



# International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 9, September 2016

## Document Modelling with Semantic Pattern- Based Topics

Rincy Joseph, Huda Noordean,

M. Tech Student, Dept. of CSE, CEMP, Alappuzha, India

Assistant Professor, Dept. of CSE, CEMP, Alappuzha, India

**ABSTRACT:** Pattern mining is an important research area in data mining and knowledge discovery. The data mining concept is used in the field of information filtering for generating user's information needs from a collection of documents. Topic modelling has become one of the most popular probabilistic text modelling techniques that has been quickly accepted by machine learning and text mining communities. The most important contribution of topic modelling is that it can automatically classify documents in a collection by choosing a number of topics that represents every document with multiple topics and their corresponding distribution. Patterns are always more discriminative than single terms for describing the documents. Selection of the most discriminative and representative patterns from the huge amount of discovered patterns becomes crucial. Topic Modelling provide a suitable way to analyse large number of unclassified text. In Maximum Pattern Based Topic Modelling (MPBTM) user information represented in terms of pattern. In MPBTM semantic features of pattern is considered in the document modelling. Since here proposed an efficient ranking method, using MPBTM by semantically analyse the pattern. Open English NLP library used for filtering semantic meanings of patterns from the collections of topics. The main features of the proposed model include: (1) Each topic is represented by patterns (2) For more information filtering, here proposed Open English 2.0NLP(3) Give more accurate document modelling method for ranking.

**KEYWORDS:** Information filtering, Topic model, Pattern mining, Maximum matched pattern, User interest model, semantic matching feature with multiple topics

### I. INTRODUCTION

To create user interest document representation by using information filtering i.e. remove or delete unwanted document and create user interested document that is main aim of information filtering. Many information filtering model are term based and pattern based string based that all are tradition model use for IF. Term based information filtering has advantage of efficient computational performance and mature theories for term weight. But it suffers from problem of Polysemy and synonymy. Here one term generates many meaning and term may be repeated in main document. Term based model have some limitation then pattern based model can be used to generate pattern based topic representation since patterns carry more semantic meaning than terms. Also, data mining has developed some techniques (i.e., maximal patterns, closed patterns and master patterns) for removing the redundant and noisy patterns from the pattern based topic representation. In pattern based approach, patterns are used by meaningfully represented topics rather than single words. Specifically, the patterns are generated from the words in the word-based topic representations of a traditional topic model such as the LDA model.

The assumption of all these data mining and text mining techniques is that the user's interest is only related to a single topic. However, in reality this is not necessarily the case. For example "Apple", in that it may be related to many topic that is price, fruit, smart phone, company location all that. Sometimes new document may arrive at that time and the user interest may change. So in this paper we are going to focus on user interest model on multiple topic rather than single topic.

Topic modelling [1], [2] has become one of the most popular probabilistic text modelling techniques, and has been quickly accepted by machine learning and text mining communities. The most inspiring contribution of topic modelling is that it automatically classifies the documents in a collection by a number of topics and represents every



# International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 9, September 2016

document with multiple topics and their corresponding distribution. There are different method are used for topic representation which are PLSA [3] and LDA [4] that generate multiple topic and distribution of topic. But this topic model suffers from two main problems. First problem is limited number of topic that is predefined which is insufficient for document representation. Second problem is, word model always generate frequent word set which some word have meaning and some are not useful for document representation. In pattern based topic model, which has been utilized in IF [5], can be considered as a "Post-LDA" model based on the patterns are generated from the topic representations of the LDA model. Patterns can represent more specific meanings than single words. By comparing the word-based topic model with pattern-based topic models, the pattern based model can be used to represent the semantic content of the user's documents more accurately than word based document. However, very often the number of patterns in some of the topics can be huge and many of the patterns are not enough to represent specific topics. We use Maximum Pattern Based Topic Model (MPBTM) to select the most representative and discriminative patterns. There are called Maximum matched Patterns which represent topics instead of using frequent Patterns. MPBTM is used for document representation and document relevance ranking. The patterns in the MPBTM are well structured so that the maximum matched patterns can be efficiently and effectively selected to represent the documents and ranking them, but it does not consider semantic meaning of pattern from the collection of topics from the document modelling. Here we propose an efficient ranking method, using MPBTM by considering semantic meaning of the pattern from collection of topics. This project proposes to overcome the limitation of existing system by using Natural Language Processing Natural language processing (NLP), i.e., the open English NLP 2.0 library used in enhanced LDA algorithm for filtering semantic meanings of patterns from the collections of topics.

