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

IJCSMC, Vol. 5, Issue. 1, January 2016, pg.116 – 120

Just In Time Retrieval System for Document Recommendation

Sayyed Shadab[1], Shaikh Asif[2], Sharma Deep[3], Mishra Vijeta[4]

[1]Student, Department of Computer Engineering, G.S.Moze College of University, India

[2]Student, Department of Computer Engineering, G.S.Moze College of University, India

[3]Student, Department of Computer Engineering, G.S.Moze College of University, India

[4]Project Guide, Department of Computer Engineering, G.S.Moze College of University, India

sayyedshadab17@gmail.com, sfshkh254@gmail.com, de.sharma@gmail.com, vijetamishra9@gmail.com

ABSTRACT - This paper addresses a diverse retrieval technique for ranking documents that are spontaneously retrieved and recommended to people during a conversation. These documents represent potentially useful information for the conversation participants [1]. In computer forensic analysis, hundreds of thousands of files are usually examined. Much of the data in those files consists of unstructured text, whose analysis by computer examiners is difficult to be performed. In this context, automated methods of analysis are of great interest [2]. A just-in-time information retrieval agent (JITIR agent) is software that proactively retrieves and presents information based on a person's local context in an easily accessible yet nonintrusive manner [4].The proposed evaluation method is shown to be reliable, and the results show that adding user initiative improves the relevance of recommendations [5].

Keywords: Document recommendation, information retrieval, keyword extraction, meeting analysis, topic modelling

1. INTRODUCTION

Keyword extraction is an important technique for document retrieval, Web page retrieval, document clustering, summarization, text mining, and so on. By extracting appropriate keywords, we can choose easily which document to read or learn the relation among documents. A popular algorithm for indexing is the tfidf measure, which extracts keywords frequently that appear in a document, but don't appear frequently in the remainder of the corpus [6].

In this paper, we propose a new method for keyword extraction that rewards both word similarity, to extract the most representative words, and word diversity, to cover several topics if necessary [7]. We define automatic key-phrase extraction as the automatic selection of important, topical phrases from within the body of a document. Automatic key-phrase extraction is a special case of the more general task of automatic key-phrase generation, in which the generated phrases do not necessarily appear in the body of the given document [16].

The focus of this paper is on formulating implicit queries to a just-in-time-retrieval system for use in meeting rooms. In this paper, we introduce a novel keyword extraction technique from ASR output, which maximizes the coverage of potential information needs of users and reduces the number of irrelevant words. Once a set of keywords is extracted, it is clustered in order to build several topically-separated queries, which are run independently, offering better precision than a larger, topically-mixed query. Results are finally merged into a ranked set before showing them as recommendations to users.

2. MATERIALS AND METHODS

Numerous methods have been proposed to automatically extract keywords from a text, and are applicable also to transcribed conversations. The earliest techniques have used word frequencies [2] and TFIDF values [3] to rank words for extraction. Alternatively, words have been ranked by counting pairwise word co-occurrence frequencies [6]. These approaches do not consider word meaning, so they may ignore low-frequency words which together indicate a highly-salient topic. For instance, the words 'car', 'wheel', 'seat', and 'passenger' occurring together indicate that automobiles are a salient topic even if each word is not itself frequent [9]. Supervised machine learning methods have been used to learn models for extracting keywords. This approach was first introduced by Turney [16], who combined heuristic rules with a genetic algorithm. Other learning algorithms such as Naïve Bayes [10], Bagging [11], or Conditional Random Fields [12] have been used to improve accuracy. These approaches, however, rely on the availability of in-domain training data, and the objective functions they use for learning do not consider the diversity of keywords.

In addition, the –NDCG measure [13] was used to measure topic diversity in the list of keywords. Afterward, the quality of implicit queries was assessed by estimating (again with human judges recruited via AMT) the relevance of the documents that were retrieved when submitting these queries to the Lucerne search engine over the English Wikipedia and merging the results as explained above.

3. CONFIGURATION MODULES

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 up to rank d in S is {Si : I : <= d}.noted as $S_{1:d}$. RBO is calculated as above equation.

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

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

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.

