Information Filtering Model Based on Topic Pattern for Document Modeling

Chinnu C. George, Abdul Ali

Abstract— In the field of machine learning and text mining topic modelling is widely used. Topic modelling generates models to discover the hidden topics in a collection of documents and each of these topics are represented by the distribution- of words. Many term-based and pattern-based approaches are there in the field of information filtering. Patterns are more discriminative than the single words. In many pattern-based methods only the presence or absence of the patterns in the documents are considered. Even if the pattern occurs multiple times in the documents to be filtered equal importance is considered. Another problem with the existing pattern-based methods is that the semantics of the terms in the patterns are not considered. Another limitation is that the distribution of the patterns is not given any importance. To deal with the above limitations and problems this paper includes a new ranking method that considers the frequency of the patterns, pattern distribution and semantic based pattern representation to estimate the relevance of the documents based on the user information needs. This helps to filter out the irrelevant documents effectively. Extensive experiments are conducted using the TREC data collection Reuters Corpus Volume 1 to evaluate the effectiveness of the proposed method .The result shows that the proposed model outperforms the pattern based topic for document modeling in information filtering.

Index Terms— Topic modelling, information filtering, user interest modeling, semantic based relevance ranking.

I. INTRODUCTION

Information filtering (IF) is a system that removes redundant or irrelevant information from document stream based on users' interest. Traditional information filtering models are based on term-based approach. The advantage of the term-based approach is its efficient computational performance, as well as mature theories for term weighting, such as Rocchio, BM25, etc [1], [2]. But these term-based approaches have certain disadvantages like the problem of polysemy and synonymy. Another approach for information filtering is phrase-based approach in which the phrases are used to represent the documents. But the phrase-based approach suffer from low frequency problem. To overcome these limitations patterns are used to represent the users interest. These patterns will provide more semantic information than the single words or terms.

Topic modelling [3],[4] is widely used to mine topics from the documents. It will identify the hidden topics and will represent each document with multiple topics. Probabilistic Latent Semantic Analysis (PLSA) [5] and LDA [3] are the two approaches in topic modeling. Probabilistic Latent Semantic Analysis [5] is a technique from the category of topic models. Its main goal is to model co-occurrence information under a probabilistic framework in order to discover the underlying semantic structure of the data. LDA [3] is a statistical topic modeling technique and is a tool which is currently used to discover hidden topics to represent the documents. It identifies the topics from the words appearing in the documents.

II. RELATED WORKS

Information filtering system gets user interest or user information needs based on the 'user profiles'. Information filtering systems expose users to the information that are more relevant to them [6].In the process of information filtering main objective is to rank the incoming documents based on its relevance. If D is the collection of incoming documents the process of information filtering is a mapping Rank(d):D-->R where rank(d) represents the relevance of the document d.

Document filtering can be considered as the document ranking process. There are several approaches to model the relevance of the documents. These include term-based model [2], pattern-based model [7] [8], a probabilistic model.

Most popular term-based models include tf*idf, Okapi BM25 and various weighting schemes for the bag of words representation [9]. These models suffer from the problem of polysemy and synonymy and have the limitation of expressing semantics. So more semantic features such as phrases and patterns are extracted to represent the documents . But the phrase-based approach suffer from low frequency problem. Pattern based approach is more effective compared to other approaches [7][8]. Pattern mining is one of the most important topic in datamining and is very widely studied over years. Many effective algorithms like Apriori, FP-tree, are developed to extract the frequent patterns. In many the number of these frequent patterns will be huge to process. So more precise or relevant patterns are discovered like closed patterns ,max pattern etc. Closed patterns [10] will be condensed representation of the frequent patterns.

Topic models techniques have been incorporated in the frame of language model and have achieved successful retrieval results[3], [11], which has opened up a new channel to model the relevance of a document. The LDA based document models are state-of-the-art topic modeling approaches . This model achieves good performance compared to other models . The authors says that the this is achived by [11],not only because of the multiple topic document model, but also because each topic in the topic model is represented by a group of semantically similar words, which solves the synonymy problem of term based document models.

Revised Version Manuscript Received on October 09, 2015.