## II. RELATED WORK

Information filtering deals with the delivery of user interested information from a collection of information. An information filtering system assists the users by filtering the data source and avoiding the irrelevant information and also delivers relevant information to the users. When the delivered information comes in the form of suggestions an information filtering system is called a recommender system. Because users have different interests the information filtering system must be personalized to accomplish the individual user's interests. This requires the gathering of feedback from the user in order to make a user profile of the preferences. Two major approaches exist for information filtering: content-based filtering and collaborative filtering system. A content-based filtering system selects items based on the correlation between the content of the items and the user's preferences, while a collaborative filtering system chooses items based on the correlation between people with similar preferences.

In Y. Cao, J. Xu' s "Adapting ranking SVM to document retrieval"[6] proposed the document filtering can be regarded as a classification task or a ranking task. Methods, such as Naive Bayes, kNN and SVM, assign binary decisions to documents (relevant or irrelevant) as a special type of classification. The relevance of a document can be modelled based on various approaches that primarily include a term-based model, a pattern-based model a probabilistic model and a language model.

In S. Robertson's "Simple BM25 Extension to Multiple Weighted Fields" [7] describes a simple way of adapting the BM25 ranking formula to deal with structured documents. In the past it has been common to compute scores for the individual field (eg: title, body) independently and then combine these scores to arrive at a final score for the document ranking. This paper propose much more intuitive alternative which weight term frequency before the nonlinear term frequency saturation function applied. In this scheme, a structured document with a title weight of two is mapped to an unstructured document which is then ranked in the usual way. Robertson demonstrates the advantage of this method with experiments on Reaturs vol1and TRECdotGor collections.

In Bastide's "Mining frequent pattern with counting inference" [8] proposes a mining frequent pattern from the pattern set. The algorithm PASCAL which introduce a novel optimization of the well known algorithm Apriori. This optimization is based on the strategy called pattern counting inference that relies on the concept of key pattern. Experimental results have shown that without accessing the database the support of frequent non-key patterns can be inferred from frequent key patterns.

## III. EXISTING SYSTEM STUDY

Topic modelling is a text modelling technique that divides the whole documents into number of topics. It can represent every document with number of topics and their corresponding distribution. PLSA and LDA are the most

# International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 9, September 2016

commonly used topic modelling technique. Patterns are used to represent the topics meaningfully than the single words through combining the topic model with pattern mining technique.

In this system Maximum Matched Pattern Based Topic Modelling is used for ranking the new incoming documents. It is an integrated data mining technique with statistical topic modelling, that is used to represent documents and document collections based on a pattern based topics. In this system we consider of users interest with multiple topics rather than a single topic. In MPBTM topic preference of each document or document collection is described by the topic distribution. In the existing system a structured pattern based topic representation in which patterns are arranged in the form of groups, called equivalence classes based on their statistical features. The Maximum matched patterns which are the largest pattern in each equivalence class that exist in the incoming documents. . Frequency of patterns in each equivalence class is the same and MMPS are calculated for relevance of new incoming documents. The maximum matched patterns are the most representative and discriminative patterns to determine the relevance of the incoming documents

## IV. PROPOSED SYSTEM METHOD

Topic modelling has become one of the most popular probabilistic text modelling techniques, and has been quickly accepted by machine learning and text mining communities. The most contribution of topic modelling is that it automatically classifies documents in a collection by a number of topics and represents every document with multiple topics and their corresponding distribution. Two representative approaches are Probabilistic Latent Semantic Analysis (PLSA) and LDA. Here we propose a promising way to meaningfully represent topics by using patterns rather than using single words through combining topic models with pattern mining techniques. Specifically, the patterns are generated from the words in the word-based topic representations of a traditional topic model such as the LDA model. LDA based on sample occurrence and co-occurrence of the words in the documents.

Maximum Matched Patterns (MPBTM), the patterns which represent user interests are not only grouped in terms of topics, but also partitioned based on equivalence classes in each topic group. The patterns in different groups or different equivalence classes have different meanings and distinct properties. The patterns in the MPBTM are well structured so that the maximum matched patterns can be efficiently and effectively selected and used to represent and rank documents. But it does not consider semantic meaning of pattern from the collection of topics while in the document ranking. . An efficient ranking method, using MPBTM considers semantic meaning of the pattern from collection of topics. This project proposes a way to overcome the limitation of existing system by using Natural Language Processing Natural language processing (NLP), i.e., the open English NLP 2.0 library used in enhanced LDA algorithm for filtering semantic meanings of patterns from the collections of topics.