3.2 Keyword Clustering:

Clusters of keywords are built by ranking keywords for each main topic of the fragment. The keywords are ordered for each topic by decreasing values of β.p (z|w) Moreover, in each cluster, only the keywords with a β.p (z|w) value higher than a threshold are kept for each topic z.

3.3 Formulation of implicit queries from conversations:

We propose a 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

3.4 Just In Time Retrieval:

Just-in-time retrieval systems have the potential to bring a radical change in the process of query-based information retrieval. Such systems continuously monitor users' activities to detect information needs, and pro-actively retrieve relevant information. To achieve this, the systems generally extract implicit queries (not shown to users) from the words that are written or spoken by users during their activities. We review existing just-in-time-retrieval systems and methods used by them for query formulation.

3.5 Ranking:

Clusters of keywords are built by ranking keywords for each main topic of the fragment. Afterward, clusters themselves are ranked based on their β_z values.

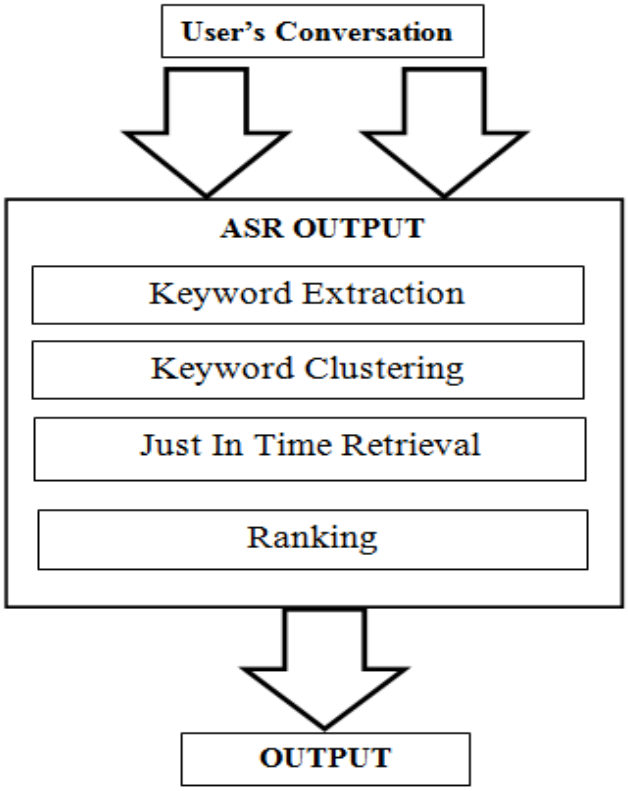

Fig. Configuration modules of JITR System

4. DESIGN AND DISCUSSION

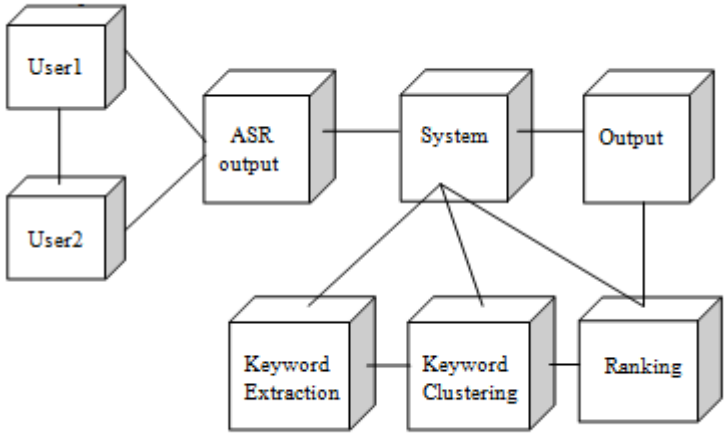

Fig. Architecture of JITR System for Document Recommendation

In this section, the diverse keyword extraction technique is compared with two state-of-the-art methods, showing that our proposal extracts more relevant keywords, which cover more topics, and are less likely to be ASR errors. Then, we compare the retrieval results of the implicit queries generated from these keyword lists, either applied entirely or by decomposing them into topically-separated queries, again showing that the lists generated by our method outperform existing methods.

