Scientific Journal of Impact Factor (SJIF): 4.14
e-ISSN (O): 2348-4470
p-ISSN (P): 2348-6406

International Journal of Advance Engineering and Research Development

Volume 3, Issue 3, March -2016

Document Recommendation for Conversation using Keyword Extraction with Clustering

Yogesh Daundkar[1], Shubham Potdukhe[2], Virendra Saini[3], Snigdha Swaroop[4], Prof. Pragati Kinage[5]
[1-5]*Department Of Computer Engineering, Indira college of engineering and management, Pune*

Abstract *— The system perform the extraction of keyword its address the problem for conversation for each short conversion section. A less number of potentially significant documents with the goal of using the information retrieved which can be recommended to participant. Using automatic speech recongnization system introduce error among them which are potentially related to various topic, even short fragment contains a variety of word. Therefore, it is complicated to infer specifically the information needs the discussion of participants. The use of topic modeling techniques and of a sub modular reward function which favors diversity in the keyword set, for making to match the potential diversity of topic and reduce ASR noise. Then, paper propose a method to derive several topically divided queries from this keyword set, in order to make the most of the chances of building at least one significant recommendation when using these queries to search over the English Wikipedia. The Fisher, AMI, and ELEA conversational corpora, rated by various human judges by using proposed methods are calculated in terms of significance with respect to conversation fragments from. The scores show that our proposal improves over previous methods that consider only word frequency or topic correspondence, and represents a promising solution for a document recommender system to be used in conversations.*

Keywords- *Document recommendation, information retrieval, keyword extraction, meeting analysis, topic modeling.*

I. INTRODUCTION

An unprecedented wealth of information, accessible as documents, databases, or multimedia resources which are used by human. For retrieving this information is conditioned by the availability of appropriate search engines, but even when these are available, users often do not begin a search, because their current activity does not allow them to do so, or because they are not aware that significant information is presented. In this paper the perspective of just-in-time retrieval, which answers this shortcoming by suddenly recommending documents that are associated to users' current activities. When these activities are mainly conversational, for instance when users take part in a meeting, their information needs can be modeled as implicit queries that are constructed in the background from the pronounced words, obtained through real-time automatic speech recognition (ASR). These implicit queries are used to retrieve and recommend documents from the Web or a local repository, which users can choose to inspect in more detail if they find them interesting. Our goal is to maintain multiple hypotheses about users' information needs, and to present a small sample of recommendations based on the most likely ones. Therefore, system goal at extracting a relevant and diverse set of keywords, cluster them into topic-specific queries ranked by importance, and present users a sample of results from these queries. The topic-based clustering decreases the chances of including ASR errors into the queries, and the diversity of keywords increases the chances that at least one of the recommended documents answers a need for information, or can lead to a useful document when following its hyperlinks.

In this paper, system introduces a novel keyword extraction technique from ASR output, which maximizes the coverage of potential information needs of users and reduces the number of inappropriate words. Once a set of keywords is extracted, it is clustered in order to construct several topically-separated queries, which are run separately, offering better precision than a larger, topically-mixed query. Results are finally merged into a ranked set before showing them as recommendations to users.

II. LITERATURE REVIEW

1) Enforcing topic diversity in a document recommender for conversations

AUTHORS: M. Habibi and A. Popescu-Belis,

This paper addresses the problem of construction summarizing, diverse and significant lists of documents, which can be recommended to the participants of a conversation to complete their information needs without distracting them. These lists are retrieved periodically by submitting various implicit queries derived from the pronounced words. Each query is related to one of the topics identified in the conversation fragment preceding the recommendation, and is submitted to a search engine over the English Wikipedia. System propose an algorithm for diverse merging of these lists, using a sub modular reward function that rewards the topical correspondence of documents to the conversation words as well as their multiplicity. They calculated the proposed method through crowd sourcing. The results show the authority of the diverse merging technique over a number of others which not enforce the diversity of topics.

@IJAERD-2016, All rights Reserved

2) A statistical approach to mechanized encoding and searching of literary information
AUTHORS: H. P. Luhn

The system proposed written communication of ideas is carried out on the basis of statistical probability in that a writer chooses that level of subject specificity and that combination of words which he feels will convey the most meaning. Since this process varies among individuals and since similar ideas are therefore relayed at different levels of specificity and by means of different words, the problem of literature searching by machines still presents major difficulties. 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 they are 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.

