Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 6.017

IJCSMC, Vol. 6, Issue. 6, June 2017, pg.103 – 111

User Document Recommendation Using Pattern Modeling

Mr. Chandrakant S. Aher[1], Prof. Rasana Sharma[2]
[1]Department of Computer Science Engineering, RKDF School of Engineering Indore, RGPV University Bhopal, MP, India
[2]Department of Computer Engineering, Central India Institute of Technology Indore, RGPV University Bhopal, MP India
1. aher_c@rediffmail.com; 2. rasna.sharma4@gmail.com

Abstract:- Topic modeling has been broadly acknowledged in the zones of machine learning and data mining, and so on. It was proposed to create statistical models to classify various topics in a gathering of documents. A crucial supposition for these methodologies is that the documents in the gathering are about one topic. Topic modeling, for example, Latent Dirichlet Allocation i.e. LDA, was proposed using K-NN classification methods to create measurable models to speak to various topics in a gathering of documents, and this has been extensively utilized as a part of the fields of information retrieval. In any case, its viability in information retrieval has not been well assessed. Patterns are more discriminative than single terms for depicting documents. Choice of the most discriminative and illustrative patterns from the huge amount of found pattern becomes critical. To manage the above said restrictions and issues, a novel information filtering model Maximum matched Pattern-based Topic Model with Dimensionality Reduction i.e. MPBTM-DR is proposed. We used K-means clustering strategies in proposed system model which incorporates user information needs that are created as far as different topics, where every topic is represented by pattern. By using K-NN classification method the patterns are created from topic models and are sorted out similarly as their statistical and taxonomic features and the most illustrative and discriminative patterns are proposed to document pertinence to the user's information needs with a specific end goal to sift through unessential documents. Experiments are conducted to discover effectiveness of "Maximum matched Pattern-based Topic Model with Dimensionality Reduction using K-means method". The results demonstrate that MPBTM-DR model fundamentally outperforms term-based model.

Keywords - Topic modeling, Pattern mining, MPBTM, MPBTM-DR, Document relevance, Information filtering, Information retrieval, Latent Dirichlet Allocation, K-NN.

I. INTRODUCTION

All data mining and text mining techniques expect that the user's interest is only related to a single topic. All but in reality, this is not necessarily the case. For instance, when a user asks for information about a product, e.g. Bmw the user does not typically mean to find documents which frequently mention the word bmw. The user likely needs 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 [1]. The most inspiring contribution of topic modeling is that it automatically classifies documents in a collection by a number of topics and represents each document with various topics and their corresponding distribution [1]. 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. Information filtering models were created utilizing a term-based methodology [2]. The benefit of the term-based methodology is its effective computational execution. Yet, term-based record representation experiences the issues of polysemy also, synonymy [1]. To conquer the impediments of term-based approaches, pattern mining based methods have been used to use patterns to speak to user's advantage and have accomplished a few enhancements in viability since pattern convey more semantic significance than terms. Moreover, a few data mining strategies have been created to enhance the quality of pattern i.e. Maximal pattern and close pattern for expelling the duplicate and noisy patterns [1]. In proposed framework a promising approaches to definitively speak to topics by patterns instead of single words through consolidating topic models with pattern mining systems. In particular, the patterns are produced from the words in the word-based topic representations of a customary topic model, for example, the lda model [9]. This guarantees that the patterns can well speak to the topic since these topic are included the words which are separated by lda taking into account sample occurrence and co occurrence of the words in the documents [9].

II. LITERATURE SURVEY

This paper deals with the discussion of the numerous papers developed at various institutes which give us idea about this topic. **By Yang Gao, Yue Xu Yuefeng Li, 2013," Pattern-based Topic Models for Information Filtering", [1]**: This paper presents an innovative model PBTM for information filtering including user interest modelling and document relevance ranking. **By N. Zhong, Y. Li, and S.T. Wu, 2012, "Effective pattern discovery for text mining", [4]:** This paper focused on relevance of a document can be modelled by a pattern-based model. **By H. D. Kim, D. H. Park, Y. Lu, and C. Zhai, 2012[6], "Enriching text representation with frequent pattern mining for probabilistic topic modelling":** This paper focused on 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. **By 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[2]:** This paper presents innovative ideas various methods of term based text mining i.e. frequent term based text mining. **By 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[3]:**This paper focused on frequent term based mining as well as frequent pattern based text mining with counting interface.

III. ANALYSIS OF EXISTING TECHNOLOGIES

There are three technical categories of model include 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 [1], [7], [9] (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 or words to speak to documents directly.

