Scientific Journal of Impact Factor (SJIF): 3.134

E-ISSN (O): 2348-4470
P-ISSN (P): 2348-6406

International Journal of Advance Engineering and Research Development

Volume 2,Issue 12,December -2015

A Survey on Keyword Extraction for Document Recommendation in Conversations

Kolhal Devyani Rajendra[1], Prof.R.B. Thakur[2]

[1]Department Of Computer Engineering, IOK college of engg. pimpale jagtap ,shikrapur, tal- shirur,dist- pune,
[2]Department Of Computer Engineering, IOK college of engg. pimpale jagtap ,shikrapur, tal- shirur,dist- pune,

Abstract *— This paper addresses the issue of keyword extraction from discussions, with the target of using these keywords to recuperate, for each short examination piece, somewhat number of possibly apropos reports, which can be recommended to individuals. Regardless, even a short piece contains a mixed bag of words, which are possibly related to a couple subjects; additionally, using a automatic speech recognition (ASR) system presents slips among them. Thusly, it is difficult to gather effectively the information needs of the talk individuals. We first propose a computation to expel conclusive words from the yield of an ASR structure (or a manual transcript for testing), which makes usage of topic showing systems and of a sub modular prize limit which backings contrasting qualities in the enchantment word set, to facilitate the potential varying characteristics of subjects and decline ASR hullabaloo. By then, we propose a system to construe different topically confined request from this definitive word set, remembering the deciding objective to enhance the conceivable outcomes of making at any rate one applicable proposition while using these inquiries to look for over the English Wikipedia. The proposed frameworks are surveyed similarly as centrality with respect to examination pieces from the Fisher, AMI, and ELEA conversational corpora, assessed by a couple of human judges. The scores exhibit that our recommendation pushes ahead over past frameworks that consider simply word repeat or subject closeness, and addresses a promising response for a report recommender system to be used as a piece of examinations.*

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

I. INTRODUCTION

Humans are encompassed by an uncommon abundance of data, accessible as records, databases, or mixed media assets. Access to this data is adapted by the accessibility of suitable web indexes, however notwithstanding when these are accessible, clients frequently don't start a pursuit, in light of the fact that their current action does not permit them to do as such, or in light of the fact that they are not mindful that applicable data is accessible. We receive in this paper the point of view of in the nick of time recovery, which replies this inadequacy by suddenly suggesting archives that are identified with clients' present exercises. At the point when these exercises are primarily conversational, for occurrence when clients take part in a meeting, their data needs can be demonstrated as understood inquiries that are built out of sight from the professed words, acquired through continuous programmed discourse acknowledgment (ASR). These certain questions are utilized to recover and suggest reports from the Web or a neighborhood storehouse, which clients can decide to investigate in more detail if they discover them intriguing.

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 around 250 words, relating to a mixed 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?

Given the potential variety of themes, strengthened by potential ASR slips or discourse disfluencies, (for example, "rush" in this illustration), our objective is to keep up different speculations about clients' data needs, and to present a little example of proposals in view of the no doubt ones. In this manner, we point at separating a pertinent and various arrangement of catchphrases, group them into theme particular questions positioned by significance, and present clients an example of results from these questions. The point based bunching abatements the possibilities of including ASR blunders into the questions, and the assorted qualities of essential words expands the possibilities that no less than one of the suggested records answers a need for data, or can prompt a helpful archive while taking after its hyperlinks. Case in point, while a strategy in view of word recurrence would recover the accompanying Wikipedia pages: 'Light', 'Lighting', and 'Light My Fire' for the aforementioned piece, clients would lean toward a set, for example, 'Lighter', "Fleece" and 'Chocolate'. Pertinence and assorted qualities can be authorized at three stages: at the point when removing the magic words; when building one or a few certain inquiries; or when re-positioning their outcomes. Relevance and diversity qualities can be upheld at three stages: while extricating the keywords; when building one or a few certain

@IJAERD-2015, All rights Reserved

questions; or when re-positioning their outcomes. The primary two methodologies are the center of this paper. Our late tests with the third one, distributed independently [1], appear that re-positioning of the consequences of a solitary certain inquiry can't enhance clients' fulfillment with the prescribed archives. Past systems for figuring certain inquiries from content depend on word recurrence or TFIDF weights to rank watchwords and afterward select the most astounding positioning ones [2], [3]. Different systems perform catchphrase extraction by utilizing topical closeness [4], [5], [6], yet don't set a topic diversity constraint.

II. LITERATURE REVIEW

1) Enforcing topic diversity in adocument recommender for conversations

AUTHORS:M. Habibi and A. Popescu-Belis

This paper addresses the problem of building concise, diverse and relevant lists of documents, which can be recommended to the participants of a conversation to fulfill their information needs without distracting them. These lists are retrieved periodically by submitting multiple 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. Author propose in this paper 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. Authors evaluate the proposed method through crowd sourcing. The results show the superiority of the diverse merging technique over several others which not enforce the diversity of topics.