Chinnu C. George, PG Scholar, Department of Computer Science, Ilahia College of Engineering, Muvattupuzha, (Kerala). India.

Abdul Ali, Assistant Professor, Department of Computer Science Department, Ilahia College of Engineering, Muvattupuzha, (Kerala). India.

Probabilistic topic modelling [12] can also extract long term user interests by analysing content and representing it in terms of latent topics discovered from user profiles. The relevant documents are determined by a user-specific topic model that has been extracted from the user's information needs [13]. Especially when information needs are sensitive to some parameters, both the topic model and the language models are very limited in representing the specificities.

III. PROPOSED METHOD

The proposed model consists of two phase: 1) training part that generates user interest model from a collection of training documents (user interest modeling) and 2) filtering part that determines the relevance of the new incoming documents based on user interest model generated during training phase (document relevance ranking). Fig.1 shows the framework of the proposed model.

3.1 User Interest modelling

To represent topics, patterns are more accurate and meaningful than words. Moreover, pattern-based representations provides more structural information which reveals the association between words. Four steps are proposed to generate the Topic based user interest model . First two steps will be discovering semantically meaningful pattern to represent topics and documents, 1) construct a new transactional dataset from the results of LDA model 2) generate pattern-based representations from the transactional dataset to represent user needs. Pattern-based topic representations may not be sufficient or accurate enough to be directly used to determine the relevance of new documents to the user interests. The document relevance ranking is done based on Maximum Matched Patterns, which are the most distinctive and representative patterns. In third step pattern equivalence classes are constructed. Finally generates the user interest model.

3.1.1 Latent Dirichlet Allocation (LDA)

LDA is a technique that automatically discovers topics that are present in the documents. In LDA, each document may be viewed as a mixture of various topics. LDA provides topic representation using word distribution and document representation using topic distribution. Topic representation means which words are important to which topics and document representation means which topics are important to which documents. LDA is a widely used topic modeling tool. Example result of LDA is shown in Table I.

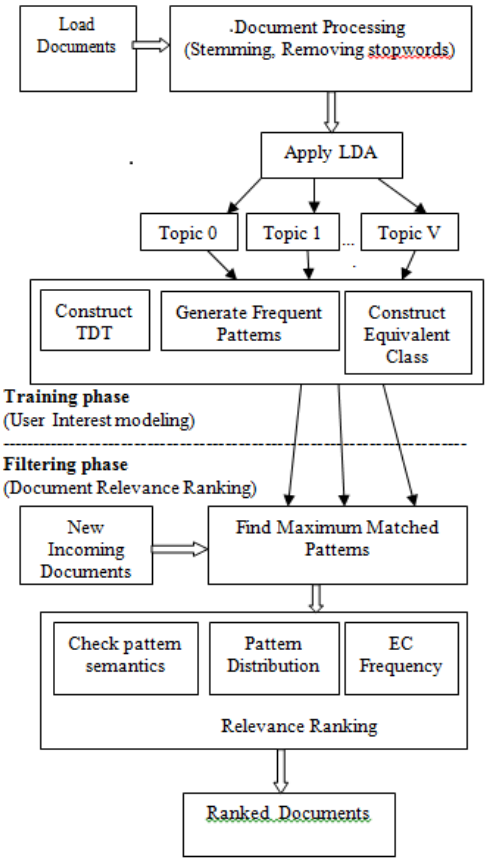


Fig. 1 Framework of the proposed model.

Table I: LDA Result, word-topic assignment

Topic Document	Z1 words	Z2 words	Z3 words
d1	*w1,w2,w3,w2,w1*	*w1,w9,w8*	*w7,w10,w10*
d2	*w2,w4,w2*	*w7,w8,w1,w8,w8*	*w1,w11,w12*
d3	*w2,w1,w7,w5*	*w7,w3,w3,w2*	*w4,w7,w10,w11*
d4	*w2,w7,w6*	*w9,w8,w1*	*w1,w11,w10*