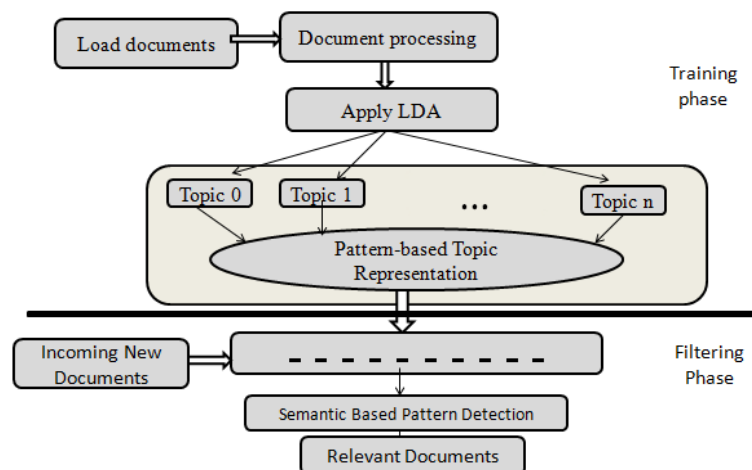


Fig. 1. Proposed system Block Diagram

# International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 9, September 2016

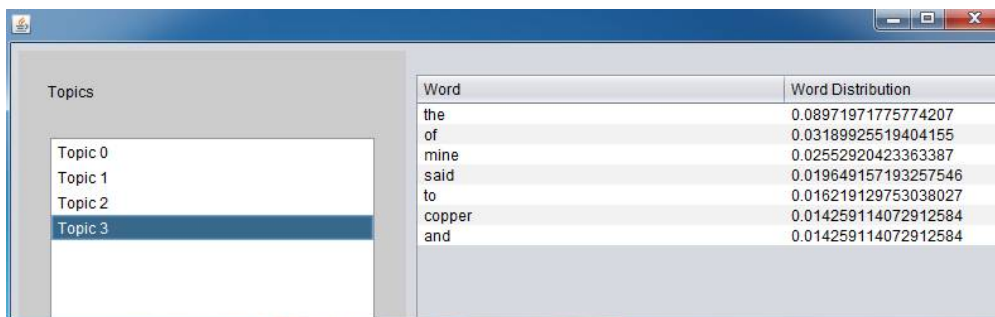
The proposed model represents topics using patterns with structural characteristics which make it possible to interpret the topics with semantic meanings. As with existing topic models, the proposed model is application independent and can be applied to various domains. Information filtering (IF) is a system that removes redundant or unwanted information from an information or document stream based on document representations which represent user's interests. The input data of IF is usually a collection of documents that a user is interested, which represent the user's long-term interests often called the user's profile. As mentioned before, user's information needs usually involve multiple topics. Hence, the proposed pattern-based topic modelling is applied to extract long-term user's interest through IF. In filtering phase the relevance of new incoming document based on semantic based topic filtering is estimated.

## A. LATENT DIRICHLET ALLOCATION

LDA is the most commonly used topic modelling algorithm that discover the hidden topics from collection of documents. Here each discovered topic is represented as distribution over words. LDA discover the hidden topics from the document set by using the word that appears in each document. Let  $D = \{d_1, d_2, \dots, d_m\}$  be the collection of documents and the total number of documents in the collection be 'm'. LDA is applied to the whole documents for dividing it into specified number of topics. The main idea behind LDA is under the assumption of each document is considered to contain multiple topics and each topic can be defined as distribution over words.

The LDA model are represented by using two levels, document level and collection level. At document level each document  $d_i$  from the document set is represented by topic distribution  $\theta_{di} = (\theta_{di,1}, \theta_{di,2}, \dots, \theta_{di,v})$ ,  $V$  is the number of topics. At collection level the document set is represented as  $D$ . Each document is represented by a probability distribution over words,  $\phi_j$  for topic  $j$ . Overall we have  $\phi = \{\phi_1, \phi_2, \dots, \phi_v\}$  for all topics. LDA model also generates the word topic assignment apart from these two levels of representation, that is the word occurrence is considered related to the topics.

The topic distribution over the whole document collection  $D$  can be calculated from the LDA model,  $\phi_D = (\phi_{D,1}, \phi_{D,2}, \dots, \phi_{D,v})$ , where  $\phi_{D,j}$  indicates the importance degree of the topic  $Z_j$  in the collection  $D$ . The most important contribution of LDA model is that the topic representation using word distribution and the document representation using topic representation. The topic representation indicates which words are important to which topic and document representation indicates which topics are important to which document. LDA can learn topics from the collection of documents and decompose the documents according to the topics. Various methods are utilized for new incoming documents to situating the content in terms of trained topics. In this paper we use a pattern based topic model to represent documents and propose an accurate ranking method that determines the relevance of new incoming documents.



