ISSN No. 0976-5697

Volume 7, No. 4, July - August 2016

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Keyword Extraction from Conversation Text Document and Recommending Document using Fuzzy Logic Based Weight Matrix Method

Snehalata M. Lad
M.E (Computer) Department of Computer Engineering,
Jayawantrao Sawant College of Engineering, Savitribai Phule Pune University,
Pune, Maharashtra, India

Abstract: This paper explores the idea of keyword extraction from conversations, the goal of using these keywords to retrieve, for each short conversation text file, a small number of possibly relevant documents, which can be recommended to the participants. However, even a short conversation contains different types of words, which are absolutely related to several topics; Therefore, it is difficult to infer precisely the information needs of the conversation participants. The existing system proposed a diverse keyword extraction technique which extracts the keyword from the meeting conversation transcripts and recommends the document to the participants. So, in this paper we first propose an algorithm to extract keywords from the output of preprocessing process where string is processed to its basic meaning by following the basic four activities. Then, we propose a feature extraction method to extract multiple topically differentiated queries from this keyword set, in order to maximize the chances of making at least one relevant recommendation to participants. The proposed methods are evaluated in terms of relevance with respect to conversation fragments from the conversation text file. The results shows that our system improves over previous methods that consider only word frequency or topic similarity, and represents a promising solution for a document recommender system to be used in conversations.

Keywords: Document recommendation, keyword extraction, feature extraction, fuzzy logic and weight matrix method.

I. INTRODUCTION

In many areas of document processing keyword extraction is an important technique such as summarization and text clustering, and retrieving text. For representing the topic and the content of the word Keywords are viewed as the words [2]. They can also be used for a divergence of language processing grind such as text categorization and information retrieval. However, most documents keywords are not provided. This is especially true for speech documents. For many purposes keywords are extracted from text. In contrast, there is less process on speech transcripts [7]. Search engines and indexes alike to quickly categorize and locate specific data based on explicitly or implicitly supplied keywords uses the keyword extraction tool which extracts the keyword automatically.

So, this paper presenting basic idea of to extracting the important keywords from conversation fragment that gives the exact meaning of the topic by using the preprocessing method. Keyword extraction from the conversation contains 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 which takes the base form of the word like going become go and so on. After preprocessing the next method is feature extraction technique, where some important features are been extracted from the conversation text document and K-Means is used for clustering which is applied after feature extraction. There are many extraction techniques such as linguistic approach, machine learning approach, statistics approach etc are used to extract the keywords [5]. But in our paper we are using the feature extraction and fuzzy logic based on weight matrix method to extract the keywords and recommending documents to the participants.

So, Eventually this paper presenting an idea of keyword extraction which ultimately uses the process by taking input as conversation text document 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 for problem statement section IV for existing system section V is for System overview and section IV is for conclusion and future scope.

II.RELATED WORKS

Different types of methods have been proposed for extracting keywords from a document, which are also applicable for transcribed conversation. The earliest technique they define word frequencies and TFIDF values to the ranking words for extracting keywords [1].

In graph based Keyword Extraction to Document Retrieval System, typically basic parts of Vector Space Machine are applied to examine semantic similarity over documents. Therefore, IF-IDF weighting is utilized instead of using performed length normalization and raw frequencies on both search queries and resulting documents. In addition, for calculating the strictness between pseudo documents (queries) and documents traditional cosine similarity is used. For clarifying the verification process, the vector space dimensionality reduction process has been omitted. It gives a better performance over frequency based system on multiple documents. While it has the limitation that the word based solution was not effective enough to capture the connective pattern of the terms in the network since it is missing the system clues associated with words steams [2].

A new method for keyword extraction that rewards both word similarities proposed in the diverse keyword extraction technique, to extract the most illustrative words, and word diversity, to cover several topics if necessary. 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, this method is proposed. This system by default uses λ=0.75 .The keyword sets that are judged most representative of the conversation fragment provided by the Diverse keyword extraction method. While setting λ for a new dataset remains an issue, and requires a small development data set[3].

Automatically extracting keywords from Document Using Conditional Random Field proposed a new method for keyword extraction based on CRF. This method are however chooses 600 academic documents in the field of economics from the database. These documents are divided into particular 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 sufficient rich linguistics features and are suitable to perform well keywords identification. Therefore, this is a very interesting work of keywords extraction from documents using CRF model. The number of the illustrated keywords of 600 documents ranges from 5 to 10 and the average of illustrated keywords is 7.83 per document.