3.1.2 Construction of Transactional Dataset

The word-topic assignment to the topic zj document di is represented as Rdi,zj We construct a transaction dataset Γj for each word-topic assignment Rdi,zj to zj , j=1,…..V and i=1,…...M where V is the number of topics and M is the number of documents. Let D={d1,……dM} be the set of document collections, the transactional dataset Γj for topic zj is defined as Γj ={ I1j, I2j, . . . ; IMj} where Iij is called topical document transaction which contains words without any duplicates. Iij contains the words which are in document di and assigned to topic zj by LDA. For example from Table I for topic z1 in document d1 word-topic assignment is <w1,w2,w3,w2,w1>after eliminating the duplicates transaction dataset Γ1 is generated for topic z1 and {w1,w2,w3} be the topical document transaction. Table II

Table-II: Transaction Datasets Generated from Table-I

	TDT Γ_1	TDT Γ_2	TDT Γ_3
1	{w1,w2,w3}	{w1,w8,w9}	{w7,w10}
2	{w2,w3}	{w1,w7,w8}	{w1,w11,w12}
3	{w1,w2,w5,w7}	{w2,w3,w7}	{w4,w7,w10,w11}
4	{w2,w6,w7}	{w1,w8,w9}	{w1,w11,w10}

3.1.4 Topic based Pattern Representation

Frequent patterns are generated during this stage. Table 3 shows the frequent patterns generated for Z2 .The frequent patterns are generated from each transactional dataset Γj. Let σ be the minimal support threshold, then an itemset X in Γj is frequent if supp(X) >= σ, where supp(X) is the support of X which is the number of transactions in Γj that contain X. The frequency of the itemset X is defined as

$$\frac{supp(X)}{|\Gamma j|}$$.

Table 3: The Frequent Patterns for Z2 σ= 2

Patterns	*supp*
{w1}, {w8}, {w1, w8}	3
{w9}, {w7}, {w8, w9}, {w1, w9}, {w1, w8, w9	2

3.1.3 Construction of Pattern Equivalence Class

The number of frequent patterns in some of the topics can be very large and many of the patterns are not discriminative enough to represent specific topics. As a result, these topic representations are not sufficient to represent the documents accurately. That means, the pattern based representation that represents the user interest is not sufficient or accurate to be used to determine the relevance of new documents. So the relevance of the new documents is estimated based on the Maximum matched pattern. These Maximum Matched patterns are the more distinctive and representative patterns. Instead of frequent patterns ,closed patterns are used for topic representation and the number of these patterns are significantly smaller than the number of frequent patterns for a dataset.

Closed Itemset [14] : For a transactional dataset, an itemset X is a closed itemset if there exists no itemset X' such that (1) X C X', (2) supp(X) = supp(X').

Generator [14] **:** For a transactional dataset G, let X be a closed itemset and T(X) consists of all transactions in G that contain X, then an itemset g is said to be a generator of X if g C X; T(g)= T(X) and supp(X)= supp(g).

Equivalence Class [14] **:** For a transactional dataset G, let X be a closed itemset and G(X) consist of all generators of X, then the equivalence class of X in G, denoted as EC(X), is defined as EC(X) =G(X) U {X}.

All the patterns in an equivalence class have the same frequency. The frequency of a pattern indicates the statistical significance of the pattern. Table 4 shows the Equivalence Classes in Z2.

Table 4: The Equivalence Classes in Z2

EC_{21}	EC_{22}	EC_{23}
{w1, w8}	{w1, w8, w9}	{w7}
{w1}	{w1, w9}	
{w8}	{w8, w9}	
	{w9}	

Each topic may be having a set of equivalent classes. These equivalent classes are used to represent the user interest model. Let E(Zi) be the set of equivalent classes of topic Zi and the user interest model for V number of topics can be represented as Um ={ E(Z1),E(Z2),……..,E(Zv)}.

3.1.5 Algorithm

Input : collection of documents D, number of topics V
Output : user interest model ,Um={ E(Z1),E(Z2),……..,E(Zv)}.
1: Load collection of input documents D.
2: Perform stemming and remove stop words.
3:Apply LDA and generate word-topic assignment for V number of topics.
4: Um = { }
5: **for** each topic Zj do
6: Construct the Transaction Dataset Γj
7: Generate all frequent patterns whose supp(X) >= σ using pattern mining techniques.
8: Construct equivalent classes E(Zj)
9:Construct user interest model Um={E(Z1),E(Z2),..,E(Zv)}
10. **End for**

3.2 Document Relevance Ranking

Relevance of the documents is estimated based on the user interest model to filter out irrelevant documents. The maximum matched pattern in the equivalent classes are used to estimate the relevance of the new incoming documents to the user interest. Based on the relevance of the documents the new documents will be ranked.

