SAVITRIBAI PHULE PUNE UNIVERSITY

A PRELIMINARY REPORT ON

DOCUMENT RECOMMENDATION IN CONVERSATIONS USING KEYWORD EXTRACTION AND CLUSTERING

SUBMITTED TOWARDS THE
PARTIAL FULFILLMENT OF THE REQUIREMENTS OF

BACHELOR OF ENGINEERING (Computer Engineering)

BY

Shruti Bhavsar	71728149D
Sanjana Khairnar	71809413B
Pauravi Nagarkar	71728292K
Sonali Raina	71728386M

Under The Guidance of

Asst. Prof. Amol Dumbare

DEPARTMENT OF COMPUTER ENGINEERING
Pimpri Chinchwad College of Engineering and Research
Ravet

Pimpri Chinchwad College of Engineering and Research
DEPARTMENT OF COMPUTER ENGINEERING

CERTIFICATE

This is to certify that the Project Entitled

Document Recommendation in Conversations using Keyword Extraction and Clustering

Submitted by

Shruti Bhavsar	71728149D
Sanjana Khairnar	71809413B
Pauravi Nagarkar	71728292K
Sonali Raina	71728386M

is a bonafide work carried out by Students under the supervision of Asst. Prof. Amol Dumbare and it is submitted towards the partial fulfillment of the requirement of Bachelor of Engineering (Computer Engineering) Project.

Asst. Prof. Amol Dumbare
Internal Guide
Dept. of Computer Engg.

Asst. Prof. Jameer Kotwal
Project Co-ordinator
Dept. of Computer Engg.

Dr. Archana Chaugule
H.O.D
Dept. of Computer Engg.

External Examiner

Dr. Harish Tiwari
Principal
PCCOER.

Abstract

This Project addresses the issue of keyword extraction from conversations, with the objective of utilizing these watchwords to recover, for every short discussion piece, a little number of conceivably pertinent reports, which can be prescribed to members. In any case, even a short piece contains a mixed bag of words, which are conceivably identified with a few themes; also, utilizing a programmed discourse acknowledgment (ASR) framework presents slips among them. Along these lines, it is hard to surmise correctly the data needs of the discussion members. We first propose a calculation to remove decisive words from the yield of an ASR framework (or a manual transcript for testing), which makes utilization of theme demonstrating methods and of a sub modular prize capacity which supports differing qualities in the magic word set, to coordinate the potential differing qualities of subjects and decrease ASR commotion. At that point, we propose a technique to infer various topically isolated inquiries from this decisive word set, keeping in mind the end goal to amplify the possibilities of making at any rate one pertinent proposal when utilizing these questions to seek over the English Wikipedia. The proposed systems are assessed as far as significance as for discussion pieces from the Fisher, AMI, and ELEA conversational corpora, appraised by a few human judges. The scores demonstrate that our proposition moves forward over past systems that consider just word recurrence or theme closeness, and speaks to a promising answer for a report recommender framework to be utilized as a part of discussions.

Acknowledgments

It gives us great pleasure in presenting the preliminary project report on ***Document Recommendation in Conversations using Keyword Extraction and Clustering*** .

We would like to take this opportunity to thank our internal guide ***Asst. Prof. Amol Dumbare*** *for giving us all the help and guidance we needed. We are really grateful to them for their kind support. Their valuable suggestions were very helpful.*

We are also grateful to ***Dr. Archana Chaugule****, Head of Computer Engineering Department, Pimpri Chinchwad College of Engineering and Research for indispensable support, suggestions.*

Shruti Bhavsar	71728149D
Sanjana Khairnar	71809413B
Pauravi Nagarkar	71728292K
Sonali Raina	71728386M

INDEX

List of Figures

List of Tables

CHAPTER 1

SYNOPSIS

1.1 PROJECT TITLE

" Document Recommendation in Conversations using Keyword Extraction and Clustering"

1.2 PROJECT OPTION

Final Year Project

1.3 INTERNAL GUIDE

Asst. Prof. Amol Dumbare

1.4 SPONSORSHIP AND EXTERNAL GUIDE

No

1.5 TECHNICAL KEYWORDS (AS PER ACM KEYWORDS)

A. Categories and Subject Descriptors:
[Information Search and Retrieval]: Search process, Query formulation.

B. General Terms: Experimentation, Measurement.

C.Keywords: Document recommendation, information retrieval, keyword extraction, meeting analysis, topic modeling.

1.6 PROBLEM STATEMENT

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.

1.7 ABSTRACT

The framework perform the extraction of keyword its address the issue for discussion for every short change segment. A less number of possibly critical archives with the objective of utilizing the data recovered which can be prescribed to member. Utilizing programmed discourse recongnization framework present blunder among them which are possibly identified with different subject, even short piece contains an assortment of word. Hence, it is confused to construe particularly the data needs the exchange of members. The utilization of point demonstrating methods and of a sub particular prize capacity which supports assorted qualities in the catchphrase set, for making to coordinate the potential differences of subject and lessen ASR clamor. At that point, paper propose a technique to infer a few topically separated inquiries from this watchword set, keeping in mind the end goal to take advantage of the odds of working no less than one noteworthy suggestion when utilizing these questions to look over the English Wikipedia. The Fisher, AMI, and ELEA conversational corpora, evaluated by different human judges by utilizing proposed techniques are figured as a part of terms of essentialness regarding discussion sections from. The scores demonstrate that our proposition enhances over past strategies that consider just word recurrence or theme correspondence, and speaks to a promising answer for a record recommender framework to be utilized as a part of discussions.

Index Terms: Document recommendation, information retrieval, keyword extraction, meeting analysis, topic modeling.

1.8 GOALS AND OBJECTIVES

1.Maximizing the coverage of all the information needs, while minimizing redundancy in a short list of documents.

2.Using the keywords to retrieve, for each short conversation fragment, a small number of potentially relevant documents, which can be recommended to participants.

1.9 EXISTING SYSTEM:

A short piece contains a mixture of words, which are possibly identified with a few subjects, in addition, utilizing a programmed discourse acknowledgment (ASR) framework presents blunders among them. In this way, it is hard to induce accurately the data needs of the discussion member.

