ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 6, June 2016

Recommendation of Conversation Documents with Keyword Based Clustering

Pallavi Gopal Patil, Prof. P. M. Yawalkar,

PG Student, Dept. of Computer Engineering, MET's Institute of Engineering, Adgaon, Nashik, Savitribai Phule Pune University, Maharashtra, India

Dept. of Computer Engineering, MET's Institute of Engineering, Adgaon, Nashik Savitribai Phule Pune University, Maharashtra, India

ABSTRACT: Recommender Systems aims to suggest items of potential interest for solving information overload and have attracted growing amounts of attention. So, there should be some system which will fulfill the information needs of participants present in conversation. To solve this problem proposed system is developed which take conversation of participants as input and give recommendation links and documents based on conversation. First keywords are extracted from input and then the keyword set is used as query and recommendations are obtained through query search over google search engine. For recommendation a paper vector concept is used which will increase the chances of relevant recommendations. Various experimental results show that proposed system outperforms the existing system in terms of precision and recall. The proposed system shows 65-70 percent of precision value against the existing system.

KEYWORDS: Document recommendation, keyword extraction, topic modeling.

I. INTRODUCTION

Recommender Systems (RSs), aim to suggest items of potential interest for solving information overload, have attracted growing amounts of attention. They have been successfully applied in many fields such as e-commerce, movies, music, e-learning, mobile service, and so on.

We all are surrounded by wealth of information which is available in the form of databases, documents, or multimedia resources. But even this possibility, access to this is conditioned by the availability of search engines. Users do not start exploring for the information because their current activity does not allow them to do the search or they are not informed that the related information is available. To solve this problem just in time retrieval system is accepted, which automatically recommend documents that are related to the current users activities. When these actions are mainly conversational, for instance when users participate in a meeting, their information needs can be formed as implicit queries that are constructed in the background from the pronounced words, received through real-time automatic speech recognition (ASR). These implicit queries are used to retrieve and recommend documents from the web or local repository [19].

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 lower the chances of counting 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 defends the diversity of topics from the conversation. While the first keyword extraction methods ignored topicality as they were based on word frequencies, more late methods have considered topic modeling factors for keyword extraction, but without clearly setting a topic diversity constraint, which is important for naturally occurring conversations [15].

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 disturbing 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:

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 6, June 2016

- While extracting queries
- Building one or several implicit queries
- Re-ranking the results of queries [17]

II. RELATED WORK

Just-in-time retrieval systems have the potential to bring advance in the process of query-based information retrieval. These systems continuously observe users activities to identify information needs, and pro-actively retrieve relevant information. To fulfill this, the systems generally extract implicit queries from the words that are written or spoken by users during their activities [19].In section B some of earlier keyword extraction techniques from a transcription or text are discussed.

A. Just-in-Time Retrieval Systems

The remembrance agent [3] performs continuous searches for information that may be relevant to user's ongoing context. It runs continuously without user intervention and uses wasted CPU cycles constructively to perform regular searching for information that might be of use in user's current situation. It displays one-line opinion along with rating marking how relevant the documents were. However, since the RA runs continuously, suggestions could immediately distract from the user's primary task if they attracted too much attention. For these reasons the RA's suggestions are kept restrained [2], [6].

Information Management Assistants system fulfill information needs by using the document text the user is manipulating and knowledge of query formulation in traditional information retrieval systems. IMAs infrastructure grant information to users without requiring explicit requests. Watson system platform is used for working of this IMA's. It gathers contextual information in the form of text of the documents, which the user is managing in order to retrieve documents from distributed information repositories [4]. IMAs provide a framework to address the problems associated with processing queries out of context. The limitation of this system is that it does not embody semantic knowledge of particular task [7].

The Implicit Queries (IQ) system [5], [9] generated context-sensitive searches by analyzing the text that a user is reading or composing. IQ automatically identified important words to use in a query using TFIDF weights.