3.2.1 Relevance Based on Pattern Semantics

Instead of simply matching the maximum matched pattern within the incoming documents the connected words of the pattern is also considered. If the connected words of the pattern is present in the document then it will be considered to estimate the count of max pattern. For example: consider the maximum matched pattern {pattern, mine, topic} instead of just taking the count of pattern, mine and topic the connected words of pattern, mine and topic will also be considered.

3.2.2 Relevance Based on Distribution of Pattern

Distribution of the maximum matched pattern in the new incoming documents is important while estimating the relevance. If the pattern is present in the title , first paragraph and last paragraph ,some additional weight should be provided while calculating the rank of the documents. Also if the pattern is distributed all over the document than just in a single paragraph then that document will be more relevant. Document segmentation is done to check the distribution of pattern in the documents.

3.2.3 Relevance Based on Equivalent Class Frequency

The count of the pattern present in the incoming document is also considered while estimating the relevance. The more the number of presence of pattern the relevance of the document is more.

Thus document relevance is estimated using the equation

$$\text{Rank(d)}=\sum_{j=1}^{V}\sum_{k=1}^{nj}(|MC_{jk}^{d}|^{5}\times\delta(MC_{jk}^{d},d)\times f_{jk}\times\vartheta_{dj})\times 0.6+0.2\times UniformDist+0.2\times ECfrequency[i] \quad (1)$$

Where V is the number of topics and *fjk* represents the frequency of equivalent class. MC_{jk}^{d} represents the maximum matched patterns to the equivalent classes ϑ_{dj} is the topic distribution. *UniformDist* represents estimated uniform distribution and ECfrequency is the Equivalent class frequency.

3.2.4 Algorithm

Input : User interest model, collection of new input documents Din

Output : ranked documents

1: Rank (d) =0
2: **for** each document d do
3: **for** each topic Zj do
4: **for** each equivalence class do
5: Scan all the equivalent classes and find out the maximum matched pattern
6: **for** each maximum matched pattern
7: Check the semantics of each term in pattern
8: Check the distribution of pattern in the document d
9:**End for**
10: Calculate the equivalent class frequency
11: Update Rank (d) using the equation (1)
13: **End for**
14: **End for**
15: **End for**

IV. EXPERIMENTAL RESULTS

We present the experimental results in terms of total weight of the ranked documents, distribution based on the top ranked document and the frequency of equivalence class.The results shows the effectiveness of new relevance ranking method .

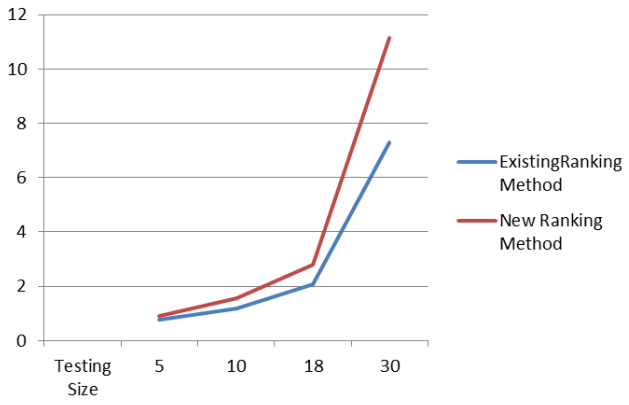


Fig 2: Comparison based on total weight

Experimental result based on total weight is shown in Fig 2 The graph shows that the proposed model with new relevance ranking method shows more weight compared to the existing pattern based model. This means that new model will retrieve more relevant documents.