Disadvantages in Existing System:

1)Each short conversation fragment contain a small number of potentially relevant documents, which can be recommended to participants.

2)Automatic speech recognition (ASR) system introduces errors.

1.10 PROPOSED SYSTEM:

Propose a calculation to separate decisive words from the yield of an ASR framework (or a manual transcript for testing), which makes utilization of theme demonstrating methods and of a sub modular prize capacity which supports differing qualities in the catchphrase set, to coordinate the potential assorted qualities of points and lessen ASR commotion. At that point, we propose a technique to infer different topically isolated questions from this magic word set, with a specific end goal to expand the shots of making no less than one important suggestion when utilizing these inquiries to pursuit over the English Wikipedia.

Advantages in Proposed System:

1)Represents a promising solution for a document recommender system to be used in conversations..

2)Formulating implicit queries to a just-in-time-retrieval system for use in meeting rooms.

1.11 RELEVANT MATHEMATICS ASSOCIATED WITH THE PROJECT

Let S is the Whole System Consist of:

S= U, D, ASR, DKE, KC, QF, O..

U = User .

U = u1, u2, ...un

D = Dataset.

D = d1,d2. . . ..dn.

ASR= Automatic Speech Recognition

DKE = Diverse keyword extraction.

KC = Keyword Clustering

QF = Query Formulation

O = Output.

Procedure:.

Keyword Extraction:

ASR: automatic speech recognition converts the speech and provides output to algorithm that extract keywords from the output of an ASR system

Selection of Configurations:

Using the rank biased overlap (RBO) as a similarity metric, based on the fraction of keywords overlapping at different ranks.

$$RBO(S,T) = \frac{1}{\sum_{d=1}^{D} \left(\frac{1}{2}\right)^{d-1}} \sum_{d=1}^{D} \left(\frac{1}{2}\right)^{d-1} \frac{|S_{1:d} \cap T_{1:d}|}{|S_{1:d} \cup T_{1:d}|}$$

Where,

RBO = rank biased overlap Sand T be two ranked lists, and Si be the keyword at rank i in S The set of the keywords upto rank d in S.

Diverse Keyword Extraction:

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

1.Used to represent the distribution of the abstract topic for each word.

2.These topic models are used to determine weights for the abstract topics in each conversation fragment represented by β_z

3.The keyword list W = w1, w2. . . . . wk. Which covers a maximum number of the most important topics are selected by rewarding diversity, using an original algorithm introduced in this section.

1.12 NAMES OF CONFERENCES / JOURNALS WHERE PAPERS CAN BE PUBLISHED

- International Journal of Advance Research in Engineering, Science Technology (IJAREST)

1.13 REVIEW OF CONFERENCE/JOURNAL PAPERS SUPPORTING PROJECT IDEA

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) 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.

3) D. Harwath and T. J. Hazen, “Topic identification based extrinsic evaluation of summarization techniques applied to conversational speech,” in Proc. Int. Conf. Acoust., Speech, Signal Process. (ICASSP), 2012, pp. 5073–5076.

4) 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.

5) B. J. Rhodes and P. Maes, “Just-in-time information retrieval agents,” IBM Syst. J., vol. 39, no. 3.4, pp. 685–704, 2000.

6)D. Traum, P. Aggarwal, R. Artstein, S. Foutz, J. Gerten, A. Katsamanis, A. Leuski, D. Noren, and W. Swartout, “Ada and Grace: Direct interaction with museum visitors,” in Proc. 12th Int. Conf. Intell. Virtual Agents, 2012, pp. 245–251.

7)A. S. M. Arif, J. T. Du, and I. Lee, “Examining collaborative query reformulation: A case of travel information searching,” in Proc. 37th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 2014, pp. 875–878.

8)A. S. M. Arif, J. T. Du, and I. Lee, “Towards a model of collaborative information

retrieval in tourism," in Proc. 4th Inf. Interact. Context Symp., 2012, pp. 258–261.

9)J. Zaino, MindMeld makes context count in search, [Online]. Available:

10)M. Habibi and A. Popescu-Belis, "Using crowdsourcing to compare document recommendation strategies for conversations," Workshop Recommendat. Utility Eval.: Beyond RMSE (RUE'11), pp. 15–20,2012.

1.14 PLAN OF PROJECT EXECUTION

Sr. No.	Month Sheduled	Phase
1	June-August	Topic Seraching
2	August-September	Topic Selection
3	August-September	Project Confirmation
4	August-September	Literature Survey
5	September-October	Requirement Analysis
6	September-October	Requirement Gathering
7	November-December	Designing
8	November-December	Designing Test
9	November-December	Database Creation
10	January-February	Coding
11	January-February	Database And Module Connectivity
12	March	Testing of Project
13	April	Result Analysis

CHAPTER 2

TECHNICAL KEYWORDS

2.1 AREA OF PROJECT

Data mining.

2.2 TECHNICAL KEYWORDS (AS PER ACM KEYWORDS)

2.2.1 A.Categories and Subject Descriptors::

[Information Search and Retrieval]: Search process, Query formulation.

2.2.2 B. General Terms:

Experimentation, Measurement

2.2.3 C.Keywords:

Collaborative search, Interactive IR, Query reformulation, Document recommendation, information retrieval,keyword extraction, meeting analysis, topic modeling