Word	Word Distribution
the	0.08971971775774207
of	0.03189925519404155
mine	0.02552920423363387
said	0.019649157193257546
to	0.016219129753038027
copper	0.014259114072912584
and	0.014259114072912584

Fig. 2. Example Results of LDA

## B. PATTERN ENHANCED LDA

Pattern based representation overcome the limitations of word based representation, which provide an accurate method for represent documents. Moreover in pattern-based representation the structural information is provided by the association among the words. In order to discover semantically meaningful pattern from the document set for representing the topics and documents, two steps are proposed:

- (1) Construct a new transactional dataset from the LDA outcomes of the document collection  $D$ .

# International Journal of Innovative Research in Computer and Communication Engineering

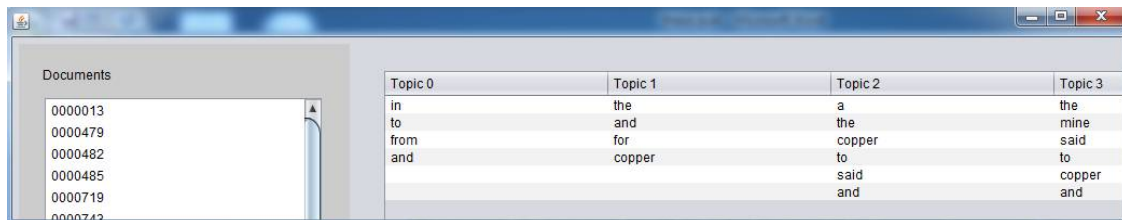
(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 9, September 2016

(2)Generate pattern based representations from the transactional dataset to represent user needs.

## 1. Construct Transactional Dataset

Let  $R_{d_i, z_j}$  represent the word-topic assignment for topic  $Z_j$  in the document  $d_i$ .  $R_{d_i, z_j}$  is a sequence of words assigned to topic  $Z_j$ . In figure 2 the whole document is divided into four topics. For applying LDA the number of topics is specified by the user. The words under each topic occurs in each document is called topical document transaction. Topical document transaction (TDT) is set of words without any duplicates. For all the word-topic assignments  $R_{d_i, z_j}$  to  $Z_j$ , we can construct a transactional dataset  $\Gamma_j$ . Let  $D = \{d_1, \dots, d_M\}$  be the original document collection, the transactional dataset  $\Gamma_j$  for topic  $Z_j$  is defined as  $\Gamma_j = \{I_{1j}; I_{2j}; \dots; I_{Mj}\}$ . Where  $I_{ij}$  is the set of words which occur in  $R_{d_i, z_j}$ .  $I_{ij}$  called a topical document transaction. For each of the topics in  $D$ , we can construct  $V$  transactional datasets ( $\Gamma_1, \Gamma_2, \dots, \Gamma_v$ ). An example of transactional dataset is illustrated in Figure 3, which is generated from Figure 2.

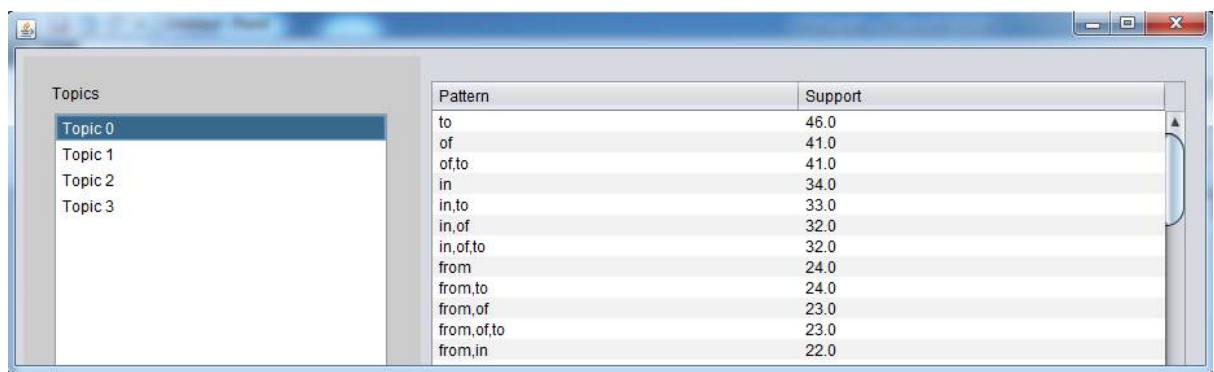


Documents	Topic 0	Topic 1	Topic 2	Topic 3
0000013	in	the	a	the
0000479	to	and	the	mine
0000482	from	for	copper	said
0000485	and	copper	to	to
0000719			said	copper
0000742			and	and