Automatic Content Linking Device (ACLD) is just-in-time retrieval system that continually retrieves items from repository & displays them to participants. AMIDA has drawbacks such as Graphical layout of user interface, Document Repository, additional functionalities such as Detecting similarities between previous discussions and current discussion would help in alerting users that they already had this discussion before [10],[12].

B. Keyword Extraction Methods

The keyword extraction algorithm when applied to a single document without using a whole database first all frequent terms were extracted and then a set is prepared which contains co-occurrences between each term and the frequent terms. The main advantage of this method is its clearness without demanding of a corpus and its high performance comparable to tfidf [1]. The limitation of this method is that it can only be used for single document [8].

D. Harwath and T. J. Hazen corelated different document summarization techniques. These techniques are especially useful for speech-based summarization where counts of common keywords can be reliably estimated over an entire document, but extracted utterance fragments with errorful transcripts may be difficult for users to read and interpret [14].

Current graph-based ranking methods for keyphrase extraction use single random walk to compute a single importance score for each word. For this reason a Topical PageRank (TPR) on word graph to compute word importance with reference to various topics is build. TPR outperforms advanced keyphrase extraction methods on two datasets under various evaluation metrics. This method considers topic information in pagerank so other graph-based ranking algorithms such as HITS can be used [12].

Conditional Random Fields (CRF) model is newest sequence labeling method, which can use the features of documents. CRF model surpasses other machine learning methods such as support vector machine; multiple linear regression models etc. in the task of keywords extraction but it does not take account of the ambiguity of the extracted keywords [11].

M. Habibi and A. Popescu-Belis, Introduces a method called Diverse Keyword Extraction which is used to derive several topically separated queries. This method with $\lambda = 0.75$ provides the keyword sets that are examined to be

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 6, June 2016

most representative of the conversation fragments. Therefore, enforcing both relevance and diversity brings an effective improvement to keyword extraction [15].

DivM method provides short, diverse relevant list of recommendations. This presents method for merging list of documents retrieved through multiple implicit queries, prepared for short conversation fragment. The goal of this method is to generate unique brief list of documents that are recommended in real time to participants. DivM method performs better than SimM which ignores diversity of topics Round-Robin merging which is used in Watson system. The merging algorithm rewards diversity by decreasing the gain of selecting documents from a list as the number of its previously selected documents increases [17].

Haifeng Liu et al. proposed a method for citation recommendation. The principle underlying this method is that, if two citing papers are significantly co-occurring with the same citing paper, they should be similar to some extent. Based on the above principle, an association mining technique is employed to attain the paper representation of each citing paper from the citation context. Then, these paper representations are pairwise compared to compute similarities between the citing papers for collaborative filtering [20].

Fanqi Meng et al. proposed a unified graph model that can easily incorporate various types of useful information (e.g., content, authorship, citation and collaboration networks etc.) for efficient recommendation. The proposed model not only allows to thoroughly explore how these types of information can be better combined, but also makes personalized query-oriented reference paper recommendation possible, which has not been explicitly addressed in the past. The experiments have demonstrated the clear advantages of personalized recommendation over non-personalized recommendation [16].

R. Thiyagarajan et al., has paid an attention to group the similar usage behavior of users using Weighted K-Means algorithm for aggregated usage profile. To evaluate the clusters quality new validating measure called MSR (Mean Square Residue) is applied. The results of this clustering approach are compared with the results of traditional clustering called K-Means. It was observed that the usage profile extracted from the MSNBC dataset using Weighted K-Means provides high quality recommendation for the given active user than results obtained by using K-Means Clustering [18].

These findings inspired us to design an innovative keyword extraction method for modeling user's information needs from conversations. As even short conversation fragments include words probably related to several topics, and the ASR transcript do addition of ambiguities, a poor keyword selection method leads to unclear queries, which often fail to catch users information needs, thus leading to low recommendation relevance and user satisfaction[19].

III. PROBLEM STATEMENT

To introduce a novel keyword extraction technique from conversation, which boosts the coverage of potential information needs of users, these keywords are clustered to build several topically-separated queries, results of these queries are merged into ranked set and finally these results are shown to user as recommendations. The main aim behind this system is to present recommendations related to user's current activity. The results are provided to users without initiation of direct search.

