Semantic Based Document Recommendation System

[1]Athulya R Krishnan, [2]Remya R
[1] PG Scholar , [2] Assistant Professor
[1]athulya1392@gmail.com , [2] remya.cep.it@gmail.com

Abstract- **The objective of a recommender system is to generate relevant recommendations for the users. It is an information filtering technique that assists users by filtering the redundant and unwanted data from a data chunk and delivers relevant information to the users. An information system is known as recommendation engine when the delivered information comes in the form of suggestions. The information filtering system must be personalized in connection with the accommodation of different user's interests. Usually recommender systems are based on the keyword search which allows the efficient scanning of very large document collections. The goal of document recommendation is entirely different from product recommendation to consumers. This paper addresses the problem of keyword extraction from conversations, and hence uses these keywords to retrieve, for each short conversation fragment, a small number of potentially relevant documents, which could then be recommended to the participants. Here analyses the problem area of the existing approaches in the keyword extraction and database search domains. The manual extraction of keywords is slow, expensive and bristling with mistakes. To overcome these scenarios here propose a new strategy, which is based on semantic meaning. It promises the solution for a document recommender system to be used in conversations.**

Index Terms—*Keyword extraction, Information filtering , Recommendation, Text mining*

I. INTRODUCTION

Information is the ultimate powerful weapon in the modern society. Every day we are overloaded with a huge amount of data in the forms of electronic newspaper articles, emails, web pages and search results etc. But we often obtain incomplete data. Usually we need to do further search activities to have the correct acquisition of information. The importance of recommender systems comes to light in this scenario. The use of dimensionality reduction is to improve the performance for a new class of data analysis software called recommender systems. It deals with the detection and delivery of information that the user is likely to find interesting or useful. It assists users by filtering the data source and deliver relevant information to the users. Recommender systems have evolved from the extremely interactive environment of the Web. One typical application of recommendation systems is to help customers find which products they would like to purchase at E-Commerce sites. In general, every recommendation system follows a specific process to produce product recommendations. It is depicted in Fig .1.

The authenticity of the information depends upon the interest of users. The recommendation system must be personalized to accommodate an individual user's interest. These systems have achieved widespread success in the E-commerce field nowadays. Suggestions for books on Amazon, or movies on Netix, are real world examples of the operation of industry-strength recommender systems. For instance, a recommender system on Amazon.com suggests books to its customers based on other books the customers have told Amazon they are interested in. Another example for recommender system is CDnow , which helps the customers choose CDs to purchase, based on other CDs the recipient has liked in the past.

The task of document recommendation system is differ from normal recommendation engines. Here we presents a document recommendation system which is based on a diverse retrieval technique for ranking documents that are spontaneously retrieved and recommended to people during a conversation. The recommended documents represent potentially useful information for the conversation participants.. The speech can be identified using an automatic speech recognition systems (ASR).The ASR system transcribes the spoken language into readable text. The implicit queries can be constructed using the pronounced words. The manual extraction of keywords is slow, expensive and bristling with mistakes. Therefore, here we use an algorithm to help people perform automatic keyword extraction from the ASR output have been proposed.