Fig. 3. Topical Document transaction

## 2. Generate Pattern based Representation

In the proposed pattern based method frequent patterns generated from each transactional dataset  $\Gamma_j$  is used to represent  $Z_j$ . Patterns is the set of related words. For a given minimal support threshold  $\sigma$ , an itemset  $X$  in  $\Gamma_j$  is frequent if and only if  $\text{supp}(X) \geq \sigma$ , where  $\text{supp}(X)$  is the support of  $X$  which is the number of transactions in  $\Gamma_j$  that contain  $X$ . Minimal support threshold is specified by the user. The itemset frequency 'X' is defined as  $\frac{\text{Supp}(x)}{|\Gamma_j|}$ .



Topics	Pattern	Support
Topic 0	to	46.0
Topic 1	of	41.0
Topic 2	of,to	41.0
Topic 3	in	34.0
	in,to	33.0
	in,of	32.0
	in,of,to	32.0
	from	24.0
	from,to	24.0
	from,of	23.0
	from,of,to	23.0
	from,in	22.0

Fig. 4. The frequent patterns for topic 0,  $\sigma = 2$

The set of all frequent pattern are represented the topic  $Z_j$ , denoted as  $X_{z_i} = \{ X_{i1}, X_{i2}, \dots, X_{imi} \}$ , where  $m_i$  is the total number of patterns in  $X_{z_i}$  and  $v$  is the total number of topics. For a minimal support threshold  $\sigma = 2$  all frequent pattern generated from Figure 3 are given in Figure 4

## PATTERN EQUIVALENCE CLASS

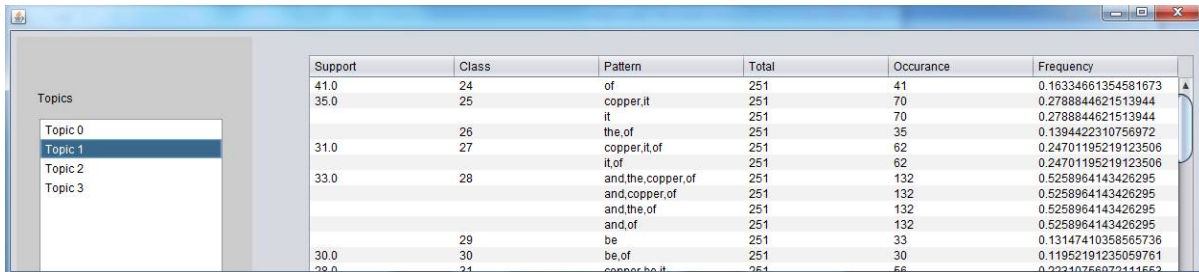
The number of frequent pattern obtained from the previous stage is considerably large and many of them are not necessarily useful. Several concise patterns have been proposed to represent useful patterns instead of frequent patterns generated from a large dataset such as maximal patterns and closed patterns. For a dataset the number of the concise patterns is significantly smaller than the number of frequent patterns generated.

Let EC1 and EC2 be two different equivalence classes of the same transactional dataset. Then  $EC1 \cap EC2 = \phi$  which means that the equivalence classes are exclusive of each other.

# International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 9, September 2016



Support	Class	Pattern	Total	Occurrence	Frequency
41.0	24	of	251	41	0.16334661354581673
35.0	25	copper,it	251	70	0.2788844621513944
		it	251	70	0.2788844621513944
	26	the,of	251	35	0.1394422310756972
31.0	27	copper,it,of	251	62	0.24701195219123506
		it,of	251	62	0.24701195219123506
33.0	28	and,the,copper,of	251	132	0.5258964143426295
		and,copper,of	251	132	0.5258964143426295
		and,the,of	251	132	0.5258964143426295
	29	and,of	251	132	0.5258964143426295
	30	be	251	33	0.1314741035955736
30.0	30	be,of	251	30	0.11952191235059761
29.0	31	copper,be,it	251	66	0.26210766072141562