IV. PROPOSED ALGORITHM

Implicit queries can be formed using two- phase approach. The first phase is the separation of keywords from the transcription of a conversation fragment for which documents must be suggested. The second phase does the clustering of the keyword set in the form of distinct topically-disjoint queries. Fig.1 shows the system architecture consisting of main blocks such as Diverse Keyword Extraction, Keyword Clustering, and Recommendation of Documents.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 6, June 2016

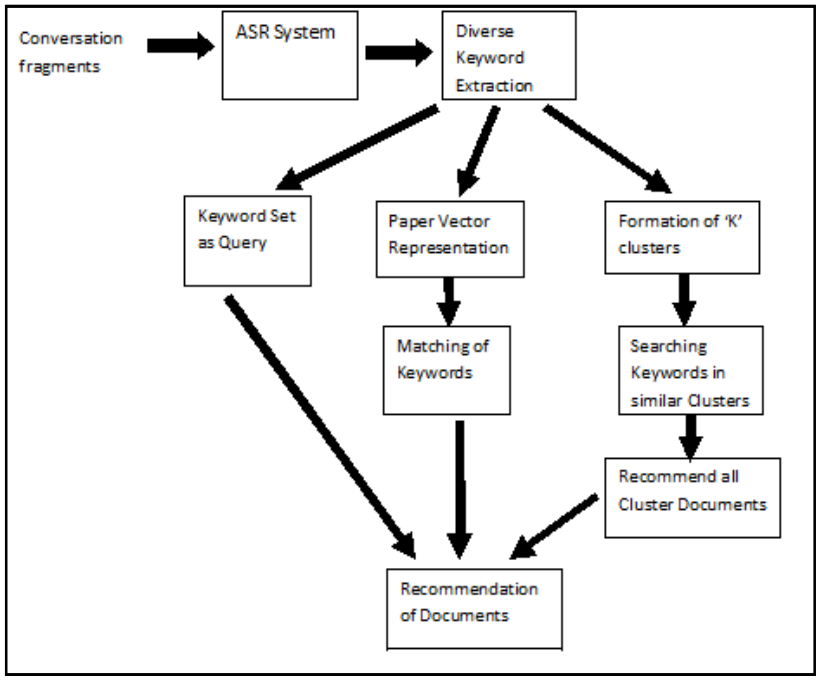

Fig 1: System Architecture

A. ASR System

The ASR System will convert the audio format of conversation fragments to text format. Further this text will be used for keyword extraction till recommendation of documents.

B. Diverse Keyword Extraction

The content words are selected as keywords by using topic modeling techniques. The advantage of this method is that, it maximizes the main topics of conversation fragments. The proposed method for Diverse Keyword Extraction proceeds in three steps:

- A topic model representation.
- Determine weights for the topics represented by βz.
- Generation of keyword list.

The distribution of each word w of the topic z is represented using topic models, which is denoted as p(z|w). The βz value is obtained by averaging all probabilities p(z|wi) of N words wi present in that fragment.

$$\beta_z = \frac{1}{N}\sum p(z|w_i) \quad (1)$$

If a conversation fragment t mentions a set of topics Z and each word w from the fragment t can evoke a subset of topics in Z, then the goal is to find a subset of k unique words, with |S| = k, which maximizes the number of covered topics. To achieve this goal the contribution of topic z with respect to each set of words is given by using a Diverse Reward Function, that gives the contribution of topic z in the keyword set S selected from fragment t.

$$r_{S,z} = \sum_{w \in S} p(z|w) \quad (2)$$

Finally, the keyword set $S\underline{C}t$, is chosen by maximizing the cumulative reward function over all the topics, formulated as,

$$R(S) = \sum_{z \varepsilon Z} \beta z \,.\, r_{S,z}^{\lambda} \quad (3)$$

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 6, June 2016

From this, the probability of selection of ASR errors is reduced because of lower βz value. To find the optimal Keyword set the greedy algorithm is used which is as follows:

1. Initially S is empty.
2. Select at each step unselected word from conversation fragment w ε t \S which has maximum similarity to main topics with respect to previously selected keywords in S.
3. Maximum similarity is find out by $h(w, S) = \sum_{z \in Z} \beta_z [p(z|w) + rS,z]\lambda$, where p(z|w) is contribution to topic z by word w.
4. Set S is updated by adding word w ε t \S to set S which maximizes h(w, S)
5. This procedure continues until reaching k keywords from the fragment t.

