International Journal of Engineering Research in Computer Science and Engineering (IJERCSE) Vol 3, Issue 2, February 2016

Community Detection with Semantic based Information Filtering

[1]Jimsy Johnson, [2], Smitha C S
[1] PG Scholar, ,[2],Assistant Professor,
College of Engineering, Perumon
[1]jimsyjohnson@gmail.com, [2] smithacscs@gmail.com

Abstract- **Topic Modelling has been widely used in the fields of machine learning, text mining etc. It was proposed to generate statistical models to classify multiple topics in a collection of document, and each topic is represented by distribution of words. Many mature term-based or pattern based approaches have been used in the field of information filtering to generate users information needs from a collection of documents. The user's interests involve multiple topics. Latent Dirichlet Allocation (LDA) was used to represent multiple topics in a collection of documents. Polysemy and synonymy are the two prominent problems in document modelling. Nowadays patterns are used for representing topics since they have more discriminative power than words for representing multiple topics in a document. But it is difficult to process the large amount of discovered patterns. So we are trying to find more efficient method for optimizing the pattern generation and trying to create a more accurate user interest modelling. Here we uses Maximum Matched Pattern based Topic Model. And the maximum matched patterns are then passed through an NLP-engine for creating synonyms of the patterns and thus more efficient search is obtained. A community for scholars is created. It is useful for doubt clearance and notifying events in a particular area.**

Index Terms— Community Detection, Pattern Mining, Topic Modelling

I. INTRODUCTION

Information Filtering is the extraction of relevant and quality information from the large number of available information based on document representation which represents user's interest. Recent years the amount of available web information increases day by day. Therefore advanced techniques are needed to understand and analyze what exactly user needs and deliver the best results based on user information needs. An Information filtering system is useful in such a way that it assists users by filtering the available data sources and delivers the relevant information to users. The ultimate aim of data mining tasks is information filtering. Topic Modelling [1] is one of the most important areas that come under the data mining. The study of topic modelling started from the need to compress large amount of data into more useful and manageable knowledge.

Topic modelling is a method for finding and tracing clusters of words in large bodies of texts. A topic modelling tools looks through the corpus for the clusters of words and groups them together by the process of similarity. A good topic model is one in which the words in it make sense. Topic modelling is mainly built for large collection of texts. It can automatically classify documents in a collection by a number of topics and every document is represented with multiple topics and their corresponding distribution. An information filtering is said to be a recommending system, when the delivered information comes in the form of suggestions since each users have different interests, and the information filtering system must be personalized to the individual user's interest.

The ultimate aim of all data mining tasks is information filtering. The word information filtering means that a system to remove redundant or unwanted information from an information stream or document using automated or computerized methods prior to presentation to a human user. The related research areas includes Information Filtering, Information filtering, Text mining and Topic modelling. All these areas are overlapped or in other words these areas are dependent on each other

The meaning of the word pattern is "a regular and intelligible form or sequence discernible in the way in which something happens or is done". When it comes to the area of text mining, pattern extraction have a great influence. In past years, text mining, topic modelling etc. are done with the help of individual words (terms). But it is not efficient, because it takes large time to work on the individual terms in a document. So the concept of term-based approach is out dated. Then a new and much more efficient method was discovered by data mining experts. Phrase based approach. Through researches it can be find that phrase based approaches can do well with modelling the document than the traditional term-based approach, since it carries more semantic information. The main advantage of using this method is that, phrases are less ambiguous and

more discriminative than individual words in describing a document. Phrase-based approach also have some disadvantages. They are, phrases have statically inferior properties, low frequency of occurrence and many phrases are redundant and noisy. So there is a need for finding new concepts for effective pattern mining techniques. As year goes several techniques were introduced by several experts. Some of them are Association rule mining, frequent item set mining, Sequential Pattern mining , matching patterns, Maximum matched patterns etc. The ultimate aim of the data mining experts are to reduce the search time of the user and to provide a better result i.e., gives the accurate information that a user wants. The process topic generation includes Dataset Preparation, Topic Generation, Construction of new datasets and the final Generate optimized Topic representations. The step wise procedure is shown in the fig: 1