The main difference 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 [1], [2], though the pattern mining and term-based methods accept that the documents inside one accumulation are around one topic and utilize patterns or terms or words to speak to documents directly.

- **Topic Modeling:** A Main Advantage of topic model is that the model can automatically categorize documents in a collection by a number of topics. Topic Model has some drawbacks like, the topic distribution and representation is inefficient due to its limited no. of dimensions [1], [3]. The topic representation is limited to distinctively represent documents which have different semantic contents [1]. Examples of Topic models are PLSA, LDA and PBTM.
- **Term-based models:** There is main advantage of term based model is that its gives Efficient computational performance. These models are suffered from the problems of polysemy and synonymy [1], [9]. Example is SVM.
- **Pattern Based Model:** The main advantage if pattern based model is that it 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. Many times the patterns are not discriminative enough to represent specific topic [1], [6]. Examples are FCP, SCP and n-Gram Model.

- **K-NN(K-Nearest Neighbor) Classification or Regression Model:** It is supervised classification or regression method that widely used for document classification. It is distance based learning method that used in machine learning techniques. It is supervised because knowledge base is present that supervising to the system [5].

IV. PROPOSED SYSTEM METHODOLOGY

In proposed system user's interest with multiple topics are considered. The proposed framework Maximum matched Pattern-based Topic Model with Dimensionality Reduction consists of topic distributions describing topic inclinations of all document or the document collection and pattern-based topic representations representing the semantic meaning of each topic [1], [9]. System proposed that a structured pattern-based topic representation in which patterns are composed into groups, called equivalence classes, based on their taxonomic and statistical features. With this organized representation, the most illustrative patterns can be identified which will benefit the filtering of relevant documents by using K-nearest neighbor classification method [1], [4]. In this system a new ranking method to determine the relevance of new documents based on the proposed framework and, especially, the structured pattern-based topic representations [1]. The maximum matched patterns, which are the longest patterns in each equivalence class that exist in the incoming documents, are used to calculate the importance of the approaching documents to the user's interest.

A. Problem Statement:

A Maximum matched Pattern-based Topic Model with Dimensionality Reduction (MPBTM-DR) generates pattern enhanced topic representations to model user's interests across multiple topics. Model selects maximum matched patterns, instead of using each discovered patterns, for evaluating the relevance of incoming documents. This model automatically creates semantic rich and discriminative representations for modeling topics and documents by combining statistical topic modeling techniques and data mining techniques using K-Nearest Neighbor (K-NN) classification clustering algorithm.

V. PROPOSED SYSTEM ARCHITECTURE (USING LDA METHOD)

A Maximum matched Pattern-based Topic Model (MPBTM) [1], [2] 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 and K-NN classification method [9].

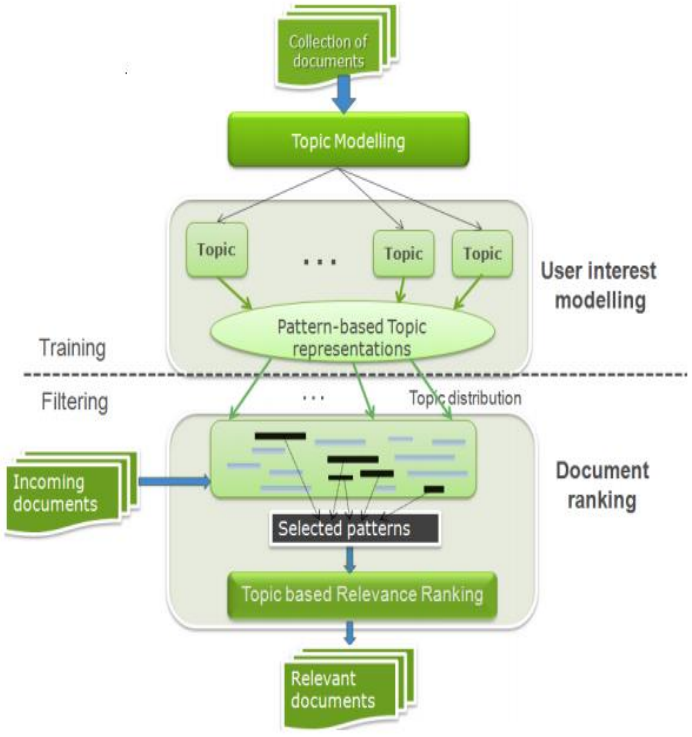

Figure 1: Proposed System Architecture (MPBTM System).

A. The system architecture is divided into following phases:

Phase 1: Training (User Interest modeling):

- Collection of documents.
- Topics modeling.
- Pattern Based Topic Representation.