C. Recommendation of Documents

The documents can be recommended to user using following methods:

1. Keywords obtained from Diverse Keyword Extraction are used as query for implicit query preparation. These keywords are compared to documents in the dataset for searching the relevant documents. The document results are prepared by selecting first d documents retrieval results of the query and are provided to user as recommendation based on conversation. Also, we will get the google links and PDFs by querying various keyword combinations over google search engine.

2. Each keyword from keyword set is compared against the paper vector constructed in background using CORA dataset. Keywords which are matched in vector are given as recommendation to user along with other papers which are related to that paper as references. Table 1 shows the relations between papers i1, i2, i3, i4, i5 using paper vector. The value 1 or 0 is used for stating whether two papers are co-occurred or not.

	I1	I2	I3	I4	I5
I1	0	1	1	0	1
I2	1	0	1	1	1
I3	1	1	0	1	1
I4	0	1	1	0	0
I5	1	1	1	0	0

Table 2: Paper vector

3. First all the documents in CORA dataset are clustered using K-means algorithm. After formation of clusters the keywords from the keyword set are compared to clusters. Cluster which is most relevant to the keywords is given as recommendation to user. The recommendation includes all the cluster documents which are relevant to keyword set.

   The k-means algorithm takes the input parameter k, and partitions a set of n objects into k clusters so that the resulting intra cluster similarity is high but the inter cluster similarity is low. Cluster similarity is measured in regard to the mean value of the objects in a cluster, which can be viewed as the clusters centroid or center of gravity.

   The k-means algorithm proceeds as follows. First, it randomly selects k of the objects, each of which initially represents a cluster mean or center. For each of the remaining objects, an object is assigned to the cluster to which it is the most similar, based on the distance between the object and the cluster mean. It then computes the new mean for each cluster. This process iterates until the criterion function converges.

D. Implementation Details

Non-Deterministic Finite Automata (NDFA): A deterministic finite automaton M is a 5-tuple, (Q,Σ,δ,q_0, F). Where,

- $Q = \{q_0, q_1, q_2, q_3, q_4, q_5, q_6, q_7, q_8\}$
- Σ={c, t, k, l ,d, pv, D, Cl, ci, R}.
- q_0 is start state.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 6, June 2016

– F= {q_8}.

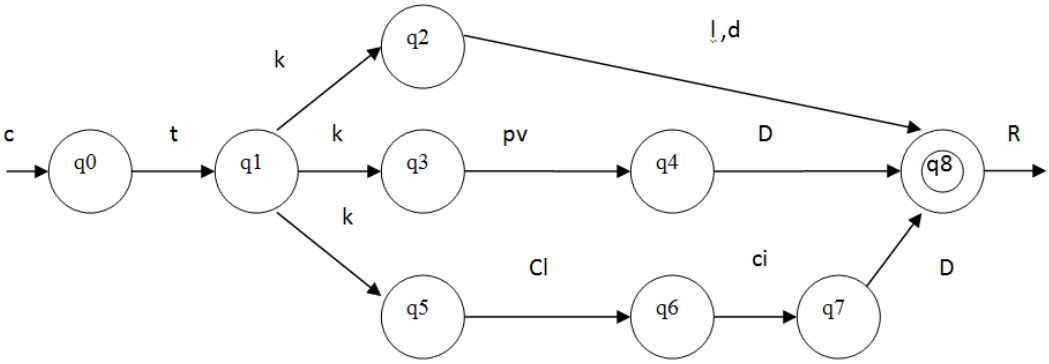

Fig 2: NDFA

Where,

q_0: ASR System Conversion.
q_1: Keyword extraction.
q_2: searching on google
q_3: paper vector calculation
q_4: Searching of keywords
q_5: cluster building
q_6: searching in cluster
q_7: recommend all cluster documents
q_8: Provide recommendations
c: conversation fragment
t: text obtained from ASR system.
k: keywords obtained from keyword extraction.
l , d: links and pdf documents
pv: paper vector
D: documents after keyword matching
Cl: clusters
Ci: matched cluster
R: recommendations

