ISSN (Online) 2393-8021
ISSN (Print) 2394-1588
International Advanced Research Journal in Science, Engineering and Technology
Vol. 2, Issue 10, October 2015

Extraction and Clustering of Keywords for Documents

Mr. Milind Hegade[1], Monika Korde[2], Monika Nawale[3], Snehal Kulkarni[4]

Professor, Department of Computer Engineering, JSPM'S BSIOTR, Pune, India [1]

Student, Department of Computer Engineering, JSPM'S BSIOTR, Pune, India [2, 3, 4]

Abstract: In this topic there are large numbers of documents which are cover more information about any topic. We are extracting one keyword from that document, when we are extracting this keyword can easily retrieve whole document. However, even a small piece contains a variety of words, which are potentially related to several topics; more- over, using an automatic speech recognition (ASR) system introduce errors among them. There for, it is difficult to infer precisely the in sequence requirements of the discussion participants. We first propose an algorithm to extract keywords from the output of an ASR system which makes use of topic modeling techniques and of a sub modular reward function which favors range in the keyword set, to match the possible range of topics and reduce ASR noise. This method is to derive many topically divided queries starting this keyword set, in organize to take full advantage of the probability of making at least one related reference when with these queries to search over the English Wikipedia. Examples like Fisher, AMI, and ELEA conversational corpora.

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

I. INTRODUCTION

Data mining is the procedure that attempts to find out patterns in large data sets. It utilizes methods at the intersection of fake aptitude, machine learning, statistics, and database systems. The overall goal of the data mining procedure is to remove in sequence from a data set and transform it into an understandable structure for further use.

Data is accessible in the form of databases, ID & multimedia resources. Access to this information is conditioned by the availability of suitable search engines. But even these are available users can not search particular information because they are not aware that relevant information is available. Just-in-time-retrieval system which is observes the current activities of users & provides relevant information. A just-in-time information retrieval agent is software that proactively retrieves and presents in sequence based on a person's local situation in an easily accessible yet nonintrusive manner.They continuously watch a person's environment and present information that may be useful without requiring any action on the part of the user.

Automatic speech recognition is the process by which a computer maps an acoustic speech indication to passage. Automatic speech appreciative is the process by which the computer maps an acoustic speech signal to some form of abstract meaning of the speech. A new method for keyword extraction from conversations is introduced, which preserves the diversity of topics.

Topic based clustering that aims only to solve the problem of grouping together articles of similar topic. News organization would like to be able to access related document with minimum effort. The topic based clustering decreases the probability of including ASR errors into the queries, and the diversity of keywords increases the probability that at least single of the recommended papers answers a need for information, or can main to a useful text when following its hyperlinks. Relevance and diversity can be enforced at three stages: when extracting the keywords; when structure one or some implicit queries; or when re-ranking their results.

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 approximately 250 words, connecting to a assorted 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?

II. IMPLEMENTATION

i. State of the art: just-in-time retrieval and keyword extraction
ii. Formulation of implicit queries from conversations
iii. Data and evaluation methods

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

Such frames persistently screen clients' workouts to separate data needs, and intellect effectively recovers applicable data. To complete this, the frames by and large focus certain questions (not indicated to clients) from the words that are composed or talked by clients amid their workouts. In this part, we study existing without a instant to replacement recovery frames and sequences utilized by them for analysis detailing. Definitely, we will contemporary our Automatic Content Linking Device (ACLD) a without a moment to replacement record suggestion frame for assemblies, for which the sequencessuggested in this paper are expected. In II-B, we talk about past necessary word mining actions from a transcription or content.

a. Query Formulation in Just-in-Time Retrieval Systems

One of the initial systems for paperauthorization, denotedto as query-free examine, was the Fixit system, an associateto anknowledgeableanalytic system for the producers of a particular company (fax machines and copiers). Fixit monitored the stateof the user's interface with the diagnostic system, in terms ofthe positions in a belief network built from the relations amongindications and errors, and ran background searches on a databankof preservationguides to provide extra support datacorrelated to the current state. The Recollection Agent, another quick in the nick of time recovery frame, is closer in idea to the frame considered in this paper. The Recollection Agent was integrated into the Emacs content tool, and ran looks at normal time intervals (like clockwork) utilizing a question that was in light of the most current words wrote by the user, for example using a structure of 20–500 words positioned by repetition. The Recollection Agent was got out to a multimodal setting under the name of Jimminy, a wearable right hand that helped users with taking records and getting to data when they couldn't use a standard PC console, e.g. while conversation about with someone else. Consuming TFIDF for significant word abstraction, Jimminy enlarged these watchwords with features from different modalities, for instance the user's position and the name of their converser(s).

