International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2014): 5.611

A Collective Study for Document Recommendation Using Textual Conversation Keywords

Snehalata M. Lad[1], Aruna Gupta[2]

[1]M.E (Computer) Department of Computer Engineering, Jayawantrao Sawant College of Engineering, Pune, India
Savitribai Phule Pune University, Pune, Maharashtra, India-411007

[2]M.E(Computer), Associate Professor, Department of Computer Engineering, Jayawantrao Sawant College of Engineering, Pune, India
Savitribai Phule Pune University, Pune, Maharashtra, India-411007

Abstract: ***Keyword Extraction is an important technique in many areas of document processing such as text clustering, text summarization, and text retrieval. Keywords are viewed as the words that represent the topic and the content of the word. This paper addresses, the new technique for keyword extraction from conversations fragment, which can be recommended to the participants to fulfill their information needs without distracting them. A short conversation fragment contains lots of words which can be related to several topics, so keywords are the index terms that contain most important information. In this paper, a survey of keyword extraction technique have been presented that can be applied to extract the keyword that uniquely identified the documents.***

Keyword: Document recommendation, information retrieval, keyword extraction, meeting analysis, topic modeling.

1. Introduction

Keyword Extraction is an important technique in many areas of document processing such as text clustering and summarization, and text retrieval. Keywords are viewed as the words that represent the topic and the content of the word. They can also be used for a divergence of language processing grind such as text categorization and information retrieval. However, most documents cannot be providing keywords. This is especially true for speech documents. There have been many purposes towards extracting the keywords for text domain. In contrast, there is less process on speech transcripts [7]. Keyword extraction from text is a tool commonly used by search engines and indexes alike to quickly categorize and locate specific data based on explicitly or implicitly supplied keywords.

So, this paper presenting basic idea of to select the words from conversation fragment that gives the exact meaning of its content. Keyword extraction from the conversation includes many processes. First process is to convert the conversation in text format; text format can be of pdf or doc file. The next process is to preprocess the document that involves removing the stop words, stem the words. After preprocessing the keywords are extracted using extraction technique. There are many extraction techniques such as statistics approach, linguistic approach, machine learning approach, etc are used to extract the keywords [5]. But in this paper we are using the fuzzy logic to extract the keywords.

So, Eventually this paper presenting an idea of keyword extraction which eventually uses the process by taking input as conversation fragment and performs the operation by using the concept and giving output as recommended document to the participants.

For further proceeding of this paper section II is dedicated for related work, section III gives system architecture and section IV is for conclusion.

2. Related Works

Number of methods has been proposed to automatically extracting keywords from a text, which are also applicable for transcribed conversation. The earliest technique has used word frequencies and TFIDF values to rank words for extraction [1].

1. Applying Graph based Keyword Extraction to Document Retrieval.
In this paper, typically basic parts of VSM are applied to calculate semantic similarity over documents. Therefore, IF-IDF weighting is adopted instead of using raw frequencies and performed length normalization on both queries and target documents. In addition, for calculating the closeness between pseudo documents (queries) and documents traditional cosine similarity is used. However, the vector space dimensionality reduction phase has been omitted to simplify the experiment process.

2. Diverse Keyword Extraction from Conversation.
In this paper, they proposed a new method for keyword extraction that rewards both word similarity, to extract the most representative words, and word diversity, to cover several topics if necessary. They propose to build a topical representation of a conversation fragment, and then to select keywords using topical similarity while also rewarding the diversity of topic coverage, inspired by some summarization methods.

3. Automatic Keyword Extraction from Document Using Conditional Random Field.
In this paper, keywords extraction based on CRF is proposed and implemented. They collected documents from database of ‗Information Center for Social Sciences of RUC', which is available at http://art.zlzx.org/. This method randomly chose 600 academic documents in the field of economics from the database. These Chinese documents are divided into 10 data sets and used 10-fold cross-validation for the CRF model. Each document includes the title, abstract,

keywords, full-text, heading of paragraph or sections, boundaries information of paragraphs or sections, references, etc. These documents have abundant rich linguistics features and are suitable to perform keywords labeling well. Therefore, this is a very interesting work of keywords extraction from documents using CRF model. The number of the annotated keywords of 600 documents ranges from 5 to 10 and the average of annotated keywords is 7.83 per document.