The proposed system is tested on The AMI Meeting Corpus which contains a total of 171 meetings. Speakers were not constrained to talk about a single topic throughout a meeting, hence these transcripts are multi-topic. Along with this Web-->KB and CORA dataset is used for giving recommendations.

All experiments will be conducted on a computer running an Intel Core 2 duo 2.93 GHz processor with 4 GB of memory and Windows 7 ultimate as the operating system. The proposed system is prototype using JAVA programming.

V. SIMULATION RESULTS

The proposed system is compared with existing in terms of recommendation of documents. The precision and recall are the evaluation metrics are used for comparison in this experiment. The comparison graph is shown in following graph. This study shows that proposed technique with clustering achieves more accurate recommendations.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 6, June 2016

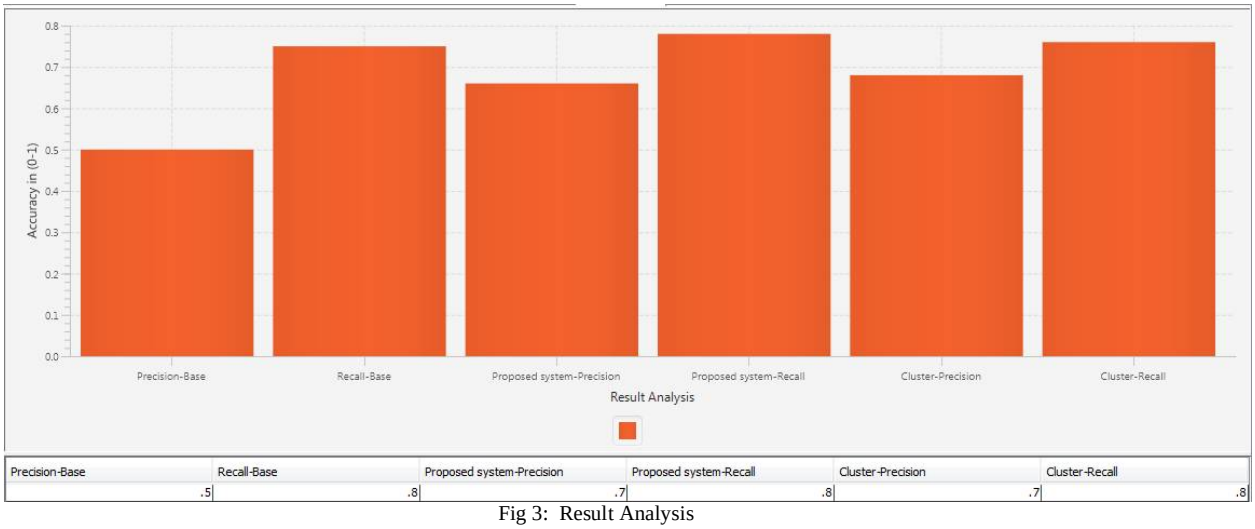

Fig 3: Result Analysis

VI. CONCLUSION AND FUTURE WORK

Just-in-time retrieval system is used to recommend the relevant documents to participants in conversational environments. From the conversation fragment implicit queries are formed to model the information needs of participants. These queries are formulated by using the extracted keywords. The proposed system uses diverse keyword extraction technique for finding the keywords. This technique covers the maximal number of important topics in a fragment. Recommendations are given to user in the form of links and PDFs by querying the keyword set over google search engine. Also a new recommendation technique is used which considers the paper vector concept. This method improves the recommendations by giving relevant recommendations documents as well as with their co-occurring documents. Various experimental results show that proposed system outperforms the existing system in terms of precision and recall. The proposed system shows 65-70 percent of precision value against the existing system.

