*Impact Factor (SJIF): 3.632*

**International Journal of Advance Research in Engineering, Science & Technology**

*e-ISSN: 2393-9877, p-ISSN: 2394-2444*

***Volume 3, Issue 7, July-2016***

# To Recommend Documents In Small Business Meetings by Extracting Keywords & Clustering them

**Manasa.N [1], Bavya S[2],SMT. Kavitha G[3]**

*[1]PG Scholar, Department of Computer Science & Engineering, UBDT College of Engineering, Davangere, India,*
*[2]PG Scholar, Department of Computer Science & Engineering, UBDT College of Engineering, Davangere, India,*
*[3]Assistant Professor, Department of Computer Science & Engineering, UBDT College of Engineering, Davangere, India*

***ABSTRACT****: To assist in small business meetings here we propose a method to recommend documents to users. Huge amount of information is present around us in the form of documents, images etc but accessibility is limited by the ease of search Engines. Sometimes we have may have a confusion about how to start something, especially searching relevant data to ease the work and make it more effective. We utilize the perception of just-in-time retrieval, which helps in instinctively recommending documents that are associated to users' present activities.*

***KEYWORDS:*** *Documents, Local Database, Extraction, Keyword, Clustering.*

## I.INTRODUCTION

The job of suggesting documents to users in small business meetings varies from the task of recommending products to consumers. As in daily life how we suggest small things to our friends like sharing knowledge or giving suggestions as applied to books, videos and the like, attempt to communicate patterns of shared taste or interest among the buying habits of individual shoppers to augment conventional search results[1,2]. But some problems are included like vary of interest and opinions however the idea of recommending can help the users in smaller way too in this type of problems.

Even a small variations in search context can weaken the effectiveness of filtering. For example, a Doctor might research in internet database on one side of a medicalcase today,. Like the advantages of performing operation on one case may seem to be dangerous for other similar case. So providing proper information is important.

Topic-based recommendation systems examine point descriptions to identify items that are of exact interest to the user. Because the details of recommendation systems differ based on the representation of items, this chapter first discusses alternative item representations. [2] Next, recommendation algorithms suited for each representation are discussed. The chapter concludes with a discussion of variants of the approaches, the strengths and weaknesses of content-based recommendation systems, and directions for future research and development.

The function of images content and metadata: In common, related images often acquire similar privacy preferences, especially when people appear in the images. Analyzing the visual content may not be sufficient to capture users' privacy preferences.[7] Tags and other metadata are indicative of the social context of the image, including where it was taken and why and also provide a synthetic description of images, complementing the information obtained from visual content analysis.

## II. RELATED WORK

Daniel Billsus and Michael J Pazzani of Rutgers University proposed a [2] system that suggest an product or information to a user based upon a explanation of the item and user's interests. These type of Content-based recommendation systems helps in recommending web pages, hotels, places ,institutes ,news articles, restaurants, television programs, and online shopping websites. Although the details of various systems differ, content-based recommendation systems have common a means of description of the items that may be recommended, a way for creating a profile of the user that describes the types of items the user likes, and a means of comparing items to the user profile to determine what to recommend. This profile is often created and updated automatically in response to feedback on the desirability of items that have been presented to the user.

The task of recommending content to professionals [3] (such as attorneys or brokers)differs greatly from the task of recommending news to casual readers. A casual reader may be satisfied with a couple of good recommendations, whereas an attorney will demand precise and comprehensive recommendations from various content sources when conducting legal research. Legal documents are intrinsically complex and multi-topic, contain carefully crafted, professional, domain specific language, and possess a broad and unevenly distributed coverage of issues.
Consequently, a high quality content recommendation system for legal documents requires the ability to detect significant topics from a document and recommend high quality content accordingly.

Prem Melville & Vikas Sindhvani al [12] They discussed the approaches for recommender systems like collaborative filtering, Content Based recommendation and hybrid approaches. Collaborative approaches only use user feedback ratings to recommend items by utilizing machine learning techniques lie k-nearest neighbor. Collaborative filtering includes two methods Neighborhood Based collaborative filtering & Model-based collaborative filtering. Content Based filtering recommends based on topic similarity for example:if the search history of user contains movies of Rajamouli,then it suggests other movies of Rajamouli. Hybrid approach uses both Collaborative and content based recommendation.

It also discuss the advantages and disadvantages of recommender systems like push attacks and Nuke attacks. Content Based are unaffected by Profile injection attacks. Both types of approaches are advantageous in their own ways.