4. An Overview of the Technique Used for Extracting Keywords from Documents.

In this paper, they focus on one speech genre — the multiparty meeting domain. Meeting speech is significantly different from written text and most other speech data. For example, there are typically multiple participants in a meeting, the discussion is not well organized, and the speech is spontaneous and contains disfluencies and ill-formed sentences. It is thus questionable whether to adopt approaches that have been shown before to perform well in written text for automatic keyword extraction in meeting transcripts.

This paper evaluates several different keyword extraction algorithms using the transcripts of the ICSI meeting corpus. Starting from the simple TFIDF baseline, they introduce knowledge sources based on POS filtering, word clustering, and sentence salience score. In addition, they also investigate a graph-based algorithm in order to leverage more global information and reinforcement from summary sentences. They have used different performance measurements: comparing to human annotated keywords using individual F-measures and a weighted score relative to the oracle system performance, and conducting novel human evaluation.

5. Automatic Key phrase Extraction via Topic Decomposition

This paper elaborated new graph-based framework, Topical Page Rank, which incorporates topic information within random walk for key phrase extraction. Experiments on two datasets show that TPR achieves better performance than other baseline methods. They also investigate the influence of various parameters on TPR, which indicates the effectiveness and robustness of the new method.

3. System Architecture

Sr. No	*Paper Name*	*Technique*	*Advantages*	*Disadvantages*
1.	Applying Graph Based Keyword Extraction to Document Retrieval.	Page Rank Algorithm, Vector Space Model (VSP).	1.Grapg-based extraction systems showed better performance over frequency based systems on multiple-theme documents.	1. The word based solution is not effective enough to capture the connective pattern of the terms in the network since it is missing the system clues associated with words stems. 2. Removimg too many terms from the text would artificially increase the value of the idf component of tf-idf as a word may be hardly ever selected as a keyword despite occurring in a large number of documents.
2.	Diverse Keyword extraction from conversation.	Diverse Keyword Extraction Algorithm.	1. Using λ=0.75the Diverse keyword extraction method provides the keyword sets that are judged most representative of the conversation fragment.	1. Setting λ for a new dataset remains an issue, and requires a small development data set.
3.	Automatic Keyword Extraction from Document Using Conditional Random Field.	Conditional Random Field Model.	1. It uses the feature of documents more sufficiently and effectively and keyword extraction can be considered as a string labeling. 2. CRF model outperforms the other machine such as support vector machine, multiple linear regression model etc. in the task of keyword extraction from academic documents.	1. No of Keywords was assigned manually and The words having similar meaning affects precision of the selected models. 2.Problem of ambiguity of the extracted keywords affects the performance of the CRF based keyword extraction.
4.	Unsupervised Approaches for Automatic Keyword Extraction Using Meeting Transcript.	Graph Based Methods.	1. Simple TFIDF based method is very competitive. 2. Adding additional knowledge such as POS and sentence salience score helps improve performance.	1. TFIDF counts the frequency for a particular word, without considering any words that are similar to it in terms of semantic meaning.
5.	Automatic key phrase extraction via topic Decomposition.	Topical page rank.	1. Documents and words can be represented by a mixture of semantic topics. 2. Topical page rank can extract key phrases with high relevance and good coverage which perform other baseline method under various evaluation metrics on two datasets.	1. Learned topics are highly dependent on the learning corpus.