It uses the component of documents more sufficiently and effectively and keyword extraction can be considered as string identification. Ambiguity problem of the keyword extraction influence the performance of the CRF based keyword extraction [4].

In [5], they focus on one speech categories the multiparty meeting domain. Meeting speech is extremely different from written text. For example, there are basically numbers of participants in a meeting, the discussion is not well arranged, and the speech is uncontrolled and contains disfluencies and ill-formed sentences. It is thus questionable whether to accept approaches that have been shown before to perform well in written text for automatic keyword extraction in meeting transcripts.

This paper evaluates several types of keyword extraction algorithms using the transcripts of the ICSI meeting corpus. Starting from the simple TF-IDF standards, They introduced knowledge sources depends on POS filtering, sentence salience score and word clustering. In addition, they also study a graph-based algorithm in order to advantage more comprehensive information and column from summary sentences. They have used different performance computation: comparing to human annotated keywords using original F-measures and a weighted score relative to the performance of oracle system, and conducting novel human evaluation.

It uses the additional knowledge such as sentence salience score which helps for improving the system performance. Without considering any words that are similar to it in terms of linguistic meaning, TFIDF counts the frequency for a particular word.

Automatic Key phrase Extraction along Topic Decomposition proposed a structure, Topical Page Rank, which organize topic information within random walk for extracting key phrases. Experiments performed on two datasets shown that TPR achieves improved performance than other standard methods. They also examine the influence of different parameters on TPR, which shown the effectiveness and robustness of the new method. Advantage of this technique is key phrases can extracted by topically Page Rank with high relevance and good coverage which perform other standard method under various evaluation metrics on two datasets[6].

III. PROBLEM DEFINATION

To cope up with ever evolving topic modeling knowledge, need to come over the hurdle of having static terms assigned to the topic. To address this problem we are using the feature extraction and fuzzy logic based weight matrix method. Enriching document recommendation in conversation is a promising approach;

We model fuzzy type search method defines process P= {$P_{r,}$ $F_{e,}$ $F_{l,}$ K, C_f} explained in detail in further section.

IV.EXISTING SYSTEM

There is a problem of extracting keywords from conversations. The existing system proposed a diverse keyword extraction technique The goal this technique is to extract the keywords from the meeting conversation transcripts and a small number of potentially relevant documents, which can be recommended to participants. However, even a short fragment contains a variety of words, which are potentially related to several topics; moreover, using an automatic speech recognition (ASR) system introduces errors among them. Therefore, it is difficult to infer precisely the information needs of the conversation participants [1].

a. just in time retrieval system:

Just-in-time retrieval systems potentially brings a radical change in the process of query-based information retrieval. For detecting the users information need such system continuously monitors the user activities, and pro-actively retrieve relevant information. To achieve this, the systems generally extract implicit queries which are not shown to users from the words that are written or spoken by users during their activities.

b . Diverse Keyword Extraction:

The diverse keyword extraction technique performs the following steps represented in the diagram.

1. Topic Modeling
2. Representation of main topics of the transcripts
3. Diverse keyword selection

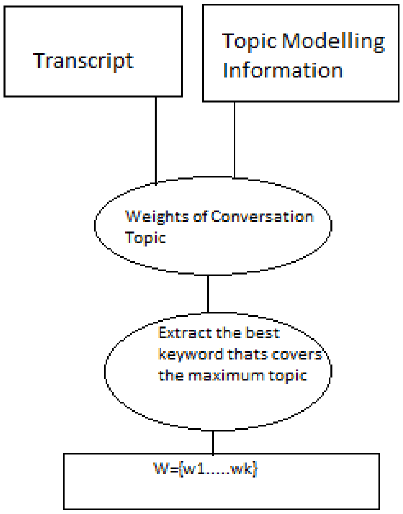
Transcript
Topic Modelling Information
Weights of Conversation Topic
Extract the best keyword thats covers the maximum topic
W={w1.....wk}

Fig1:Diverse Keyword Extraction Method

V. OVERVIEW OF PRAPOSED SYSTEM

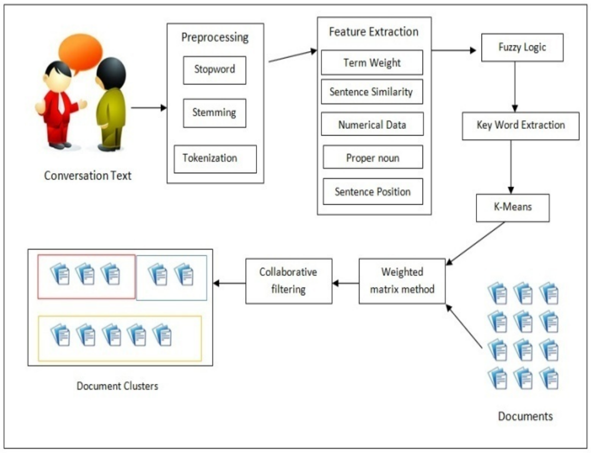

Fig 2.System Architecture

1. Preprocessing.

Preprocessing performs following four important activities:

Step 1: Sentence segmentation perform boundary detection and separating source text into sentence.
Step 2: Special Character and special symbol are replaced with empty character in input document file.
Step 3: Afterward, Stop Words are removed, stop words are the words which seem repeatedly in document but provide fewer meaning in recognizing the important content of the document such as ‘a’, ‘an’, ‘the’, etc..
Step 4: Removing prefixes and suffixes of each word to bring to its base form the word steaming process is performed. For Example goes remain go, going remain go, etc.

Algorithm: Preprocessing

```
Step 0: Start
Step 1: Read string
Step 2: Divide string into records on space and store in a vector V
Step 3:Remove Special symbol
Step 4: Identify stopwords
Step 5: Remove Stopwords
Step 6: Identify steaming substring
Step 7: Replace substring to desired string
Step 8: Concatenate string
```

2 .Feature Extraction

In this process, an attribute vector of feature can be used to characterize each sentence of the document. for attempting the signified data used for their task these features are used as an attribute. We are concentrating on five features for each sentence. Each feature produces a value between ‘0’ and ‘1’. There are five features as follows:

I. Term Weight

For calculating the rank of sentence the frequency of term incidences within a document has frequently been used. The score of a sentence can be intended as the sum of the score of words in the sentence. The score of significant score w_i of word i can be intended by the traditional tf_idf method as follows .We applied this method to tf-isf (Term frequency, Inverse sentence frequency).

ALGORITHM: TERM WEIGHT

```
Step 0: Start
Step 1: Read String
Step 2: Divide main string into words on space and store it in a Vector V
Step 3: Identify the duplicate words in the vector and Remove them
Step 4: for i=0 to N(where N is defined as length of V)
Step 5: for i^th word of N check for its frequency
Step 6: Add frequency in List called L
Step 7: End of for
Step 8: return L
Step 9: Stop
```

II. Sentence to Sentence Similarity

This feature is a similarity surrounded by sentences. For each sentence S, the cosine similarity measure computes the similarity between S and each other sentence with a resulting value between 0 and 1. In sentence si and sj the term weight wi and wj of term t to n are denoted as the vector. The similarity of each sentence couple is intended based on similarity.

ALGORITHM: NOUN DETECTION

```
Step 0: Start
Step 1: Read string
Step 2: Divide main string into words on space and store in a vector V
```

Step 3: Identify the duplicate words in the vector and remove them
Step 4: **for** i=0 to N (Where N is defined as length of V)
Step 5: for i^{th} word of N check for its occurrence in dictionary
Step 6: if present then return
Step 7: else return false
Step 8: Stop

III**.** Numerical Data

The number of numerical data in sentence, sentence that holds numerical data is important and it is most probably included in the document summary. The score for this feature is intended as the proportion of the number of numerical data that arise in sentence over the sentence length.

IV**.** Proper Noun

The sentence that controls more proper nouns (name entity) is necessary and it is most apparently included in the document summary. The score for this feature is proposed as the proportion of the number of proper nouns that arise in sentence over the sentence length.

V**.** Sentence Position

Sentence location in text gives the rank of the sentences ,whether it is the first 5 sentences in the article. This feature can contain several substances such as the location of a sentence in the document, section, and paragraph, etc., suggested the first sentence is highest ranking. The score for this feature: we consider the first 5 sentences in the paragraph. This feature score is intended as the succeeding equation.

3. Fuzzy Logic

The classification of text is based on extraction method of sentence selection. Sentence weighting is one of the methods to get the appropriate sentences to consign some numerical measure of a sentence for the Summary and then select the best ones. Therefore, for acquiring the significant sentences, features score of each sentence that we termed in the prior section are used. In this section, we use method to extract the essential sentences: text classification based on fuzzy logic method. The system consist of the following core Steps:

Step 1: In the Fuzzyfier, crisp inputs are taken, for producing the result of the feature extraction.

Step 2: After Fuzzification, the interpreted engine refers to the rule base containing fuzzy IF-THEN rules.