Sangeetha. J et.al [4] To provide security for the information, automated annotation of images are introduced which aims to create the meta data information about the images by using the novel approach called Semantic annotated Markovian Semantic Indexing(SMSI) for retrieving the images. The proposed system automatically annotates the images using hidden Markov model and features are extracted by using color histogram and Scale-invariant feature transform (or SIFT) descriptor method. After annotating these images, semantic retrieval of images can be done by using Natural Language processing tool namely Word Net for measuring semantic similarity of annotated images in the database. Experimental result provides better retrieval performance when compare with the existing system.

Using social media we are able to communicate with lot of people. Facebook is most popular example of social media which enable us to communicate with lot of people. In which peoples have opportunities to meet new peoples, friends and communicate with each other.[5] In this paper author concentrated on Social media, content sharing sites, Privacy, Meta data. We propose a two-level [6] framework which according to the user's available history on the site, determines the best available privacy policy for the user's images being uploaded. Our solution relies on an image classification framework for image categories which may be associated with similar policies, and on a policy prediction algorithm to automatically generate a policy for each newly uploaded image, also according to users' social features.

## III. PROPOSED SYSTEM

Here propose an well-organized way for document recommendation system for user using the conversational data. Textfile of informal data is given as input. These informal data is partitioned into m clusters. Clusters contain numerous numbers of keywords including surplus words. Using Worddictionary only important and useful topic related keywords are extracted.

Keywords are graded based on their no of occurrences or weights. By selecting maximum ranked keyword document recommendation method will be achieved.

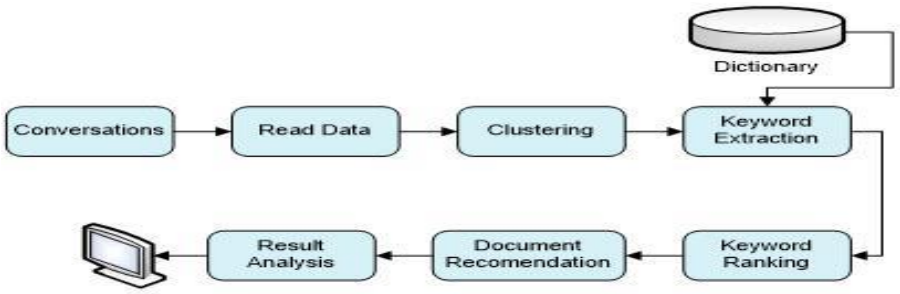


***Figure1: Architecture of proposed system***

**A. Keyword Extraction**

Term Frequency - Inverse Document Frequency or simply TF-IDF, weights

It is used to determine the term that explains a particular document within a quantity. It does this by weighting the term positively for the number of times the term occurs within the specific document, while also weighting the term negatively relative to the number of documents which contain the term. Consider term k and document d ∈ D, where k appears in p of P documents in D. The TF-IDF function is of the form:

T F IDF (k, d, p, N) = T F (k, d) × IDF (p, P) There are many possible TF and IDF functions.

Almost, practically any function could be used for the TF and IDF. Regularly-used functions include:

1 if t ∈ d
0 else
1 if word=k
TF(k,d)=∑

When the TF-IDF function is run against all terms in all documents in the document corpus, the words can be ranked by their scores. A higher TF-IDF score indicates that a word is both important to the document, as well as relatively uncommon across the document corpus. This is often interpreted to mean that the word is significant to the document, and could be used to accurately summarize the document. TF-IDF provides a good heuristic for determining likely candidate keywords, and it (as well as various modifications of it) have been shown to be effective after several decades of research. Several different methods of keyword extraction have been developed since TF-IDF was first published in 1972, and many of these newer methods still rely on some of the same theoretic backing as TF-IDF. Due to its effectiveness and simplicity, it remains in common use today.

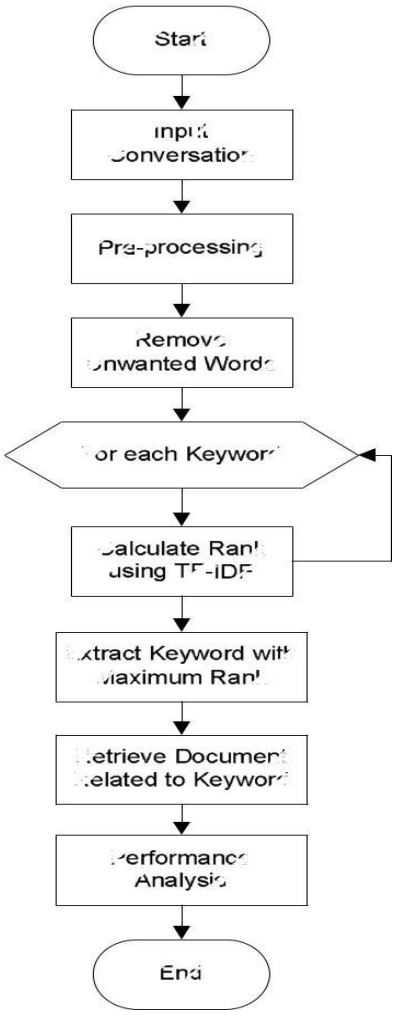


***Figure1: Flowchart of Extraction & Recommendation Process***

***B.* Clustering**

The diverse set of extracted keywords is considered to represent the possible information needs of the participants to a conversation, in terms of the notions and topics that are mentioned in the conversation. 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, as follows. Clusters of keywords are built by ranking keywords for each main topic of the fragment.