CHAPTER 3

INTRODUCTION

3.1 PROJECT IDEA

- Humans are encompassed by an uncommon abundance of data, accessible as records, databases, or mixed media assets. Access to this data is adapted by the accessibility of suitable web indexes, however notwithstanding when these are accessible, clients frequently don't start a pursuit, in light of the fact that their current action does not permit them to do as such, or in light of the fact that they are not mindful that applicable data is accessible. We receive in this paper the point of view of in the nick of time recovery, which replies this inadequacy by suddenly suggesting archives that are identified with clients' present exercises. At the point when these exercises are primarily conversational, for occurrence when clients take part in a meeting, their data needs can be demonstrated as understood inquiries that are built out of sight from the professed words, acquired through continuous programmed discourse acknowledgment (ASR). These certain questions are utilized to recover and suggest reports from the Web or a neighborhood storehouse, which clients can decide to investigate in more detail if they discover them intriguing. The center of this paper is on figuring verifiable questions to a without a moment to spare recovery framework for utilization in meeting rooms. Conversely to unequivocal talked inquiries that can be made in business Web crawlers, our in the nick of time recovery framework must develop certain questions from conversational information, which contains a much bigger number of words than a question. For example, in the illustration examined in Section V-B underneath, in which four individuals set up together a rundown of things to help them get by in the mountains, a short piece of 120 seconds contains around 250 words, relating to a mixed bag of areas, for example, 'chocolate', 'gun', or 'lighter'. What might then be the most supportive 3–5 Wikipedia pages to prescribe, and how might a framework focus them? Given the potential variety of themes, strengthened by potential ASR slips or discourse disfluencies, (for example, "rush" in this illustration), our objective is to keep up different speculations about clients' data needs, and to present a little example of proposals in view of the no doubt ones. In this manner, we point at separating a pertinent

and various arrangement of catchphrases, group them into theme particular questions positioned by significance, and present clients an example of results from these questions. The point based bunching abatements the possibilities of including ASR blunders into the questions, and the assorted qualities of essential words expands the possibilities that no less than one of the suggested records answers a need for data, or can prompt a helpful archive while taking after its hyperlinks. Case in point, while a strategy in view of word recurrence would recover the accompanying Wikipedia pages: 'Light', 'Lighting', and 'Light My Fire' for the aforementioned piece, clients would lean toward a set, for example, 'Lighter', "Fleece" and 'Chocolate'. Pertinence and assorted qualities can be authorized at three stages: at the point when removing the magic words; when building one or a few certain inquiries; or when repositioning their outcomes.

3.2 MOTIVATION OF THE PROJECT

- Sentence clustering is an important for text processing and can also be used in more general text mining task. In previous system clustering perform on document level and give the text. It is motivated us and now we performing a sentence level text extraction using clustering algorithm. By clustering the sentences of that document we would intuitively expect at least one of the clusters to be closely related to the concept described by the query term; however, other cluster may contain information pertaining to the query in some way hitherto unknown to us and in such a case we would have successfully mined new sentence. The main aim of sentence clustering is to present the most important sentence in text document will be extracted while keeping main content of document. This process is significantly different from that of human based text summarization since human can capture and relate deep meanings and themes of text document while automation of such skill is very difficult to implement.

CHAPTER 4

PROBLEM DEFINITION AND SCOPE

4.1 PROBLEM STATEMENT

Problem statement defines difficulties in existing system and how to overcome it and which algorithm are uses. In Existing system performing fuzzy relational algorithm for sentence level clustering and form clusters at sentence level. There is drawback in existing system it is overlapping of sentences and at time they used only one Keyword there have some limitations. They work on single document.

In proposed system we overcome all the problems by using hierarchical fuzzy relational clustering algorithm it is also reduce overlapping and use multiple keywords. It counts the number of keyword which is present in the sentences. The hierarchical clustering algorithm works on multiple documents.

The main objective of proposed system is 1.The project work on text document for sentence level clustering after keyword extraction. Implementing page rank algorithm on the text document which is based on keywords. 2. Formation of sentence level clustering on the basis of keyword by using hierarchical fuzzy relational clustering algorithm and give the sentence. 3. Developing a pattern as per hierarchical fuzzy relational clustering algorithm with keyword and sentences and give the final output is analytical cluster Membership value and related sentence as output.

4.2 GOALS AND OBJECTIVES

- Maximizing the coverage of all the information needs, while minimizing redundancy in a short list of documents.
- Using the keywords to retrieve, for each short conversation fragment, a small number of potentially relevant documents, which can be recommended to participants.

4.2.1 Statement of scope

- In this section we explain the future scope of our project. The future scope gives the scope of our project in future. Those works are not done in our project that work will be done in future.

The future work may include:

1.We optimize a time in future by using optimization technique in our project we require 90 sec for one keyword searching in future reduce that time 40% by using optimization technique.

2. When user enter statement at that time stemword and stopword will be removed and keyword will be remain only

4.3 METHODOLOGIES OF PROBLEM SOLVING AND EFFICIENCY ISSUES

- State of the art: just-in-time retrieval and keyword extraction

  Just-in-time retrieval systems have the potential to bringa radical change in the process of query-based informationretrieval. Such frameworks persistently screen clients' exercises to distinguish data needs, and genius effectively recover applicable data. To accomplish this, the frameworks by and large concentrate certain questions (not indicated to clients) from the words that are composed or talked by clients amid their exercises. In this segment, we survey existing without a moment to spare recovery frameworks and routines utilized by them for inquiry detailing. Specifically, we will present our Automatic Content Linking Device (ACLD) a without a moment to spare record suggestion framework for gatherings, for which the routines proposed in this paper are expected. In II-B, we talk about past essential word extraction procedures from a transcript or content.

- Keyword Extraction Methods

  Various strategies have been proposed to consequently remove pivotal words from a content, and are relevant additionally to interpreted discussions. The most punctual procedures have utilized word frequencies and TFIDF qualities to rank words for extraction. On the other hand, words have been positioned by checking pairwise word co-event frequencies. These methodologies don't consider word significance, so they may overlook low-recurrence

words which together demonstrate an exceedingly notable subject. For example, the words 'auto', 'wheel', 'seat', and "traveler" happening together demonstrate that autos are a notable theme regardless of the fact that every word is not itself incessant. To enhance over recurrence based strategies, a few approaches to utilize lexical semantic data have been proposed. Semantic relations between words can be acquired from a physically developed thesaurus, for example, Word Net, or from Wikipedia, or from a naturally assembled thesaurus utilizing idle subject displaying strategies, for example, LSA, PLSA, or LDA. For example, pivotal word extraction has utilized the recurrence of all words having a place with the same WordNet idea set, while the Wikifier framework depended on Wikipedia connections to register another substitute to word recurrence. 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 scoring function. 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-document information extraction. Formulation of implicit queries from conversations We propose a two-stage approach to the formulation ofimplicit queries. The first stage is the extraction of keywordsfrom the transcript of a conversation fragment for which documentsmust be recommended, as provided by an ASR system. These keywords should cover as much aspossible the topics detected in the conversation, and if possibleavoid words that are obviously ASR mistakes. The secondstage is the clustering of the keyword set in the form of severaltopically-disjoint queries