2)A statistical approach to mechanized encoding andsearching of literary information

AUTHORS:H. P. Luhn

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 understandingand summarization

AUTHORS:S. Ye, T.-S. Chua, M.-Y. Kan, and L. Qiu

We argue that the quality of a summary can be evaluated based on how many concepts in the original document(s) that can be preserved after summarization. Here, a concept refers 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, Author propose 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 2005 and 2006 evaluations as well as follow-on tests._ 2007 Elsevier Ltd. All rights reserved.

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.

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 by 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. SURVEY OF PROPOSED SYSTEM

The proposed methods are evaluated in terms of relevance with respect to conversation fragments from the Fisher, AMI, and ELEA conversational corpora, rated by several human judges. 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.

IV. Mathematical Model

Let S is the Whole System Consist of

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

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.

Procedure:

Keyword Extraction:
ASR: automatic speech recognition converts the speech and provides output to algorithm that extract keywords from the output of anASR system

Selection of Configurations:
Using the rank biased overlap (RBO) as a similarity metric, based on the fraction of keywords overlapping at different ranks.

$$RBO(S,T) = \frac{1}{\sum_{d=1}^{D} (\frac{1}{2})^{d-1}} \sum_{d=1}^{D} (\frac{1}{2})^{d-1} \frac{|S_{1:d} \cap T_{1:d}|}{|S_{1:d} \cup T_{1:d}|}$$

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 . RBO is calculatedas above Equ.

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,

1. Used to represent the distribution of the abstract topic 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 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.

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.

@IJAERD-2014, All rights Reserved

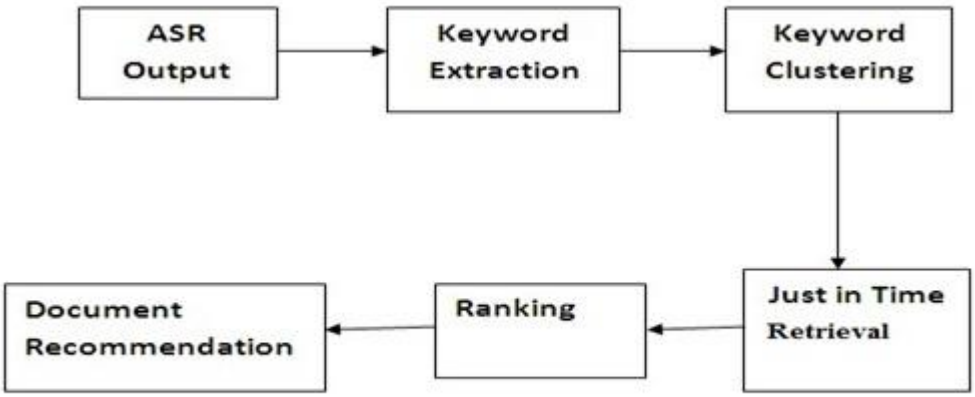
Formulation of implicit queries from conversations:
We propose a two-stage approach to the formulation of implicit queries. The first stage is the extraction of keywords from the transcript of a conversation fragment for which documents must be recommended, as provided by an ASR system

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.

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

V. SYSTEM ARCHITECTURE



.

VI. CONCLUSION AND FUTURE WORK

We have considered a specific type of without a moment to spare recovery frameworks proposed for conversational situations, in which they prescribe to client archives that are important to their data needs. We concentrated on displaying the client data needs by getting verifiable questions from short discussion pieces. These questions are in light of sets of pivotal words separated from the discussion. We have proposed a novel different pivotal word extraction strategy which covers the maximal number of vital themes in a piece. At that point, to lessen the boisterous impact on questions of the blend of themes in a decisive word set, we proposed a grouping system to isolate the arrangement of catchphrases into littler topically-autonomous subsets constituting understood inquiries.

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 raters recruited via the Amazon Mechanical Turk crowd sourcing platform. The experiments showed that the diverse keyword extraction method provides on average the most representative keyword sets, with the highest -NDCG value, and leading–through multiple topically-separated implicit queries–to the most relevant lists of recommended documents. Therefore, enforcing both relevance and diversity brings an effective improvement to keyword extraction and document retrieval. The keyword extraction method could be improved by considering n-grams of words in addition to individual words only, but this requires some adaptation of the entire processing chain.

@IJAERD-2014, All rights Reserved

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.

[5] A. Csomai and R. Mihalcea, "Linking educational materials to encyclopedic knowledge," in *Proc. Conf. Artif. Intell. Educat.: Building Technol. 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.

AUTHORS

Kolhal Devyani Rajendra,Pursuing M.E. in Computer Engineering at IOK college of engg. pimpale jagtap ,shikrapur, tal- shirur,dist- pune.

@IJAERD-2014, All rights Reserved