ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Special Issue 10, March 2016

# KEYWORD EXTRACTION AND CUMULATION FOR SCRIPT SANCTION IN CONFABULATION

*P.S.Indhumathi **Mrs.A.Kavidha Inst.Of Road and Technology,Erode

**ABSTRACT- The problem of keyword extraction from conversations, with the goal of using these keywords to retrieve, for each short conversation fragment, it contains a variety of words. Using an Automatic Speech Recognition (ASR) system introduces errors among them. An algorithm to extract keywords from the output of an ASR system, which makes use of topic modeling techniques and of a sub modular reward function which favors diversity in the keyword set, to match the potential diversity of topics and reduce ASR noise. An Just In Time Retrieval(JIT) system, which displays whether the keyword is present in a document and how many times it is present in a document, it shows the position of the keyword. Then, a method Incremental Key Index Fast Search(IKIFS) is to extract the documents based upon the filtering option. If the user need on particular format, this system will derived only that selected files. It provides high performance, high availability, and easy scalability.**

***General Terms: Keyword Extraction, Topic Modeling, Information Retrieval, Document Recommendation.***

## 1. INTRODUCTION

The problem of keyword extraction from conversations, with the goal of using these keywords to retrieve, for each short conversation fragment, 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. An algorithm to extract keywords from the output of an ASR system (or a manual transcript for testing), which makes use of topic modeling techniques and of a sub modular reward function which favors diversity in the keyword set, to match the potential diversity of topics and reduce ASR noise.

An algorithm to extract keywords from the output of an JIT Retrieval system (or a manual transcript for testing), which displays whether the keyword is present in a document and how many times it is present in a document, it shows the position of the keyword. Then, a method IKIFS(Incremental Key Index Fast Search) is to extract the documents based upon the filtering option ,Not all the documents will be derived only the user required document will be extracted,if the user need on particular format (Pdf,text,Word etc) this system will derived only that selected files. Then the extracted documents will be stored in MONGO DB. It is a document database that provides high performance, high availability, and easy scalability.

### Definitions and concepts

Humans are surrounded by an unprecedented wealth of information, available as documents, databases, or multimedia resources. Access to this information is conditioned by the availability of suitable search engines, but even when these are available, users often do not initiate a search, because their current activity does not allow them to do so, or because they are not aware that relevant information is available.

The perspective of Just-In-Time Retrieval,which answers this shortcoming by spontaneously recommending documents that are related to users'

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Special Issue 10, March 2016

current activities. When these activities are mainly conversational, for instance when users participate in a meeting, their information needs can be modeled as implicit queries that are constructed

in the background from the pronounced words, obtained through real-time automatic speech recognition (ASR). These implicit queries are used to retrieve and recommend documents.

**Diverse Keyword Extraction**

To take advantage of topic modeling techniques to build a topical representation of a conversation fragment, and then select content words as keywords by using topical similarity, while also rewarding the coverage of a diverse range of topics, inspired by recent summarization methods. The benefit of diverse keyword extraction is that the coverage of the main topics of the conversation fragment is maximized. Moreover, in order to cover more topics, the proposed algorithm will select a smaller number of keywords from each topic. The words which are in reality ASR noise can create a main topic in the fragment,then the algorithm will choose a smaller number of these noisy keywords compared to algorithms which ignore diversity.

**Keyword Clustering**

The diverse set of extracted keywords is considered to represent the possible information needs of the participants to a conversation, in terms of the notions and topics that are mentioned in the conversation. To maintain the diversity of topics embodied in the keyword set, and to reduce the noisy effect of each information need on the others, this set must be split into several topically disjoint subsets. Each subset corresponds then to an implicit query that will be sent to a document retrieval system. These subsets are obtained by clustering topically-similar keywords.

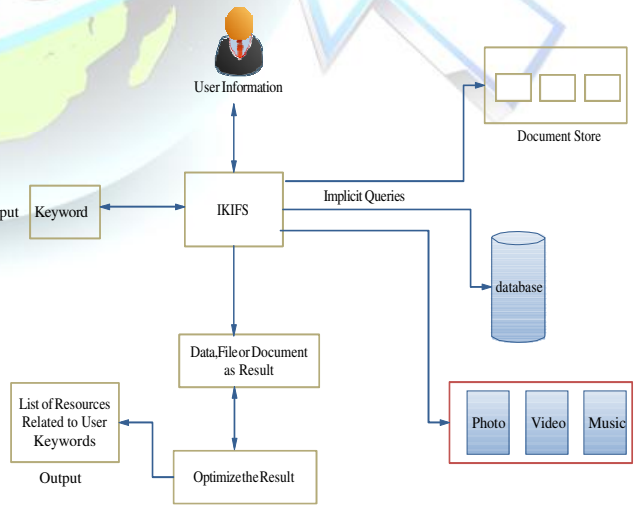
**From Keywords to Document Recommendations**

One implicit query can be prepared for each conversation fragment by using as a query all keywords selected by the diverse keyword extraction technique. However, to improve the retrieval results, multiple implicit queries can be formulated for each conversation fragment, with the keywords of each cluster.The recommendation lists were prepared by selecting the first document retrieval results of each implicit query and then ranking documents based on the topical similarity of their corresponding queries to the conversation fragment.

