ISSN:0975-9646

Document Recommendation Using Keyword Extraction for Meeting Analysis

Kumodini V. Tate[#1], Bhushan R. Nandwalkar[*2]

[#] *PG Student , Department of Computer Engineering, Savitribai Phule Pune University*
Late G. N. Sapkal College of Engineering, Anjneri, Nasik, India

**Assistant Professor, Department of Computer Engineering, Savitribai Phule Pune University*
Late G. N. Sapkal College of Engineering, Anjneri, Nasik, India

***Abstract*— The purpose of meetings is to assist direct communication between participants. Manuscript plays an important part in meetings. Manuscripts contain details that are currently discussed, but they are not necessarily at hand. In this topic the technique known as keyword extraction and clustering is overviewed which naturally recommend the documents that are related to users' current activities for an on-going discussion. When users join in a meeting, their informational necessities can be displayed as keywords that can be extracted from text based conversations and documents. These keywords then ordered into subsections and can be matched to recommend relevant document to the user. The importance of the proposed method can be measured by comparing the method with Fisher manual transcripts and AMI ASR transcripts.**

***Keywords*— Document recommendation, information retrieval, keyword extraction, meeting analysis, topic modelling.**

I. INTRODUCTION

Unlimited amount of information is present with the humans in the form of documents, databases, or multimedia. This information is accessed using suitable search engines, but even when search engines are available, users always do not go for a search, because they are very much busy in their current activity. This system makes use of Just- In Time Information-Retrieval system.

Just-In-Time-Information [1] is software that retrieves & presents information based on person's local environment. It continuously guards person's environment & presents information that is useful to user. When the users activities are written in text in online textual chat based meeting, users information needs can be built as implicit queries constructed in the background from the words chat in real time chat based conversation.

These implicit queries are used for recommended information retrieval in the form of documents from web or local knowledgebase. Afterwards user can refer those documents of their interest. This concept focuses the formulation of implicit queries to a just-in-time-retrieval system for use in real-time textual conversation on chat window. In opposite to queries made on search engines, this system constructs implicit queries from textual conversational input, having more numbers of words than actual query as an input to the system. Therefore, goal of this system is extraction of relevant & diverse set of keywords & to cluster the extracted keywords as per topic-specific queries ranked by importance & sample of results from the queries is presented to user. This system introduces novel keyword extraction technique from textual chat conversation output maximizing the coverage of potential information needs of user & reduces the numbers of irrelevant words. Once keyword set is extracted then next phase is clustering of keywords to construct several topically disjoint queries run separately give better precision than large topically ad joint query. Recommendations to users are the results finally merged into a ranked set.

The goal of keyword extraction from conversations is to provide a set of words that are representative of the semantic content of the conversation. Therefore the aim is to find set of keywords, clustering of keywords and present result of this query to users in the form of documents. Mainly topic-based clustering is used to reduce the chances of including ASR errors into the queries. The focus of this is on formulating implicit queries to a just-in-time-retrieval system for use in meeting rooms. It is important that the keyword set preserves the diversity of topics from the conversation. While the first keyword extraction methods ignored topicality as they were based on word frequencies, more recent methods have considered topic modeling factors for keyword extraction, but without specifically setting a topic diversity constraint, which is important for naturally occurring conversations [3].

Consider scenario of meeting where documents related to meeting discussion are already informed to participants of meeting. Due to some of the reasons participants does not have sufficient time to search that contents on the internet or on any other source of information. During meeting to find information related to some point is very difficult without interrupting the discussion flow. This problem occurs most of the time in meeting. To fulfill the information needs of participants some systems must be developed which will take conversation as query and give related documents to that without the direct interaction of participants to the system. Relevance and diversity of documents can be modeled at three levels:

- While extracting queries
- Building one or several implicit queries
- Re-ranking the results of queries[4]

II. Related Work

Just-in-time document retrieval systems have been designed to recommend to their users documents which are potentially relevant to their activities, e.g. individual users authoring documents or browsing various repositories, or small groups holding business or private meetings (Hart and Graham, 1997; Rhodes and Maes, 2000; Popescu-Belis et al., 2008). When using a document recommender system, people are generally unwilling to examine a large number of recommended documents, mainly because this would distract them from their main activity. Several solutions to this problem have been proposed.