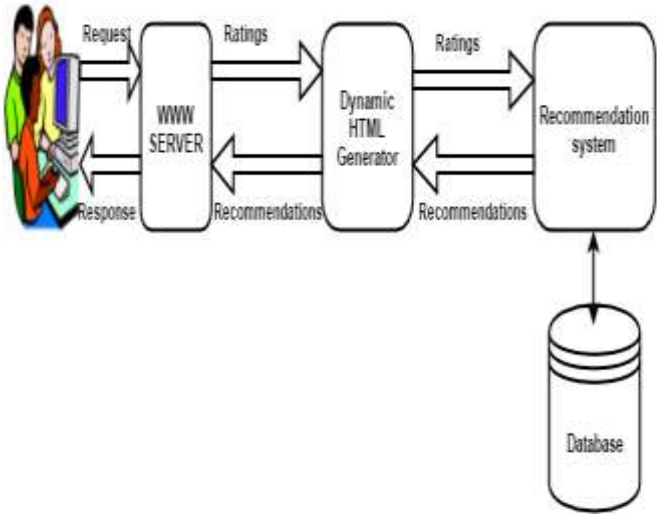


Fig.1 Architecture of recommendation System

Information Retrieval using only keywords is not usually very efficient. .The existing approaches does not consider the semantics of keywords hence it cannot ensure the relevance of recommended documents. To enforce the relevance and diversity of documents the system works through 5 stages. The conversation data may consist of multiple topics hence topically the model the data. Keywords are extracted from these multiple topics. Hence it can ensure the diversity of documents. To improve the retrieval performance the seed query obtained fro extracted keywords are reformulated by using the synonym detection. By expanding search query we can make match with additional documents. The reformulated query set is clustered into different clusters. Finally measure the similarity between the query and matched documents. Based on the similarity we ranking the results and finally recommend the documents with high ranking.

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, then query is reformulated using query expansion , 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

II. REALATED WORK

Just-In-Time Information Retrieval (JITIR) agent that proactively retrieves and presents information based on a person’s local context in an accessible yet non-intrusive manner. JITIRs (pronounced “jitter”) continuously watch a person’s environment and present information that may be useful without any explicit action required on the part of the user. There are several types of recommender systems available which are helpful in various scenarios. The amazon.com recommendation system [1] uses item to item collaborative technique that can be used for Ecommerce websites. They use recommendation algorithms to personalize the online store for each customer. The suggestions must depend upon the customer’s behavior. Here the suggestions are built by the conversation rating and the click-through process. The most similar match for a given item is determined, and the proposed algorithm builds a similar table by finding items that customers tend to purchase together. The main issue with this method is that it cannot provide suggestions to new items.

Another system is Remembrance Agent[2],[3] is a software which augments human memory by displaying a list of documents which might be relevant to the users current context. It runs without user intervention. It continuously monitors the user activities and identifies the information needs. For continuous monitoring, it generates the explicit queries from words that are written or spoken by the user. Based on the explicit query generated, relevant information is extracted as suggestions. The suggestions are presented in the form of a one line summary to the users. Jay Budzik et al., [4] propose a Watson just-in-time-retrieval system is a recommender system that assists the users by finding the relevant documents while browsing web or writing. The Watson just in system is efficient than the remembrance agent system. Because it taking the advantage of structure of written text in addition to word frequency. The Watson system uses Information Management Assistant System (IMA) observes users interaction with everyday applications and anticipate their information needs using a task model of hand. IMAs then automatically fulfill these needs using the text of the document the user is manipulating. ,Watson ventures to detect conceptually atomic, lexically regular structures in the document. The generated query is ranked by using a term weighting algorithm. Based on the formulated query, the relevant documents are retrieved and the search engine results are clustered using an incremental algorithm to avoid the redundancy of pages.

The tourist information retrieval system uses collaborative filtering technique [5][6].Collaborative filters predict someone’s personal preferences for information and/or products by keeping track of their likes and dislikes, and then connecting that information with a database of other peoples’ preferences to check for matches, and to make predictions. In Query reformulation, web searchers frequently modify their queries to obtain better results. Collaborative Tourism Information Search supports the searching for travel-related information in both standalone (a single user) mode and collaborative mode (multiple users).They proposed a system that uses remotely-located collaboration technique and the participants in the team could communicate with each other by sending instant text messages.

David Traum et al.,[7][8] presents Ada and Grace, a twin virtual guide that are used in the Museum of Science, Boston, to interact directly with museum visitors. .The quality of a museum visitors experience depends upon a well informed guide or interpreter. Virtual guides designed to engage visitors in an interactive and increase their knowledge and promote excitement about museum content are used. In order to provide efficient recommendations to the visitors, large amounts of data was aggregated so as to promote the efficient identification and recognition of the questions that are asked frequently by the visitors. The virtual guide interacts with visitors using natural language input and produce output rather than the traditional menu driven approach because it makes the interface more user friendly. To interact with the system, an operator presses an push-to-talk button and speaks into a microphone. An audio acquisition client sends these captured audio into automatic speech recognition (ASR) module. The ASR module convert the audio into text and is sent to the Language Understanding (LU) module to understand the language. Dialogues are then analyzed by the Dialogue Management (DM) module which later processes the responses to be forwarded to the user. The main inconvenience of this system is that it cannot provide recommendations for a new question, which is not in the domain specific library.

