document in a standard format.
K (i, j) → The list of keywords or index terms
where,
'i' represents the current keyword of jth document.
f → The linking function.

Tokenization Group of token based key-word recommender system uses same tokenization methodology as used in token based key-word recommender system.

B. *POS Tagging*

Part-of-speech tagging (POS tagging or POST) is also called grammatical tagging. It is the process of identification of words as nouns, verbs, adjectives, adverbs, etc. It marks a word in a text as corresponding to a particular part of speech. It is based on both its meaning as well its context(semantics), i.e. relationship with adjacent and related words in a phrase, sentence or paragraph. Keyword Recommender interfaces modified versions of OpenNLP POS tagger. OpenNLP POS tagger consist of coded abbreviations conforming to the scheme of the *Penn TreeBank (*linguistic corpus developed by the University of Pennsylvania).

Example - "It iss a dazzling element built"
POS tagging : "It/PRP is/VBD a/DT dazzling /JJ element /NN built /VBN"

C. *REMOVING STOP WORDS*

These are some of the common short function words, such and, then, a, an, is, then, etc. These words are filtered out prior to key-word extraction from the document in order to improve the efficiency of the system. The Stop words are removed based on the POS tagging done by the POS tagger.

D. *Context based Grouping and Group Filtering*

The system also implements Grouping or shallow parsing which is an improved version of chunking. It is based on POS tagging done on the tokens of the document.

1. Each token is grouped individually with its predecessor and successor token. The number of tokens in the group or chunk can be specified by the user. The system allows minimum two and maximum four tokens in a group. However this can be increased.
2. Once grouping is done, each token occurrence with the associated POS tags in each group is counted and compared.
3. For each token, the groups in which it appears with the most frequent POS tag is considered. Rest of the groups for that particular token are discarded.
4. As a result the context of a word in the text is given suitable weightage for better recommendation of key-words.

E. *STEMMING*

Group of token based key-word recommender system interfaces same stemming tools as used in token based key-word recommender system.

F. *Frequency Counting*

After stemming is done, minimised version of P (j) called PShort(j) or PS (j) is produced, which is further used for frequency counting. PS(j) contains first twenty most frequent groups as recommended key-words.

4.2. CASE STUDY AND RESULTS OF GROUP OF TOKEN BASED KEY-WORD RECOMMENDER SYSTEM

The above automatic extraction process is applied to a set of 30 input research papers P (j). The key-words thus recommended by the system are then compared with the author's list of keywords. The following tables illustrates the comparison statistics of input research paper entitled – "KNOWLEDGE RETRIEVAL USING HYBRID SEMANTIC WEB SEARCH" -

TABLE 1 LIST GIVEN BY AUTHOR

Key-word given by Author	**Number of times keyword**		**Frequency in body of paper**
	Appears in main title	**Appears in sub-title**	
Knowledge Retrieval	1	3	22
Semantic Web Search	1	0	13
User Profiling	0	2	17

TABLE 2 LIST PRODUCED BY THE SYSTEM

Frequent words suggested by the key-word recommender	**Frequency of word**	**Is it a key-word given by author?**	**Appears in Main Title?**
Semantic Web	61		
Knowledge Retriev	22	YES	YES
User Profil	17	YES	
Web Search	14		YES

International Confer	10		
Search Engine	8		
Retrieval Semant	7		
Data Inform	7		
Online Ontolog	6		
Conference Comput	5		
Computer Inform	5		
Information Sci	5		
Science Icci	5		
Information Knowledge	5		
Information Retriev	5		
Rdb Rdf	5		
Information Technolog	4		
Paper Pres	4		
Relational Database	4		
Clustering Us	4		

Observations made from the above case study –

- 66.66% of author's key-words are present in main title.
- 42.85% of recommended keywords are present in author's list of keywords.
- 100% of recommended keywords appear in main title.

4.3. COMPARISION OF KEY-WORDS IN GROUP OF TOKEN BASED KEY-WORD RECOMMENDER SYSTEM

Keywords recommended by the system K (i, j) are then compared with the author's given list of keywords KA (i, j) and the matching percentage is indicated. After repeating the above steps for a large number (30 at present) of cases, the efficacy of the key-word recommender system is established. The system recommends typical groups of tokens as key words to the author of a research paper.

