International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 5 Issue: 4 402 – 406

Survey Paper on Pattern-Enhanced Topic Model for Data Filtering

Chandrakant S. Aher
Computer Science and Engineering , RKDF SOE, Indore
RGPV University Bhopal, India
e-mail: aher_c@rediffmail.com

Dr. Rekha Rathore
Computer Science and Engineering , RKDF SOE, Indore
RGPV University Bhopal, India
e-mail: rekharathore23@gmail.com

***Abstract*—** The machine learning & text mining area topic modeling has been extensively accepted etc. To generate statistical model to classify various topics in a collection of documents topic modelling was proposed. A elementary presumption for those approaches is that the documents in the collection are all about one topic. To represent number of topics in a collection of documents, Latent Dirichlet Allocation (LDA) topic modelling technique was proposed, it is also used in the fields of information retrieval. But its effectiveness in information filtering has not been well evaluated. Patterns are usually thought to be more discriminating than single terms for demonstrating documents. To discovered pattern become crucial when selection of the most representative and discriminating patterns from the huge amount. To overcome limitations and problems, a new information model approach is proposed. Proposed model includes user information important to generate in terms of various topics where each topic is represented by patterns. Patterns are generated from topic models and are organized in terms of their statistical and taxonomic features and the most discriminating and representative patterns are proposed to estimate the document relevant to the user's information needs in order to filter out irrelevant documents. To access the propose model TREC data collection and Reuters Corpus vol. 1 are used for performance

Keywords- *Topic model, information filtering, pattern mining, relevance ranking, user interest model*

*****

I. INTRODUCTION

All data mining and text mining techniques assume that the user's interest is only related to a single topic. But in reality, this is not necessarily the case. For instance, when a user asks for information about a product, e.g. "CAR", the user does not typically mean to find documents which frequently mention the word "CAR". The user probably wants to find documents that Contain information about different aspects of the product, such as location, price, and servicing. This means that a user's interest usually involves multiple aspects relating to multiple topics. The most inspiring contribution of topic modeling 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. The topic-based representation generated by using topic modeling can conquer the problem of semantic confusion compared with the traditional text mining techniques. Topic modeling needs improved modeling users interests in terms of topics' interpretations.

II. LITERATURE SURVEY

A. Survey Of Verious Existing Technologies

This paper deals with the discussion of the numerous papers developed at various institutes which give us idea about this topic.

i. *By Yang Gao, YueXuYuefeng Li, 2013," Pattern-based Topic Models for Information Filtering":*
This paper presents an innovative model PBTM for information filtering including user interest modelling and document relevance ranking.

ii. *By N. Zhong, Y. Li, and S.T. Wu,2012, "Effective pattern discovery for text mining":*
The relevance of a document can be modelled by a pattern-based model.

iii. *By H. D. Kim, D. H. Park, Y. Lu, and C. Zhai, 2012, "Enriching text representation with frequent pattern mining for probabilistic topic modelling":*
Frequent patterns are pre-generated from the original documents and then inserted into the original documents as part of the input to a topic modelling model .The resulting topic representations contain both individual words and pre-generated patterns.

B. Survey Of Various Information Filtering Models

Three technical categories of model includes term-based methods [1][2], pattern mining methods[3][4] and topic modeling methods. For each class, a few strategies were chosen as the standard models. For the topic modeling category, three topic modeling methods are chosen as baseline models, PLSA [5] (Probabilistic Latent Semantic Analysis) word and LDA [6],[7],[8] (Latent Dirichlet Allocation) word ,

PBTM (Pattern-based Topic Model). For the pattern mining category, the baseline models incorporate frequent closed patterns (FCP), frequent sequential closed patterns (SCP) and phrases (n-Gram)[9]. The third category includes the classical term-based methods SVM [10] (Support Vector Machine). An important distinguish between the topic modeling technique and different techniques is that, the topic modeling methods consider numerous topics in each document collection and use patterns (e.g. PBTM and MPBTM) or words (e.g. LDA word) to represent the topics, though the pattern mining and term-based methods accept that the documents inside one accumulation are around one topic and utilize patterns or terms/words to speak to documents directly.

i. Topic Modelling

The study of topic modelling started from the need to compress large data into more
useful and manageable knowledge. Firstly, Latent semantic analysis (LSA) [Deerwesteret al., 1990] uses a singular value decomposition of the matrix of a collection, forming a reduced linear subspace that captures the most significant features of the collection. Then, another remarkable step is Probabilistic LSA (PLSA) model [Hofmann, 1999] which is a generative data model that can provide a solid statistical foundation. In the statistical mixture model, each word in a document as a sample from a mixture model, where the mixture components are multinomial random variables that can be viewed as representations of topics. Thus each independent document is represented by a list of mixing proportions of latent topics, and each of topic is represented by mixing words of Multinomial random variable.