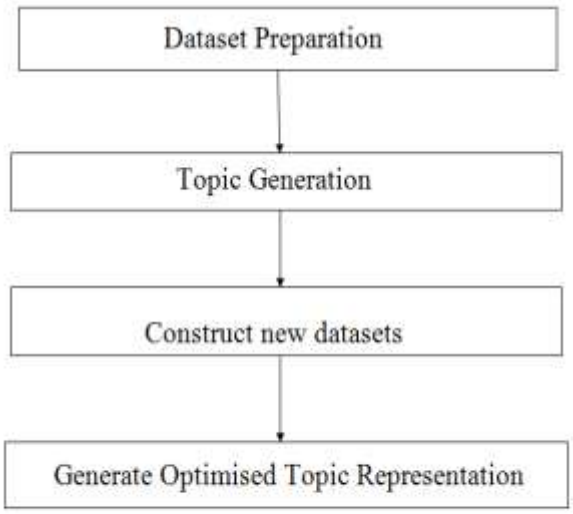

Fig : 1

II. RELATED WORKS

In many years a variety of efficient algorithms have been developed for mining frequent patterns and some of them are Apriori, Prefix Span, and FP-tree which are found more efficient .But normally, the number of returned patterns is huge because if a pattern is frequent, then each of its sub patterns is frequent too. Thus, selecting reliable patterns [14] is always very crucial. For example, a number of condensed representations of frequent itemsets have been proposed such as closed itemsets [15], maximal itemsets [16], free itemsets [17], disjunction-free itemsets [18] etc. The primary purpose of these condensed representations is to enhance the efficiency of using the generated frequent itemsets without losing any information. Among these proposed itemsets, frequent closed patterns show great potential for representing user profiles and documents. That is mainly because for a given support threshold, all closed patterns contain sufficient information about all that is involved in all corresponding frequent patterns. Wang et al. [19] proposed the TFP algorithm to extract the top-k most representative closed patterns by pattern length that no less than min instead of traditional support confidence criteria. In addition, closed patterns stand on the top of the hierarchy induced by each equivalence class, allowing the algorithm to informatively infer the supports of frequent patterns. So, in this paper, we intend to utilize the hierarchical structure of patterns based on equivalence class partitions to represent user profiles creatively. Therefore, people more often try to to extract more semantic features (such as phrases and patterns) to represent a document in many applications. Data mining techniques were applied to text mining and classification by using word sequences as descriptive phrases (n-Gram) from document collections [12], [13]. But the performance of n-Gram is restricted due to the low frequency of phrases. Pattern mining has been extensively studied for many years. A variety of efficient algorithms such as Apriori, Prefix Span, and FP-tree have been proposed and extensively developed for mining frequent patterns more efficiently. But normally, the number of returned patterns is huge because if a pattern is frequent, then each of its sub patterns is frequent too. Thus, selecting reliable patterns [14] is always very crucial. For example, a number of condensed representations of frequent itemsets have been proposed such as closed itemsets [15], maximal itemsets [16], free itemsets [17], disjunction-free itemsets [18] etc. The primary purpose of these condensed representations is to enhance the efficiency of using the generated frequent itemsets without losing any information. Among these proposed itemsets, frequent closed patterns show great potential for representing user profiles and documents. That is mainly because for a given support threshold, all closed patterns contain sufficient information about all that is involved in all corresponding frequent patterns. Wang et al. [19] proposed the TFP algorithm to extract the top-k most representative closed patterns by pattern length that no less than min instead of traditional support confidence criteria. In addition, closed patterns stand on the top of the hierarchy induced by each equivalence class, allowing the algorithm to informatively infer the supports of frequent patterns. So, in this paper, we intend to utilize the hierarchical structure of patterns based on equivalence class partitions to represent user profiles creatively.

III. LATENT DIRICHILET ALLOCATION (LDA) ALGORITHM

Topic Modelling algorithms are a set of algorithms that mines the hidden thematic structure of a

document. To develop new ways to search, such algorithms are used

Latent Dirichlet Allocation (LDA)[20] algorithm is a well known algorithm used in the field of topic modelling.LDA is a technique that automatically finds topics in a collection of documents. In LDA, each document can be viewed as a mixture of topics which splits each word with certain probabilities. Suppose we have a set of documents and each document has a fixed number of topics to be discovered and let it be K.