We have shown that sub-modularity naturally arises in document summarization. Not only do many existing automatic summarization methods correspond to sub-modular function optimization, but also the widely used ROUGE evaluation is closely related to sub-modular functions [14]. In the context of text mining, to discover keywords or relation of keywords are important topics. The general purpose of knowledge discovery is to extract implicit, previously unknown, and potentially useful information from data. Our algorithm can be considered as a text mining tool in that it extracts important terms even if they are rare [6].

5. CONCLUSION

We proposed a diverse merging technique for combining lists of documents from multiple topically separated implicit queries, prepared using keyword lists obtained from the transcripts of conversation fragments [1]. A great advantage of the system described here would be the assurance that the idea content of information was being left intact. The ideas of an author would not be narrowed, biased, or distorted through the intervention of an interpreter [2]. In conclusion, just-in-time information retrieval agents are systems that proactively provide information based on a person's local context in a general, task-independent way. They are related to search engines, alarms, and news-clipping software, but differ from each [4].

We also proposed an unsupervised clustering-based key-phrase extraction algorithm. This method group's candidate terms into clusters and identify the exemplar terms. Then key-phrases are extracted from the document based on the exemplar terms. The clustering based on term semantic relatedness guarantees the extracted key-phrases have a good coverage of the document. Experiment results show the method has a good effectiveness and robustness, and outperforms baselines significantly [15]. 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.

ACKNOWLEDGEMENTS

The author's thanks for the acknowledgements from the anonymous reviewers for their precise comments and insightful remarks that improved the quality and clarity of our submission.

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] B. J. Rhodes and P. Maes, *"Just-in-time information retrieval agents,"* IBM Syst. J., vol. 39, no. 3.4, pp. 685–704, 2000.

[5]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.

[6] Y. Matsuo and M. Ishizuka, *"Keyword extraction from a single document using word co-occurrence statistical information,"* Int. J. Artif. Intell. Tools, vol. 13, no. 1, pp. 157–169, 2004.

[7] M. Habibi and A. Popescu-Belis, *"Diverse keyword extraction from conversations,"* in Proc. 51st Annu. Meeting Assoc. Comput. Linguist., 2013, pp. 651–657.

[8] G. Salton, C.-S. Yang, and C. T. Yu, *"A theory of term importance in automatic text analysis,"* J. Amer. Soc. Inf. Sci., vol. 26, no. 1, pp. 33–44, 1975.

[9] A. Nenkova and K. McKeon, *"A survey of text summarization techniques,"* in Mining Text Data, C. C. Aggarwal and C. Zhai, Eds. New York, NY, USA: Springer, 2012, ch. 3, pp. 43–76.

[10] E. Frank, G. W. Paynter, I. H. Witten, C. Gutwin, and C. G. Nevill- Manning, "*Domain-specific key-phrase extraction,"* in Proc. 16th Int. Joint Conf. Artif. Intell. (IJCAI'99), 1999, pp. 668–673.

[11] A. Hulth, *"Improved automatic keyword extraction given more linguistic knowledge,"* in Proc. Conf. Empir. Meth. Nat. Lang. Process. (EMNLP'03), 2003, pp. 216–223.

[12] 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.

[13] C. L. A. Clarke, M. Kolla, G. V. Cormack, O. Vechtomova, A. Ashkan, S. Büttcher, and I. MacKinnon, *"Novelty and diversity in information retrieval evaluation,"* in Proc. 31st Annu. Int. ACM SIGIR Conf. Res.

Develop. Inf. Retrieval, 2008, pp. 659–666.

[14] H. Lin and J. Bilmes, *"A class of submodular functions for document summarization,"* in Proc. 49th Annu. Meeting Assoc. Comput. Linguist. (ACL), Portland, OR, USA, 2011, pp. 510–520.

[15] Z. Liu, P. Li, Y. Zheng, and M. Sun, *"Clustering to find exemplar terms for keyphrase extraction,"* in Proc. Conf. Empir.Meth. Nat. Lang. Process. (EMNLP'09), 2009, pp. 257–266.

[16] P. Turney, "*Learning to extract key-phrases from text National Research Council Canada (NRC)*," Tech. Rep. ERB-1057, 1999.