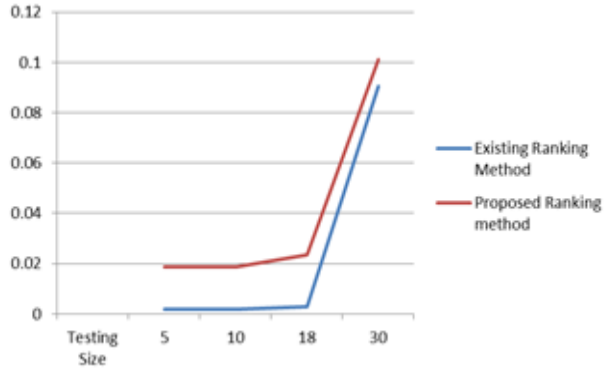
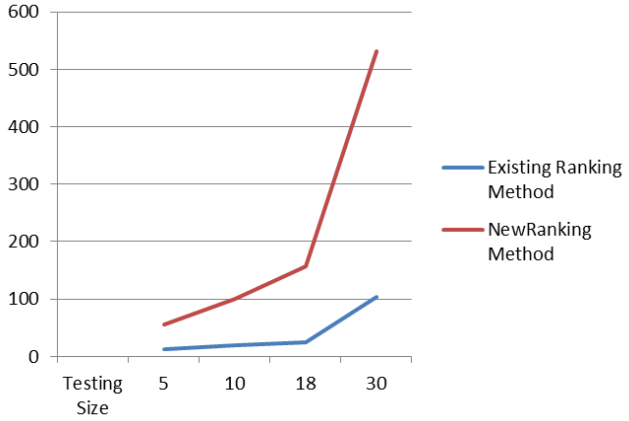


Fig 3: Comparison based on uniform distribution

Fig 3 shows the results based on uniform distribution. Uniform distribution for comparison is measured based on the top ranked document. The proposed model shows higher distribution means this document covers the topic more. Experimental result based on frequency of equivalent class is shown in Fig 4. The graph shows that the proposed model with new relevance ranking method will detect documents with meaning of the words in pattern.



V. CONCLUSION

In this paper we have presented a pattern based topic model for information filtering including user interest modelling and document relevance ranking. The proposed model generates user interest model based on multiple topics. The model consists of two phases a training phase and a document filtering phase. The proposed model generates a user interest model and based on that the relevance estimation of the new documents is performed. Many limitations of the existing pattern-based topic model is resolved by the proposed relevance ranking method. In the new ranking method meaning of the terms in the pattern and the frequency of the pattern and the distribution of the patterns in the new documents are also considered. In the future, we can select more discriminative and precise patterns for representing topics and document relevance.

ACKNOWLEDGMENT

The authors wish to thank the Management, the Principal and Head of the Department (CSE) of ICET for the support and help in completing the work.

REFERENCES

1. 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.
2. 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.
3. X. Wei and W. B. Croft, “LDA-based document models for ad-hoc retrieval,” in Proc. 29th Annu. Int. ACM SIGIR Conf. Res. Develop. Inform. Retrieval, 2006, pp. 178–185.
4. D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent dirichlet allocation,” J. Mach. Learn. Res., vol. 3, pp. 993–1022, 2003.
5. T. Hofmann, “Probabilistic latent semantic indexing,” in Proc. 22nd Annu. Int. ACM SIGIR Conf. on Res. Develop. Inform. Retrieval, 1999, pp. 50–57.R.
6. Hanani, U., Shapira, B., and Shoval, P. (2001). Information filtering: Overview of issues, research and systems. User Modeling and User-Adapted Interaction, 11(3):203–259.
7. 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.
8. 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.
9. 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.G.H.
10. 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.
11. 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
12. C. Wang and D. M. Blei, “Collaborative topic modeling for recommending scientific articles,” in Proc. 17th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., 2011, pp. 448–456.
13. Y. Zhang, J. Callan, and T. Minka, “Novelty and redundancy detection in adaptive filtering,” in Proc. 25th Annu. Int. ACM SIGIR Conf. Res. Develop. Inform. Retrieval, 2002, pp. 81–88.
14. Gao, Y., Xu, Y., and Li, Y. (2013a). Pattern-based topic models for information filtering. In Proceedings of International Conference on Data Mining Workshop SENTIRE, ICDM’2013. IEEE.J.R. Quinlan, C4.5: Programs for Machine Learning. Morgan Kaufman, 1993.