- Diverse Keyword Extraction

  We propose to take advantage of topic modeling techniquesto build a topical representation of a conversation fragment,and then select content words as keywords by using topicalsimilarity, while also rewarding the coverage of a diverserange of topics, inspired by recent summarization methods. The benefit of diverse keyword extraction is thatthe coverage of the main topics of the conversation fragmentis maximized. Moreover, in order to cover more topics, theproposed algorithm will select a smaller number of keywordsfrom each topic. This is desirable for two reasons. thiswill lead to more dissimilar implicit queries, thus increasing thevariety of retrieved documents. and, if words which arein reality ASR noise can create a main topic in the fragment,then the algorithm will choose a smaller number of these noisykeywords compared to algorithms which ignore diversity.

- Keyword Clustering

  3The diverse set of extracted keywords is considered to representthe possible information needs of the participants to aconversation, in terms of the notions and topics that are mentionedin the conversation. To maintain the diversity of topicsembodied in the keyword set, and to reduce the noisy effect ofeach information need on the others, this set must be split intoseveral topically-disjoint subsets. Each subset corresponds thento an implicit query that will be sent to a document retrievalsystem.

4.4 APPLICATIONS

- Office Meetings.
- Lectures.

4.5 HARDWARE RESOURCES REQUIRED

Sr. No.	Parameter	Minimum Requirement
1	SYSTEM	Pentium IV 2.4 GHz.
2	HARD DISK	40 GB.
3	FLOPPY DRIVE	1.44 Mb.
4	MONITOR	15 VGA Color.
5	MOUSE	Touch Pad.
6	RAM	512 Mb.

Table 4.1: Hardware Requirements

4.6 SOFTWAREWARE RESOURCES REQUIRED

Sr. No.	Parameter	Minimum Requirement
1	OPERATING SYSTEM	Windows 7 and above.
2	CODING LANGUAGE	JAVA and Android
3	IDE	Eclipse Kepler.
4	DATABASE	SQLYog community/XAMPP Server..
5	Web ServerE	Apache Tomcat.

Table 4.2: Software Requirements

CHAPTER 5

PROJECT PLAN

5.1 PROJECT ESTIMATES

We are using waterfall model for our project estimation.

5.1.1 Reconciled Estimates

5.1.1.1 Cost Estimate

Not applicable

5.1.1.2 Time Estimates

Approximately 11 months

5.1.2 Project Resources

Windows , eclipse IDE, 2.93 GHZ cpu speed, 8 GB RAM, High speed internet connection.

5.2 RISK MANAGEMENT W.R.T. NP HARD ANALYSIS

5.2.1 Risk Identification

For risks identification, review of scope document, requirements specifications and schedule is done. Answers to questionnaire revealed some risks. Each risk is categorized as per the categories mentioned in [**?**]. Please refer table 5.1 for all the risks. You can refereed following risk identification questionnaire.

1. Have top software and customer managers formally committed to support the project?
Ans-Not apllicaable.
2. Are end-users enthusiastically committed to the project and the system/product to be built?

Ans-Not known at this time.

3. Are requirements fully understood by the software engineering team and its customers?

Ans-Yes

4. Have customers been involved fully in the definition of requirements?

Ans-Not applicable

5. Do end-users have realistic expectations?

Ans-Not applicable

6. Does the software engineering team have the right mix of skills?

Ans-yes

7. Are project requirements stable?

Ans-Not applicable

8. Is the number of people on the project team adequate to do the job?

Ans-Not applicble

9. Do all customer/user constituencies agree on the importance of the project and on the requirements for the system/product to be built?

Ans-Not applicable

5.2.2 Risk Analysis

The risks for the Project can be analyzed within the constraints of time and quality

ID	Risk Description	Probability	Impact		
			Schedule	Quality	Overall
1	Choosing images for password	Low	Low	High	High
2	Generating pass-matrix	High	Low	High	High
2	Generating horizontal and vertical access control	High	High	High	High

Table 5.1: Risk Table

Probability	Value	Description
High	Probability of occurrence is	$> 80\%$
Medium	Probability of occurrence is	$261 - 79\%$
Low	Probability of occurrence is	$< 20\%$

Table 5.2: Risk Probability definitions [**?**]

5.2.3 Overview of Risk Mitigation, Monitoring, Management

Following are the details for each risk.

Risk ID	1
Risk Description	Description 1
Category	Development Environment.
Source	Software requirement Specification document.
Probability	Low
Impact	High
Response	Mitigate
Strategy	Strategy
Risk Status	Occurred

Risk ID	2
Risk Description	Description 2
Category	Requirements
Source	Software Design Specification documentation review.
Probability	Low
Impact	High
Response	Mitigate
Strategy	Better testing will resolve this issue.
Risk Status	Identified

Risk ID	3
Risk Description	Description 3
Category	Technology
Source	This was identified during early development and testing.
Probability	Low
Impact	Very High
Response	Accept
Strategy	Example Running Service Registry behind proxy balancer
Risk Status	Identified

5.3 PROJECT SCHEDULE

5.3.1 Project task set

Major Tasks in the Project stages are:

- Task 1: Choosing Project Area and planning.
- Task 2: selecting paper and literature survey.
- Task 3: project designing
- Task 4: Implementation
- Task 5: Execution

5.4 TEAM ORGANIZATION

The manner in which staff is organized and the mechanisms for reporting are noted.