ii. Latent dirichlet Allocation

Latent dirichlet Allocation in [6], the development of successful application there was various latent model element for discrete data. Documents are modelled via a hidden Dirichlet random variable that specifies a probability distribution on a latent low measurement topic space. The distribution over words of an unseen document was a continuous mixture over document area and a discrete mixture over the document area and a discrete mixture over all possible subjects. The generative nature of LDA makes it was easy to utilize as a module in more complex architectures and expand it in different directions. If the categorization variable of LDA is treated as a latent variable they obtain a mixture of LDA models, a useful model for situation in the document bunch not only according to their topic overlap, but along the proportion as well. In 2003, the development of successful application there was various latent model elements for discrete data. In this method they report a new model for gathered of discrete data that provided full generative probabilistic semantics for documents. Documents are modelled via a hidden Dirichlet random variable that specifies a probability distribution on a latent low measurement topic space. In [6],utilize as a module in more complex architectures and expand it in different directions. If the categorization variable of LDA is treated as a latent variable they obtain a mixture of LDA models, a useful model for situation in the document bunch not only according to their topic overlap, but along the proportion as well.

iii. Topical n-Gram (TNG)

The TNG model automatically and simultaneously discovers topics and extracts topically relevant phrases. It has been seamlessly integrated into language modeling based IR tasks [Wang et al., 2007]. Compared with word representation, phrases are more discriminative and carry more concrete semantics. Since phrases are less ambiguous than words, they have been widely explored as text representation for text retrieval, but little research shows significant effectiveness improvements. The likely reasons for the discouraging performance include:

1) Low occurrences of phrases in relevant documents;

2) Lack of flexible number of words for a set of discovered phrases, which restricts the Semantic expression.

The topic representation using word distribution and the document representation using topic distribution are the most important contributions provided by LDA. The topic representation indicates which words are important to which topic and the document representation indicates which topics are important for a particular document. Given a collection of documents, LDA can learn topics and decompose the documents according to the topics. Furthermore, for a new-coming document, vibration inference can be utilized to situate its content in terms of the trained topics. However, single word-based topic representations contain ambiguous semantics. Thus, the TNG improves the LDA model by expanding word-based topic representation to phrase-based, which enhances the explicit semantics of topics. However, TNG suffers from the low occurrence problem and fails to significantly improve LDA.

iv. *Topic Model Labeling***:**

Word-based multinomial distribution is used to represent topics based on the statistical topic model, but it works less well on explicitly interpreting the semantics of the topics. Normally words with high probability of a topic tend to suggest the meaning of the topic, but single words have the problems of polysemy and synonymy. Thus, people tend to label topics with semantic phrases. The general processes are normally conducted in two steps [Zhai, 2008].

First, a set of candidate phrases are generated, either by parsing the text collection or using statistical measures such as mutual information. Second, these candidate phrases are ranked based on a probabilistic measure, which indicates how well a phrase can characterize a topic model. Finally, a few

top-ranked phrases would be chosen as labels for a topic model. The selected labels can be diversified though eliminating redundancy. Two popular methods are normally used. In the first, phrases are simply ranked based on the likelihood of the phrase given in the topic model. Intuitively this would give meaningful phrases with high probabilities according to the word distribution of the topic model to be labeled. In the second method, phrases are ranked based on the expectation of the mutual information between a word and the phrase taken under the word distribution of the topic model. This second method is shown to be better than the first because it would favors a phrase that has an overall similarity to the high probability words of the topic model. Furthermore, a topic can also be labeled with respect to an arbitrary reference/context collection to enable an interpretation of the topic in different contexts. But one drawback of the existing topic models is that the labeled representations are heavily restricted to candidate resources. If the candidate resources cannot cover the meaning of topics, the topic will be unavoidably mislabeled. There is other research on topic semantics interpretation. For example, [Magatti et al., 2009] presented a method to choose the most agreed labels to represent topics according to the similarity measures between given topics and known hierarchies, and specific labeling rules. [Chang et al., 2009] presented an experiment with human subjects and demonstrated that real world task performance by topic modelling is the most convincing method for evaluation.

v. *Term-based models:*

The popular term-based models include tf*idf, Okapi BM25 and various weighting schemes for the bag of words representation [Li and Liu, 2003, Robertson et al.,2004]. Term-based models have an unavoidable limitation on expressing semantics and problems of polysemy (the presence of multiple meanings for one word) and synonymy (multiple words with the same meaning). Therefore, people turn to extracting more semantic features (such as phrases, n-grams, or the phrases from the ontology-based user profiles) to represent a document. Although phrases are less ambiguous than single words, there is little research [Turney, 2000] showing effective improvement on text-based applications. The likely reason is that, although phrases have superior semantic qualities, they have inferior statistical qualities which means the frequency of occurrence in a document of phrases is usually lower than individual words. Hence, the feature vectors of documents are extracted with pattern-based representations, and corresponding approaches ([Liet al., 2014, Wu et al., 2006, Zhong et al., 2012]) were proposed to increase the effectiveness in IF area. The relevance of documents can also be determined by similarities between user's interests and documents , which are represented by vectors of features. The commonly used similarity measures are cosine similarity and Kullback -Leibler distance.