Average observations from 30 research papers –

- 54.75% of the author's keywords are present in main title.
- 70.26% of words in main title are recommended keywords.
- 61.18% of words present in author's list of keywords are recommended keywords.

4.4. *COMBINATION OF RESULTS OF TOKEN BASED AND GROUP OF TOKEN BASED KEY-WORD RECOMMENDER SYSTEM*

Token based and Group of Token based Key-word Recommender systems are used in combination for enhanced recommendations. The following tables illustrate the combination statistics of example input research paper entitled – "KNOWLEDGE RETRIEVAL USING HYBRID SEMANTIC WEB SEARCH" –

Key-word given by Author	Recommended Words and Frequency in Token based Recommender System		Recommended Words and Frequency in Group of Tokens based Recommender System (combination if 2 words)		Recommended Words and Frequency in Group of Tokens based Recommender System (combination if 3 words)	
Knowledge Retrieval	Knowledge	68	Knowledge retriev	22	Knowledge retrieval semant	3
	Retrieve	39				
Semantic Web Search	Semantic	80	Semantic web	61	Semantic web search	13
	Web	85	Web search	14		
	Search	37				
User Profiling	User	32	User profil	17	Clustering user profil	3
	Profiling	19				

V. CONCLUSION AND FUTURE SCOPE

Combination of the two modules Token based and Group of Token based key-word recommender systems can be developed, at the cost of little additional system time. The above combination will improve the overall efficiency of the recommendation system.

Reverse engineering of the key-word recommender system could be used to enhance the chance of getting the desired page in search whereby increasing the efficiency of search engine.

Besides, business documents like tender are also often voluminous. As the result, the selection process based on the document contents is time consuming. The suitable modification of the recommender system developed in the present project can help in deriving the gist of information useful in Management Information System (MIS).

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 characterisation. *Natural Language Engineering*, 5(1):17–44.

[3] Ken Barker and Nadia Cornacchia. 2000. Using noun phrase heads to extract document keyphrases. In *Canadian Conference on AI*.

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

[5] Peter D. Turney. 2000. Learning algorithms for keyphrase extraction. *Information Retrieval*, 2(4):303–336.

[6] Marko Balabanovik And Yoav Shoham (Acm), March 1997/Vol. 40, No. 3. "Content-Based, Collaborative Recommendation".

[7] David A Hull, Rank Xerox Research Centre, 38240, Meylan, France, June 7, 1995. "Stemming Algorithms – A Case Study For Detailed Evaluation".

[8] Sugandha Dani And Dr. C. S. Warnekar, (Ijca), (2011). "Improvising Search Engine By Prioritizing Query String".

[9] Lobur, M. Cad Dept., Lviv Polytech. Nat. Univ., Lviv, Ukraine Romanyuk, A. ; Romanyshyn, M. (Cadsm), 23-25 Feb. 2011, . "Using Nltk For Educational And Scientific Purposes".

[10] Imad A. Al-Sughaiyer, Ibrahim A. Al-Kharashi (Jasict) Vol.55, Issue3, January (2011). "Arabic Morphological Analysis Techniques: A Comprehensive Survey"

[11] Second International Workshop on Education Technology and Computer Science (ETCS), 2010, pages 673 – 675.

[12] W. John Wilbur, Karl Siroktin, Journal Of Information Science February 1992 Vol. 18 No. 1 45-55. "The Automatic Identification Of Stop Words".

[13] Tarik Noryusliza Abdullah, Rosziati Ibrahim, (Iccis) @ 2012 Ieee, "Knowledge Retrieval Using Hybrid Semantic Web Search"

[14] Ju Jiehui, "Chinese Search Engines' PageRank Algorithm and Implementation". Computer Engineering and Design, 2007,28 (7) Page 1632 - 1635.

[15] Tolga Aydn, January, 2009, "Modeling Interestingness Of Streaming Association Rules As A Benefit Maximizing Classification Problem"