Pattern Enhanced Topic Model for Document Modelling and Ranking in Information Filtering

Sreerekha R S, Smitha E S

Abstract—Topic modelling is widely used in the field of machine learning and text mining. 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, pattern-based and topic modelling based approaches are used in the field of information filtering. There is a Maximum Matched pattern based topic model for information filtering. In this model the user interest are generated in terms of multiple topics and the most discriminative and representative patterns are used to estimate the document relevance to the users information needs. But this method consider only the presence or absence of the pattern in the document to estimate the relevance of the documents. Another problem with existing method is that it does not consider the distribution of patterns in the incoming documents. To deals with the above problems this paper proposed a new ranking method which consider the number of matched patterns and the location of sentences which contain matched patterns in the incoming documents to estimate the relevance of that document based on the user information needs. This helps to find the most relevant documents effectively. Extensive experiments are conducted using the TREC data collection Reuters Corpus Volume 1 to evaluate the effectiveness of the proposed method.

Index Terms— Document Relevance Ranking, Information Filtering, Pattern Mining, Topic Modelling, User interest Modelling.

1 INTRODUCTION

Information Filtering is the largely evolving area to manage large information flows. Information filtering (IF) system remove unwanted information from an information or document stream based on the interest of user [1]. This process includes two components: User interest Modelling (user profiling) component and filtering component. The user interest modelling component identify the user's information needs and build a user Profile. The filtering model filtered the incoming documents from the document collection, based on the constructed user profile. i.e., the success of the Information Filtering system will depend on the constructed user profile. To construct an accurate user profile is an important task in Information Filtering.

Traditional information filtering models use the term-based [2], phrase based [3] and pattern based [4] [5] approaches for building user profiles and representing documents. All these approaches have some limitations and these are based on the assumption that user interest is about single topic and documents contain single topic. But really user's interest is dynamic and documents contain multiple topics. So topic modelling is used to model user interest and documents in a collection. Topic modelling [6], 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) [7] and LDA [6] are the two approaches in topic modelling. But these topic model suffers from two problems. First problem is the limited number of predefined topic which are insufficient for document representation.Second problem is that these models are word based. To overcome all the problems in the topic modelling techniques, patterns are used to represent the user interest since patterns are more semantic than words. But number of patterns in some of the topic can be huge and many of the patterns are not discriminative to represent specific topics. So we use most discriminative and representative pattern to model the user interest and representing the documents in information Filtering.

The filtering component filter out the irrelevant documents which is done based on the created user profile. The filtering is done by finding the ranks of the incoming documents based on the matching of user profile with the incoming documents. For finding the ranks of the document we consider the number of times a matched pattern appear in a document and also consider the location or position of the matched pattern in the document which is not yet considered to finding the ranks of the documents.

2 LITERATURE REVIEW

In Information Filtering system user needs are obtained from the user profiles. The main aim of information filtering system is to mapping form space of incoming documents to space of user relevant documents [1]. Document filtering is a task associated with ranking the document based on their relevance to the user profile. There are various approaches used in the Information Filtering for modelling user interest and the documents. These approaches include term based [2] [8], pattern based [3] [5] and topic modelling based [6] methods.

The popular term based models include tf*idf [9] [10], BM25 [8] and various weighting schemes for bag of words representation. The advantage of term based approach is its efficient computational performance. But the term based representation suffers the problem of polysemy and synonymy and have some limitation in expressing semantics. To avoid these limitations the combinations of single words (phrases) can be used. There exists a work [3] which explain how to use phrases (sequential patterns) for modelling an Information Filtering system. Phrases have more semantic meaning than words but it has low occurrence in a document. So various pattern mining technique have been studied in many years. So variety of efficient algorithms such as Apriori [10], PrefixSpan [4] etc. have been developed for frequent patterns. These algorithms return huge number of patterns. So to find the most representative pattern for representing a document is very difficult task. Thus improved representations of patterns such as maximal pattern [12], closed pattern [4] etc. have been proposed. Among these, Frequent closed item set have the high reliability to represent user profiles and documents. There is an effective pattern discovery technique [13] 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.