Susan Dumais et al.,[9][10] suggest Implicit Query (IQ) prototype is a system which automatically generates context-sensitive searches based on a users current computing activities. This method demonstrates an IQ system running during reading or composing email. The system analyzes the email message and delivers the results to users. The important words are extracted using TF-IDF weights.TF-IDF stands for term frequency-inverse document frequency, is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus. In real world the existing recommendation systems facing several issues.

III. PROBLEM ANALYSIS

In real world the existing document recommendation systems facing several issues. Mainly the existing system is based on keyword based technique. Keyword search is a type of search that looks for matching documents that contain one or more words specified by the user. The problem of keyword extraction from conversations, with the goal of using these keywords to retrieve, for each short conversation fragment, a small number of potentially relevant documents, which can be recommended to participants. However, even a short fragment contains a variety of words, which are potentially related to several topics. Moreover, using an automatic speech recognition (ASR) system introduces errors among them. Therefore, it is difficult to infer precisely the information needs of the conversation participants. To overcome these scenarios here propose a new strategy, which is based on semantics. The topic-based clustering decreases the chances of including ASR errors into the queries, and the diversity of keywords increases the chances that recommended documents answers a need for information. The new strategy can eliminate irrelevant documents. It promises the solution for a document recommender system that can be used in conversations. As a result, the drawbacks of the existing system are eliminated and the quality of the recommended documents is boosted to a much higher level.

IV. SYSTEM DESCRIBTIONS

Proposes a document recommendation system that provides diverse and relevant lists of documents, which can be recommended to the participants of a conversation to fulfill their information needs without distracting them. It is very effective for the participants of a meeting. The conversations of meetings are captured by using automatic speech recognition system (ASR). We propose a three-stage approach to the formulation of implicit queries (Fig.2). In first stage the multiple topics are extracted using topic modeling. Then the keywords are extracted using diverse keyword extraction algorithm. To improve the performance of recommendation query is expanded by detecting their synonyms.

A. Topic modeling

A topic model is a type of statistical model for discovering the abstract "topics" that occur in a collection of documents. Intuitively, given that a document is about a particular topic, one would expect particular words to appear in the document more or less frequently. There are several ways for topic modeling. To extract several topics from the manuscript we use Latent Dirichlet allocation (LDA) which is the simplest topic model, In this thesis LDA algorithm is implemented using mallet tool kit. The first step in mallet tool kit is to import the text document into mallet internal format. To implement the topic modeling process run the mallet. batch file and creating two output files which is key index and composition file. The key index file contains a "key" consisting of the top k words for each topic. Hence the important topics from a document are obtained by using LDA algorithm.

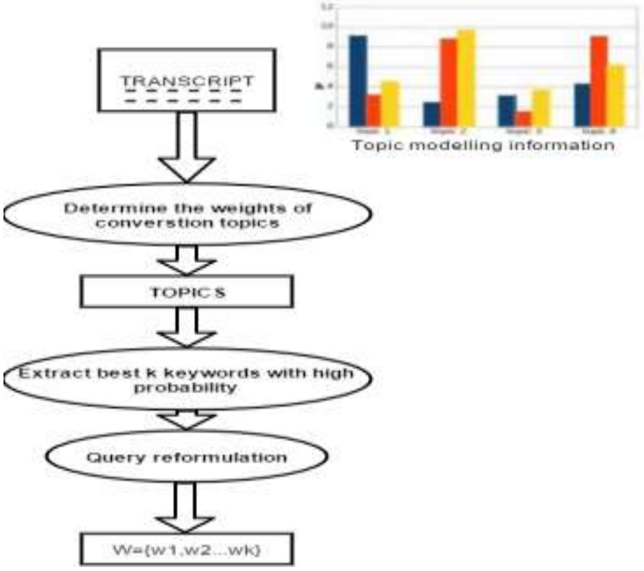

TRANSCRIPT
Determine the weights of conversation topics
TOPICS
Extract best k keywords with high probability
Query reformulation
W={w1,w2...wk}
Topic modelling information

Fig.2. Keyword extraction method from transcripts.

