ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 2, February 2016

Real-Time Document Recommendations Based on User Conversation

Nilesh Avinash Joshi

M.E Student, Department of Computer Engineering, MCOERC, Nasik, Savitribai Phule, Pune University Maharashtra, India

ABSTRACT: Document Recommendation system is the technique which provides necessary document to user in the context of his interest, in spite of having the search on search engine. This Dissertation work proposed a real time document recommendation system that depends on keyword extraction. The system is designed basically for a chat conversion .It can be deployed on local network wherein new documents are recommended based on conversation. Also the data can be analyzed to improve the domain knowledge and relevancy of documents retrieved. This dissertation work going to improve an answering system so that user can handle multiple requests effectively and accurately. This technique uses real time database which is provided by admin to the system. Effective Techniques for keyword extraction and clustering will make the system to provide more relevant recommendations.

KEYWORDS: Document Recommendations: information retrieval: keyword extraction: meeting analysis; topic modeling;

I. INTRODUCTION

Unlimted amount of information is present with the humans in the form of documents, databases, or multimedia. This information is accessed using suitable search engines, but even when search engines are available, users always do not go for a search, because they are very much busy in their current activity. This system makes use of Just- In Time Information-Retrieval system. Just-In-Time-Information is software that retrieves & presents information based on persons local environment. It continuously watches person's environment & presents information that is useful to user. When the users activities are written in text in online textual chat based meeting, users information needs can be built as implicit queries constructed in the background from the words chat in real time chat based conversation.

These implicit queries are used for recommended information retrieval in the form of documents from web or local knowledgebase. Afterwards user can refer those documents of their interest. This concept focuses the formulation of implicit queries to a just-in-time-retrieval system for use in real-time textual conversation on chat window. In opposite to queries made on search engines , this system constructs implicit queries from textual conversational input, having more numbers of words than actual query as an input to the system. Therefore, goalof this system is extraction of relevant & diverse set of keywords & to cluster the extracted keywords as per topic-specific queries ranked by importance & sample of results from the queries is presented to user. This system introduces novel keyword extraction technique from textual chat conversation output maximizing the coverage of potential information needs of user & reduces the numbers of irrelevant words. Once keyword set is extracted then next phase is clustering of keywords to construct several topically disjoint queries run separately give better precision than large topically adjoint query. Recommendations to users are the results finally merged into a ranked set.

1. *Need of the System*

In this present era, though people are geographically apart from each other, but they have come electronically closer to each other. Textual chat by fingers has the same importance as that of the oral conversation. Due to expansion of organization, institutes many organizational activities like meetings, conferences are going through emails, social sites, private sites of organization by means of keyboards of computers, keypads of the smart phones. For example when users attend meeting, conference users information needs can be modeled as implicit queries that are created in the background from the typed words by the users on chatting app , obtained through real-time automatic textual chat conversion.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 2, February 2016

II. RELATED WORK

Existing techniques of keyword Extraction

A. Methods

Various methods [1] of locating and determining keywords have been used, both Individually and in concert. in spite of their differences, a large amount of methods have the equal purpose and try to do the same thing: using some heuristic (such as distance between words, frequency of word use, or predetermined Word relationships), locate and de_ne a set of words that accurately convey themes or describe information contained in the text.

B. Word Frequency Analysis

In [3] A lot before time work troubled the frequency of term usage in the content, except the majority of this work focused on defining keywords in relation to a single document. during 1972, the thought of statistically analyzing the frequency of keyword usage surrounded by a document in relative to multiple other documents became more common .This method, recognized as Term Frequency - Inverse Document Frequency or purely[4] TF-IDF, weights known term to conclude how well the term describes an individual paper inside a corpus. It does this by weighting the term positively for the number of times the term occurs within the specified document, while also weighting the term negatively relative to the number of documents which contain the term. Consider term t and document d 2 D, where t appears in n of N documents in D. The TF-IDF as follows:

TFIDF(t, d, n,N) = TF(t, d) IDF(n,N) ...(1)

C. Word Co-Occurrence Relationships

In [4] While many methods of keyword extraction rely on word frequency (either within the document, within the corpus, or some combination of these), various possible problems have been pointed out with these metrics ,Including reliance on a corpus, and the assumption that a good keyword will appear frequently within the document but not within other documents within the corpus. These techniques to do not try to monitor any kind of relationship among words in a document.