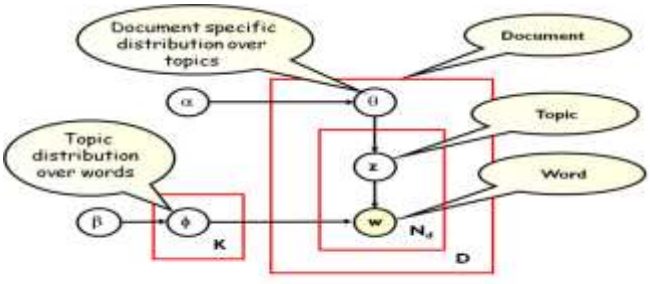

Fig2 :LDA

IV. PATTERN BASED TOPIC MODELLING

There are many ways are for doing topic modelling. Most important among them are term-based approaches, phrase-based approaches, and pattern-based approaches. Term-based approaches are the most primitive type of document modelling method. Term-based approaches have some advantages as well as some disadvantages also. The advantages are efficient computational performance and mature theories for term-weighting. The disadvantages are Polysemy and synonymy. After that Phrase-based methods are introduced. But it also has some disadvantages. So to overcome all the limitations of these methods pattern-based approaches are introduced. Patterns are considered to be more discriminative than single terms in representing multiple topics in a document. But here also one problem is there. ie; the number of patterns discovered from a single document may be very large. So it is difficult to handle this huge amount of discovered patterns. Here we are using patterns for topic modelling. Nowadays there is many types of patterns are there. Some of them are structural patterns, matching patterns, maximum matched patterns etc. The latest among them is maximum matching pattern based topic modelling. The main features of this topic modelling method are: (i) User information needs are generated in terms of multiple topics. (ii) Each topic is represented by patterns. (iii) Patterns are generated from multiple topics. (iv) The most discriminative and representative patterns called Maximum Matched Patterns(MPBTM) are used to estimate the document relevance to the user's information needs in order to filter out irrelevant documents.

This model generates more semantic topic representation to avoid the ambiguity problem. The patterns are generated from the word based topic representation of a traditional topic model like LDA. This assures that patterns can well represent topics because these patterns are made up with words which are extracted by LDA based on sample occurrence of words in the document.

V. COMMUNITY DETECTION WITH SEMANTIC BASED INFORMATION FILTERING

Community Detection means that creating a community for scholars. Since our primary motive is topic modelling or document modelling, First of all model all the available documents with the help of pattern mining algorithms. Here the input is the Routers Corpus. Then the documents are presented for the pre-processing stage. This stage may include stop-word removal and stemming. After the pre-processing stage the documents are then presented for the next phase. i.e... Pattern Generation Phase. The detailed architecture of the proposed scheme is given below.

The pattern generation stage consists of two steps. Construct a transactional dataset and generate pattern enhanced representation

A. Construction of Transactional Dataset

Table 1
Transactional Datasets Generated from Table 1
(Topical Document Transaction (TDT))

T	TDT	TDT	TDT
1	$\{w_1, w_2, w_3\}$	$\{w_1, w_8, w_9\}$	$\{w_7, w_{10}\}$
2	$\{w_2, w_4\}$	$\{w_1, w_7, w_8\}$	$\{w_1, w_{11}, w_{12}\}$
3	$\{w_1, w_2, w_5, w_7\}$	$\{w_2, w_3, w_7\}$	$\{w_4, w_7, w_{10}, w_{11}\}$
4	$\{w_2, w_6, w_7\}$	$\{w_1, w_8, w_9\}$	$\{w_1, w_{11}, w_{10}\}$
	Γ_1	Γ_2	Γ_3