***C.* Keyword to Document Recommendation**

As a first idea, one implicit query can be prepared for each conversation fragment by using as a query all keywords selected by the diverse keyword extraction technique. However, to improve the retrieval results, multiple implicit queries can be formulated for each conversation fragment, with the keywords of each cluster from the previous section, ordered as above (because the search engine used in our system is not sensitive to word order in queries). In experiments with only one implicit query per conversation fragment, the document results corresponding to each conversation fragment were prepared by selecting the first document retrieval results of the implicit query. The recommendation lists were prepared by selecting the first document retrieval results of each implicit query and then ranking documents based on the topical similarity of their corresponding queries to the conversation fragment.

## IV. RESULT AND DISCUSSION

The experiment is tested on 50 queries taken from twitter, the conversations are given in the form of text files. By using language model and clustering, keywords are extracted. Then, each keyword is ranked based on its frequency in the database. Finally most ranked keyword is chosen as keyword for document recommendation. The analysis table is shown in table 1.

| | Number of Queries | Keyword Relevancy | irrelevant |
|---|---|---|---|
| Existing System | 50 | 80% | 20% |
| Proposed Method | 50 | 88% | 12% |

***Table 1: Result Analysis***

## V. CONCLUSION

Our present goals are to practice explicit queries, and to grade document results with the aim of increasing the exposure of all the information requirements, while decreasing redundancy in a shortlist of documents. In our proposed system. We have considered a retrieval systems projected for informal environments, in which they suggest to users documents that are appropriate to their information wants. Enforcing both significance and variety brings an successful progress to keyword extraction & document retrieval.

## REFERENCES

[1] Khalid Al-Kofahi, Peter Jackson, Mike Dahn*, Charles Elberti, William Keenan, John Duprey.A "Document Recommendation System Blending Retrieval and Categorization Technologies".
[2] Michael J. Pazzani and "Daniel Billsus, Content-based Recommendation Systems "..
[3] Qiang Lu and Jack G. Conrad, "Bringing Order to Legal Documents An Issue-based Recommendation System via Cluster Association". Thomson Reuters Corporate Research & Development when this work was conducted.
[4] Sangeetha. J 1, Kavitha R ," An Improved Privacy Policy Inference over the Socially Shared Images with Automated Annotation Process ", / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 6 (3) , 2015, 3166-3169.
[5] Aishwarya Singh, Bhavesh Mandalkar, Sushmita Singh , Prof. yogesh Pawar, "A Survey on User-Uploaded Images Privacy Policy Prediction Using Classification and Policy Mining",International Journal of Innovative Research in Computer and Communication Engineering .Vol. 3, Issue 9, September 2015.
[6] X. Liu, X. Zhou, Z. Fu, F. Wei, and M. Zhou, "Exacting social events for tweets using a factor graph," in Proc. AAAI Conf. Artif. Intell., 2012, pp. 1692–1698.
[7] A. Cui, M. Zhang, Y. Liu, S. Ma, and K. Zhang, "Discover breaking events with popular hashtags in twitter," in Proc. 21st ACM Int. Conf. Inf. Knowl. Manage., 2012, pp. 1794–1798.
[8] A. Ritter, Mausam, O. Etzioni, and S. Clark, "Open domain event extraction from twitter," in Proc. 18th ACM SIGKDD Int. Conf. Knowledge Discovery Data Mining, 2012, pp. 1104–1112.

[9] X. Meng, F. Wei, X. Liu, M. Zhou, S. Li, and H. Wang, "Entitycentric topic-oriented opinion summarization in twitter," in Proc. 18th ACM SIGKDD Int. Conf. Knowledge Discovery Data Mining, 2012, pp. 379–387.
[10] X. Wang, A. McCallum, and X. Wei, "Topical n-grams: Phrase and topic discovery, with an application to information retrieval," in Proc. IEEE 7th Int. Conf. Data Mining, 2007, pp. 697–702.
[11] L. Ratinov and D. Roth, "Design challenges and misconceptions in named entity recognition," in Proc. 13th Conf. Comput. Natural Language Learn., 2009, pp. 147–155.
[12] Prem Melville and Vikas Sindhwani,"Recommender Systems" IBM T.J. Watson Research Center, Yorktown Heights, NY 10598