VII. ACKNOWLEDGEMENTS

The author is thankful to MET's Institute of Engineering Bhujbal Knowledge City Nashik, HOD of computer department, guide, parents and friends for their blessing, support and motivation behind this work.

REFERENCES

1. G. Salton and C. Buckley, Term-weighting approaches in automatic text retrieval, In Information Processing & Management. Journal, vol. 24, no. 5, pp. 513-523, 1988.
2. B. Rhodes and T. Starner, Remembrance Agent: A continuously running automated information retrieval system, in Proc. 1st Int. Conf. Pract. Applicat. Intell.Agents Multi Agent Technol., London, U.K., 1996, pp. 487-495.
3. B. J. Rhodes, The wearable Remembrance Agent: A system for augmented memory, In Personal Technol., vol. 1, no. 4, pp. 18-224, 1997.
4. Budzik, J., and Hammond, K. Watson: Anticipating and Contextualizing Information Needs, In Proceedings of the ASIS 1999 Annual Conference. Information Today, Inc., Medford NJ, 1999.
5. M. Czerwinski, S. Dumais, G. Robertson, S. Dziadosz, S. Tiernan, and M. Van Dantzich, Visualizing implicit queries for information management and retrieval, in Proc. SIGCHI Conf. Human Factors Comput. Syst. (CHI), 1999, pp. 560-567.
6. B. J. Rhodes and P. Maes, Just-in-time information retrieval agents, In IBM Syst. J., vol. 39, no. 3.4, pp. 685-704, 2000.
7. A. J. Budzik and K. J. Hammond, User interactions with everyday applications as context for just-in-time information access, in Proc. 5th Int. Conf. Intell. User Interfaces (IUI00), 2000, pp. 44-51.
8. Y. Matsuo and M. Ishizuka,Keyword extraction from a single document using word co-occurrence statistical information, in Int. J. Artif. Intell. Tools, vol. 13, no. 1, pp. 157-169, 2004.
9. S. Dumais, E. Cutrell, R. Sarin, and E. Horvitz, Implicit queries (IQ) for contextualized search, in Proc. 27th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 2004, pp. 594-594.
10. 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.
11. 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.
12. Z. Liu, W. Huang, Y. Zheng, and M. Sun, Automatic keyphrase extraction via topic decomposition, in Proc. Conf. Empir. Meth. Nat. Lang. Process. (EMNLP10), 2010, pp. 366-376.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 6, June 2016

13. Popescu-Belis, M. Yazdani, A. Nanchen, and P. N. Garner, A speech-based just-in-time retrieval system using semantic search, in Proc. Annu. Conf. NorthAmer. Chap. ACL (HLT-NAACL), 2011, pp. 80-85.
14. 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.
15. M. Habibi and A. Popescu-Belis, Diverse keyword extraction from conversations, in Proc. 51st Annu. Meeting Assoc. Comput. Linguist. 2013, pp. 651-657.
16. F. Meng, D. Gao, W. Li, X. Sun, and Y. Hou, A unified graph model for personalized query-oriented reference paper recommendation , in Proc. 22nd ACM Int. Conf. Inf. Knowl. Manage. , 2013, pp. 1509-1512.
17. 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.
18. R. Thiyagarajan, K. Thangavel, R. Rathipriya, Recommendation of Web Pages using Weighted K-Means Clustering, in International Journal of Computer Applications (0975-8887), Volume 86, No. 14, January 2014, pp. 44-48.
19. M. Habibi and A. Popescu-Belis, Keyword Extraction and Clustering for Document Recommendation in Conversations, in IEEE transaction on Audio, Speech, and language processing , Vol. 23, No. 4 , 2015, pp. 746-759.
20. Haifeng Liu, Xiangjie Kong, Xiaomei Bai, Wei Wang,Teshome Bekele, and Feng Xia, Context-Based Collaborative Filtering for Citation Recommendation, in IEEE Access , Vol. 3, 2015, pp. 1695-1703.