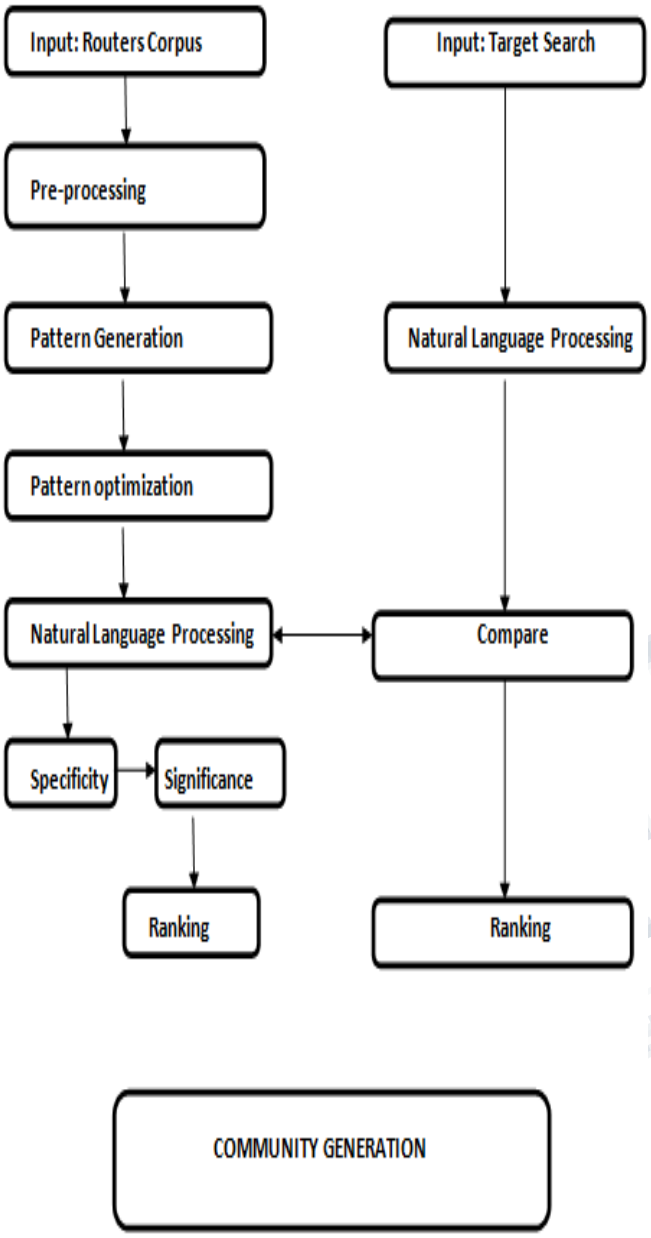

Fig 3: Proposed Architecture

is essential to construct a transactional data set from each topic-word based representation. Take a simple example, consider a collection of data D = $\{d_1,d_2,d_3,d_4\}$ be a small collection of 4 documents with 12 words and assume the documents in D involves 3 topics Z_1,Z_2,Z_3.

The transactional dataset Γ_j for the topics in D, we can construct V transactional dataset (Γ_1, Γ_2,….. Γ_v).

B. Generate Pattern Enhanced Representation

The most basic idea of the pattern based approach is the use of frequent pattern mining algorithms. Frequent pattern mining algorithms are used for mining frequent patterns from a collection of document. Frequent patterns have found broad applications in the field of association rule mining, clustering, indexing etc.. Frequent patterns carry strong association between items and carry more semantic meanings of the data. To find patterns in a database is the most basic operation behind all data mining tasks. So pattern mining algorithms have been developed to work on the databases where the longest patterns are relatively short. Apriori like algorithms are used at the initial times for pattern mining. But on later, it is found that Apriori algorithms are inadequate on dataset with long patterns. Apriori employs a bottom up search mechanism. It includes a phase for finding patterns called frequent item set. A frequent item set is a set of items appearing together in a number of database records meeting a user specified threshold. Since apriori algorithms are inefficient for pattern mining another algorithm named FP-tree algorithm is used, which is more efficient than the existing ones. Here a threshold is set by the user and further processing is done with the help of this.