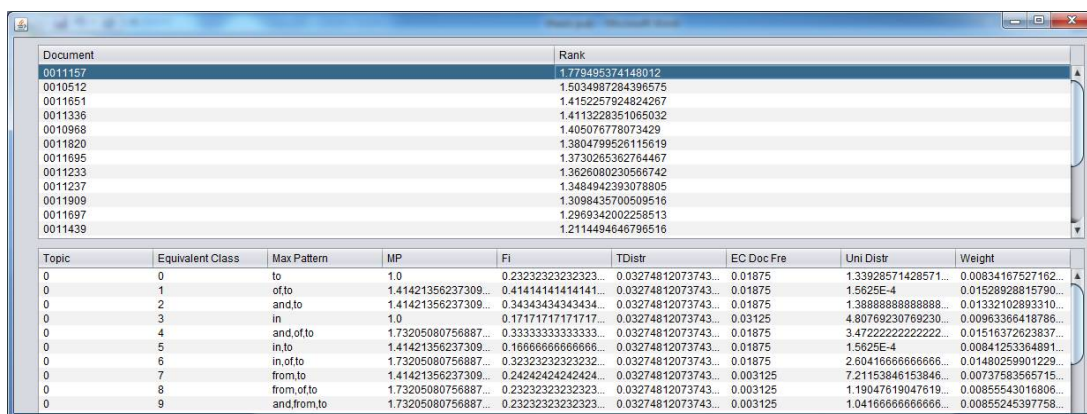
Fig. 5. The equivalence classes in topic 1

There are two pertaining parts used in the proposed model. The training part used to generate user interest model from collection of training document and filtering part determine the relevance of new incoming document. In the proposed model check the semantic meaning of pattern by using open NLP library.

### C. TOPIC-BASED DOCUMENT RELEVANCE RANKING

The patterns that appearing in one equivalence class are same depending on their statistical significance. The difference among them is their size. If longer pattern and shorter pattern that appear in same equivalence class of a document simultaneously, then the shorter one become insignificant since it also covered by longer one and it has the same statistical significance of the longer one.

A new ranking method is proposed to filter out irrelevant document from a set of document based on the users information needs. In this paper, for a new incoming document *d*, the basic way to determine the relevance of '*d*' based on the users interest is firstly to identify maximum matched pattern in *d* which match some patterns in the topic based users interest model and estimate the relevance of '*d*' based on the users topic interest distribution and also considering the semantic features of pattern. Here we not only consider the maximum matched patterns but also consider semantic meaning of pattern for estimate the relevance of new documents. Semantic meaning of pattern is computed by using open NLP library.



Document	Rank
0011157	1.779495374148012
00109512	1.5034987284396575
0011651	1.4152257924824267
0011336	1.4113228351065032
0010968	1.405076778073429
0011820	1.3804799526115619
0011695	1.3730265362764467
0011233	1.3626080230566742
0011237	1.3484942393078805
0011909	1.3098435700509516
0011697	1.2969342002258513
0011439	1.2114494646796516

Topic	Equivalent Class	Max Pattern	MP	Fi	TDistr	EC Doc Fre	Uni Distr	Weight
0	0	to	1.0	0.23232323232323...	0.03274812073743...	0.01875	1.33928571428571...	0.00834167527162...
0	1	of,to	1.41421356237309...	0.41414141414141...	0.03274812073743...	0.01875	1.5625E-4	0.01528928815790...
0	2	and,to	1.41421356237309...	0.34343434343434...	0.03274812073743...	0.01875	1.38888888888888...	0.01332102893310...
0	3	in	1.0	0.17171717171717...	0.03274812073743...	0.03125	4.80769230769230...	0.0096366418786...
0	4	and,of,to	1.73205080759887...	0.33333333333333...	0.03274812073743...	0.01875	3.47222222222222...	0.01516372623837...
0	5	in,to	1.41421356237309...	0.16666666666666...	0.03274812073743...	0.01875	1.5625E-4	0.00841253364891...
0	6	in,of,to	1.73205080759887...	0.32323232323232...	0.03274812073743...	0.01875	2.60416666666666...	0.01480259901229...
0	7	from,to	1.41421356237309...	0.24242424242424...	0.03274812073743...	0.003125	7.21153846153846...	0.00737583565715...
0	8	from,of,to	1.73205080759887...	0.23232323232323...	0.03274812073743...	0.003125	1.19047619047619...	0.0085543018806...
0	9	and,from,to	1.73205080759887...	0.23232323232323...	0.03274812073743...	0.003125	1.04166666666666...	0.0085245397758...

Fig. 6. Document Relevance Ranking

The significance of one pattern is determined not only from their statistical significance, but also its size is considered, since the size of the pattern indicates the specificity level. In Figure 6 shows the ranked document from a document collection. Total weight is referred to as rank of the document. Equivalence class document frequency is computed by dividing the minimum occurrence of a pattern in each equivalence class by the total number of words in each document.