## 2.PROPOSED SYSTEM

The Multi viewpoint it is used to find the similarity between the multiple Documents. These are used to Compare the Documents and find the Similarities .To Compare the documents first we need to submit two Documents and then document has been compared by line by line And show the similarity in that line what it have compared. CSMTP (CONCEPT BASED SIMILARITY MEASURE FOR TEXT PROCESSING).These Algorithm Selects the terms from the testing documents Generates the terms from the documents, selects the appropriate feature and Calculates the similarity measure based on the term and its respective concepts. It is compare the two Document to provide matching result and clustering the files.

**System Architecture**




## 3.JUST-IN-TIME RETRIEVAL

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Special Issue 10, March 2016

Just-in-time retrieval systems have the potential to

bring a radical change in the process of query-based information retrieval. Such systems continuously monitor users' activities to detect information needs, and pro-actively retrieve relevant information. To achieve this, the systems generally extract implicit queries from the words that are written or spoken by users during their activities. In this section, we review existing just-in-time-retrieval systems and methods used by them for query formulation. In particular, we will introduce our Automatic Content Linking Device (ACLD) a just-in-time document recommendation system for meetings, for which the methods proposed in this paper are intended.

## 4.QUERY FORMULATION IN JUST-IN-TIME RETRIEVAL SYSTEMS

One of the first systems for document recommendation, re-ferred to as query-free search, was the Fixit system [9], an assis-tant to an expert diagnostic system for the products of a specific company (fax machines and copiers). Fixit monitored the state of the user's interaction with the diagnostic system, in terms of the positions in a belief network built from the relations among symptoms and faults, and ran background searches on a data-base of maintenance manuals to provide additional support in-formation related to the current state.

Other real-time assistants are conversational: they interact with users to answer their explicit information needs or to provide recommendations based on their conversation. For instance, Ada and Grace[1] are twin virtual museum guides [17], which interact with visitors to answer their questions, suggest exhibits, or explain the technology that makes them work. A collaborative tourist information retrieval system [18], [19] interacts with tourists to provide travel information such as weather conditions, attractive sites, holidays, and transporta-tion, in order to improve their travel plans. Christo Ananth et al. [7] proposed a system in which the cross-diamond search algorithm employs two diamond search patterns (a large and small) and a halfway-stop technique. It finds small motion vectors with fewer search points than the DS algorithm while maintaining similar or even better search quality. The efficient Three Step Search (E3SS) algorithm requires less computation and performs better in terms of PSNR. Modified objected block-base vector search algorithm (MOBS) fully utilizes the correlations existing in motion vectors to reduce the computations. Fast Objected - Base Efficient (FOBE) Three Step Search algorithm combines E3SS and MOBS. By combining these two existing algorithms CDS and MOBS, a new algorithm is proposed with reduced computational complexity without degradation in quality.

The keyword extraction method proposed here accounts for a diversity of hypothesized topics in a discus-sion, and is accompanied by a clustering technique that formu-lates several topically-separated queries.

## 5. KEYWORD EXTRACTION METHOD

Numerous methods have been proposed to automaticallyextractkeywordsfromatext,andareapplicablealsototranscribed conversations. The earliest techniques have used word frequen-cies [2] and TFIDF values [3], [23] to rank words for extrac-tion. Alternatively, words have been ranked by counting pair-wise word co-occurrence frequencies [24]. These approaches do not consider word meaning, so they may ignore low-frequency words which together indicate a highly- lient topic.For instance, the words 'car', 'wheel', 'seat', and 'passenger' occur-ring together indicate that automobiles are a salient topic even if each word is not itself frequent [25].

## 6.CONCLUSION

We have considered a particular form of just-in-time re-trieval systems intended for conversational environments, in which they recommend to users documents that are relevant to their information needs. 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 fragment. Then, to reduce the noisy 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.The diverse keyword extraction technique with existing methods, based on word frequency or topical similarity, in terms of the representativeness of the keywords and the relevance of retrieved documents. Our current goals are to process also explicit queries, and to rank document results with the objective of maximizing the coverage of all the information needs.

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Special Issue 10, March 2016

**REFERENCES**

[1]. M. Habibi and A. Popescu-Belis, "Enforcing topic diversity in a document recommender for conversations," in Proc. 25th Int. Conf.Comput. Linguist. (Coling), 2014, pp. 588–599.
[2]. H. P. Luhn, "A statistical approach to mechanized encoding and searching of literary Information," IBM J. Res. Develop., vol. 1, no. 4, pp. 309–317, 1957.
[3].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.
[4].A. Csomai and R. Mihalcea, "Linking educational materials to encyclopedic knowledge," in Proc. Conf. Artif. Intell. Educat.: Building Technol. Rich Learn. Contexts That Work, 2007, pp. 557–559.
[5].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), 2012, pp. 5073– 5076.
[6] . B. J. Rhodes and P. Maes, "Just-in-time information retrieval agents,"IBM Syst. J., vol. 39, no. 3.4, pp. 685–704, 2000.
[7]. Christo Ananth, A.Sujitha Nandhini, A.Subha Shree, S.V.Ramyaa, J.Princess, "Fobe Algorithm for Video Processing", International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering (IJAREEIE), Vol. 3, Issue 3,March 2014 , pp 7569-7574
[8]. 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.
[9].T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality,"in Proc. 26th Annu. Conf. Neural Inf. Process. Syst. (NIPS),2013, pp. 3111–3119.
[10]. M. Habibi and A. Popescu-Belis, "Diverse keyword extraction fromconversations," in Proc. 51st Annu. Meeting Assoc. Comput. Linguist.,2013, pp. 651–657
[11]. T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimationof word representations in vector space," in Preprint arXiv:1301.3781,2013.