Volume 6, Issue 1, January 2016 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Document Recommendation in Text Based Conversations for Effective Communication: A Review

Wani H. Bisen, Antara Bhattacharya

Computer Science & Engineering, RTMNU, Nagpur, Maharashtra,
India

Abstract— the purpose of meetings is to facilitate direct communication between participants. Document plays an important role in meetings. Documents contain facts that are currently discussed, but they are not necessarily at hand. In this paper the method known as keyword extraction and clustering is overviewed which spontaneously recommend the documents that are related to users' current activities for an ongoing discussion. When users participate in a meeting, their informational needs can be modelled as keywords that can be extracted from text based conversations and documents. These keywords then organized into subgroups and can be matched to recommend relevant document to the user. This method uses Natural Language Processing for extraction of the keywords. Clustering of the keywords is form by using method known as hierarchical clustering. The similarity between both the keywords is measured by using Euclidean distance. The relevance of the proposed method can be measured by comparing the method with Fisher manual transcripts and AMI ASR transcripts.

Keywords— Document recommendation, keyword extraction, natural language processing, chunking, hierarchical clustering,

I. INTRODUCTION

This paper explores an approach to design a document recommender system intended for use in meetings. Humans are surrounded by an unrivalled wealth of information, available as documents, databases, or multimedia resources. Access to this information is ordered by the availability of suitable search engines, but even when these are available, users often do not initiate a search, because the current activity of users does not allow them to do so, or because they are not aware that relevant information is available. Documents often contain facts that are currently discussed, but they are not necessarily at hand. If the documents were available in a document management system, participants could search for them. Usually participants of a meeting do not have the time to perform such queries often during a meeting. Therefore if a system can design that will provide relevant documents for an ongoing discussion would be very helpful. Keyword extraction plays very important role in document retrieval. The goal of keyword extraction from texts is to provide a set of words that are representative of the semantic content of the texts.

II. SYSTEM DESCRIPTION

Document Recommender is a system that could provide relevant documents for an ongoing discussion, intended for use in meetings. This system can be use for business to business communication easy and more productive. M. Habibi and A. Popescu-Belis have discussed the problem of keyword extraction from conversations. The author addresses this problem of keyword extraction from conversations. The goal of the system was to retrieve keywords, for each short conversation fragment, a small number of possibly relevant documents, which can be recommended to participants. However, the author proposed the method known as just-in-time retrieval which is a query based retrieval. The method used diverse keyword extraction for extracting keywords, ranking keywords and topical similarities for measuring similarities between the keywords. Evaluating the relevance of recommendations produced by such method is a challenging task.

In this paper the method named as Keyword extraction from conversations is proposed with the goal of using the keywords to retrieve, for each short conversation fragment, a small number of possibly relevant documents, which can be recommended to participants. This method spontaneously recommends the documents that are related to users' current activities. Activities are mainly conversational. Conversation fragment contains variety of words which are potentially related to several topics. When users participate in a meeting, their information needs can be modelled as keywords that can be extracted from text based conversation and documents. These keywords then organized into subgroups and can be matched to recommend relevant document to the user. Keyword extraction method uses Natural Language Processing. NLP is the field concerned with human (language) and computer interaction. The subsets of the keywords are obtained by using hierarchical clustering. Hierarchical clustering algorithms are either top-down (Divisive) or bottom-up (Agglomerative). The recommendation lists were prepared by ranking the documents and measuring the similarity based on the Euclidean distance of the corresponding keywords extracted from conversation fragment and documents.