For an incoming document *d*, the relevance ranking of *d* denoted as  $Rank_E(d)$ , is estimated by the following equation:

$$Rank_E(d) = \sum_{j=1}^v \sum_{k=1}^n |MC_{jk}^d|^{0.5} * \delta(MC_{jk}^d, d) * f_{jk} * v_{D,j} \quad (1)$$



# International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 9, September 2016

Where  $v$  is the total number of topics,  $MC_{jk}^d$  is the maximum matched patterns to equivalence class  $EC_{jk}$ ,  $k = 1, \dots, n_j$  and  $f_{j1}, \dots, f_{jn_j}$  is the corresponding statistical significance of the equivalence classes,  $v_{D,j}$  is the topic distribution, and

$$\delta(x,d) = \begin{cases} 1 & \text{if } x \in d \\ 0 & \text{otherwise} \end{cases}$$

## E. ALGORITHMS

The proposed model is implemented by using two algorithms: *user profiling*, that is used to generating user interest model and *document filtering* is used to estimate the relevance of new incoming document. By using user profiling algorithm, users information needs are represented by using pattern based topic representation.

### Algorithm 1. User Profiling

**Input:** collection of positive training document  $D$ ; minimum support  $\sigma_j$  as threshold for topic  $Z_j$ ; number of topics  $v$

**Output:**  $U_E = \{ E(Z_1), \dots, E(Z_v) \}$

1: Generate topic representation  $\phi$  and word-topic assignment  $z_{d,i}$  by applying LDA to  $D$

2:  $U_E = \emptyset$

3: **for** each topic  $Z_j \in [Z_1, Z_v]$  **do**

4: Construct transactional dataset  $\Gamma_j$  based on  $\phi$  and  $z_{d,i}$

5: Construct user interest model  $x_{z_j}$  for topic  $Z_j$  using a pattern mining technique so that for each pattern  $X$  in  $X_{z_j}$ ,  $\text{supp}(X) > \sigma_j$

6: Construct equivalence class  $E(Z_j)$  from  $X_{z_j}$

7:  $U_E = U_E \cup \{ E(Z_j) \}$

8: **end for**

### Algorithm 2. Document Filtering

**Input:** user interest model  $U_E = \{ E(Z_1), \dots, E(Z_v) \}$ , a list of incoming document  $D_{in}$

**Output:**  $\text{rank}_E(d)$ ,  $d \in D_{in}$

1:  $\text{rank}(d) = 0$

2: **for** each  $d \in D_{in}$  **do**

3: **for** each topic  $Z_j \in [Z_1, Z_v]$  **do**

4: **for** each equivalence class  $EC_{jk} \in E(Z_j)$  **do**

5: scan  $EC_{k,j}$  and find maximum matched pattern  $MC_{jk}^d$  which exists in  $d$

6: update  $\text{rank}_E(d)$  using equation(1)

7:  $\text{rank}(d) := \text{rank}(d) + |MC_{jk}^d|^{0.5} * f_{jk} * v_{D,j} * \text{uniform distribution} * \text{equivalent class frequency}$

8: **end for**

9: **end for**

10: **end for**

## V. EXPERIMENTAL RESULTS

In the proposed system the new incoming documents are ranked by using semantic pattern based detection method as shown in Fig.1. In the proposed system initially apply LDA to the set of loading documents. In Fig. 2 shows the results after applying LDA to the documents. LDA finds the hidden topics from the document set. The next step is to determine the topical document transaction (TDT) as in Fig. 3. TDT is the words under each topic that occurs in each of document. Next step is to determine the frequent pattern under each of the topic. Frequent pattern is the pattern that satisfies the minimal support threshold value. Fig. 4 shows the frequent patterns for topic 0. The next step is to

# International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 9, September 2016

determine the equivalent classes in each topic as shown in Fig.5. Equivalent classes are the super class of the frequent pattern. Finally the new incoming documents are ranked depending on the equivalent class as shown in Fig.6.

In training stage, for different document collection the number of topics involved in the collections is different. Therefore the selection of appropriate number of topics in the LDA stage is important. In the proposed method maximum matched pattern is increased since the rank of document is also increases.