CHAPTER 6

SOFTWARE REQUIREMENT SPECIFICATION (SRS IS TO BE PREPARED USING RELEVANT MATHEMATICS DERIVED AND SOFTWARE ENGG. INDICATORS IN ANNEX A AND B)

6.1 INTRODUCTION

6.1.1 Purpose and Scope of Document

Humans are encompassed by an uncommon abundance of data, accessible as records, databases, or mixed media assets. Access to this data is adapted by the accessibility of suitable web indexes, however notwithstanding when these are accessible, clients frequently don't start a pursuit, in light of the fact that their current action does not permit them to do as such, or in light of the fact that they are not mindful that applicable data is accessible. We receive in this paper the point of view of in the nick of time recovery, which replies this inadequacy by suddenly suggesting archives that are identified with clients' present exercises. At the point when these exercises are primarily conversational, for occurrence when clients take part in a meeting, their data needs can be demonstrated as understood inquiries that are built out of sight from the professed words, acquired through continuous programmed discourse acknowledgment (ASR). These certain questions are utilized to recover and suggest reports from the Web or a neighborhood storehouse, which clients can decide to investigate in more detail if they discover them intriguing.

6.1.2 Overview of responsibilities of Developer

1.To have understanding of the problem statement.
2.To know what are the hardware and software requirements of Proposed system.
3.To have understanding of proposed system.
4.To do planning various activities with the help of planner.
5.Designing,programming,testing etc.

6.2 USAGE SCENARIO

6.2.1 User profiles

The user classes will be User and management system (Admin).

6.2.2 Use-cases

All use-cases for the software are presented. Description of all main Use cases using use case template is to be provided.

6.2.3 Use Case View

Use Case Diagram. Example is given below

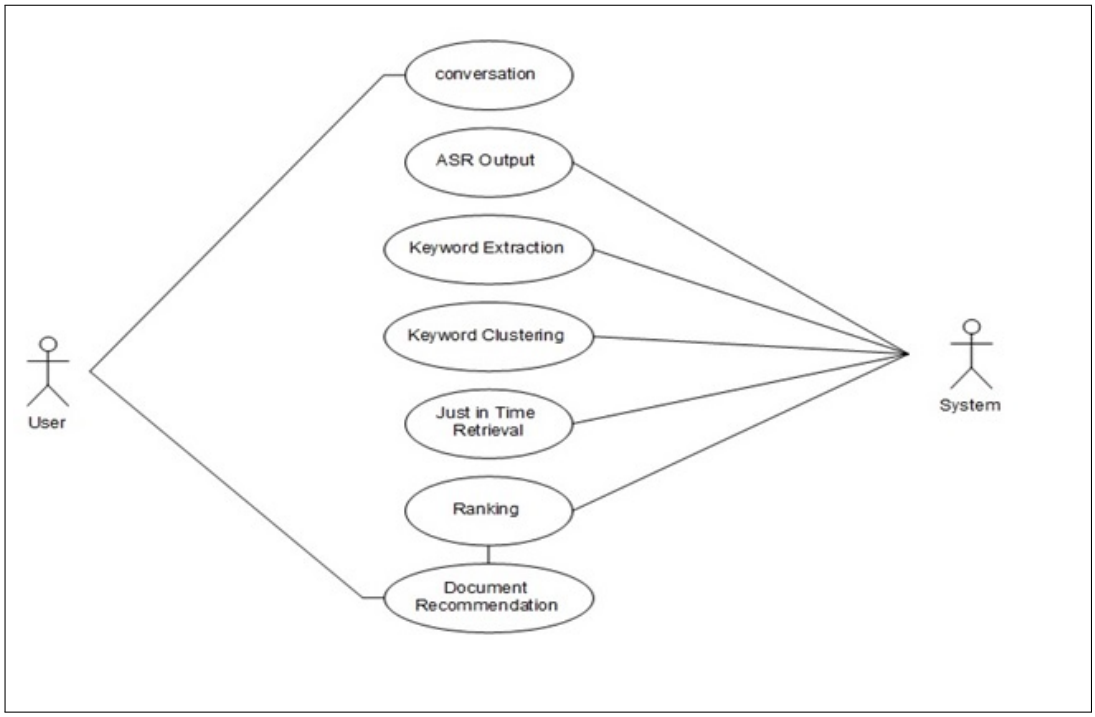


Figure 6.1: Use case diagram

6.3 DATA MODEL AND DESCRIPTION

6.3.1 Data Description

Data objects that will be managed/manipulated by the software are described in this section. The database entities or files or data structures required to be described. For data objects details can be given as below

6.3.2 Data objects and Relationships

let X be the user object and Y be the system object.