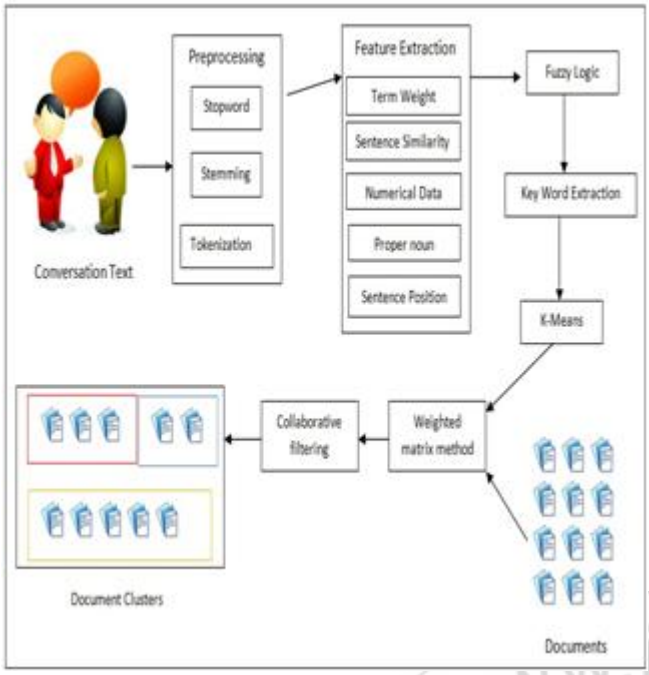

Figure: System Architecture.

4. Conclusion

As this complete paper narrate different methodology on keyword extraction, but none of the methodology are seems to be perfect. So, this paper as bit introduces an idea of extracting keywords which is generated by using fuzzy logic and K-means algorithm on the conversation fragments.

References

[1] Maryam Habibi and Andrei Popescu-Belis, ―Keyword Extraction and Clustering for Document Recommendation in Conversations*”* *IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 23, NO. 4, APRIL 2015.*

[2] Youngsam Kim, Munhyong Kim1, Andrew Cattle, Julia Otmakhova, ―Applying Graph-based Keyword Extraction to Document Retrieval” *International Joint Conference on Natural Language Processing, pages 864–868, Nagoya, Japan, 14-18 October 2013.*

[3] M. Habibi and A. Popescu-Belis, ―Diverse keyword extraction from conversations, ” in *Proc. 51st Annu. Meeting Assoc. Comput. Linguist.*, 2013, pp. 651–657.

[4] C. Zhang, H. Wang, Y. Liu, D. Wu, Y. Liao, and B. Wang, ―Automatic keyword extraction from documents using conditional random fields, ”*J. Comput. Inf. Syst.*, vol. 4, no. 3, pp. 1169–1180, 2008.

[5] Menaka S, Radha N ―An Overview of Techniques Used for Extracting Keywords from Documents” *International Journal of Computer Trends and Technology (IJCTT) – volume 4 Issue 7, pp2321-2325, July 2013.*

[6] Z. Liu, W. Huang, Y. Zheng, and M. Sun, ―Automatic keyphrase extraction via topic decomposition, ” in *Proc. Conf. Empir. Meth. Nat.Lang. Process. (EMNLP’10)*, 2010, pp. 366–376.

[7] F. Liu, D. Pennell, F. Liu, and Y. Liu, ―Unsupervised approaches for automatic keyword extraction using meeting transcripts, ” in *Proc. Annu. Conf. North Amer. Chap. ACL (HLT-NAACL)*, 2009, pp.620–628.

Author Profile

Snehalata M. Lad currently pursing M.E. (Computer Engineering) from Department of Computer Engineering, Jawantrao Sawant College of Engineering, Pune, India. Savitribai Phule Pune University, Pune, Maharashtra, Pune-411007. He received her B.E. (Information Technology) Degree from Bharat Ratna Indira Gandhi Collage Of Engineering, Solapur. Maharashtra, India. Solapur University, Maharashtra, India

Aruna Gupta M.E. (Computer), Associate Professor, Department of Information Technology, Jawantrao Sawant College of Engineering, Pune, India. Savitribai Phule Pune University, Pune, Maharashtra, India-411007. She is awarded with the degree of B.E (Computer) and M. E (Computer).She has around 10 to 12 years of teaching and industrial Experience. She guided many students for the dissertation. She has published many national and international journals in this domain also. Her Research area includes Network Security and WSN.