Table 2

The Frequent Patterns for Z_2, $\sigma = 2$

Patterns	*supp*
$\{w_1\},\{w_8\},\{w_1,w_8\}$	3
$\{w_9\},\{w_7\}\{w_8,w_9\},\{w_1,w_9\},\{w_1,w_8,w_9\}$	2

For a given minimal support threshold σ, an item set X in Γ_j is frequent if supp(X) >= σ, where supp(X) is the support of X which is the number of transactions in Γ_j that contain X.

VI. NLP ENHANCED PATTERN MINING

The main disadvantages of almost all pattern mining algorithm is polysemy and synonymy. In order to handle the problem of synonymy we are using NLP, i.e., Natural Language Processing. By using NLP, it can retrieve more relevant documents.

Consider the scenario in which, the query that a user input is ‘data mining’. The resultant documents consist of documents which contains the term ‘data mining’. It does not retrieve documents which contains ‘text mining’ or something like that. With the use of NLP, it tries to generate all patterns at the time of document modelling. So more efficient search is obtained. Now day large types of NLP-engines are available. Here in this work, Stanford NLP is used. There are many advantages are there for using Stanford NLP. It provides a Stanford toolkit for natural language processing and it is an open source NLP technology.

In semantic based pattern mining there are mainly two parts are there, Information Retrieval and Information Filtering. The NLP is applied to both stages. In information Retrieval stage it is applied just after the pattern optimization stage and at the information filtering stage it is applied at the 'search stage'. At the information filtering stage , when the user input with a query term, the scenario is that documents with the exact term should be retrieved. But there may be other relevant documents are there with the synonymous of the query term. By adding NLP to the current system, it can handle this problem to a great extend.

VII. COMMUNITY DETECTION

As we see earlier community detection is performed with the help of clustering algorithms. Here community is created on the basis of the documents uploaded by the scholars. Since we are performing user interest modelling, we can also get the details of each user. At the time of registration the details of each scholar is saved and community generation is performed with the help of this. The attributes which are used for community generation are name of the scholar, area of interest, document id etc. At the time of community generation, the documents with similar contents are clustered or grouped together and community is generated between those scholars. So the members in a particular community can be communicate with each other and also can clear doubt about that area. Since a single author can upload papers on different areas. There is a chance of being a scholar in more than one community.

VIII. K-MEANS CLUSTERING ALGORITHM

As we see, a clustering algorithm is used for generating community. Clustering algorithms are used for grouping a collection of documents based on some criteria. There are many clustering algorithms are used now a days. Here k-means clustering algorithm is used. It partitions 'n' objects into 'k' clusters in which each object belongs to the cluster with the nearest man. This produces exactly k-different clusters of greatest possible distinction. The best number of clusters *k* leading to the greatest separation (distance) is not known as a priori and must be computed from the data. The objective of K-Means clustering is to minimize total intra-cluster variance, or, the squared error function:

Algorithm:

1. Clusters the data into *k* groups where *k* is predefined.
2. Select *k* points at random as cluster centres.
3. Assign objects to their closest cluster centre according to the *Euclidean distance* function.
4. Calculate the centroid or mean of all objects in each cluster.
5. Repeat steps 2, 3 and 4 until the same points are assigned to each cluster in consecutive rounds.

There are some advantages are there for using K-means clustering algorithm. They are (i) If variables are huge , then K-means most of the times computationally faster than hierarchical clustering. (ii) K-Means produce tighter clusters than hierarchical clustering, especially if the clusters are globular.

IX. CONCLUSION

Topic modelling approaches generally have a sound statistical foundation. The model can be applied to an arbitrary set of documents to learn a set of latent topics, each of which is represented by a word distribution and where each document is represented by topic distribution. However, the single words hardly satisfy the needs of semantic representations at topic level. Hence, proposes a pattern-based topic model which automatically generates discriminative and semantic rich representations for modelling topics and documents by combining topic modelling techniques and data mining techniques. Such a combination allows benefits from statistical latent topics and matching semantically related patterns. Here, all the research was conducted using content-based analysis. There are two types of user input data: user profiles and user generated queries in information filtering and information retrieval, respectively. By incorporating NLP to the system, tries to avoid the problem of Polysemy to a great extent. Since user interest model is generated, a community can be formed so that scholars in a particular area can communicate.

