Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 5.258

IJCSMC, Vol. 5, Issue. 3, March 2016, pg.655 – 658

Document Recommendation for Conversation Based on Keyword Extraction and Clustering

G.Subhashini[1], D.Jayakumar[2]
B.E Student, Department of Computer Science and Engineering, IFET College of Engineering
Subhashini24g@gmail.com
jayakumar1988@hotmail.com

Abstract- Through this project we are extracting appropriate keyword from conversation input. Extracted keywords are matched with available documents. Finally, we recommend appropriate documents to the participants for reference. It also represents the problem faced during keyword extraction in conversation using automatic speech recognition (ASR) system which brings errors in result. In order to overcome this problem, proposed work introduces method to retrieve multiple queries from keyword set, in ordered to maximize the chances of making 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. The scores show that our proposal improves over previous methods that consider only word frequency or topic similarity and represents a promising solution for a document recommender system to be used in conversations.

I. INTRODUCTION

This document is a template. An electronic copy can Humans are encompassed by abundance of data, accessible as records, data stores, or mixed media sources of information. Access to this data is adapted by the accessibility of suitable web indexes, however when these are accessible, clients frequently don't start a search action, in light of the reality that their current action does not permit them to do as such, or in light of the reality that they are not mindful that applicable data is accessible. We suggest a novel technique in this paper of suggesting archives just-in-time that is identified with clients' present work. At the point when these tasks are primarily conversational, for instance when clients take part in a meeting, their data needs can be understood by the keywords present in speech, acquired through continuous automatic speech recognition(ASR) engine. These certain implicitly generated questions are utilized to recover and suggest reports from the Web or a localStorehouse, which clients can decide to, investigate in

more detail if they discover them intriguing. The focus of this paper is on figuring framework for utilization in meeting rooms/conferences where information is to be fetched in-time to assist involved people in better understanding of the topic. This framework will be including speech to text translation, extraction of words from this text, formulating implicit queries from the words and fetching documents from the available storehouses relating to the words in the queries.

II. SURVEY RESULTS

Maryam Habibi and Popescu-Belis (2014) has discussed an algorithm for diverse merging of these lists, using a sub modular reward function that rewards the topical similarity of documents to the conversation words as well as their diversity. We evaluate the proposed method through crowdsourcing. The results show the superiority of the diverse merging technique over several others which not enforce the diversity of topics. Gerad salton and Christopher Buckley(1988) shows that This article summarizes the insights gained in automatic term weighting and provides baselines single-term-indexing models with which others more elaborate content analysis procedures can be compared. Text indexing system based on the assignment of appropriately weighted single terms produce retrieval results that are superior to those obtainable with other more elaborate text representations.These results depend crucially on the choice of effective term weighting systems. H.Luhn shows that a statistical approach to this problem will be outlined and the various steps of a system based on this approach will be described. Steps include the statistical analysis of a collection of documents in a field of interest, the establishment of a set of "notions" and the vocabulary by which theyare expressed, the compilation of a thesaurus-type dictionary and index, the automatic encoding of documents by machine with the aid of such a dictionary, the encoding of topological notations (such as branched structures), the recording of the coded information, the establishment of a searching pattern for finding pertinent information, and the programming of appropriate machines to carry out a search.

III. EXISTING SYSTEM

The problem of keyword extraction from conversations, with the goal of using these keywords to retrieve, for each short conversation fragment, a small number of potentially relevant documents, which can be recommended to participants. However, even a short fragment contains a variety of words, which are potentially related to several topics; moreover, using an automatic speech recognition (ASR) system introduces errors among them. Therefore, it is difficult to infer precisely the information needs of the conversation participants

IV. MODULE DESCRIPTION

1.SPEECH TO TEXT CONVERSION

It performs the task of conversion from speech to text according to user conversation input. It takes audio file as input and produces text file as output. After conversion, output will be stored in separate text file. The required validation will be performed before making the conversion.

2. KEYWORD EXTRACTION

It can be performed after the speech to text conversion module which depends on it for input. For keyword extraction, first we have to remove all stop words from the input .Next, find the frequency and degree for each word in order to extract keyword from the input file.To improve over frequency-based methods, several ways to use lexical semantic information have been proposed. Semantic relations between words can be obtained from a manually-constructed thesaurus such as WordNet, or from Wikipedia, or from an automatically-built thesaurus using latent topic modelling techniques such as LSA, PLSA, or LDA. For instance, keyword extraction has used the frequency of all words belonging to the same WorldNet concept set [4], while the Wilier system [5] relied on Wikipedia links to compute another substitute to word frequency. Hazen also applied topic modelling techniques to audio files [26]. In another study, he used PLSA to build a thesaurus, which was then used to rank the words of a conversation transcript with respect to each topic using a weighted point-wise mutual information scoring function [27]. Moreover, Harwich and Hazen utilized PLSA to represent the topics of a transcribed conversation, and then ranked words in the transcript based on topical similarity to the topics found in the conversation [6]. Similarly,

Harwath et al. extracted the keywords or key phrases of an audio file by directly applying PLSA on the links among audio frames obtained using segmental dynamic time warping, and then using mutual information measure for ranking the key concepts in the form of audio file snippets [28]. A semi-supervised latent concept classification algorithm was presented by Celikyilmaz and Hakkani-Tur using LDA topic modeling for multi-docu- mint information extraction [29].