B. Keyword Extraction Algorithm

Propose to take advantage of topic modeling techniques to build a topical representation of a conversation fragment, and then select content words as keywords by using topical similarity [11]. The benefit of diverse keyword extraction is that the coverage of the main topics of the conversation fragment is maximized. When a conversation fragment is considered for keyword extraction, its topics are weighted, each by which is obtained by averaging over all probabilities P(Z |w_i) of the N words w_i spoken in the fragment.

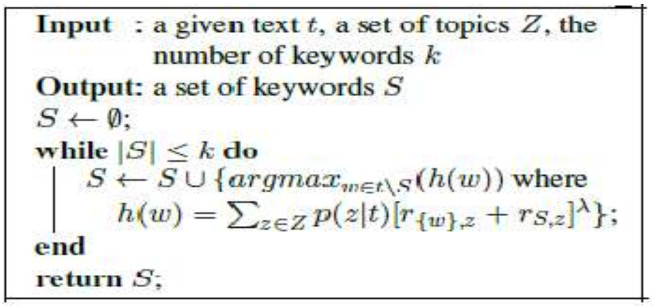
$$\beta_z = \frac{1}{N} \sum_{1 \leq i \leq N} P(z|wi) \;\ldots\ldots\ldots\ldots\ldots \quad (1)$$

We first introduce rS,z, the topical similarity with respect to topic z of the keyword set S selected from the fragment t, defined as follows:

$$r_{s,Z} = \sum_{w \in Z} P(z|wi) \,.\, P(z|t) \;\ldots\ldots\ldots\ldots (2)$$

We use a reward function for each topic, where P(z|t) is the importance of the topic and λ is a parameter between 0 and 1:

$$f\colon r_{S,z} \;\rightarrow\; P(z|t).\, r_{S,z}{}^{\lambda} \;\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3)$$



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

Algorithm: Keyword Extraction Algorithm

If λ = 1, the reward function is linear and only measures the topical similarity of words with the main topics of *t*. When 0 < λ < 1, as soon as a word is selected from a topic, other words from the same topic start having diminishing gains. This procedure continues until reaching keywords from the conversational fragment .

C. Query Expansion

To improve the performance of our system we reformulate the seed query obtained from keyword extraction algorithm. Today search is performed by searching the exact keywords entered by the user. But this may not result in the effective search because user may not know exact keywords. For example-search for data mining may not result in documents related to knowledge discovery, classification and outliers because these documents may not contain keyword data mining. In this work the query expansion implements by Finding synonyms of words, and searching for the synonyms as well.

D. Clustering

To maintain the diversity of topics embodied in the keyword set, and to reduce the noisy effect of each information need on the others, this set must be split into several topically-disjoint subsets. Each subset corresponds then to an implicit query that will be sent to a document retrieval system. These subsets are obtained by clustering topically-similar keywords. In this thesis the clustering is done based on affinity propagation algorithm. Affinity propagation is a new algorithm that takes as input measures of similarity between pairs of data points and simultaneously considers all data points as potential exemplars. Real-valued messages are exchanged between data points until a high-quality set of exemplars and corresponding clusters gradually emerges. In our system clusters of keywords are built by ranking keywords for each main topic of the fragment.

