KEYWORD EXTRACTION AND CLUSTERING FOR DOCUMENT RECOMMENDATION USING JUST IN TIME RETRIEVAL TECHNIQUE

Mr.R.Janarthanan[1] B.E.,M.E.,MBA.,(Ph.D)

HOD/CSE

hodcsetjsengg@gmail.com

Anitha.v[2] anithaalwin18@gmail.com

Jansirani.s[3] elshajansimman33@gmail.com

Kanimozhi.m[4] kanimozhimari40@gmail.com

[2][3][4] UG students

Department of Computer science and engineering

TJS engineering college

Abstract:

This paper addresses the problem of keyword extraction from conversations, with the goal of using this 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 error among them. Therefore, it is difficult to inter precisely the information needs of the conversation participants. We first proposed an algorithm to extract keywords from the output of an ASR system(or a manual transcript for testing), which make use of topic modeling techniques and of a sub modular reward function which favors diversity in the keyword set to match potential diversity of topics an reduce ASR noise. Then, we propose a method to derive multiple topically separated queries from this keyword set, in order to maximize the chances of making at least one relevant recommendation when using these queries to search over the English Wikipedia. The proposed methods are evaluated in terms of relevance with respect to conversation fragments from the FISHER, AMI and ELEA conversational corpora rated by several human judges. The scores show that our proposal improves over previous methods that consider only word frequency are topic similarity, and represents a promising solution for a document recommender system to be used in conversations.

INTRODUCTION

The focus of this paper is system IKIFS (Incremental Key Index Fast Search) and UFS (User Frequency Suggestion). We aim at extracting some keywords, cluster them into topic-specific queries ranked by importance the diversity of keywords increases the chances that at least one of the recommended documents answers a need for information. This paper addresses the problem of keyword extraction from conversations, with the goal of using these keywords to retrieve, for each short conversation fragment, which can be recommended to participants. We first proposed 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. Then, we propose a method IKIFS 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. Then the extracted documents will be stored in MONGO DB. It is a document database that provides high performance, high availability, and easy scalability.

RELATED WORKS:

In the existing system Just-in-time retrieval systems have the potential to bring a radical change in the retrieval.JIT process of query-based information 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 highlights the position of the keyword.It retrieves the document based on the query,a particular part of a document is retrieved to answered that query.In this section. It reads the full document only a part relevant to a query will be displayed, we review existing just-in-time-retrieval systems and methods used by them for query formulation.

LIMITATIONS:

- This system will extract only a particular part of the document based on query request.
- Time consumption will be high
- Relevant document will not be displayed.
- It will cover a few key word from Each topic.
- It retrieves both relevant and irrelevant part of the documents
- It displays more uncertain words in the documents and show ambiguous questions

PROPOSED SYSTEM:

We have proposed a IKIFS(Incremental Key Index Fast Search) to derived a documents based upon the user requirements once the keyword is given for a search this system will ask for filtering option here the user needs to filter what type of file is to be extracted like pdf or Word or video or music etc.This filtering option is used to extract the file in a particular format so that search will be fast and then time consumption will be less. the user will upload a new file which user wants to add in a storage area before uploading a file user needs to assigned a key for a particular file so that by using that key particular file will be retrieved again. This file will be uploaded in a MONGO DB.It is a cross-platform, document database that provides high performance, high availability, and easy scalability.It focuses on flexibility, power, speed, and ease of use.