Phase 2: Filtering and Topic Distributions:

- Relevant Documents are filtered.

Phase 3: Document Ranking:

- Selected Patterns Classification and clustering using K-NN Algorithm:- The relevant documents easily collected by using K-NN classification method.
- Topic Based Document Relevance Ranking.
- Relevant Documents

B. Pattern Enhanced LDA (Latent Dirichlet Allocation):

A pattern is usually defined as a set of related terms or words. Patterns carry more semantic meaning and are more understandable than individual words. The idea of the pattern-based representations starts from the knowledge of frequent pattern mining [9]. It plays an essential role in many data mining tasks directed toward finding interesting patterns in datasets. Pattern-based representations are more meaningful and more accurately represent topics than word-based representations. Moreover, pattern-based representations contain structural information which can reveal the relationship between words [9], [8]. In order to search semantically meaningful patterns to represent topics and documents, two steps are proposed: firstly, construct a new transactional dataset from the LDA results of the document collection D secondly, generate pattern-based representations from the transactional dataset to speak user needs of the collection D. Topic modeling algorithms are used to discover a set of hidden topics from collections of documents, where a 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) [9], [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 = {d1, d2, ---, dn} 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.

C. K-Nearest Neighbor(K-NN)Classification Algorithms:

K-NN classification algorithm is widely used as supervised learning algorithms for making a classification of relevant information's. The k-NN classification algorithms are tries to partition and classify a set of points in set (clusters) such that the point in each set tends to be near each other. From fig. 1 proposed architecture shows that after completion of phase 1 and 2, there is a third phase i. e. document ranking in which the relevant documents easily collected and classified by using K-NN classification method.

VI. ALGORITHMS

The proposed IF model can be formally described in three algorithms:

- K-NN classification Algorithm(i.e. while ranking relevant document are easily collected and classified).
- User Profiling Algorithm (i.e. generating user interest models).
- Document filtering Algorithm (i.e. relevance ranking of incoming documents).

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

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 € $[Z_1, Z_V]$ **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: $U_E := U_E \cup \{E(Z_j)\}$
8: **end for**

B. Algorithm 2: Document Filtering

Input: user interest model $U_E = \{E(Z_1), \text{-----}, E(Z_V)\}$,
A list of incoming document D_{in}
Output: rank $_E(d)$, d € D_{in}

1: rank (d) := 0
2: **for** each d € D_{in} **do**
3: **for** each topic Z_j € $[Z_1, Z_V]$ **do**
4: **for** each equivalence class EC_{jk} € $E(Z_j)$ **do**
5: Scan $ECk_{,j}$ 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.

C. Algorithm 3: K-NN Classification:-

The K-NN is a non parametric algorithm that used in pattern recognition phase of machine learning used for the purpose of regression and classification of required information patterns. Generally k-nn used in two cases i. e. regression and second is classification. In both cases i/p is consist of the k closest training examples in the feature space. And the o/p is depends on whether k-nn is used for which purpose i. e. for regression or classification. In k-nn classification o/p is a class membership an object is classified by a majority vote of its neighbors with the object being assigned to the class most common among its k nearest neighbors where 'k' is a positive integer value that typically small integer. If k=1, then the object is simply assigned to the class of that single nearest neighbor. And in case of k-nn classification and regression phase the k-nn algorithm is used to calculating continuous variables. In case of k-nn regression the output is the property value for the object and this value is the average of the values f its k-nearest neighbors.

Input: -k (k is the no. of classes in the data),

-Training sets: {xi, where **i=1,2,..,n}**

1. Initially compute the distances: In the initial step compute the Euclidean or Mahalanobis distances from the query examples to the labeled examples.

2. Label the examples: Give the order to the labeled examples by using increasing order of their distances.

3. Find the 'K' heuristically: Then find a heuristically optimal number 'k' of nearest neighbors, based on root-means-square deviation error (RMSE) this is done by using cross validation method.

4. Calculate an inverse distances: Then finally just calculate an inverse distances that weighted average with the k-nearest multivariate neighbors.

VII. MATHEMATICAL MODEL

S = {D, DR, WTA, TD, FP, EC, F, U, O}

Where

D : Collection documents.
DR : Collection of documents after dimension reduction techniques.
WTA : Word topic assignment.
TD : Transactional dataset
FP : Frequent pattern.
EC : Equivalence class
F : Function.
U : User.
O : output i.e. recommended documents.

