# A Review on Modelling Methods for Information Filtering

**[1]Sreerekha R S, [2] Smitha E S**
[1]M.Tech Student, [2]Associate Professor
[1]Department of Computer science and Engineering,
[1]LBS Institute of Technology For women, Thiruvananthapuram, India

***Abstract— Information Filtering (IF) is the rapidly evolving area to manage large information flows. User modelling and filtering are the main components of Information Filtering system. The filtering of documents that is relevant to the particular user is depend upon the user model or user profile. To build an accurate user profile is the key task. There are various approaches used for modelling user profile and documents in an Information Filtering system. These approaches include term based, phrase based, pattern based and topic modelling based models. This paper describes these various methods or techniques used for Information Filtering, especially for document modelling and user modelling.***

***Index Terms— Information Filtering (IF), Pattern Mining, Topic Modelling***

## I. INTRODUCTION (HEADING 1)

The amount of information is increasing in every year. So we want a system to identify the needs of user and to deliver the information based on user's needs. Information Filtering is such systems which remove unwanted information from document stream based on some representations. These document representations developed from the user's interest. There are mainly two components associated with Information Filtering system: user model component and filtering component. The first one gather user information needs and construct user model (user profile). The second one match the user profile with the represented data items and decide the information which is relevant to the user. The successes of the system depend on the ability of the constructed profile. Build a good profile is the main problem to achieve good performance in Information Filtering systems. This review mainly focused on the different statistical approaches used for explicit user profiling and filtering of documents based on these profiles.

The research work [1] describes that the main objective of Information Filtering system is to remove irrelevant data from incoming streams of data items. According to that work there are two major different filtering approaches: Content Based Filtering (CBF) and Collaborative Filtering (CF) [1]. According to them the filtering based on content i.e., user areas of interest is content based filtering and filtering which uses the known preferences of group of users to make predictions of the unknown preferences of other users is collaborative filtering [1]. The work [1] describes that there are different components or models associated with IF systems. The core models are user model and filtering model. They describes that the user model explicitly or implicitly gather user information and their information needs and then construct a user profile based on these information. This constructed profile can be used as input to the filtering model. The filtering model matches the user profile with the represented data extracted from the incoming document stream and then filters out the irrelevant document. The basic decision mechanism can be binary (i.e. relevant or irrelevant) or probabilistic [1].

The work [1] also describes about the various techniques or concepts for implementing the components of IF systems. The two main concepts are: IF system based on statistical concept and IF system based on knowledge based concept [1]. Statistical concept implemented the user profile is a weighted vector of index terms. Filtering component implement a statistical algorithm (cosine measure between user profile and document vector) that compute the similarity of vector of term that represent the data item being filtered to the user profile. Filtering system follow the knowledge based concept utilize artificial intelligence techniques to represent user profile and implement filtering model [1]. User profiling or modelling is a critical task in filtering systems because the main goal of filtering system is to evaluate the relevance of data items according to the model or profile [1]. Inaccuracy of a given model leads to wrong filtering results. So in this paper we describe the different approaches used for modelling documents and user profiles.

## II. LITERATURE SURVEY

Vector space model is the most common method used in Information Filtering for achieving filtering task based on document representations. In vector space model each document is represented by a vector with an n-dimensional space. Each dimension corresponds to distinct type of representations such as term, phrase or pattern. A given document vector has associated with a weight which indicate the importance of that document. According to weighting function a list of documents are sorted based on their relevance ranking.

**Term based Approach**

Term weighting scheme, which has been used to convert the documents as vector in the term space. The popular term weighting model include tf*idf, weighting scheme for the bag of words representation.

*Term frequency Inverse Document frequency:* The work [2] describes the document could be represented as term vectors of the form $D = (ti, tj... tp)$ and a typical query Q might be formulated as $Q = (q_a, q_b ...qr)$. They also describe the representation of document D and query Q. According to them if *wdk* represent weight of term in document D, then term vector for document D and query Q can be written as $D = (t0, wd0; t1, wd1; ....; tr, wdr)$, $Q = (q0, wq0; q1, wq1; ....; qr, wqr)$ [2]. According to the work [2] the similarity between the query and document is calculated by cosine similarity measure. The work [2] also describes, the term frequency (tf) and inverse document frequency (idf), which are the two factors used as the part of the term weighting system. According to the work [2] the term frequency is measuring the frequency of occurrence of terms in the document and an inverse document frequency factor (idf), varies inversely with the number of documents n to which a term is assigned in number of documents [2]. This is the most basic and effective term based knowledge extraction method widely used in Information Filtering.