vi. *Collaborative topic modeling***:**

"Collaborative topic modeling for recommending scientific articles," proposed Probabilistic topic modelling can also extract long term user interests by analyzing content and representing itin 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 [30]. These topic model based applications are all related to long-term user needs extraction and related to the task of this paper. But, there is a lack of explicit discrimination in most of the language model based approaches and probabilistic topic models. This weakness indicates that there are still some gaps between the current models and what we need to accurately model the relevance of a document.

vii. *Frequent pattern mining***:**

"Frequent pattern mining: Current status and future directions," Data Min. Knowle.Discov., vol. 15, no. 1, pp. 55–86, 2007. Wang et al. [27] proposed the TFP algorithm to extract the top-k most representative closed patterns by pattern length that no less than minimal 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.

III. SUMMERY OF IF MODELS

Table 1: Summery of various Information Filtering Models

Model	Advantages	Disadvantages	Representative Approaches
1.Topic Model	1) A model can automatically categorize documents in a collection by a no. of topics.	1) The topic distribution and representation is inefficient due to its limited no. of dimensions. 2) The topic representation is limited to distinctively represent documents which have different semantic contents.	PLSA,LDA ,PBTM.
2.Term Based Model	1) Efficient computational performance.	1) This models are suffered from the problems of polysemy and synonymy.	SVM
3.Patternbased model	1) The model is used to represent semantic contents of the user documents more accurately.	1) The no. of patterns in some of the topics can be huge. 2) Many times the patterns are not discriminative enough to represent specific topic.	FCP,SCP, n-Gram Model.

IV. PROPOSED SYSTEM METHODOLOGY

In proposed system user's interest with multiple topics are considered. The proposed model Maximum matched Pattern-based Topic Model consists of topic distributions describing topic preferences of each document or the document collection and pattern-based topic representations representing the semantic meaning of each topic. Here proposed that a structured pattern-based topic representation in which patterns are organized into groups, called equivalence classes, based on their taxonomic and statistical features. With this structured representation, the most representative patterns can be identified which will benefit the filtering of relevant documents. In this system a new ranking method to determine the relevance of new documents based on the proposed model and, especially the structured pattern-based topic representations. The Maximum matched patterns, which are the largest patterns in each equivalence class that exist in the incoming documents, are used to calculate the relevance of the incoming documents to the user's interest.

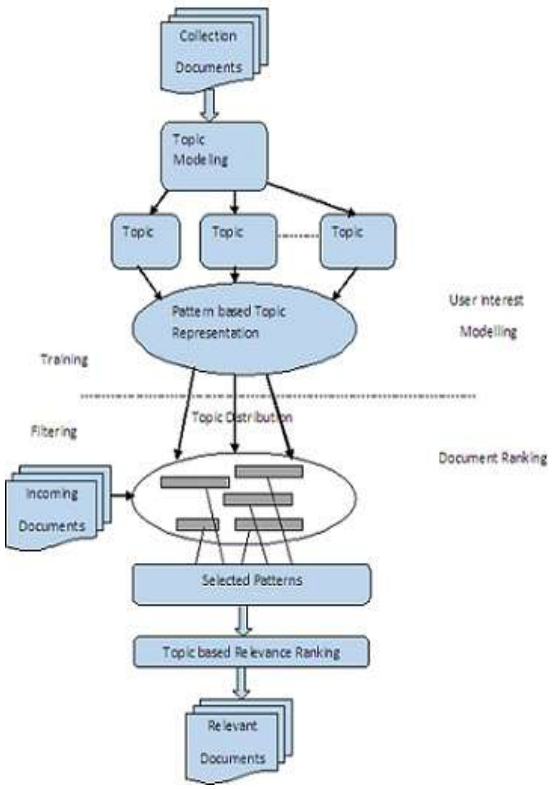

Figure 1: Proposed System Architecture (MPBTM System).