Step 3: In the last step, we get the final sentence score. In interpreted engine, Definition of fuzzy IF-THEN rules is the most important parts. According to our features criteria the most necessary sentences are extracted from these rules. Example of IF-THEN rules are stated below. IF (NoWordInTitle > 0.81) and (Sentence Length > 0.81) and (TermFreq>0.81) and (SentencePosition > 0.81) and (SentenceSimilarity > 0.81) and (NoProperNoun > 0.81) and (NoSemanticWord > 0.81) and (NumbericalData > 0.81) THEN (Sentence is important). After this process all the document sentences are ranked in a descending order according to their scores. A set of uppermost score sentences are extracted as a document summary.

4. K-Means

Partitioning a group of data points into a small number of clusters is nothing but a clustering. For example, the items in a supermarket are clustered in a group (butter, cheese and milk are grouped in dairy products). Of course this is a qualitative kind of partitioning. A quantitative approach would be to measure certain features of the products, say percentage of milk and others, and products with high percentage of milk would be grouped together. In general, we have n data points X_i, i=1...n that have to be partitioned in k clusters. The goal is to assign a cluster to each data point. K-means is a clustering method that aims to find the positions μ_i, i=1...k of the clusters that minimize the distance from the data points to the cluster.

5. Collaborative Filtering

Collaborative filtering (CF) is one of the techniques, which is used by some recommender systems. In general, collaborative filtering is the filtering process for information or patterns using techniques involving collaboration among multiple agents, viewpoints, data sources, etc. Very large data sets involved in the application of collaborative filtering. collaborative filtering method is generally applied on Many different kinds of data including: sensing and monitoring data, such as in mineral exploration, environmental sensing over large areas or multiple sensors; financial data, such as financial service institutions that integrate many financial sources; or in electronic commerce and web applications where the focus is on user data, etc.

.Collaborative filtering algorithms often require (1) participation of active user (2) an easy way to represent interested user to the system, and (3) algorithms that are able to match people with similar interests.

Typically, the workflow of a collaborative filtering system is:

1. By rating items, user expresses his or her preferences (e.g. books, movies or CDs) of the system. These ratings can be viewed as an approximate representation of the user's interest in the corresponding domain.

© 2015-19, IJARCS All Rights Reserved

2. The system matches this user's ratings against other users' and finds the people with most "similar" tastes.

3. With similar users, the system recommends items that the similar users have rated highly but not yet being rated by this user (presumably the absence of rating is often considered as the unfamiliarity of an item)

A key problem of collaborative filtering is how to combine and weight the preferences of user neighbors. Sometimes, users can immediately rate the recommended items. As a result, the system gains an increasingly accurate representation of user preferences over time.

Mathematical Model

1. S= {} be as system for Keyword Extraction and document recommendation

2. Identify Input as $C=\{C_1, C_2, C_3\ldots..C_n\}$

Where C_n= Conversation Text

3. Identify D as Output i.e. Document Clusters

$S= \{C_n, D\}$

4. Identify Process P

$S= \{C_{n,}, P, D\}$

$P= \{P_{r,} F_{e,} F_{l,} K, C_f\}$

Where P_r = Preprocessing

F_e = Feature Extraction

F_l = Fuzzy Logic

K= K-Means

C_f = Collaborative Filtering

5. $S = \{C_n, P_{r,} F_{e,} N_{d,} N_{u,} F_{l,} C_{f,} D\}$

The union of all subset S gives the final proposed system.

VI.CONCLUSION

In the need of having better topic modeling search and user experience, various techniques related to keyword extraction, indexing and searching are re- viewed in this paper. Numbers of researchers have focused on specific areas of keyword extraction. Some efforts are made towards considering topic modeling search as a homogeneous system and developing frameworks as a whole. Though weight matrix based approach is a promising approach; there is scope for more research in improving keyword extraction technique, users search experience, query formation and processing leveraging the domain knowledge and ontology. Existing search engines can be used in more constructive way to help enrich the document base. We plan to come up with a framework which touches each aspect of topic modeling and tries to improve the performance, user experience and the recommending document itself using weight matrix..

VII. ACKNOWLEDGEMENT

It is a pleasure for me to thank many people who have supported me for completing this paper work. Firstly I would like to thank my Guide, Prof.Aruna Gupta for her support during the entire work. She highlighted the key areas in this topic that helped to get the right content in the topic. I am thankful to our HOD, Principle, Teacher and college to for providing us the facilities and their guidance.

VIII. REFFERENCES

[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. She 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.