Topic Modelling, the most popular text mining technique, which has been quickly accepted by machine learning and text mining. The topic modelling techniques automatically classifies the documents in a collection by number of topics and every document is represented by multiple topics. The topic models techniques opened a new channel to model the relevance of the document. The Latent Dirchilet Allocation (LDA) [6] based document models are state-of-the-art topic modelling approaches. This model achieves good performance compared to other models. So LDA is the most commonly used tool and algorithm for topic modelling. But some problems or limitations associated with topic modelling since all these approaches have been developed based on words which is not sufficient to represent a document.

3 Proposed MODel

The proposed model consists of two phases: User interest modelling phase and Filtering phase (document filtering or Ranking). The first phase generate a user profile from the collection of training documents. The second phase determines the relevance of incoming documents based on the user profile constructed in the first phase. Fig 1 shows the architecture of proposed model.

3.1 User Interest Modelling

The process which is used to generate a topic based user interest model consists of five steps. In the first step we use LDA to generate topic based representation of the documents. The next two steps are used for discovering semantically meaningful patterns for representing topics and documents. In second step we have to construct transactional dataset for each topic from the result of LDA and in third step generate pattern based representation from the transactional dataset for representing user needs. In the fourth step pattern equivalence class are generated, these equivalence class contains the most representative patterns for representing the topics. Finally the patterns in the equivalent class are used to represent the user interest or user profile.

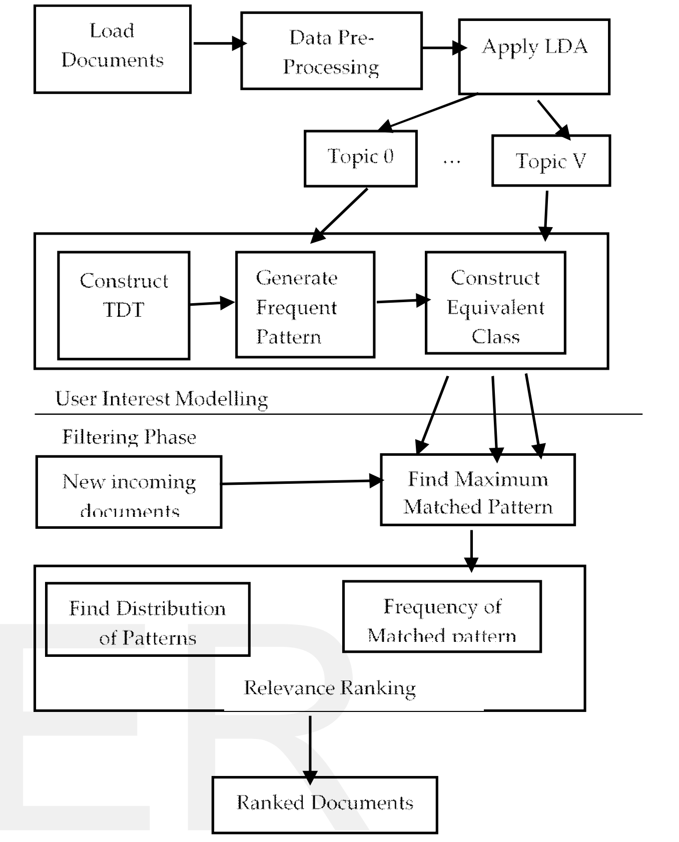

Fig 1: System Architesture

3.1.1. Latent Dirichlet Allocation (LDA)

LDA is the most commonly used topic modelling technique and currently used tool for topic modelling [6] [14]. The idea behind LDA [6] is to model documents as arising from multiple topics. It is a technique that automatically discovers topics that are contained in the document. LDA provide topic representation by using words and document representation by using topics. The representation of results of LDA model is in two level: Document level and collection level [14]. At document level document is represented by topic distribution and collection level, collection of documents D is represented by set of topics which are represented by word distribution. LDA also represent result in the form of word topic assignments.