b. Keyword Extraction Methods

Differentschemes have been suggested to subsequentlyeliminatekey words from a content, and are relatedalso to interpreted discussions. The most punctual processes have used word frequencies and TFIDF qualities to rank words for abstraction. On the further side, words have been located by checking pairwise word co-event frequencies. These methods don't reflect word significance, so they may overlook low-recurrence words which together demonstrate an extremelynotable subject. Semantic relations between terms can be developed from a really developed dictionary, for example, Word Net, or from Wikipedia, or from a logicallygathereddictionarydeveloping idle issuesschemes, for example, LSA, PLSA, or LDA. Hazen also usefulissuemodeling techniques to audio records. In alternative learning, heused PLSA to build a dictionary, which was then used to gradethe terms of a discussiontext with respect to eachtopic using a weighted point-wise commondatacountingpurpose.

ii. Formulation of Implicit Queries from conversations

We advise a two- phasetechnique to the formulation ofimplicit queries. The first phase is the extraction of keywordsfrom the record of a discussionpart for which documentsmust be suggested, as provided by an ASR system. These keywords should cover as much aspossible the areasnoticed in the discussion, and if potentialkeep away from words that are clearly ASR errors. The secondstage is the clustering of the keyword set in the form of some topically-disjoint queries

a. Diverse Keyword Extraction

We adviseto takeadvantage of topic modeling techniquesto build a topical representation of a discussionpart,and then select content words as keywords by using relevantrelationship, while also fulfilling the reporting of a variousrange of subjects, motivated by recent summarization methods. The benefit of diverse keyword extraction is that the coverage of the main subjects of the discussion part is maximized.

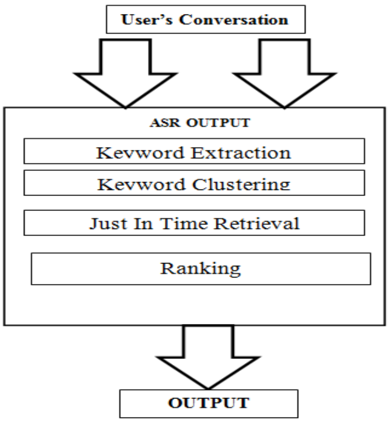

Fig. 1 The three steps of the proposed keyword extraction method

Additionally, in order to cover more subjects, thesuggestedalgorithm willchoice a smaller number of keywordsfrom each subject. This is required for two reasons. Thiswill lead to more different implicit queries, thus increasing themultiplicity of recovered documents. and, if words which arein actuality 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 overlook mixture.

Algorithm 1: Diverse keyword extraction.

Input: a given text t, a set of topics Z, the number of keywords k

Output: a set of keywords S

$S \leftarrow \emptyset$

While $|S| \leq k$ **do**

$S \leftarrow S \cup \{argmax_{w \in t \setminus S}(h(w,S))$ where $h(w,S) = \sum_{z \in Z} \beta_z [p(z|w) + r_{S,z}]^{\lambda}$;

end

return S

The benefit of diverse keyword extraction is thatthe coverage of the main topics of the conversation fragmentis

ISSN (Online) 2393-8021
ISSN (Print) 2394-1588
International Advanced Research Journal in Science, Engineering and Technology
Vol. 2, Issue 10, October 2015

maximized.The future method for diverse keyword extraction proceeds in three steps,

1. Used to represent the division of the abstract subject 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 keywordlist W = {w1, w2……wk}. which covers a maximum number of the most important topics are preferred by rewarding range, using an unique algorithm introduced in this part.

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 overlapSand 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 is {Si : I : <= d}.noted as $S_{1:d}$. RBO is calculated as above Equ.

a. Keyword Clustering

The different set of extracted keywords is measured to denotethe possible information needs of the applicants to adiscussion, in terms of the ideas and topics that are declaredin the discussion. To maintain the variety of topicsalive in the keyword set, and to decrease the noisy result ofeach data need on the others, this set must be divided intoseveral topically-disjoint subsets. Each subset corresponds thento an implicit query that will be sent to a document recovery system. These subsets are obtained by clustering topically-similarkeywords, as follows.