D. Using a Document Corpus

In [4][1]Single effort at with this additional data utilizes a Markov Chain which take to evaluate every word in the corpus of all documents. This technique defines a Markov Chain for document d and term t with two states (C, T) where the probability of transitioning from C to T is the probability to the given term was observed in documents d out of all documents (effectively the number of times that t occurs in d divided the number of times t occurs in all documents), while the probability of moving from T to C is the probability that the term was examined every one of terms in d (the amount of times t occurs in d divided by the number of term occurrences in d. abstractly, if two terms appear at the similar state by means of related regularity, they are related.

E. An overview of Query Clustering

In [6] the literature enunciates that query clusters play a crucial role in cataloguing the concepts behind queries and the association among them. It has been recommended to group the queries into query taxonomies and a term vector becomes the manifestation of each query. A standard term frequency and inverse document frequency (tf-idf) schema, where the collection term for each query is retrieved from the top-N documents selected are the two parameters employed to calculate each component of the vector. A hierarchical agglomerative clustering technique (HAC) combined with a partitioning technique to generate a multi-way-tree cluster hierarchy is drawn on to estimate the query taxonomy. This

Technique is recommended to Frequently Asked Question (FAQ) identification and query filtering. With reference to the users query and click-through data, Beeferman and Berger [BB00] applied a content-ignorant approach and a graph based iterative hierarchical agglomerative clustering (HAC) method to cluster both the URLs and queries. This technique has been applied and recommended queries using the real users click data in the search engine Lycos[1].

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 2, February 2016

III. PROPOSED SYSTEM AND IMPLEMENTATION DETAILS

The advantage of DKE is to cover the main topics of the conversation fragment is maximized. In additional, in order to coverage large topics, the proposed algorithm will select a smaller number of keywords from each. Let z1,z2z7 be the topics discussed by user1,user2 & so on. Let P(Z—wi) be the probability of specific word w spoken in topic Z. Let z be the weight of the topic Z.

A. *Algorithm*

1. *Calculate the probability of word P(Z—wi) for topic z.*
2. *Determine weight of each topic z=(1/N).P(Z/wi) 1¡=i¡=N where N is number of words in topic.*
3. Extract best k keywords that cover all main topics with high probability.
4. Required feature of keyword extracted

Following chat windows describes the example of one manufacturing Industries suffering quality problem whose branches are located worldwide. Thus outcome of algorithm is set of keywords product, quality and xyz.

Fig 1 Diverse keyword Extraction Algorithm

2. *Keyword Clustering*

To maintain the diversity of topics embodied in the keyword set, and to reduce the noisy effect of each information need when running the queries. It is proposed to split this set into several topicallydisjoint subsets, corresponding to implicit queries to the retrieval system. To enhance topically based similar searching First, keywords are ranked within each topic by decreasing values of z. p(z—w).

Fig 2 Chat Window

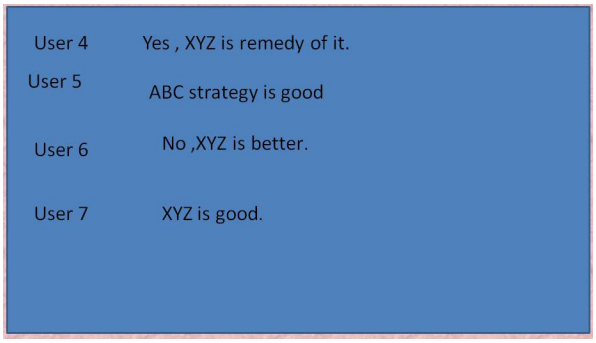

Fig 3 Chat Window

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 2, February 2016

Moreover, in each cluster, only the keywords with a z p(z—w) value higher than a threshold (0.01 in the current setting) are kept for each topic z. keywords with high value of p(z—w) (i.e. much representative in topic) will be order in sequence raise in the cluster of topic z and these keywords will be selected from the topics with high value of z. Then, clusters are ranked according to the z of the topic. SVM algorithm clusters data with no priori of knowledge. SVM provides very efficient mechanism to construct a separating hyper plane surrounded by thickest margin using a set of training of data. Prediction is made according to some measures of distance between test data Fig.4. graph showing words (Quality, xyz) vertical hyper plane are outcome of clustering and word product is bypassed. Thus system will present quality related document of xyz strategy only to user. Scale of Graph x axis=keywords extracted by diverse keyword extraction algorithm y axis=Weight of word extracted Red Arrow= hyper plane of SVM Blue Arrow= Weights of keyword extracted