For instance, the Watson document recommender system (Budzik and Hammond, 2000), designed for reading or writing activities, clustered the document results and selected from each cluster the best representative to generate a list of recommendations. Clustering results is not suitable for our application where the mixture of topics in a single query will degrade the document results aimed to be clustered (Bhogal et al., 2007; Carpineto and Romano, 2012), and consequently may have a damaging effect on the clusters' representatives. The second part of the method, which selected the best representative of the clusters in the final document list can be helpful; however, its effectiveness relies on having clusters with the same level of importance (Wu and McClean, 2007).

Many studies in information retrieval addressed the problem of diverse ranking, which can be stated as a tradeoff between finding relevant versus diverse information (Robertson, 1997). The existing diverse ranking proposals differ in their diversifying policies and definitions, which can be categorized into implicit methods (Carbonell and Goldstein, 1998; Zhai et al., 2003; Radlinski and Dumais, 2006; Wang and Zhu, 2009) or explicit ones (Agrawal et al., 2009; Carterette and Chandar, 2009; Santos et al., 2010; Vargas et al., 2012). The implicit approaches assume that similar documents will cover similar aspects of a query, and have to be demoted in the ranking to promote relative novelty and reduce overall redundancy.

In one of the earliest approaches, Carbonell and Goldstein (1998) introduced Maximal Marginal Relevance (MMR) to re-rank documents based on a tradeoff between the relevance of document results and relative novelty as a measure of diversity. MMR was also used by Radlinski and Dumais (2006) to re-rank results from a query set which is generated for a user query and represents a variety of potential user intents.

Instead of implicitly accounting for the aspects covered by each document, another option is to explicitly model these aspects within the diversification approach. Agrawal et al. (2009) introduced a sub modular objective function to minimize the probability of average user dissatisfaction by assuming taxonomy of information and modeling user query aspects at the topical level of this taxonomy. Alternatively, Santos et al. (2010) proposed another sub modular objective function to maximize coverage and minimize redundancy with respect to query aspects modeled in a keyword-based representation form instead of a predefined taxonomy.

In our case, the recommender system for conversational environments requires diversity in the results of multiple topically-separated queries, rather than of a single ambiguous query. Therefore, a new approach will be proposed, and will be compared in particular to a version of the explicit diversification approach (Santos et al., 2010) adapted to our problem.

III. Existing System

In the existing system, human are enclosed by an unmatched wealth of information, available as documents, databases, or multimedia resources. Access to this information is hardened by the accessibility of suitable search engines.

A. Drawbacks of Existing System

In general, users participate in a meetings, their information wants to be modeled as implicit queries that are constructed in the background from the pronounced words, obtained during manual recognition. These explicit queries are used to get back and suggest documents from the Web or a local repository, which users can choose to inspect in more detail if they find them.

IV. Proposed System

Then to overcome the existing system drawbacks propose a method to obtain multiple topically separated queries from this keyword set, in order to make best use of the chances of making at least one appropriate recommendation when using these queries to search. The proposed methods are evaluated in terms of significance with respect to conversation rated by several human judges. The scores show that our proposal improves more than previous methods that consider only word frequency or topic similarity, and represents a capable solution for a document recommended system to be used in conversations.

B. Algorithm

1. Keyword Extraction.

We advise to take advantage of topic modelling techniques to build a topical representation of a discussion part, and then select content words as keywords by using relevant relationship, while also fulfilling the reporting of a various range of subjects, motivated by recent summarization methods. The benefit of diverse keyword extraction is that the coverage of the main subjects of the discussion part is maximized.

The benefit of diverse keyword extraction is that the coverage of the main topics of the conversation fragment is maximized. The future method for diverse keyword extraction proceeds in three steps,

1. Used to represent the division of the abstract subject for each word.
2. These topic models are used to determine weights for the abstract topics in each conversation fragment represented by
3. The keyword list W = {w1, w2……wk}. Which covers a maximum number of the most important topics is

preferred by rewarding range, using a unique algorithm introduced in this part.