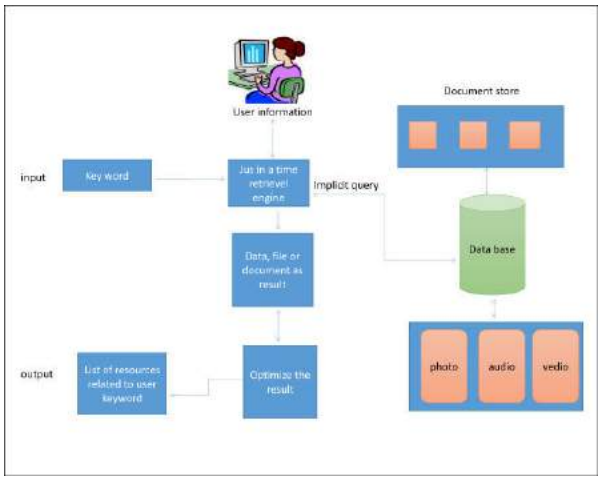
It works on concept of collection and document. Database is a physical container for collections. Each database gets its own set of files on the file system. A single MongoDB server typically has multiple databases.Collection is a group of MongoDB documents. It is the equivalent of an RDBMS table. A collection exists within a single database. Collections do not enforce a schema. Documents within a collection can have different fields. Typically, all documents in a collection are of similar or related purpose.A document is a set of key-value pairs. Documents have dynamic schema. Dynamic schema means that documents in the same collection.

MODULES:

- Keyword Extraction
  - How to give keyword to search
- Implicit Query Construction
  - Fetching from different sources like related data and intensive.
- just-in-time Retrieval
  - It can be a controller which call the other modules

- Results Reranking(Optimization) Or Document Recommender
  - Consolidation from different modules.

ARCHITECTURE DIAGRAM:



ALGORITHM:

IKIFS (Incremental Key Index Fast Search)

```
Key wordWrite()
   throws FileIOException
 Database Connection= keyword() + path();
intcs = conection.updataQuery(key , fpath);
foreachavailablefile: targetfile
    do Write(path, E, count);
  while (key <uKeyword)
     replied = receivefile(path);
    if (afile== keywore) break;
repliedfile.add(file);
 if (afile.size<getcount)
   throw new Exception();
int check = checksum(keyword, path);
return new keywordWrite (filepath FP, Keyword E);
```

DESCRIPTION:

Fisher, AMI & ELEA (Emergent LEader Analysis) are the Corporation. These corporations examined our proposals via Amazon Mechanical Turk platform. They compare our keyword relevance & Put some calculations on the average & Quality of implicit queries are analyzed here.

We compared 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. These were judged by human raters recruited via the Amazon Mechanical Turk crowdsourcing platform.

Diversity Keyword Extraction algorithm

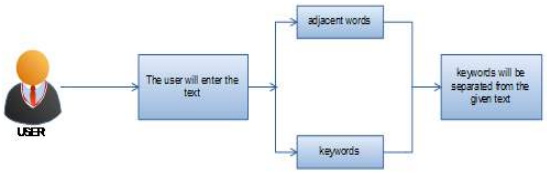
In Diversity algorithm, it checks for the whole document, whenever we are searching for the particular or Individual keyword. If we want to search another keyword. it will search the whole document one more time. This is the major drawback in diversity algorithm. The keyword extraction method could be improved by considering no of words addition to individual words only, but this requires some adaptation of the entire processing chain.

So we added our concept of Mongo- DB, this concept helps to add additional attributes according to the future requirements. So we replaced the Diversity Keyword Extraction algorithm by **IKIFS (Incremental Key Index Fast Search)** algorithm. It is to extract the documents based upon the filtering option,not all the documents will be derived only the user

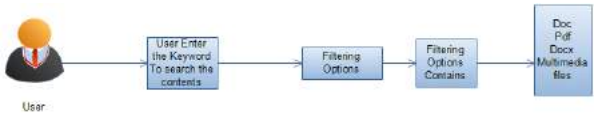
required document will be extracted, if the user need on particular format (Pdf,text,Word etc..) this system will be derived only that selected files

FUNCTIONALITY DESCRIPTION:

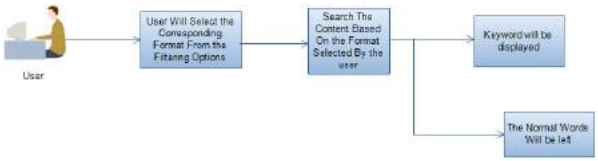
1. Keyword Extract



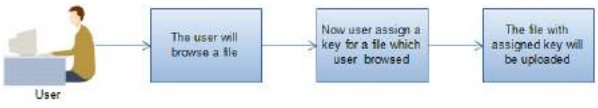
2. Filtering System



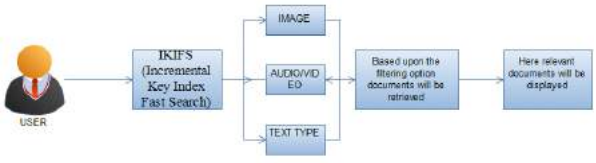
3. Inspection System



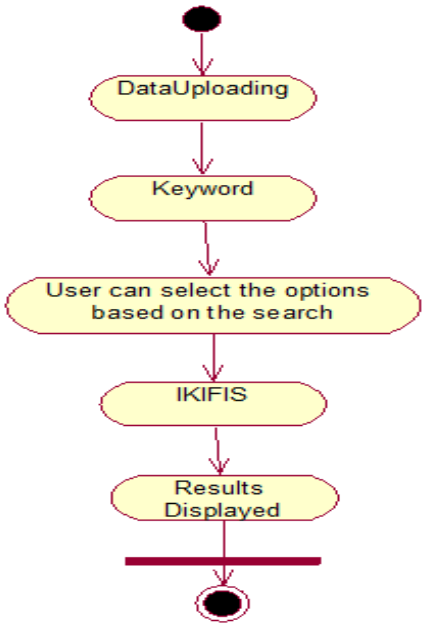
4. Personalized Search



5. Search Based on Type



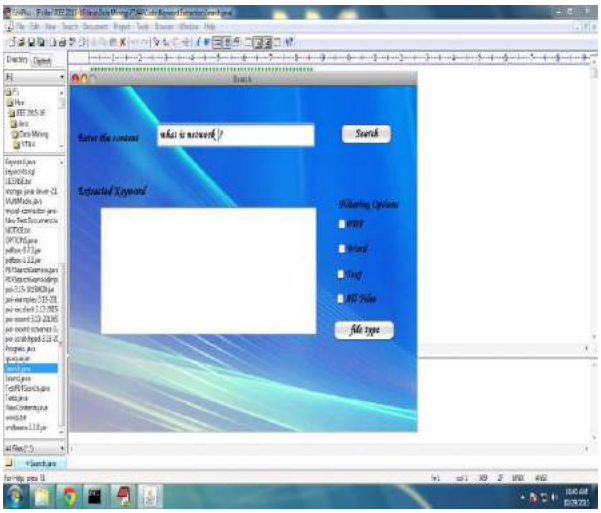
ACTIVITY DIAGRAM:

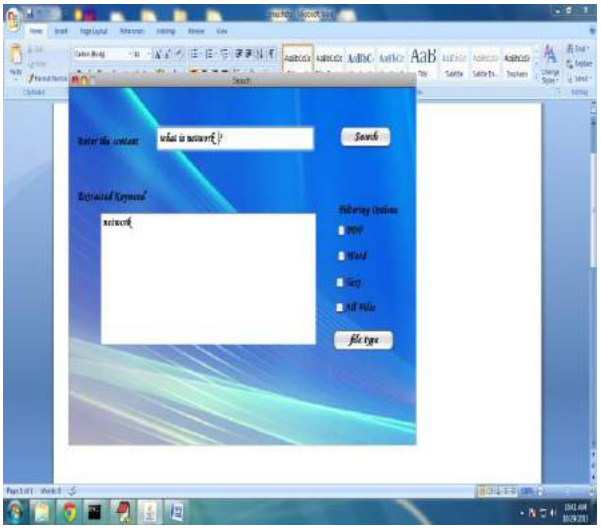


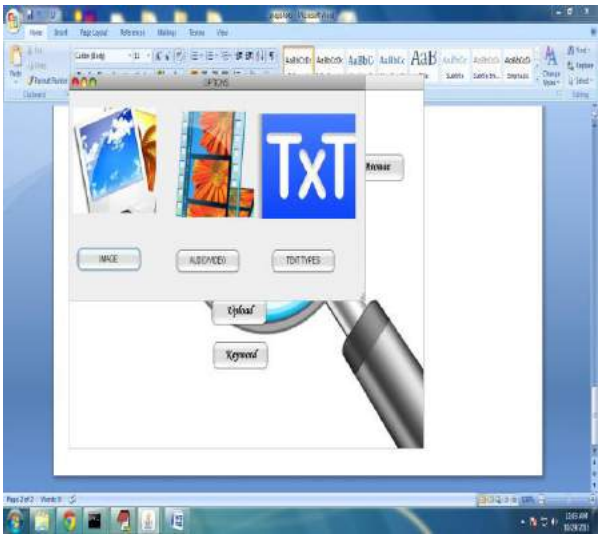
ADVANTAGES OF THE SYSTEM

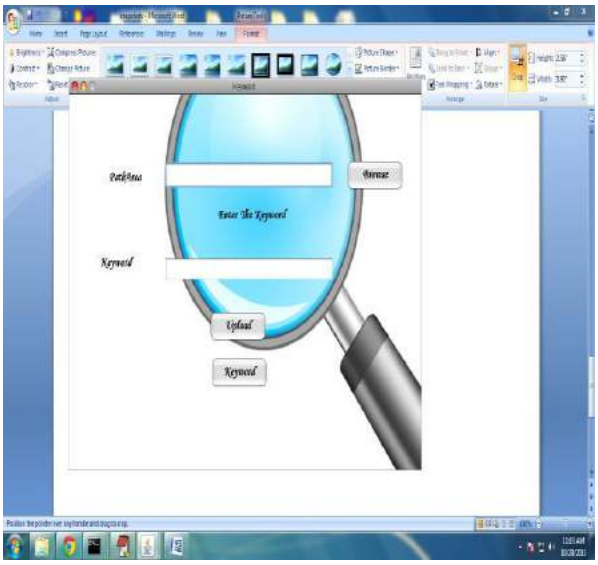
- This clustering decreases the chances of errors into the queries.
- Quick response to the user.
- Content of the will be displayed

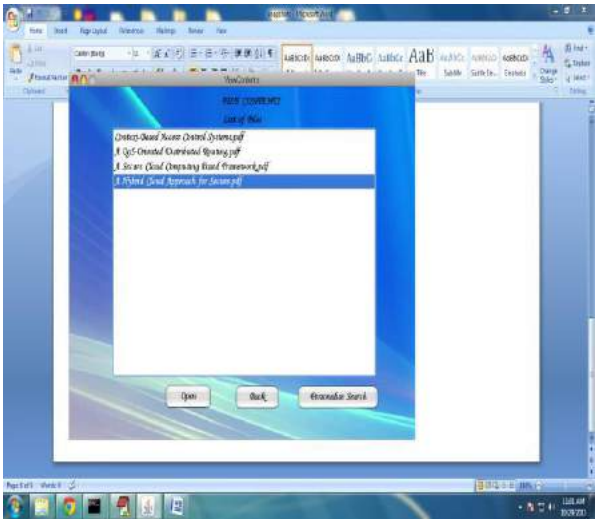
SAMPLE SCREENS:











CONCLUSION:

We have considered a particular form of just-in-time retrieval systems intended for conversational environments, in which they recommend to users' documents that are relevantto their information needs. We focused on modeling the users' information needs by deriving implicit queries from short conversation fragments. These queries are based on sets ofkeywords extracted from the conversation. We have proposed 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.

FUTURE ENHANCEMENTS

We have considered a particular form of just-in-time retrieval systems intended for conversational environments, in which they recommend to users' documents that are relevant to their information needs. We focused on modeling the users' information needs

by deriving implicit queries from short conversation fragments. These queries are based on sets of keywords extracted from the conversation. We have proposed 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.

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] 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 conceptlatticefortextunderstandingandsummarization," 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.: Building Technol.RichLearn.ContextsThatWork,2007,pp.557–559.

[6] D.HarwathandT.J.Hazen,"Topic identification based extrinsic evaluation of summarization techniquesappliedtoconversationalspeech," in Proc. Int. Conf. Acoust., Speech, Signal Process. (ICASSP), 2012, pp. 5073–5076.