**Keyword Extraction and Clustering for Document**

**Recommendation in Conversations**

**ABSTRACT**

This paper addresses the problem of keyword extractionfrom conversations, with the goal of using these keywords toretrieve, for each short conversation fragment, a small numberof potentially relevant documents, which can be recommended toparticipants. However, even a short fragment contains a varietyof words, which are potentially related to several topics; moreover,using an automatic speech recognition (ASR) system introduceserrors among them. Therefore, it is difficult to infer preciselythe information needs of the conversation participants. We firstpropose an algorithm to extract keywords from the output of anASR system (or a manual transcript for testing), which makes useof topic modeling techniques and of a submodular reward functionwhich favors diversity in the keyword set, to match the potentialdiversity of topics and reduce ASR noise. Then, we proposea method to derive multiple topically separated queries from thiskeyword set, in order to maximize the chances of making at leastone relevant recommendation when using these queries to searchover the English Wikipedia. The proposed methods are evaluatedin terms of relevance with respect to conversation fragments fromthe Fisher, AMI, and ELEA conversational corpora, rated by severalhuman judges. recommendersystem to be used in conversations.

**EXISTING SYSTEM**

The problem of keyword extractionfrom conversations, with the goal of using these keywords toretrieve, for each short conversation fragment, a small numberof potentially relevant documents, which can be recommended toparticipants. However, even a short fragment contains a varietyof words, which are potentially related to several topics; moreover,using an automatic speech recognition (ASR) system introduceserrors among them. Therefore, it is difficult to infer preciselythe information needs of the conversation participants

**PROPOSED SYSTEM:**

we proposea method to derive multiple topically separated queries from thiskeyword set, in order to maximize the chances of making at leastone relevant recommendation when using these queries to searchover the English Wikipedia. The proposed methods are evaluatedin terms of relevance with respect to conversation fragments fromthe Fisher, AMI, and ELEA conversational corpora, rated by severalhuman judges. The scores show that our proposal improvesover previous methods that consider only word frequency or topicsimilarity, and represents a promising solution for a document recommendersystem to be used in conversations.

**MODULE DESCRIPTION:**

# Number of Modules:

After careful analysis the system has been identified to have the following modules:

1.Document recommendation

2.Information retrieval

3.Keyword extraction

4. Meeting analysis,

5. Topic modeling.

**Document recommendation**

As a first idea, one implicit query can be prepared for eachconversation fragment by using as a query all keywords selectedby the diverse keyword extraction technique. However, to improvethe retrieval results, multiple implicit queries can be formulatedfor each conversation fragment, with the keywords ofeach cluster from the previous section, ordered as above (becausethe search engine used in our system is not sensitive toword order in queries).

Just-in-time retrieval systems have the potential to bringa radical change in the process of query-based informationretrieval. Such systems continuously monitor users’ activitiesto detect information needs, and pro-actively retrieve relevantinformation. To achieve this, the systems generally extractimplicit queries (not shown to users) from the words thatare written or spoken by users during their activities. In thissection, we review existing just-in-time-retrieval systems andmethods used by them for query formulation. In particular, wewill introduce our Automatic Content Linking Device (ACLD), a just-in-time document recommendation system formeetings, for which the methods proposed in this paper areintended. In II-B, we discuss previous keyword extractiontechniques from a transcript or text.

**Information retrieval**

The Watson just-in-time-retrieval system assisted userswith finding relevant documents while writing or browsing theWeb. Watson built a single query based on a more sophisticatedmechanism than the Remembrance Agent, by taking advantageof knowledge about the structure of the written text, e.g. by emphasizingthe words mentioned in the abstract or written withlarger fonts, in addition to word frequency. The Implicit Queries(IQ) system generated context-sensitive searches byanalyzing the text that a user is reading or composing. IQ automaticallyidentified important words to use in a query usingTFIDF weights. Another query-free system was designed forenriching television news with articles from the Web . Similarlyto IQ or Watson, queries were constructed from the ASRusing several variants of TFIDF weighting, and considering alsothe previous queries made by the system.Other real-time assistants are conversational: they interactwith users to answer their explicit information needs or toprovide recommendations based on their conversation. Forinstance, Ada and Grace1 are twin virtual museum guides which interact with visitors to answer their questions, suggestexhibits, or explain the technology that makes them work. Acollaborative tourist information retrieval system interacts with tourists to provide travel information such asweather conditions, attractive sites, holidays, and transportation,in order to improve their travel plans. MindMeld2 is acommercial voice assistant for mobile devices such as tablets,which listens to conversations between people, and showsrelated information from a number of Web-based informationsources, such as local directories. MindMeld improves theretrieval results by adding the users’ location information tothe keywords of conversation obtained using an ASR system.As far as is known, the system uses state-of-the-art methods forlanguage analysis and information retrieval