E. Similarity Measurement

The similarity between keyword set and documents are measured by using cosine similarity measurement. Cosine similarity is a measure of similarity between two vectors of an inner product space that measures the cosine of the angle between them. The cosine of 0^0 is 1, and it is less than 1 for any other angle. It is thus a judgment of orientation and not magnitude: two vectors with the same orientation have a cosine similarity of 1, two vectors at 90^0 have a similarity of 0, and two vectors diametrically opposed have a similarity of -1, independent of their magnitude. Cosine similarity is particularly used in positive space, where the outcome is neatly bounded in [0,1].The cosine similarity between two vectors (or two documents on the Vector Space) is a measure that calculates the cosine of the angle between them. This metric is a measurement of orientation and not magnitude, it can be seen as a comparison between documents on a normalized space because we are not taking into the consideration only the magnitude of each word count of each document Cosine Similarity will generate a metric that says how related are keyword set and documents by looking at the angle instead of magnitude. On basis of similarity measurement the documents are ranked. The documents with high ranking are offered as the recommendation to corresponding manuscript.

V. CONCLUSION

Due to the overload of information on the World Wide Web, the necessity of recommender systems to generate efficient solutions have evolved. In such circumstances the need for effective information retrieval and implementation of filtering tools have became essential for easy access of relevant information Recommender Systems (RS) are software tools and techniques that providing suggestions for items to be of use to a user. Recommender systems have become extremely common in recent years, and are applied in a variety of applications. We consider a specific form of just-in-time retrieval systems that are mainly used for conversational environments where users are recommended different documents that are relevant to their information queries. We focused on the generation of implicit queries from the conversations using keyword extraction algorithm which covers the maximal number of important topics in a fragment. To improve the performance of search the keyword set are reformulated using synonym detection. The effect of noise on queries of keyword set topic mixtures, the keywords are clustered into smaller topically independent subsets. The subsets compose the implicit queries. The similarity between the documents and keyword set is finally measured. The need for maximizing the coverage of all information needs results in document results. It minimizes the redundancy in a short list of documents. The integration of these schemes to a working prototype will guarantee that the users will find valuable documents in no time and with little effort. It does not interrupt the conversation flow and ensures the usability of the system. In the future, the system is aimed to adapt to real life meetings

REFERENCES

[1] Greg Linden, Brent Smith, and Jeremy York,"Amazon.com Recommendations :Item-to-Item Collaborative Filtering IEEE Computer Society,2008.

[2] 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. 487495.

[3] B. J. Rhodes and P. Maes," Just-in-time information retrieval agents",IBM Syst. J., vol. 39, no. 3.4, pp. 685704, 2000.

[4] J. Budzik and K. J. Hammond, "User interactions with everyday applications as context for just-in-time information access", in Proc. 5th Int. Conf. Intell. User Interfaces (IUI00), 2000, pp. 4451.

[5] A. S. M. Arif, J. T. Du, and I. Lee," Towards a model of collaborative information retrieval in tourism", in Proc. 4th Inf. Interact. Context Symp., 2012, pp. 258261.

[6] A. S. M. Arif, J. T. Du, and I. Lee," Examining collaborative query reformulation: A case of travel information searching", in Proc. 37th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval,2014, pp.875878.

[7] D. Traum, P. Aggarwal, R. Artstein, S. Foutz, J. Gerten, A. Katsamanis,A. Leuski, D. Noren, and W. Swartout, "Ada and Grace: Direct interaction with museum visitors", in Proc. 12th Int. Conf. Intell. Virtual Agents, 2012, pp. 245251.

[8] David Traum, William Swartout," Ada and Grace : Toward Realistic and Engaging Virtual Museum Guides", Springer- Verlag Berlin Heidelberg 2010

[9] S. Dumais, E. Cutrell, R. Sarin, and E. Horvitz," Implicit queries (IQ) for contextualized search", in Proc. 27th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 2004, pp. 594594.

[10] M. Czerwinski, S. Dumais, G. Robertson, S. Dziadosz, S.Tiernan, and M. Van Dantzich, "Visualizing implicit queries for information management and retrieva"l, in Proc. SIGCHI Conf. Human Factors Comput. Syst. (CHI), 1999, pp. 560567.

[11] Maryam Habibi and Andrei Popescu-Belis, "Keyword Extraction and Clustering for Document

Recommendation in Conversations", VOL.23,NO.4,IEEE 2015