Let, $D= \{D_1; D_2; :::; D_n \}$

$DR= \{DR_1; DR_2; :::; DR_n \}$

$WTA= \{W\,TA_1; W\,TA_2; :::; W\,TA_n \}$

$TD= \{T\,D_1; T\,D_2; :::; T\,D_n \}$

$FP= \{FP_1; FP_2; :::; FP_n \}$

$EC= \{EC_1; EC_2; :::; EC_n \}$

The functions used are as below:

1. F1=Data dimensionality reduction.
2. F2=Word topic assignment.
3. F3=Construct transaction dataset.
4. F4=Frequent pattern.
5. F5=Construct equivalence class.
6. F6=Recommendation.
7. F7=Display Output.

The input and output given to the functions are shown in table 1.

TABLE I
INPUT/OUTPUT TO THE FUNCTIONS.

Function	Input	Output
F1	D	DR
F2	DR	WTA
F3	WTA	TD
F4	TD	FP
F5	FP	EC
F6	EC	R
F7	R	O

The dataset $D_1:::: D_n$ contains documents. Documents are then processed with LDA and MPBTM-DR. Finally we get recommended documents as an output.

VIII. RESULT DISCUSSION

When the experiment is performed in existing system the accuracy is not so high as compared to the proposed system. In this the accuracy is calculated using precision and recall, F1 measure and MAP. Table 2: and figure 2: shows the accuracy by calculating performance measure of proposed system using precision and recall, F1 measure and MAP for various types of query documents.

TABLE II
PERFORMANCE MEASURE USING PRECISION AND RECALL

Query Document	Precision	Recall	F Measure	MAP
D1	**0.8**	**0.7**	0.74666667	0.74833148
D2	**0.7**	**0.6**	0.64615385	0.64807407
D3	**0.8**	**0.9**	0.84705882	0.84852814
D4	**0.9**	**0.6**	**0.72**	0.73484692

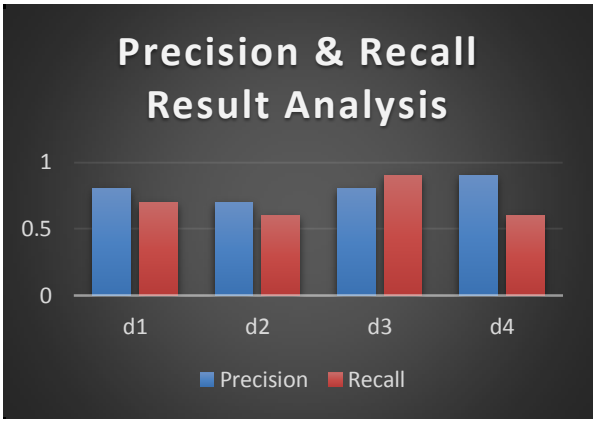

Figure 2: Comparison of Precision and recall.

When we executed the proposed system experiment against existing system to check for the calculation of accuracy with respect to F1 measure, we analyze that new proposed system is better and accurate than old system. Table 3: and figure 3: show the comparison between old system and newly proposed system for performance measuring with respect to F1 measure.

TABLE 3:
PERFORMANCE MEASURE OF F1 MEASURE USING OLD SYSTEM WITH NEW PROPOSED SYSTEM
.

No of Topics	F1 Base paper	F1 Contribution
3	**0.436**	0.746666667
5	**0.457**	0.646153846
10	**0.46**	0.847058824
5	**0.433**	0.72

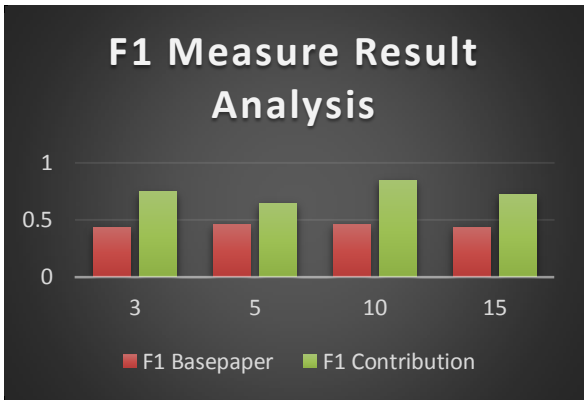

Figure 3: Comparison of Old base paper system with new proposed system with respect to F1 Measure.

IX. CONCLUSION AND FUTURE SCOPE

This paper presents a new unique MPBTM Architecture for pattern enhanced topic model for information filtering with user interest modeling and document relevance ranking using K-NN classification method. By using K-NN classification method 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.

This research solely focuses on recommending documents to the user as per user's interest. For the future work system can also give recommendation to the user to recommend online documents by using user logs.

References

[1] Yang Gao,Yue Xu 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. Bayardo Jr, "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 forclosed itemset 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.