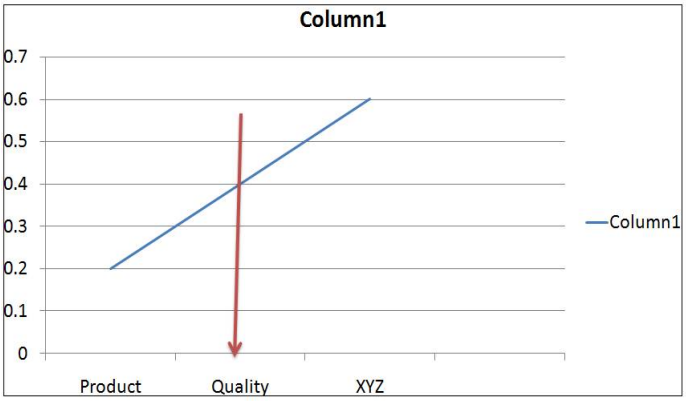

Fig.4. graph showing words (Quality xyz) vertical hyper plane are outcome of clustering and word product is bypassed. Thus system will present quality related document of xyz strategy only to user.

Scale of Graph
x axis=keywords extracted by diverse keyword extraction algorithm

y axis=Weight of word extracted

Red Arrow= hyper plane of SVM
Blue Arrow= Weights of keyword extracted

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 2, February 2016

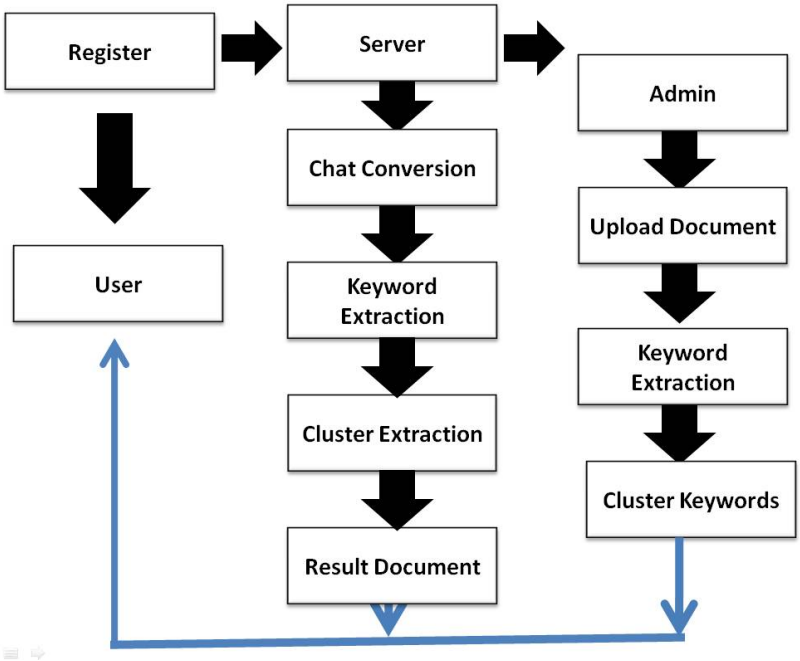

Fig 5. Architectural Block Diagram

IV. RESULT

Thus system will retrieve quality documents pertaining XYZ strategy & presented to user1.

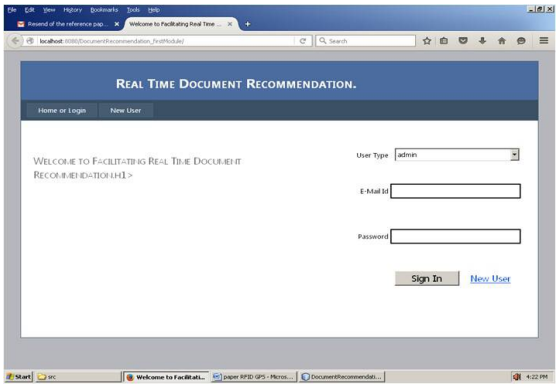

Fig 6. Resgistration This registration window is for registering the new user in chat conversation .sign in is mandatory for user authentication.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 2, February 2016

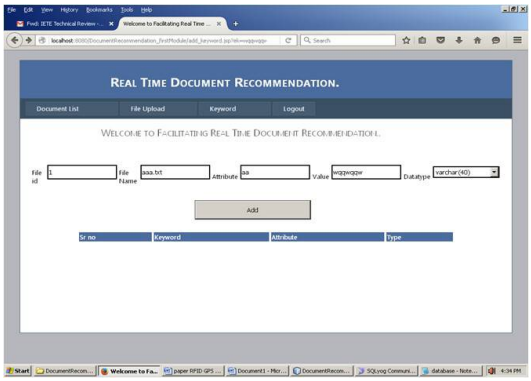

Fig 8. list of Documents The list of documents shows the present set documents in the present knowledgebase.