2. Metaphone Rules.

So to extract the relevant keywords from transcript, we are applying the Metaphone rules (shown in table I)to encode the large word into small size and by using these rules we can save our lots of time to search large words easily. For instance, "enough" word is converted like "enf".Because when we pronounce enough, 'gh' pronounced as 'F'. Benefit of using Metaphone rules we reduced processing time.

TABLE I Metaphone Rules

1.	Drop duplicate adjacent letters, except for C.
2.	If the word begins with 'KN', 'GN', 'PN', 'AE', 'WR', drop the first letter.
3.	Drop 'B' if after 'M' at the end of the word.
4.	'C' transforms to 'X' if followed by 'IA' or 'H' (unless in latter case, it is part of '-SCH-', in which case it transforms to 'K'). 'C' transforms to 'S' if followed by 'I', 'E', or 'Y'. Otherwise, 'C' transforms to 'K'.
	'D' transforms to 'J' if followed by 'GE', 'GY', or 'GI'. Otherwise, 'D' transforms to 'T'.
6.	Drop 'G' if followed by 'H' and 'H' is not at the end or before a vowel. Drop 'G' if followed by 'N' or 'NED' and is at the end.
7.	'G' transforms to 'J' if before 'I', 'E', or 'Y', and it is not in 'GG'. Otherwise, 'G' transforms to 'K'.
8.	Drop 'H' if after vowel and not before a vowel.
9.	'CK' transforms to 'K'.
10.	'PH' transforms to 'F'.
11.	'Q' transforms to 'K'.
12.	'S' transforms to 'X' if followed by 'H', 'IO', or 'IA'.
13.	T' transforms to 'X' if followed by 'IA' or 'IO'. 'TH' transforms to '0'. Drop 'T' if followed by 'CH'.
14.	'V' transforms to 'F'.
15.	'WH' transforms to 'W' if at the beginning. Drop 'W' if not followed by a vowel.
16.	'X' transforms to 'S' if at the beginning. Otherwise, 'X' transforms to 'KS'.
17.	Drop 'Y' if not followed by a vowel.
18.	'Z' transforms to 'S'.
19.	Drop all vowels unless it is the beginning.

Our Proposed System architecture is shown below in figure1.

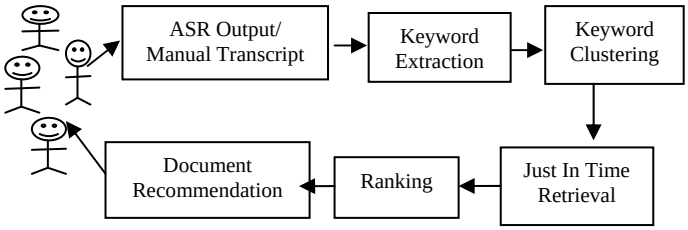

Fig1.Proposed System Architecture

C. Keyword Clustering

The different set of extracted keywords is measured to denote the possible information needs of the applicants to a discussion, in terms of the ideas and topics that are declared in the discussion. To maintain the variety of topics alive in the keyword set, and to decrease the noisy result of each data need on the others, this set must be divided into several topically-disjoint subsets. Each subset corresponds then to an implicit query that will be sent to a document recovery system. These subsets are obtained by clustering topically-similar keywords, as follows. Clusters of keywords are constructed by ranking keywords for each main topic of the fragment. The keywords are ordered for each topic by decreasing values of .p(z|w) Moreover, in each cluster, only the keywords with a .p(z|w) value higher than a threshold are kept for each topic z.

D. From Keywords to Document Recommendations

As a first impression, one implicit query can be arranged for each discussion part by using as a query all keywords special by the various keyword removal techniques. However, to improve the retrieval results, multiple implicit queries can be formulated for each discussion part, with the keywords of each cluster from the before fragment. In tests with only one implicit query per discussion fragment, the document results parallel to each discussion fragment were arranged by selecting the first document retrieval results of the implicit query.