3.1.2. Construction of Transactional Dataset

The R_{d_i}, Z_j represent word-topic assignment to topic Z_j in document d_i. Then we have to construct transactional dataset [14] Γ_j for each word-topic assignment R_{d_i}, Z_j to Z_j, j=1…V and i=1…M where V is the number of topics and M is the number of documents. Let D = $\{d_1, \ldots \ldots d_M\}$ be the set of document collections, the transactional dataset Γ_j for topic Z_j is defined as $\Gamma_j = \{ I_{1j}, I_{2j}, \ldots ; I_{Mj} \}$ where I_{ij}, is called topical document transaction which contains words without any duplicates [15]. I_{ij} , contains the words which are in document di and assigned to topic Z_i by LDA.

3.1.3. Pattern based representation for Topics

In this stage, the frequent patterns are generated for representing the topics. Frequent patterns are patterns that occurs frequently in a document. The frequent patterns are generated from each transactional dataset Γ_i. Let σ be the minimal support threshold, then an item set X in Γ_i is frequent if supp(X) >= σ, where supp(X) is the support of X which is the number of transactions in Γ_i that contain X.

3.1.4. 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 [15]. 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. All the patterns in an equivalence class have the same frequency. The frequency of a pattern indicates the statistical significance of the pattern.

3.1.5. Topic Based User interest Modelling

For a collection of documents, the user's interest can be represented by the patterns in the topic of that collection. Each topic may be having set of equivalent classes. These equivalent classes are used to represent the user interest model. Let $E(Z_1)$ be the set of equivalent classes of topic Z_i and the user interest model for V number of topics can be represented as Um ={ $E(Z_1), E(Z_2), \ldots \ldots \ldots, E(Z_V)$}.

```
Algorithm: User Interest Modelling
Input:  Collection of documents D and number of topics V
Output: User interest model:
Load the collection of documents D
Perform Data Pre-processing
Apply LDA to generate word-topic assignment to V
Um = { }
For each topic $Z_i$ do
    Construct Transactional Dataset $\Gamma_i$
    Generate frequent patterns for each topic from the transactional dataset
    Construct Equivalence class for each topic $E(Z_i)$
    Construct User Interest Model Um= { $E(Z_1), E(Z_2), \ldots \ldots \ldots, E(Z_V)$}.
End for
```

3.2 Document Filtering

In this phase Filtering of the document is takes place. This stage filter out the irrelevant documents. The relevance of the document is estimated based on the user's information needs. For a new incoming document, to determine the relevance of the document firstly to identify the maximum matched patterns in the document which match some patterns in the topic based user interest model. Then we have to find the relevance of the document based on user's topic interest distributions, significance of matched patterns, frequency of matched pattern in the document and the distribution of matched pattern in the document.

Instead of simply matching the maximum matched pattern within the incoming document we consider the number of times the matched pattern appeared in the document. In this we not only consider the frequency but also consider the synonyms of the words in the maximum matched pattern. The more the number of presence of pattern the relevance of the document is more. Another parameter which is considered to find the relevance of the document is the distribution of matched pattern in the document. 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. Text summarisation [16] techniques are used to find the pattern distribution in the document.

Thus document relevance is estimated using the equation

$$Rank(d) = \left(\sum_{i=1}^{V} \sum_{k=1}^{nj} |MC_{ik}^{d}|^{.5} \ * \ f_{ik} * \vartheta_{D,i}\right) + 0.2^* \text{ distribution } +0.2^*\text{EC frequency} \qquad (1)$$

V -number of topics, f_{ik} - Frequency of equivalent class, MC_{ik}^{d} - Maximum Matched Pattern, $\vartheta_{D,i}$ - Topic Distribution of topic j, distribution is the estimated uniform distribution, EC Frequency –frequency of matched patterns.

```
Algorithm:  Document Filtering
Input: User interest Model, a list of testing documents
Output: Ranked documents
Rank (d) =0
for each document d do
    for each topic Zj
        for each equivalence class do
            Scan all the equivalent classes and find Maximum matched pattern
            for each Maximum Matched pattern
                Check the distribution of pattern in document d
                Calculate the Equivalent class Frequency of document d
            end for
            update rank using the equation above equation
        end for
    end for
end for
```

4 EXPERIMENTAL RESULT

We present the experimental results in terms of rank of the ranked document and the frequency of the matched pattern in the incoming document. For evaluating this model, we use the Reuters Corpus Volume1 dataset. The results shows the effectiveness of new ranking method.