3) Document concept lattice for text understanding and summarization
AUTHORS: S. Ye, T.-S. Chua, M.-Y. Kan, and L. Qiu

In this paper having concept of referring to an abstract or concrete entity or its action often expressed by diverse terms in text. Summary generation can thus be considered as an optimization problem of selecting a set of sentences with minimal answer loss. In this paper, a document concept lattice that indexes the hierarchy of local topics tied to a set of frequent concepts and the corresponding sentences containing these topics. The local topics will specify the promising sub-spaces related to the selected concepts and sentences. Based on this lattice, the summary is an optimized selection of a set of distinct and salient local topics that lead to maximal coverage of concepts with the given number of sentences. Our summarizer based on the concept lattice has demonstrated competitive performance in Document Understanding Conference.

4) Linking educational materials to encyclopedic knowledge
AUTHORS: A. Csomai and R. Mihalcea,

This paper describes a system that automatically links study materials to encyclopedic knowledge, and shows how the availability of such knowledge within easy reach of the learner can improve both the quality of the knowledge acquired and the time needed to obtain such knowledge. The disadvantage of Automatic speech recongnization system introduces errorin linking educational material to encyclopedia knowledge.

5) Remembrance Agent: A continuously running automated information retrieval system.
AUTHORS: B. Rhodes and T. Starner

The Remembrance Agent (RA) is a program which augments human memory displaying a list of documents which might be relevant to the user's current context. Unlike most information retrieval systems, the RA runs continuously without user intervention. Its unobtrusive interface allows a user to pursue or ignore the RA's suggestions as desired.

III. PROPOSED SYSTEM

The proposed system having two-stage approach to the formulation of implicit queries. The first stage is the extraction of keywords from the transcript of a conversation fragment for which documents must be recommended, as provided by an ASR system. These keywords should cover as much as possible the topics detected in the conversation, and if possible avoid words that are obviously ASR mistakes. The second stage is the clustering of the keyword set in the form of several topically-disjoint queries. Proposed system take advantage of topic modeling techniques to construct a topical representation of a conversation fragment, and then select content words as keywords by using topical similarity, while also rewarding the coverage of a diverse range of topics, inspired by recent summarization methods. The advantage of diverse keyword extraction is that the coverage of the main topics of the conversation fragment is maximized. Moreover, in order to cover more topics, the proposed algorithm will select a lesser number of keywords from each topic. Clusters of keywords are building by ranking keywords for each main topic of the fragment. Recommendation lists were prepared by selecting the first document retrieval results of each implicit query and then ranking documents based on the topical similarity of their corresponding queries to the conversation fragment. This is a baseline algorithm, but these describe an improved ranking method in.

IV. MATHEMATICAL MODEL

Let S is the Whole System Consist of

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

@IJAERD-2016, All rights Reserved

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.

Output: The output will be the response of the user query

V. SYSTEM ARCHITECTURE

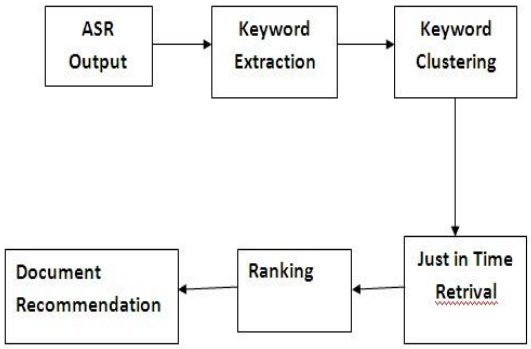

Fig1.System architecture

VI. CONCLUSION

We have considered a particular form of just-in-time retrieval systems intended for conversational environments, in which they recommend to users documents that are relevant to their information needs. We focused on modeling the user's 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. Our current goals are to process also explicit queries, and to rank document results with the objective of maximizing the coverage of all the information needs, while minimizing redundancy in a short list of documents. Integrating these techniques in a working prototype should help users to find valuable documents immediately and effortlessly, without interrupting the conversation flow, thus ensuring the usability of our system. In the future, this will be tested with human users of the system within real-life meetings.

VII. 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.

@IJAERD-2016, All rights Reserved

[5] A. Csomai and R. Mihalcea, "Linking educational materials to encyclopedic knowledge," in *Proc. Conf. Artif. Intell.Educat.: BuildingTechnol. Rich Learn. Contexts That Work*, 2007, pp. 557–559.

[6] 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.

[7] A. Popescu-Belis, E. Boertjes, J. Kilgour, P. Poller, S. Castronovo, T.Wilson, A. Jaimes, and J. Carletta, "The AMIDA automatic contentlinking 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, "Aspeech-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 running automated information retrieval system," in *Proc. 1st Int. Conf.Pract. Applicat.Intell. Agents Multi Agent Technol.*, London, U.K.,1996, pp. 487–495.

@IJAERD-2016, All rights Reserved