E. Advantages of Proposed System

To maintain multiple hypotheses about user's information need. To present a small sample of recommendations based on the most likely ones. Retrieving of documents by keyword query is faster and Clustering of documents by multi-key word similarity.

V. Result and Analysis

We conduct experiment on five transcripts. The results are based on the number of words in the transcripts. With different transcripts we obtained different number of keyword. So, the results depend on number of words in the transcript.

Here in fig 2. we input a transcript of different word length to existing as well as proposed system and observed the keywords extracted by both the system, Where we consider the keywords extracted in percentage(%) .We noticed that proposed system extracts more keywords for different transcript input.

TABLE II Accuracy Computation

	Existing Accuracy(%)	Proposed Accuracy (%)
1	78	85
2	75	79
3	72	78
4	55	65
5	60	68

The fig 2 shown below shows the keyword extraction accuracy by the Proposed and existing system, where we consider the different transcripts. From the fig 2 we can see that results remain constant.

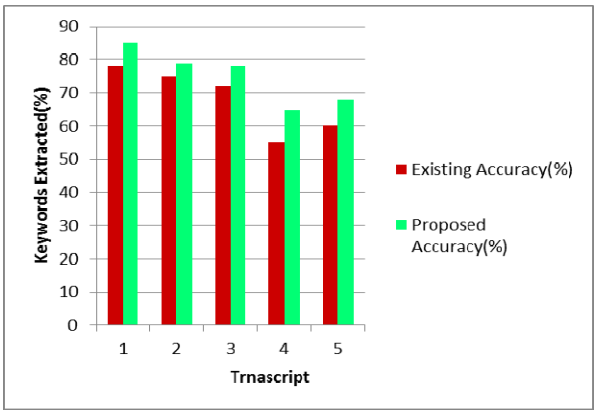

Fig2.Graphical Representation of Keyword Extraction From different transcripts

VI. CONCLUSION

We have considered a particular form of Metaphone rules intended for conversational environments, in which they recommend to users documents that are relevant to their information needs. We focused on modelling the user's needs. We focused on modelling the user's information needs by deriving implicit queries from short discussion fragments. These queries are based on sets of keywords extracted from the conversation. We have proposed a novel diverse keyword extraction technique which covers the maximal number of important topics in a part. Then, to reduction the loud effect on queries of the mixture of topics in a keyword set, we proposed a clustering technique to divide the set of keywords into smaller topically-independent subsets constituting implicit queries.

REFERENCES

[1] Maryam Habibi and Andrei Popescu-Beli,"Keyword extraction and clustering for Documents Recommendation in Conversation" *in Proc. IEEE/ACM transaction on audio, speech, and language processing,* VOL.23, NO.4 April 2015.

[2] M. Habibi and A. Popescu-Belis, "Enforcing topic diversity in a document recommender for conversations," *in Proc. 25th Int. onf.Comput.Linguist. (Coling)*, pp. 588–599, 2014.

[3] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Inf. Process. Manage. J.*, vol. 24, no. 5, pp. 513–523, 1988.

[4] S. Ye, T.-S. Chua, M.-Y.Kan, and L. Qiu, "Document concept lattice for text understanding and summarization," *Inf. Process. Manage.*, vol.43, no. 6, pp. 1643–1662, 2007.

[5] A. Csomai and R. Mihalcea, "Linking educational materials to Encyclopedic knowledge," *in Proc. Conf. Artif.Intell.Educat.:BuildingTechnol. Rich Learn. Contexts That Work*, 2007, pp. 557–559.

[6] D. Harwath and T. J. Hazen, "Topic identification based extrinsic evaluation of summarization techniques applied to conversational speech,"*inProc. Int. Conf. Acoust.*, Speech, Signal Process. (ICASSP) ,pp. 5073–5076,.

[7] A. Popescu-Belis, E. Boertjes, J. Kilgour, P. Poller, S. Castronovo, T.Wilson, A. Jaimes, and J. Carletta, "The AMIDA automatic content linking device: Just-in-time document retrieval in meetings," *in Proc.5th Workshop Mach. Learn. Multimodal Interact. (MLMI)*, 2008,pp.272–283.