A Maximum matched Pattern-based Topic Model (MPBTM) generates pattern enhanced topic representations to model user's interests across multiple topics. Model selects maximum matched patterns, instead of using all discovered patterns, for estimating the relevance of incoming documents. This model automatically generates discriminative and semantic rich representations for modeling topics and documents by combining statistical topic modeling techniques and data mining techniques using LDA method.

i. THE SYSTEM ARCHITECTURE IS DIVIDED INTO FOLLOWING PHASES:-

Phase 1: Training(User Interest modeling):

1. Collection of documents.
2. Topics modeling.
3. Pattern Based Topic Representation.

Phase 2: Filtering and Topic Distributions:

Phase 3: Document Ranking:

1. Selected Patterns Classification.
2. Topic Based Document Relevance Ranking.
3. Relevant Documents

ii. LATENT DIRICHLET ALLOCATION (LDA):

Topic modeling algorithms are used to discover setoff hidden topics from collections of documents; whereas topic is represented as a distribution over words. Topic models provide an interpretable low-dimensional representation of documents (i.e. with a limited and manageable number of topics). Latent Dirichlet Allocation (LDA) [11] is a typical statistical topic modeling technique and the most common topic modeling tool currently in use. It can discover the hidden topics in collections of documents using the words that appear in the documents. Let D = fd1; d2,---,dMg be a collection of documents. The total number of documents in the collection is M. The idea behind LDA is that each document is considered to contain multiple topics and each topic can be defined as a distribution over a fixed vocabulary of words that appear in the documents.

iii. ALGORITHMS

The proposed IF model can be formally described in two algorithms: User Profiling (i.e. generating user interest models) Algorithm and Document Filtering (i.e.relevance ranking of incoming documents) Algorithm. The former generates pattern-based topic representations
to represent the user's information needs. The latter ranks the incoming documents based on the relevance of the documents to the user's needs.

A. Algorithm 1:User Profiling:-

Input: a collection of positive training documents D;
minimum support σ_j as threshold for topic Z_j ;number of topics V

Output: $U_E = \{E(Z_1),-----,E(Z_V)\}$

1: Generate topic representation ϕ and word-topic assignment $Z_{d,\,i}$ by applying LDA to D
2: $U_E := \phi$
3: **for** each topic $Z_j \in [Z_1, Z_V]$ **do**
4: Construct transactional dataset Γ_j based on ϕ and

$Z_{d,i}$
5: Construct user interest model $\mathbf{X}_{Zj}$ for topic Z_j using a pattern mining technique so that for each pattern X in $\mathbf{X}_{Zj}$, supp(X) $>\sigma_j$
6: Construct equivalence class $E(Z_j)$ from $\mathbf{X}_{Zj}$
7: $U_E := U_E U \{E(Z_j)\}$
8: **end for**

B. Algorithm 2: Document Filtering
Input: user interest model $U_E = \{E(Z_1),-----, E(Z_V)\}$, a list of incoming document D_{in}
Output: $rank_E(d)$, $d \in D_{in}$
1: 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^d_{jk} which exists in d
6: update $rank_E(d)$ using Equation 3:
7: $rank(d) := rank(d) + |MC^d_{jk}|^{0:5 \times} f_{jk} \times V_{D,j}$
8: **end for**
9: **end for**
10: **end for**

V. CONCLUSION

This paper presents a new unique MPBTM Architecture for pattern enhanced topic model for information filtering with user interest modeling and document relevance ranking. The proposed MPBTM system produces the pattern enhanced topic representations to model user's interests across multiple topics. In the filtering stage of MPBTM Architecture, instead of using all discovered patterns, the MPBTM system selects maximum matched patterns for estimating the relevance of incoming documents. The proposed approach incorporates the semantic structure from topic modeling and the specificity as well as the statistical significance from the most representative patterns. In order to perform the task of information filtering the proposed system has been designed by using the TREC and RCV1 collections systems. In comparison with the state-of-the-art system, the proposed system shows excellent results on document modeling with relevance ranking.

REFERENCES

[1] Yang Gao,YueXu and Yuefeng Li," Pattern-based Topics for Document Modeling in Information Filtering", This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.1109/TKDE.2014. 2384497, IEEE Transactions on Knowledge and Data Engineering.

[2] 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.

[3] 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.

[4] H. Cheng, X. Yan, J. Han, and C.-W.Hsu, "Discriminative frequent pattern analysis for effective classification," in IEEE 23rd International Conference on Data Engineering, ICDE'2007.IEEE, 2007, pp.716–725.

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

[6] 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.

[7] M. J. Zaki and C.-J. Hsiao, "CHARM: An efficient algorithm forcloseditemset mining." in SDM, vol. 2, 2002, pp. 457–473.

[8] Y. Xu, Y. Li, and G. Shaw, "Reliable representations for association rules," Data & Knowledge Engineering, vol. 70, no. 6, pp. 555–575,2011.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] Y. Gao, Y. Xu, Y. Li, and B. Liu, "A two-stage approach for generating topic models," in Advances in Knowledge Discovery and Data Mining, PADKDD'13. Springer, 2013, pp. 221–232.

[14] 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.

[15] J. Mostafa, S. Mukhopadhyay, M. Palakal, and W. Lam, "A multilevel approach to intelligent information filtering: model, system, and evaluation," ACM Transactions on Information Systems (TOIS), vol. 15, no. 4, pp. 368–399, 1997.