Clusters of keywords are constructed 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 thana threshold are kept for each topic z.

b. From Keywords to Document Recommendations

As a first impression, one implicit query can be arranged for eachdiscussionpart by using as a query all keywords specialby the various keyword removaltechnique. However, to improvethe retrieval results, multiple implicit queries can be formulatedfor each discussionpart, with the keywords ofeach cluster from the beforefragment. In tests with only one implicit query per discussionfragment, the document results parallel to each discussionfragment were arranged by selecting the first documentretrieval results of the implicit query.

iii. Data and evaluation methods

Our proposals were tested on three conversational corpora, the Fisher Corpus, the AMI Meeting Corpus, and the ELEA Corpus. The significance of the keywords was assessed by designing association task and averaging some judgments obtained by mobsourcing this assignment through the AmazonMechanical Turk (AMT) stage. In addition, the –NDCGmeasure was used to determine topic range in the catalogof keywords. Afterward, the quality of implicit queries was assessedby estimating (again with human judges recruited viaAMT) the significance of the papers that were retrieved whensubmitting these queries to the Lucene search engine over theEnglish Wikipedia and merging the results as explained above.Here, the conversational data came only from the ELEA Corpus,which offers clearer criteria for assessing the significance of referencesthan the Fisher and AMI Corpora. We now describethe three corpora and the data extracted from them, aswell as the evaluation methods for each task .

a. Conversational Corpora Used for Experiments

The Fisher Corpus contains about 11,000 topic-labeledtelephone conversations, on 40 pre-selected topics (one per conversation).

In our experiments, we used the manual suggestion transcripts available with the corpus. We created a topic model using the Mallet implementation of LDA, over two thirdsof the Fisher Corpus, given the enough number of single-topicdocuments, fixing the number of abstract topics at 40. The remainingdata was used to build 11 fakediscussion fragments for testing, by concate 11 timesthree remains about three dissimilar topics. The AMI Meeting Corpus contains discussion on manipulativeremote controls, in sequence of four scenario-based meetingseach, for a total of 171 meetings. Speakers were not constrainedto talk about a single topic throughout a meeting, hencethese transcripts are multi-topic4. Since the number of meetingsin the AMI Corpus is not large enough for building topicmodels with LDA, we used a subset of the English Wikipediawith 124,684 articles. Following several previous studies, we fixed the number of topics at 100.We chosen for trying 8 conversation fragments, each2–3 minutes long, from the AMI Corpus. We used both manualand ASR transcripts of these fragments. The ASR transcriptswere generated by the AMI real-time ASR system for meetings, with an average word error rate (WER) of 36%.

b. Evaluation Protocol and Metrics

We designed comparison tasks to evaluate the relevance ofextracted keywords and of recommended documents with respectto each discussion fragment. For the former evaluation,we compared the relevance (or representativeness) of twolists of keywords extracted from the same conversation fragmentby two different extraction methods. We displayed thetranscript of the fragment to a human subject in a web browser,followed below it by several control questions about its content,and then by two lists of keywords (typically, nine keywordsin our experiments).The subjects had to read the conversationtranscript, answer the control questions, and then decidewhich keyword cloud better represented the content of the discussionfragment.Withoutthe control questions or the discussion transcript. A similarmethod was applied to compare recommended documents, exceptthat two lists of retrieved documents (typically, with sevenitems each) are shown instead of word clouds, and their potentialutility as recommendations to the discussion participantsis compared.

III. 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 users information needs by deriving implicit queries from short discussion 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 importanttopics in a part. Then, to reduction the loud 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.

ACKNOWLEDGMENT

Dairazalia Sanchez-Cortes and the Idiap Social Computing group for access to the ELEA Corpus. We also identify the strange commentator for their selective comments and perspective comments that enhanced the excellence and cleanness 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] 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 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,”inProc. 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,” inProc. 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.
[11] Maryam Habibi and Andrei Popescu-Beli,”Keyword extraction and clustering for Documents Recommendation in Conversation” in Proc.IEEE/ACM transaction on audio, speech, and language processing, VOL.23, NO.4 Aprial 2015.