Experimental result based on rank 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.

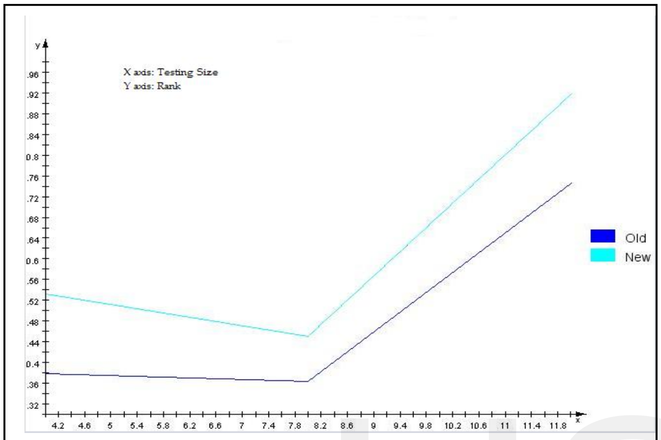

Fig 2: Comparison between Testing size and Rank

Experimental result based on frequency of equivalent class is shown in Fig 3. The graph shows that the proposed model with new relevance ranking method will detect documents with meaning of the words in pattern.

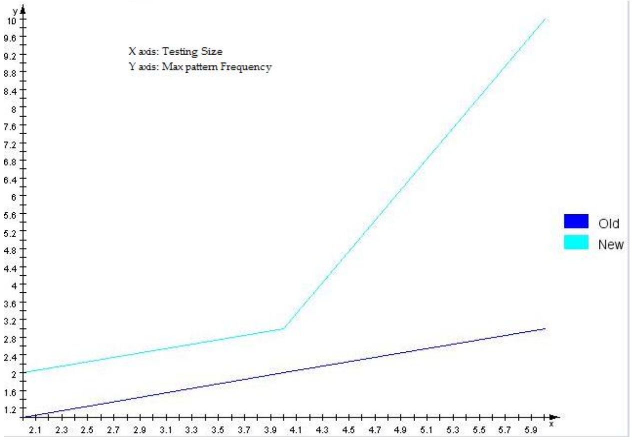

Fig 3: Comparison between Testing size and Pattern frequency

5. CONCLUSION

In this paper we have presented a pattern based topic modelling for Information Filtering. This model includes the user interest modelling and the document relevance ranking. User interest is represented in terms of multiple topics. Topics are represented in terms of patterns. The most reliable and representative patterns are used for representing the user interest. The document ranking is done based on the matching of these patterns in the document. This new ranking method consider the frequency of matched pattern and the distribution of patterns in the document. In future, this model will be applied to some other content based systems such as content based recommender systems.

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):203-259

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

[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]] 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.

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

[7] T. Hofmann, "Probabilistic latent semantic indexing," *in Proc. 22nd Annu. Int. ACM SIGIR Conf. on Res. Develop. Inform. Retrieval*, 1999, pp. 50–57.R.

[8] 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.

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

[10] R. Baeza-Yates and B. Ribeiro-Neto, Modern Information Retrieval.Addison Wesley, 1999.

[11] 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.

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

[13] 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.

[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.

[15] Yang Gao, Yue Xu, and Yuefeng Li,"Pattern based topics for Document Modelling in Information Filtering," IEEE Trasactions on know.june 2015.

[16] Vishal Gupta,Gurpreet Singh Lehal, "A survey of Text Summarization Extractive Techniques,"Journal of Emerging Technologies in Web Intelligence,2010