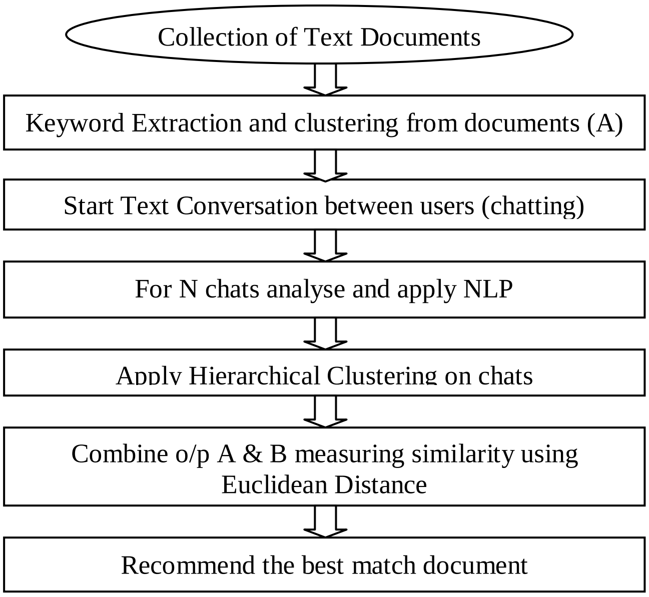
Collection of Text Documents
Keyword Extraction and clustering from documents (A)
Start Text Conversation between users (chatting)
For N chats analyse and apply NLP
Apply Hierarchical Clustering on chats
Combine o/p A & B measuring similarity using Euclidean Distance
Recommend the best match document

Fig. 1 Flow chart of Proposed Method

III. THEORETICAL FOUNDATION

The Theoretical foundation consists of various keyword extraction techniques, clustering techniques and similarity measurement techniques.

A. Keyword Extraction Techniques

Keyword extraction techniques use Natural Language Processing (NLP). It is field concerned with human (language) and computer interaction. NLP is used to extract keywords.

NLP consists of following major tasks.

1) Part of speech tagging (POS): In corpus linguistics, part-of-speech tagging or grammatical tagging is the process of making up a word in a text as particular part of speech which is corresponding to that word, based on its definition as well as its context. Part of speech tagging tells that whether the words are noun, verb, adjectives, etc.

For example:

John	saw	the	saw
PN	V	Det	N

Pos is Useful for subsequent syntactic parsing and word sense disambiguation. Many words, especially which are common ones that can serve as multiple parts of speech. For example, the word “book” can be noun (“the book on the table’) or verb (“to book a flight”).

2) *Chunking*: Chunking is the method of partial parsing. Chunks are those regions of the text which does not overlap. Each chunk contains head, with the possible addition of some preceding function word and modifiers For example:

[walk] [straight past] [the lake]

Chunks are non-recursive, i.e. they do not contain another chunks of the same category.

In the proposed research Stanford NLP and Riwordnet is used for keyword extraction. Stanford NLP Group is the effective combination of sophisticated and deep linguistic modelling and data analysis with innovative probabilistic and machine learning approaches to NLP. Riwordnet provides support for access to the Wor*A chunker assigns a partial synthetic structure to a sentence.* dNet ontological database.

B. Clustering Techniques

Hierarchical clustering is the method which works by grouping keywords into a hierarchy. It requires measure of similarity between groups of data points. These Hierarchical clustering algorithms are either top-down (Divisive) or bottom-up (Agglomerative).

1) *Agglomerative:* Agglomerative clustering uses bottom up order to cluster the keyword. Each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy.

```
Algorithm: Step 1] Place each data point into singleton group
           Step 2] Repeat: iteratively merge
           Step 3] Until: all the data are merged into a single cluster
```

2) *Divisive:* Divisive clustering uses top down order to cluster the keywords. All the observations start in one cluster, and splits are performed recursively as one moves down the hierarchy.

```
Algorithm: Step 1] Put all objects in one cluster
           Step 2] Repeat until all clusters is singletons
           Step 3] Choose a cluster to split
           Step 4] Repeat the chosen cluster with the sub-clusters
```

C. Similarity Measurement Techniques

1) Euclidean Distance:

It is the distance between two points in Euclidean space. It is the distance measure, used to measure the similarity between keywords from documents and conversation. if, p = (p1,p2,p3….pn) and q = (q1,q2,q3….qn) are two points then, distance from p to q or from q to p is calculated as, d(p,q) = d(q,p) = $\sqrt{(q1 - p1)}$)^2

IV. COMPARISON WITH PREVIOUS METHODS

When we compare natural language processing with other document retrieval techniques such as just-in-time retrieval technique which based on implicit queries, we found that evaluating the relevance of recommendations produced by such a system is a challenging task. This technique uses diverse keyword extraction, ranking of keywords and topical similarities. On the other hand NLP uses:

[1] Automatic summarization: This is the technique which produces a readable summary of a chunk of text

[2] Conference resolution: Given a sentence or larger chunk of text, determine which words refer to the same objects

[3] Machine Translation: automatically translates text from one human language to another

[4] Morphological segmentation: separate words from individual morphemes and identify the class of morphemes. While comparing the two nlp techniques i.e Stanford nlp and riwordnet, it is found that the Stanford nlp is slow and inaccurate method of keyword extraction while riwordnet possesses accuracy as well as speed.

V. ADVANTAGES AND DISADVANTAGES

A document recommender system provides suggestions for potentially relevant documents within conversation, such as business meetings. It can be used as a virtual secretary. We extract keywords from the document so that they can be organized and can be used while recommending a relevant document to the user. We use natural language processing for the extraction of the keywords from the document. The natural language processing is the leading field which based on in human (language) computer interaction. NLP relieves burden of syntax.NLP provides various applications like automatic summarization, conference resolution, morphological segmentation, machine translation, etc. While discussing its disadvantage, in part of speech tagging, it is difficult to settle on a single “correct” set of tags. In part of speech tagging it is difficult to identify whether “fire” is an adjective or a noun in “the big green fire truck.” Natural language processing is unpredictable and it requires clarification dialogue.

VI. CONCLUSION

The Document Recommender is a system that could provide relevant documents for an ongoing discussion. The scope of the system is to let its user have a list of documents that are automatically recommended by the system. The current goal of proposed research is to maximize 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.

ACKNOWLEDGMENT

We thank Stanford NLP Group and Riwordnet for access to the text database. We also thank IEEExplore.org for providing references.

REFERENCES

[1] Habibi and A. Popescu-Belis, “*Keyword Extraction and clustering for document recommendation in conversations*, ”, IEEE/ACM transactions on audio, speech, and language processing, VOL. 23, NO. 4, pp. 746-759, April 2015 .

[2] 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.

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

[4] D. F. Harwath, T. J. Hazen, and J. R. Glass, “Zero resource spoken audio corpus analysis,” in *Proc. Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2013, pp. 8555–8559.

[5] 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.

[6] 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.

[7] 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.

[8] A. Celikyilmaz and D. Hakkani-Tur, “Concept-based classification for multi-document summarization,” in Proc. Int. Conf. Acoust., Speech, Signal Process. (ICASSP), 2011, pp. 5540–5543.

[9] T. J. Hazen, “Topic identification,” in *Spoken Language Understanding:* Systems for Extracting Semantic Information from Speech, G. Tur and R. De Mori, Eds. New York, NY, USA: Wiley, 2011, ch. 12, pp. 319–356, 1em plus 0.5em minus 0.4em.

[10] Z. Liu, W. Huang, Y. Zheng, and M. Sun, “Automatic keyphrase extraction via topic decomposition,” in Proc. Conf. Empir. Meth. Nat. Lang. Process. (EMNLP’10), 2010, pp. 366–376.

[11] 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.

[12] K. Riedhammer, B. Favre, and D. Hakkani-Tur, “A keyphrase basedapproach to interactive meeting summarization,” in *Proc. IEEE Spoken* Lang. Technol. Workshop (SLT’08), 2008, pp. 153–156.

[13] 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.

[14] 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.

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