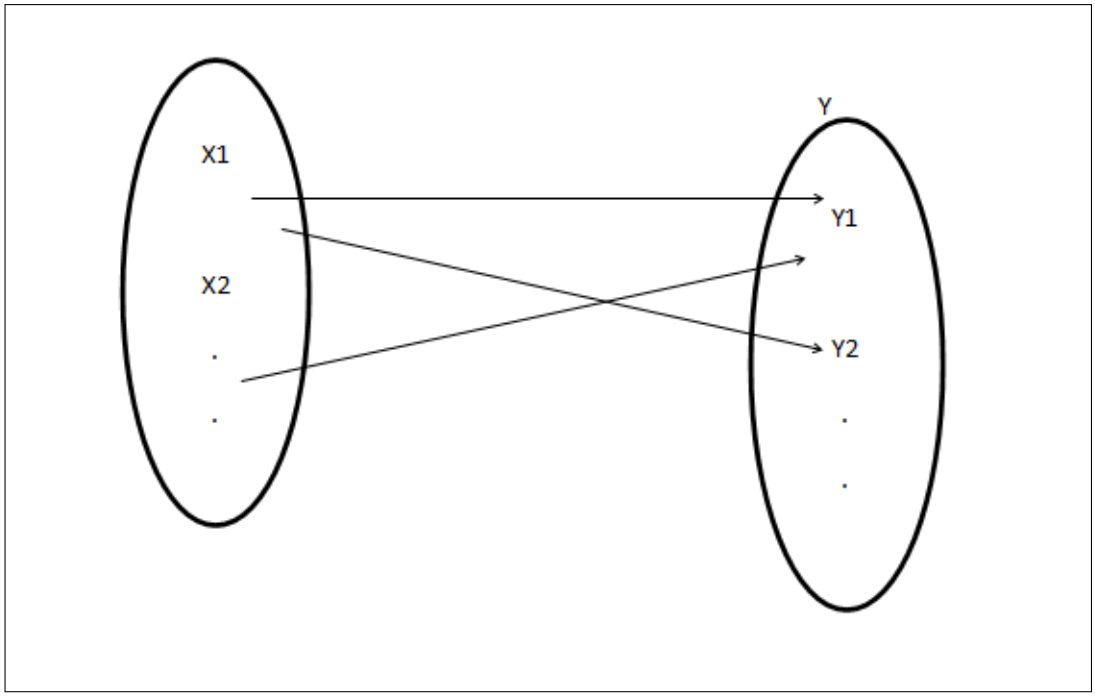
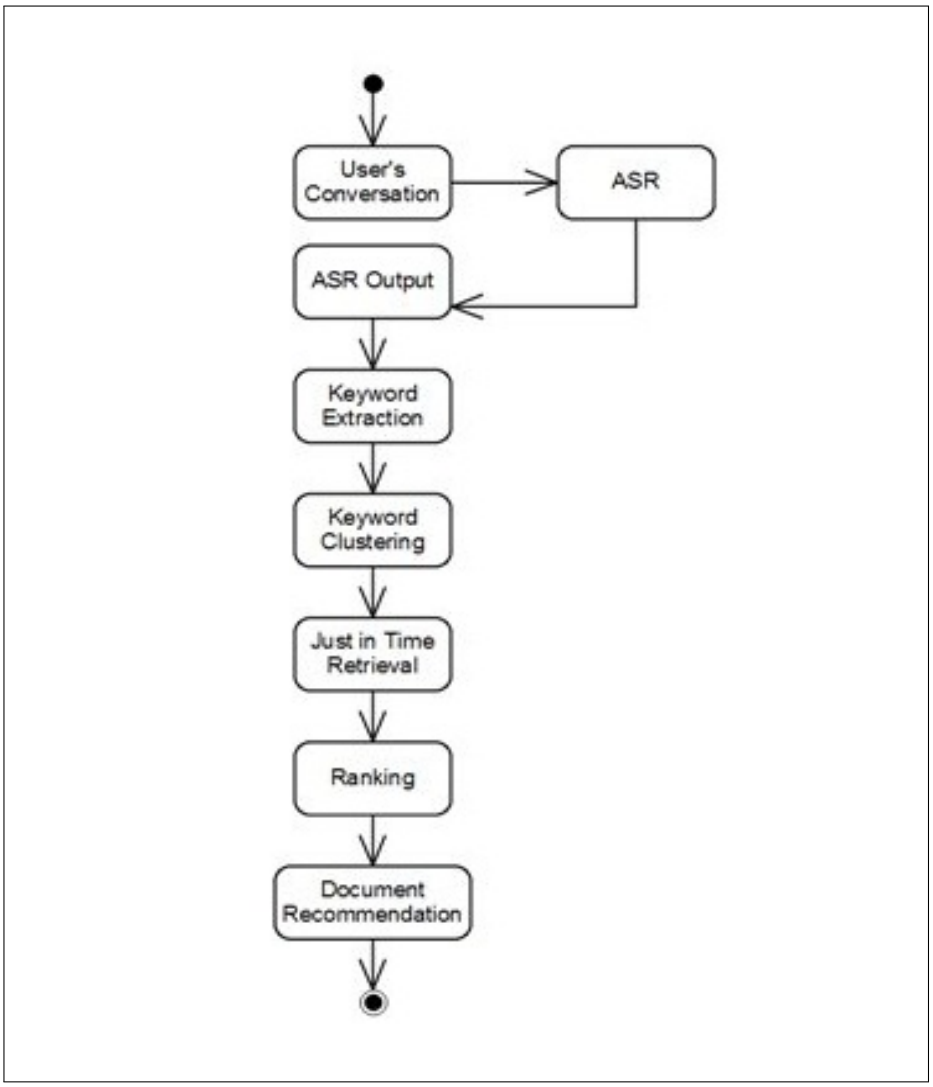


Figure 6.2: Use case diagram

6.3.3 Activity Diagram:

- The Activity diagram represents the steps taken.



6.3.4 Non Functional Requirements:

6.3.4.1 Performance Requirements:

- Every module of the system should work efficiently.
- The system should perform fast.
- System can produce results fasteron4GB of RAM.
- It may take more time for peak loads at main node.

6.3.4.2 Safety and security Requirements:

- The java application will not affect other applications on user's system.

6.3.5 Design Constraints

1. Apache Tomcat webserver.
2. MySQL as a database

6.3.6 Software Interface Description

The software interface to the outside world is very good and user friendly
The requirements for interfaces are Windows os 12.04,8 GB RAM etc
We are designing the user interface in JSP, HTML and CSS technology.

CHAPTER 7

DETAILED DESIGN DOCUMENT USING APPENDIX A AND B

7.1 INTRODUCTION

Humans are encompassed by an uncommon abundance of data, accessible as records, databases, or mixed media assets. Access to this data is adapted by the accessibility of suitable web indexes, however notwithstanding when these are accessible, clients frequently don't start a pursuit, in light of the fact that their current action does not permit them to do as such, or in light of the fact that they are not mindful that applicable data is accessible. We receive in this paper the point of view of in the nick of time recovery, which replies this inadequacy by suddenly suggesting archives that are identified with clients' present exercises. At the point when these exercises are primarily conversational, for occurrence when clients take part in a meeting, their data needs can be demonstrated as understood inquiries that are built out of sight from the professed words, acquired through continuous programmed discourse acknowledgment (ASR). These certain questions are utilized to recover and suggest reports from the Web or a neighborhood storehouse, which clients can decide to investigate in more detail if they discover them intriguing.

The center of this paper is on figuring verifiable questions to a without a moment to spare recovery framework for utilization in meeting rooms. Conversely to unequivocal talked inquiries that can be made in business Web crawlers, our in the nick of time recovery framework must develop certain questions from conversational information, which contains a much bigger number of words than a question. For example, in the illustration examined in Section V-B underneath, in which four individuals set up together a rundown of things to help them get by in the mountains, a short piece of 120 seconds contains around 250 words, relating to a mixed bag of areas, for example, 'chocolate', 'gun', or 'lighter'. What might then be the most supportive 3–5 Wikipedia pages to prescribe, and how might a framework focus them?