In future a feedback mechanism can also be included and so as to improve the efficiency of the searching mechanism. Since the most important problem in the field of topic modelling is polysemy and synonymy, there should be more efficient methods must be discovered to avoid this.

REFERENCES

[1] Blei, D. M. (2012). Probabilistic topic models. Communications of the ACM, 55(4):77 84.
.

[2] J. Mostafa, S. Mukhopadhyay, M. Palakal, and W. Lam, "A multilevel approach to intelligent information filtering: Model, system, and evaluation," ACM Trans. Inform. Syst., vol. 15, no. 4, pp. 368– 399, 1997

[3] S. E. Robertson and I. Soboroff, "The TREC 2002 filtering track report," in Proc. TREC, 2002, vol. 2002, no. 3, p. 5.

[4] Y. Cao, J. Xu, T.-Y. Liu, H. Li, Y. Huang, and H.-W. Hon, "Adapting ranking svm to document retrieval," in Proc. 29th Annu. Int. ACM SIGIR Conf. Res. Develop. Inform. Retrieval, 2006, pp. 186–193.

[5] F. Beil, M. Ester, and X. Xu, "Frequent term-based text clustering," in Proc. 8th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., 2002, pp. 436–442.

[6] S.-T. Wu, Y. Li, and Y. Xu, "Deploying approaches for pattern refinement in text mining," in Proc. 6th Int. Conf. Data Min., 2006, pp. 1157–1161

[7] N. Zhong, Y. Li, and S.-T. Wu, "Effective pattern discovery for text mining," IEEE Trans. Knowl. Data Eng., vol. 24, no. 1, pp. 30–44, Jan. 2012.

[8] J. Lafferty and C. Zhai, "Probabilistic relevance models based on document and query generation," in Language Modeling for Information Retrieval. New York, NY, USA: Springer, 2003, pp. 1–10.

[9] L. Azzopardi, M. Girolami, and C. Van Rijsbergen, "Topic based language models for ad hoc information retrieval," in Proc. Neural Netw. IEEE Int. Joint Conf., 2004, vol. 4, pp. 3281–3286.

[10] S. Robertson, H. Zaragoza, and M. Taylor, "Simple BM25 extension to multiple weighted fields," in Proc. 13th ACM Int. Conf. Inform. Knowl. Manag., 2004, pp. 42–49.

[11] Y. Cao, J. Xu, T.-Y. Liu, H. Li, Y. Huang, and H.-W. Hon, "Adapting ranking svm to document retrieval," in Proc. 29th Annu. Int. ACM SIGIR Conf. Res. Develop. Inform. Retrieval, 2006, pp. 186–193.

[12] X. Li and B. Liu, "Learning to classify texts using positive and unlabeled data," in Proc. Int. Joint Conf. Artif. Intell., 2003, vol. 3, pp. 587–592.

[13] J. F€urnkranz, "A study using n-gram features for text categorization," Austrian Res. Inst. Artif. Intell., vol. 3, no. 1998, pp. 1–10, 1998.

[14] W. B. Cavnar and J. M. Trenkle, "N-gram-based text categorization," Ann Arbor MI, vol. 48113, no. 2, pp. 161–175, 1994.

[15] Y. Xu, Y. Li, and G. Shaw, "Reliable representations for association rules," Data Knowl. Eng., vol. 70, no. 6, pp. 555–575, 2011

[16] 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.

[17] R. J. Bayardo Jr, "Efficiently mining long patterns from databases," in Proc. ACM Sigmod Record, 1998, vol. 27, no. 2, pp. 85–93.

[18] J.-F. Boulicaut, A. Bykowski, and C. Rigotti, "Free-sets: A condensed representation of boolean data for the approximation of frequency queries," Data Min. Knowl. Discov., vol. 7, no. 1, pp. 5– 22, 2003.

[19] A. Bykowski and C. Rigotti, "Dbc: A condensed representation of frequent patterns for efficient mining," Inform. Syst., vol. 28, no. 8, pp. 949–977, 2003.

[20] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," J. Mach. Learn. Res., vol. 3, pp. 993–1022, 2003.