**Keyword extraction**

These findings motivated us to design an innovative keywordextraction method for modeling users’ information needsfrom conversations. As mentioned in the introduction, sinceeven short conversation fragments include words potentiallypertaining to several topics, and the ASR transcript adds additionalambiguities, a poor keyword selection method leads tonon-informative queries, which often fail to capture users’ informationneeds, thus leading to low recommendation relevanceand user satisfaction. The keyword extraction method proposedhere accounts for a diversity of hypothesized topics in a discussion,and is accompanied by a clustering technique that formulatesseveral topically-separated queries.

**Meeting analysis**

when usersparticipate in a meeting, their information needs can be modeledas implicit queries that are constructed in the backgroundfrom the pronounced words, obtained through real-time automaticspeech recognition (ASR). These implicit queries are usedto retrieve and recommend documents from the Web or a localrepository, which users can choose to inspect in more detail ifthey find them interesting.The focus of this paper is on formulating implicit queries to ajust-in-time-retrieval system for use in meeting rooms. In contrastto explicit spoken queries that can be made in commercialWeb search engines, our just-in-time-retrieval system mustconstruct implicit queries from conversational input, which containsa much larger number of words than a query.

**Topic modeling**

keyword extraction has used the frequency of allwords belonging to the same WordNet concept set whilethe Wikifier system relied on Wikipedia links to computeanother substitute to word frequency. Hazen also applied topicmodeling techniques to audio files. In another study, heused PLSA to build a thesaurus, which was then used to rankthe words of a conversation transcript with respect to eachtopic using a weighted point-wise mutual information scoringfunction Moreover, Harwath and Hazen utilized PLSAto represent the topics of a transcribed conversation, and thenranked words in the transcript based on topical similarity tothe topics found in the conversation Similarly, Harwathet al. extracted the keywords or key phrases of an audio fileby directly applying PLSA on the links among audio framesobtained using segmental dynamic time warping, and thenusing mutual information measure for ranking the key conceptsin the form of audio file snippets A semi-supervised latentconcept classification algorithm was presented by Celikyilmazand Hakkani-Tur using LDA topic modeling for multi-documentinformation extraction

**System Configuration:**

HARDWARE REQUIREMENTS:

Hardware - Pentium

Speed - 1.1 GHz

RAM - 1GB

Hard Disk - 20 GB

Key Board - Standard Windows Keyboard

SOFTWARE REQUIREMENTS:

Operating System : Windows

Technology : Java and J2EE

Web Technologies : Html, JavaScript, CSS

IDE : My Eclipse

Web Server : Tomcat

Database : My SQL

Java Version : J2SDK1.5

**CONCLUSION**

We have considered a particular form of just-in-time retrievalsystems intended for conversational environments, inwhich they recommend to users documents that are relevantto their information needs. We focused on modeling the users’information needs by deriving implicit queries from shortconversation fragments. These queries are based on sets ofkeywords extracted from the conversation. We have proposeda novel diverse keyword extraction technique which covers themaximal number of important topics in a fragment. Then, toreduce the noisy effect on queries of the mixture of topics ina keyword set, we proposed a clustering technique to dividethe set of keywords into smaller topically-independent subsetsconstituting implicit queries.We compared the diverse keyword extraction techniquewith existing methods, based on word frequency or topicalsimilarity, in terms of the representativeness of the keywordsand the relevance of retrieved documents. These were judgedby human raters recruited via the Amazon Mechanical Turkcrowdsourcing platform. The experiments showed that the diversekeyword extraction method provides on average the mostrepresentative keyword sets, with the highest -NDCG value,and leading–through multiple topically-separated implicitqueries–to the most relevant lists of recommended documents.Therefore, enforcing both relevance and diversity brings aneffective improvement to keyword extraction and documentretrieval.