Given the potential variety of themes, strengthened by potential ASR slips or discourse disfluencies, (for example, "rush" in this illustration), our objective is to keep up different speculations about clients' data needs, and to present a

little example of proposals in view of the no doubt ones. In this manner, we point at separating a pertinent and various arrangement of catchphrases, group them into theme particular questions positioned by significance, and present clients an example of results from these questions. The point based bunching abatements the possibilities of including ASR blunders into the questions, and the assorted qualities of essential words expands the possibilities
that no less than one of the suggested records answers a need for data, or can prompt a helpful archive while taking after its hyperlinks. Case in point, while a strategy in view of word recurrence would recover the accompanying Wikipedia pages: 'Light', 'Lighting', and 'Light My Fire' for the aforementioned piece, clients would lean toward a set, for example, 'Lighter', "Fleece" and 'Chocolate'. Pertinence and assorted qualities can be authorized at three stages: at the point when removing the magic words; when building one or a few certain inquiries; or when re-positioning their outcomes.

7.2 ARCHITECTURAL DESIGN

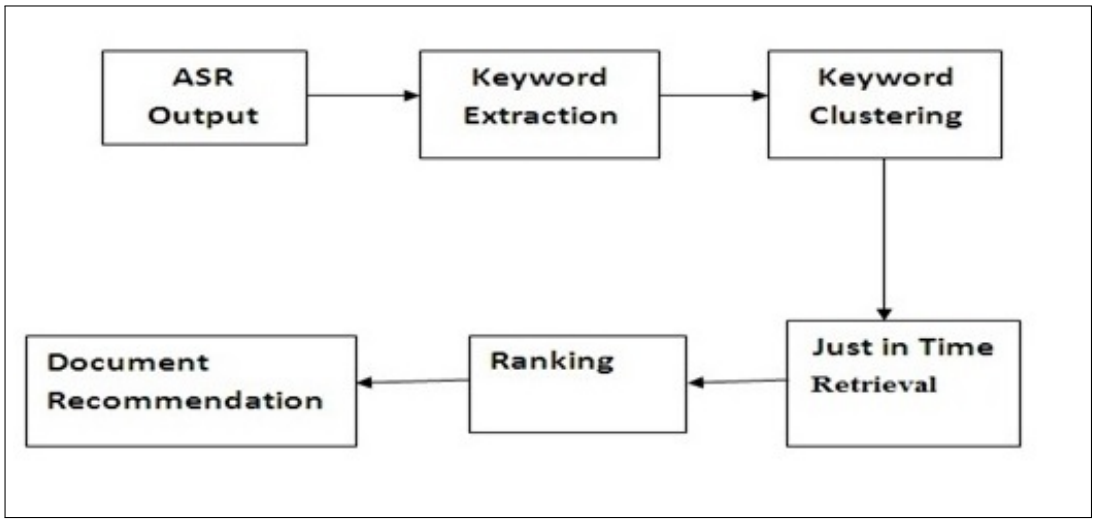


Figure 7.1: Architecture diagram

7.2.1 Data Flow Diagram

1.The DFD is also called as bubble chart. It is a simple graphical formalism that can be used to represent a system in terms of input data to the system, various processing carried out on this data, and the output data is generated by this system.

2.The data flow diagram (DFD) is one of the most important modeling tools. It is used to model the system components. These components are the system process, the data used by the process, an external entity that interacts with the system and the information flows in the system.

3.DFD shows how the information moves through the system and how it is modified by a series of transformations. It is a graphical technique that depicts information flow and the transformations that are applied as data moves from input to output.

4.DFD is also known as bubble chart. A DFD may be used to represent a system at any level of abstraction. DFD may be partitioned into levels that represent increasing information flow and functional detail.