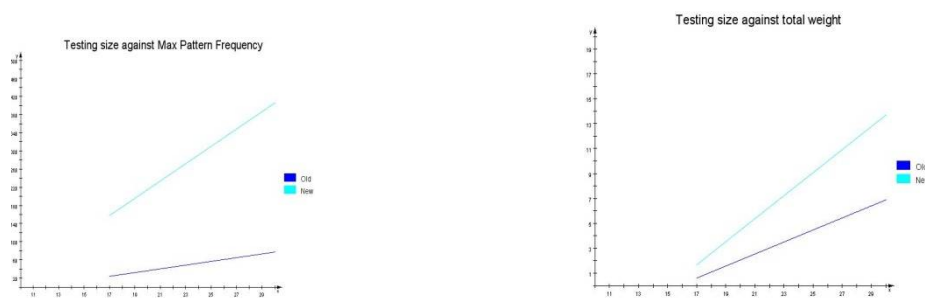


Fig. 7. Comparison of proposed MPBTM with existing method

Fig.7 shows difference among the proposed MPBTM with existing MPBTM. Testing size indicates the total number of documents to be tested. In the proposed method maximum matched pattern will be increased. For example , 17 document is tested for ranking ,then the maximum matched pattern in the old method will be 28 and new method have 118. By using proposed method rank of the testing document will be increased. From the figure we can see proposed model is best among all models.

## VI. CONCLUSION

This presents an innovative pattern enhanced topic model for information filtering including user interest modelling and document relevance ranking. The proposed MPBTM model generates pattern enhanced topic representations to model user's interest's across multiple topics. In the filtering stage, the MPBTM selects maximum matched patterns, instead of using all discovered patterns, for estimating the relevance of incoming documents. The proposed approach incorporates the semantic structure from topic modelling and the statistical relevant method from the most representative patterns. The proposed model introduces the Open English NLP 2.0 on the enhanced LDA models for retrieving semantic meaning of patterns.

The proposed model automatically generates discriminative and semantic rich representations for modelling topics and documents by combining statistical relevant topic modelling techniques and data mining techniques. The technique not only can be used for information filtering, but also can be applied to many content-based feature extraction and modelling tasks. Proposed model demonstrates excellent strength on document modelling relevance ranking.

## REFERENCES

1. X. Wei and W. B. Croft, "LDA-based document models for ad-hoc retrieval," in proceedings of the 29th annual International ACM SIGIR conference on Research and Development in Information Retrieval. ACM, 2006, pp. 178–185.
2. C. Wang and D. M. Blei, "Collaborative topic modeling for recommending scientific articles," in Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2011, pp. 448–456.
3. T. Hofmann, "Probabilistic latent semantic indexing," in Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 1999, pp. 50–57.
4. D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," the Journal of Machine Learning Research, vol. 3, pp. 993–1022, 2003.
5. Y. Gao, Y. Xu, and Y. Li, "Pattern-based topic models for information filtering," in Proceedings of International Conference on Data Mining Workshop SENTIRE, ICDM'2013. IEEE, 2013.
6. Y. Cao, J. Xu, T.-Y. Liu, H. Li, Y. Huang, and H.-W. Hon, "Adapting ranking svm to document retrieval," in Proceedings of the 29<sup>th</sup> Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 2006, pp. 186–193.





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 9, September 2016**

7. S. Robertson, H. Zaragoza, and M. Taylor, "Simple BM25 extension to multiple weighted fields," in Proceedings of the thirteenth ACM International Conference on Information and Knowledge Management. ACM, 2004, pp. 42–49.
8. Y. Bastide, R. Taouil, N. Pasquier, G. Stumme, and L. Lakhal, "Mining frequent patterns with counting inference," ACM SIGKDD Explorations Newsletter, vol. 2, no. 2, pp. 66–75, 2000.
9. F. Beil, M. Ester, and X. Xu, "Frequent term-based text clustering," in Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2002, pp. 436–442.
10. S.-T. Wu, Y. Li, and Y. Xu, "Deploying approaches for pattern refinement in text mining," in 6th International Conference on Data Mining, ICDM'06. IEEE, 2006, pp. 1157–1161.
11. H. Cheng, X. Yan, J. Han, and C.-W. Hsu, "Discriminative frequent pattern analysis for effective classification," in IEEE 23<sup>rd</sup> International Conference on Data Engineering, ICDE'2007. IEEE, 2007, pp.716–725.
12. J. Han, H. Cheng, D. Xin, and X. Yan, "Frequent pattern mining: current status and future directions," Data Mining and Knowledge Discovery, vol. 15, no. 1, pp. 55–86, 2007.
13. R. J. Bayardo Jr, "Efficiently mining long patterns from databases," in ACM Sigmod Record, vol. 27, no. 2. ACM, 1998, pp. 85–93.