The advantage of term based approach is its efficient computational performance. But the term based representation suffers the problem of polysemy and synonymy, the polysemy is a word has multiple meanings and the synonymy is multiple words have the same meaning. All the term based representations are based on terms that is to improve the statistics of a single term and do not focus on improving the semantic accuracy. To avoid these limitations the combinations of single words (phrases) can be used.

**Phrase based approach**

To overcome the disadvantages of term based approach frequent sequential patterns or phrases used as profile descriptors of documents. The works [3] explain a method used to extract descriptive frequent sequential pattern. The proposed method in that work uses a pattern taxonomy extraction model and this model is tested in Information Filtering system. Pattern based model uses frequent sequential patterns to represent documents instead of the keyword based concept used in traditional document representation model. The work [3] describes is a tree like structure which is called pattern taxonomy that illustrates the relationship between patterns extracted from a text collection. This include a pruning step to eliminate the meaningless pattern i.e., that work uses closed sequential pattern for developing pattern taxonomy model.

Phrase based approaches are more discriminative and should carry certain semantic meaning. But the phrases have low frequency of occurrence in a document. To overcome the limitations of term based and phrase based approaches, pattern mining based approaches are proposed in which patterns are used to represent user's interest.

**Pattern based approach**

Patterns are item sets, subsequences or substructures that appear in dataset with frequency no less than a user specified threshold [4]. The most basic algorithm used in the frequent pattern mining is the Apriori algorithm [4]. The work [4] describes that the frequent pattern mining algorithms are majorly classified into three categories. 1) Candidate generation approach 2) without candidate generation and 3) vertical layout approach.

The candidate generation approach included in the Apriori based algorithms. According to the work [5] the apriori algorithms are based on apriori property associated with association rule mining. According to them the apriori property is that the sub patterns in a frequent pattern are also frequent. There are two limitations associated with apriori algorithm. It generates huge number of candidate sequences and repeatedly scanning the database [4]. Frequent patterns are more semantic than phrases but the number of returned pattern is huge, i.e., large numbers of frequent patterns are returned. So there is more attempt to select the most reliable and concise patterns [4]. A number of condensed representations such as closed item set [6] and maximal item set of frequent item set [7] have been proposed.

The work [8] describes a two stage Information Filtering model which combines the merits of term based and pattern based approaches. According to them information overload and mismatch are two fundamental problems affecting the effectiveness of Information Filtering systems. The two stage filtering model proposed by them describes a new model to reduce these problems in Information Filtering. They describes that the first stage is supported by a novel rough analysis model which efficiently removes large number of irrelevant documents. After the first stage the small numbers of relevant documents remain as the input to the second stage. The second filtering stage effectively rank the incoming documents according to specific information needs of a user and fetches the top ranking documents for a user. The underlying concept of this two stage model proposed by them is rough set decision rule based filtering. Based on the rough set theory, the decision rule for the partitioning of the incoming document stream is divided into positive, boundary and negative regions [8].

There is an effective pattern discovery technique [9] proposed for discovering patterns and explain how effectively use the discovered patterns. In that work they describe that the text mining is the discovery of interesting knowledge in the text documents. According to them finding accurate knowledge in the text documents is the main issue in text mining. They also describe the issues regarding the effectiveness of pattern based approaches: low frequency and misinterpretation. The proposed effective pattern discovery techniques [9] solve these problems. This technique first calculates the discovered specificities of patterns and then evaluates term weights according to distribution of terms in the discovered patterns for solving the misinterpretation problem. The work [6] also considers the influence of patterns from the negative training examples to find ambiguous (noisy) patterns and try to reduce their influence in the low frequency problem [9]. The process of updating ambiguous pattern can be referred as pattern evolution. Thus the proposed technique [9] uses two processes, pattern deploying and pattern evolving, to refine the discovered patterns in text documents and to solve the problem of low frequency and misinterpretation in text mining [9].

**Topic based Modelling**

The techniques mentioned above are developed under the assumption that user's interest and the documents in the collection related to single topic. But really user interest is dynamic and collections of documents include multiple topics. So topic modelling techniques are used for Information Filtering.

A topic model based approach [10] is proposed for user modelling and document modelling. In that work they use the combination of topic modelling and pattern mining techniques. According to them topic modelling compress large data into more useful and manageable knowledge and topic models provide a convenient way to analyze large of unclassified text. Topic modelling has the advantage of represent the documents with the discovered topic. But there is some problems such as word ambiguity and semantic adherence associated with it. To avoid these problems the topic modelling is combined with pattern mining [10].