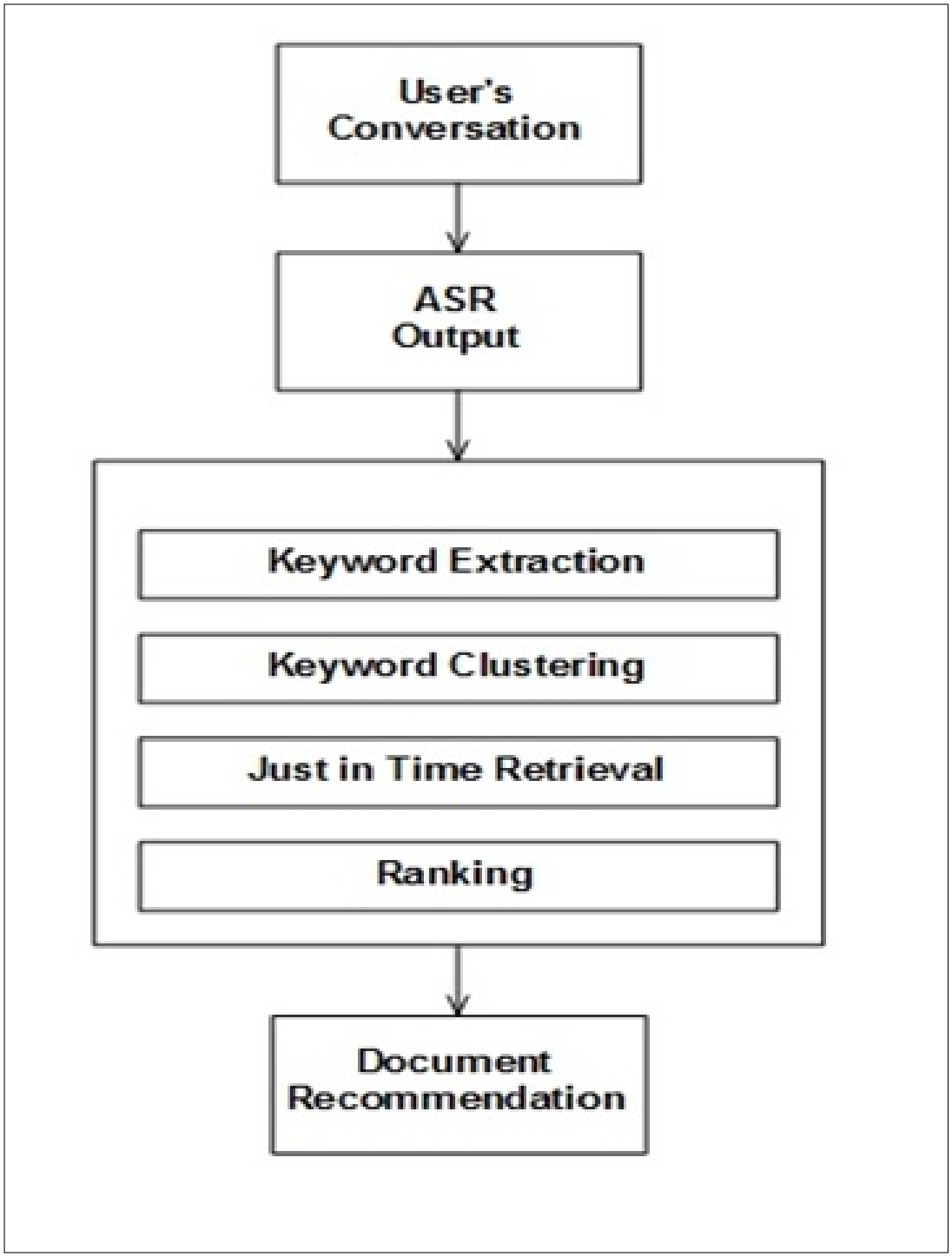


Figure 7.2: Architecture diagram

7.3 DATA DESIGN (USING APPENDICES A AND B)

A description of all data structures including internal, global, and temporary data structures, database design (tables), file formats.

7.3.1 Internal software data structure

Data processing query for generating pass values.

7.3.2 Global data structure

No global data structure used

7.3.3 Temporary data structure

Processing Voice.

7.3.4 Database description

Processed and Analyzed data

7.4 COMPONENT DESIGN

Class diagrams, Interaction Diagrams, Algorithms. Description of each component description required.

7.4.1 Class Diagram

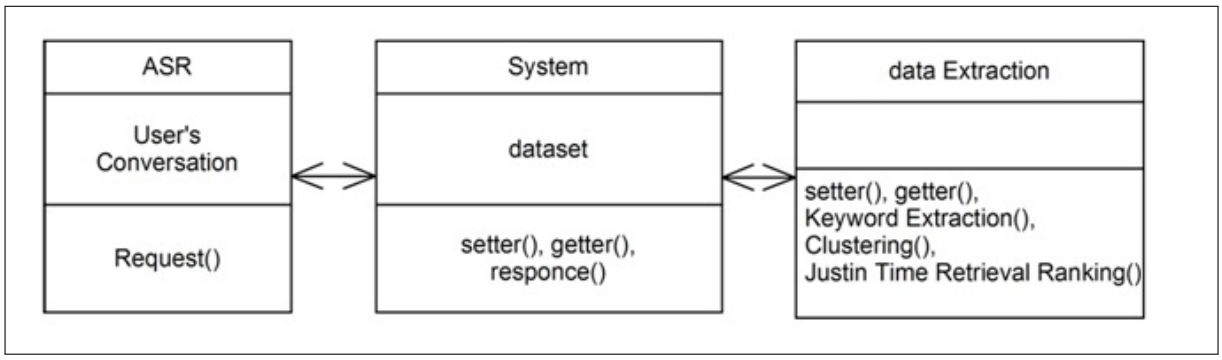


Figure 7.3: Class Diagram

CHAPTER 8

SUMMARY AND CONCLUSION

We have considered a specific type of without a moment to spare recovery frameworks proposed for conversational situations, in which they prescribe to client's archives that are important to their data needs. We concentrated on displaying the client's data needs by getting verifiable questions from short discussion pieces. These questions are in light of sets of pivotal words separated from the discussion. We have proposed a novel different pivotal word extraction strategy which covers the maximal number of vital themes in a piece. At that point, to lessen the boisterous impact on questions of the blend of themes in a decisive word set, we proposed a grouping system to isolate the arrangement of catchphrases into littler topically-autonomous subsets constituting understood inquiries.

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 crowd sourcing platform. The experiments showed that the diverse keyword extraction method provides on average the most representative keyword sets, with the highest -NDCG value, and leading–through multiple topically-separated implicit queries–to the most relevant lists of recommended documents. Therefore, enforcing both relevance and diversity brings an effective improvement to keyword extraction and document retrieval. The keyword extraction method could be improved by considering n-grams of words in addition to individual words only, but this requires some adaptation of the entire processing chain.

CHAPTER 9

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) 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.

3) D. Harwath and T. J. Hazen, “Topic identification based extrinsic evaluation of summarization techniques applied to conversational speech,” in Proc. Int. Conf. Acoust., Speech, Signal Process. (ICASSP), 2012, pp. 5073–5076.

4) 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.

5) B. J. Rhodes and P. Maes, “Just-in-time information retrieval agents,” IBM Syst. J., vol. 39, no. 3.4, pp. 685–704, 2000.

6)D. Traum, P. Aggarwal, R. Artstein, S. Foutz, J. Gerten, A. Katsamanis, A. Leuski, D. Noren, and W. Swartout, “Ada and Grace: Direct interaction with museum visitors,” in Proc. 12th Int. Conf. Intell. Virtual Agents, 2012, pp. 245–251.

7)A. S. M. Arif, J. T. Du, and I. Lee, “Examining collaborative query reformulation: A case of travel information searching,” in Proc. 37th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 2014, pp. 875–878.

8)A. S. M. Arif, J. T. Du, and I. Lee, “Towards a model of collaborative information retrieval in tourism,” in Proc. 4th Inf. Interact. Context Symp., 2012, pp. 258–261.

9)J. Zaino, MindMeld makes context count in search, [Online]. Available: http://semanticweb.com/mindmeld-makes-context-countsearch b42725 2014

10)M. Habibi and A. Popescu-Belis, “Using crowdsourcing to compare document recommendation strategies for conversations,” Workshop Recommendat. Utility Eval.: Beyond RMSE (RUE’11), pp. 15–20,2012.