3.DATA CLUSTERING

Clustering is the process of making a group of similar data . While doing cluster analysis, we first partition the set of data into groups based on data similarity and then assign the labels to the groups.The main advantage of clustering over classification is that, it is adaptable to changes and helps single out useful features that distinguish different groups.

4.DOCUMENT RECOMMENDATION

As a first idea, one implicit query can be prepared for each conversation fragment by using as a query all keywords selected by the diverse keyword extraction technique. However, to improve the retrieval results, multiple implicit queries can be formulated for each conversation fragment.By using multiple implicit queries, Suitable document will be recommended to user based on their conversional data.

V. PROPOSED SYSTEM

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. The scores show that our proposal improves performance over previous methods that consider only word frequency or topic similarity, and represents a promising solution for a document recommender system to be used in conversations.

VI. SYSTEM ARCHITECTURE

Figure 1 shows user conversation recorded by ASR system reduces noises from it and output of ASR send further. Then various keywords are extracted by keyword extraction method. Then clustering of diverse keywords is done. Further these keywords are identified or matched by implicit queries. By ranking of documents, the recommendation of related document can be easily made.

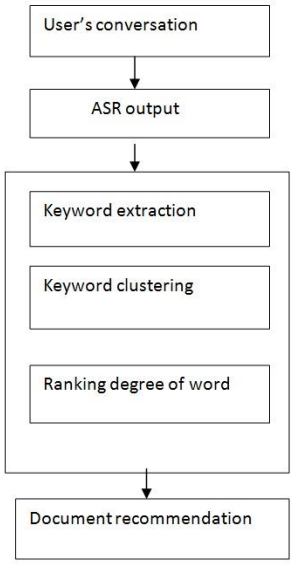

Figure 1.SYSTEM ARCHITECTURE

VII. CONCLUSIONS

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 a novel diverse keyword extraction technique which covers the maximal number of important topics in a fragment. Then, to reduce the noisy effect on queries of the mixture of topics in a keyword set, we proposed a clustering technique to divide the set of keywords into smaller topically-independent subsets constituting implicit queries. 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 ratters recruited via the Amazon Mechanical platform. The experiments showed that the diverse keyword extraction method provides on average the most representative keyword sets for recommending documents. Therefore, enforcing both relevance and diversity brings an effective improvement to keyword extraction and document retrieval.

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 concept lattice for text understanding and summarization," *Inf. Process. Manage.*, vol.43, no. 6, pp. 1643–1662, 2007..
[5] A. Csomai and R. Mihalcea, "Linking educational materials to ency-clopedic knowledge," in Proc. Conf. Artif. Intell. Educat.: Building Technol. Rich Learn. Contexts That Work, 2007, pp.557–559.
[6] D.HarwathandT.J.Hazen, "Topic identification based xtrins I evaluation of summarization techniques appliedtoconversationalspeech," in Proc. Int. Conf. Acoust., Speech, Signal Process. (ICASSP), 2012, pp. 5073–5076.
[7] 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.
[8] A. 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. North Amer. Chap. ACL (HLT-NAACL), 2011, pp. 80–85.
[9] P. E. Hart and J. Graham, "Query-free information retrieval," Int. J. Intell. Syst. Technol. Applicat., vol. 12, no. 5, pp. 32–37, 1997.
[10] B. Rhodes and T. Starner, "Remembrance Agent: A continuously run- ning automated information retrieval system," in Proc. 1st Int. Conf. Pract. Applicat. Intell. Agents Multi Agent Technol., London, U.K., 1996, pp. 487–495.
[11] B. J. Rhodes and P. Maes, "Just-in-time information retrieval agents," IBM Syst. J., vol. 39, no. 3.4, pp. 685–704, 2000.
[12] B. J. Rhodes, "The wearable Remembrance Agent: A system for augmented memory," Personal Technol., vol. 1, no. 4, pp. 218–224, 1997.
[13] 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 (IUI'00), 2000, pp. 44–51.
[14] 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.
[15] 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.