The research work [10] describes that the topic modelling algorithms are used to discover set of hidden topics from collections of documents, where a topic is represented as a distribution over words. According to them LDA (Latent Dirchilet allocation) [11] is the most commonly used statistical topic modelling technique. LDA can discover the hidden topics in collections of documents using the words that appear in the documents [10]. The idea behind LDA is that each document contains multiple topics and each topic can be defined as a distribution over a fixed vocabulary of words that appear in the documents [10].

In LDA the topic representation is based on words. But the word based representations do not provide the sufficient information to determine the relevance of documents. To deal with these problems some researchers proposed a model [10] by combining text mining especially pattern mining and topic modelling. According to them pattern based representations are more meaningful and more accurate to represent topics. To discover semantically meaningful and efficient patterns to represent topics they propose two steps: construct a new transactional dataset from the LDA result of the document collection D and generate pattern based representations from the transactional dataset to represent the user needs and apply it to Information Filtering systems.

## III. CONCLUSION

These reviews mainly focus on the user modelling and document modelling techniques used in the field of Information Filtering. The earlier techniques based on term based, phrase based and pattern based approaches. The fundamental assumption for all these approaches is that the documents in the collection are all about one topic. However in reality, user's interest can be diverse and the documents in the collection often involve multiple topics. So topic modelling techniques are combined with data mining techniques especially pattern mining to generate discriminative and semantic representation for modelling topics and documents.

## IV. ACKNOWLEDGEMENT

We are greatly indebted to our principal Dr. K. C. RAVEENDRANATHAN, Dr. SHREELEKSHMI R, Professor, Head of the Department of Computer Science and Engineering, Mr. MANOJ KUMAR G., Associate Professor, Department of Computer Science and Engineering, LBS Institute of Technology for Women who has been instrumental in keeping my confidence level high and for being supportive in the successful completion of this paper. We would also extend our gratefulness to all the staff members in the Department; also thank all my friends and well-wishers who greatly helped me in my endeavor. Above all, we thank the Almighty God for the support, guidance and blessings bestowed on us, which made it a success.

## REFERENCES

**[1]** Hanani, U, Shapira,B., and Shoval,P. (2001). Information Filtering: Overview of issues, research and systems. User Modelling and User-Adapted Interaction, 11(3):203-259J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

**[2]** Salton, G., and Buckley, C. 1998.Term Weighting approaches in Automatic Text Retrieval. Inf. Process.Manage.24 (5):513-523.

**[3]** S.-T. Wu, Y.Li, Y. Xu, B.Pham, and P. Chen,"Automatic Pattern Taxonomy Extraction for Web mining," in Proc.IEEE/WIC/ACM Int. Conf. Web Intell., 2004, pp.242-248.

**[4]** J.Han, H.Cheng, D.Xin, and X.Yan,"Frequent Pattern Mining:Current status and Future Directions" Data Min.Knowl.Discov., vol.15, no.1, pp.55-86, 2007.

**[5]** R. Agrawal and R. Srikant, Fast Algorithms for Mining Association Rules in Large Databases, Proc. 20th Intl Conf. Very Large Data Bases (VLDB 94), pp. 478-499, 1994.

**[6]** Pasquier N, Bastide Y, Taouil R, Lakhal L (1999) Discovering frequent closed item sets for association rules. In: Proceeding of the 7th international conference on database theory (ICDT'99), Jerusalem, Israel, pp 398–416.

**[7]** Bayardo RJ (1998) efficiently mining long patterns from databases. In: Proceeding of the 1998 ACM-SIGMOD international conference on management of data (SIGMOD'98), Seattle, WA, pp 85–93

**[8]** Y. Li, X. Zhou, P. Bruza, Y. Xu, and R.Y. Lau, A Two-Stage Text Mining Model for Information Filtering, Proc. ACM 17th Conf.Information and Knowledge Management (CIKM 08), pp. 1023-1032,2008.

**[9]** N. Zhong, Y. Li, and S.-T. Wu, Effective pattern discovery for text mining, IEEE Trans. Knowl. Data Eng., vol. 24, no. 1, pp. 3044, Jan.2012.

**[10]** Y. Gao, Y. Xu, and Y. Li, Pattern-based topic models for Information Filtering, in Proc. Int. Conf. Data Min. Workshop SENTIRE, 2013, pp.921928.

**[11]** D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," J. Mach. Learn. Res., vol. 3, pp. 993–1022, 2003.