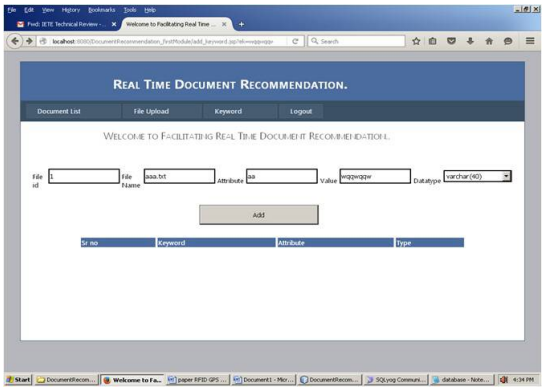

Fig 9. Add documents In add administrator adds the documents to knowledgebase recommended by user conversation which are not present in current knowledgebase, administrator adds such documents from internet source.

V. CONCLUSION AND FUTURE WORK

We have considered a particular form of just-in-time retrieval systems intended for textual chat 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 conversation fragments.

ACKNOWLEDGMENT

I am thankful to Dr. N.A. Deshpande (ME Coordinator), , MCOERC, Nasik for giving her precious time and guideline during this paper also for her expert guidance and continuous encouragement throughout this paper. I would like to express deepest appreciation towards Dr. V. H. Patil, HOD and vice Principal, MCOERC, Nasik and Prof. Dr. G. K. Kharate, Principal MCOERC, Nasik.

REFERENCES

1. Maryam Habibi and Andrei Popescu-Belis, Key-word Extraction and Clustering for Document Rec-ommendation in Conversations, IEEE transactions on audio, speech and language processing,2015
2. M. Habibi and A. Popescu-Belis, Enforcing topic diversity in adocument recommender for conver-sations, in Proceedings of the 25th International Conference on Computational Linguistics (Coling), 2014,pp. 588599.
3. R. L. T. Santos, C. Macdonald, and I. Ounis, Ex-ploiting query reformulations for Web search result diversification, in Proceedingsof the 19th Int. Conf. on the World Wide Web, 2010, pp. 881890.
4. H. P. Luhn, A statistical approach to mechanized encoding and searching of literary information, IBM Journal of Research and evelopment, vol. 1, no. 4, pp. 309317, 1957.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 2, February 2016

5. H. P. Luhn, A statistical approach to mechanized encoding and searching of literary information, IBM Journal of Research and evelopment, vol. 1, no. 4, pp. 309317, 1957.
6. G. Salton and C. Buckley, Term-weighting ap-proaches in automatic text retrieval, Information Processing and Management Journal, vol. 24, no. 5, pp. 513523, 1988.
7. S. Ye, T.-S. Chua, M.-Y. Kan, and L. Qiu, Document concept lattice for text understanding and summa-rization, Information Processing and Management, vol. 43, no. 6, pp. 16431662, 2007.
8. A. Csomai and R. Mihalcea, Linking educational materials to encyclopedic
9. knowledge, in Proceedings of the 2007 conference on Artificial Intelligence in Education: Building Technology Rich Learning Contexts That Work, 2007, pp. 557559.
10. D. Harwath and T. J. Hazen, Topic identification based extrinsic evaluation of summarization tech-niques applied to conversational speech, in Pro-ceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2012, pp. 50735076. Popescu-Belis, E. Boertjes, J. Kilgour, P. Poller, S. Castronovo, T. Wilson, A. Jaimes, and J. Car-letta, The AMIDA automatic content linking de-vice: Just-in-time document retrieval in meetings, in Proceedings of the 5th Workshop on Machine Learning for Multimodal Interaction (MLMI), 2008, pp. 272283.

BIOGRAPHY

JOSHI NILESH AVINASH, Department of Computer Engineering, MCOERC, Nasik, Savitribai Phule, Pune University, Maharashtra, India