ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 4, Issue 1, pp: (189-194), Month: January - March 2016, Available at: www.researchpublish.com

Pattern Enhanced Topic Model

Tincy Chinnu Varghese[1], Smitha C Thomas[2]

M.G University, Mount Zion College of Engineering, Pathanamthitta, India

Abstract: **In the field of Information Filtering we have many term-based or pattern –based methods for generating user's needed information from a set of documents .A general thinking is that documents in a particular collection is related to only a single topic. But in real life user's interest is different and documents in a set or collection includes multiple topics. Most commonly used topic modelling method is Latent Dirichlet Allocation (LDA) which generates a structural model to represent multiple topics in a set of documents. Patterns generally are more descriptive and efficiently used in real time applications. So to select most descriptive and efficient patterns from the discovered set of patterns here a Maximum matched Pattern-based Topic Model is introduced. It helps us to get the relevant document according to user needs by filtering out unwanted documents.**

Keywords: **Topic Model, Information Filtering, Pattern mining, relevance ranking, user interest model.**

I. INTRODUCTION

Information filtering (IF) is a way of removing unwanted information from a set of information or documents on the basis of user needs. Many traditional models include Term-based way of filtering documents or information, which faces the monosemy and antonym .To overcome this a Pattern-based model was introduced which gave a more effective result. This model took into consideration the user's interest on the particular search..For improving the quality of patterns used in the search data mining techniques like maximal patterns, closed patterns, master patterns was included. Hence it helped to remove the unwanted and noisy patterns in the search.

Topic modeling is commonly used now a day in the field of machine learning and text mining. It classifies the documents in a set by a given number of topics and presents each document as multiple topics and related distribution. Here introduces a efficient way to represent the topics as patterns than single term words, because patterns can describe a user's needs or interest more than a single word. By using patterns we are able to get the documents filter effectively and efficiently upto to an extent than before .But achieving the best pattern from the huge collection is a crucial thing .For this a new topic model called Maximum matched pattern based model (MPBTM) is introduced.

Contributions of MPBTM for information filtering include:

1) Considers user's interest with multiple topics than one topic based on the thinking that user's information needs are diverse.

2) Integrates data mining techniques with statistical topic modeling techniques to get documents and collection of documents.

3) In the structured pattern-based representation of topics, the patterns are divided into groups called equivalence classes on the basis of structural and similar features Each group have words with same frequency and similar meanings

4) A new approach called Ranking method for determining the relevance of new documents is introduced.

II. EXISTING SYSTEM

In the information filtering technique the aim is to perform mapping from a set of incoming documents to a user relevant document. Let us denote the set of incoming documents as D ,the mapping rank be: D->R, such that rank(d) gives the relevant document .As in the traditional models like term-based have limitations in expressing semantics and also monosemy and antonym. As the number of returned patterns is large selecting reliable patterns is very hard.. Probabilistic topic modeling is another model which helps to extract long term user's needs by verifying content and by representing it

ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 4, Issue 1, pp: (189-194), Month: January - March 2016, Available at: www.researchpublish.com

latent topics which are discovered from user profiles. Lack of explicit discrimination in most of the languages model based approaches and probabilistic topic models .This problems are driven out by labelling topic techniques which considers phrases instead of words for information filtering. In this model n-gram structure is included along with latent topic variables for generating topic relevant phrases. Here it faces the low frequency problem. To overcome all the problems in the topic modeling techniques we introduce the maximum matched pattern based topic model which also have the relevance ranking mechanism which generate an efficient and descriptive patterns from a huge document and also rank the maximum matched patterns

III. PROPOSED SYSTEM

LATENT DIRICHLET ALLOCATION:

Generally topic modeling techniques are used for discovering a set of hidden documents from a group of collection ,where topics are a set of words. Latent Dirichlet Allocation is a commonly used topic modeling now a days for information filtering. It can discover the hidden topics in a set of documents .Let document be taken as D = $\{d_1, d_2, d_3, d_4\}$ .The basis goal of LDA is that every documents consist of a number of topics and every topics consist of a number of words. Mainly the LDA process consists of two parts. The document level and collection level .At the document level, each document say $d_i$is represented by set of topics such as $\boldsymbol{\theta_{di}} = (\boldsymbol{\vartheta_{di,1}}, \boldsymbol{\vartheta_{di,2}}, \ldots \ldots \boldsymbol{\vartheta_{di,V}})$,V is the number of topics in the document. At collection level these each topics is represented by set of words such as ϕ_j, for topic j ,such as $\phi = \{\phi_1, \phi_2, \ldots \ldots \phi_V\}$.

Take an example of a set of documents D =$\{d_1, d_2, d_3, d_4\}$ be a set of some documents with 12 words in each documents. Let us divide the documents in D into three topics such as Z_1, Z_2, Z_3.The table below shows the topic and word distributions in the set of documents.

TABLE I
EXAMPLE RESULTS OF LDA: WORD-TOPIC ASSIGNMENTS

Topic		Z_1		Z_2		Z_3
Document	$\vartheta_{d,1}$	words	$\vartheta_{d,2}$	words	$\vartheta_{d,3}$	words
d_1	0.6	w_1, w_2, w_3, w_2, w_1	0.2	w_1, w_9, w_8	0.2	w_7, w_{10}, w_{10}
d_2	0.2	w_2, w_4, w_4	0.5	w_7, w_8, w_1, w_8, w_8	0.3	w_1, w_{11}, w_{12}
d_3	0.3	w_2, w_1, w_7, w_5	0.3	w_7, w_3, w_3, w_2	0.4	w_4, w_7, w_{10}, w_{11}
d_4	0.3	w_2, w_7, w_6	0.4	w_9, w_8, w_1	0.3	w_1, w_{11}, w_{10}

PATTERN ENHANCED LDA:

Pattern based model contains structural information which gives us an idea of the relation between words. To discover similar meaning patterns to define a topic and document two steps are considered. First ,To construct a new transactional dataset from the document set based on LDA model. Second, To generate pattern-based topics from this transactional dataset to present user's interest.

Construct a transactional dataset:

We construct a set of words from each word-topic assignment Rdi, Zj.Here from the example discussed above let us take the word-based assignment for $\boldsymbol{Z_1}$in document $\boldsymbol{d_1}$. $Rdi, Z_1 = \langle w_1, w_2, w_3, w_2, w_1 \rangle$. Let $\boldsymbol{I_{ij}}$ be a set of words which occur in di, Z_j , Iij ={w|w ∈ Rdi, Zj }, i.e. $\boldsymbol{I_{ij}}$ contains the words which are in document $d_i$and assigned to topic $\boldsymbol{Z_j}$ by LDA. $\boldsymbol{I_{ij}}$, called a Topical document transaction, which is a set of words without any duplicates. From all the word-topic assignments R*di,Zj* to $\boldsymbol{Z_j}$, i=1,. . .M, we can construct a transactional dataset $\boldsymbol{T_j}$. Let D = $\{d_1, \ldots \ldots d_M\}$ be the original document collection, the transactional dataset $\boldsymbol{T_j}$ for topic $\boldsymbol{Z_j}$ is defined as $\boldsymbol{T_j}$ = $\{I_{1j}, I_{2j}, \ldots \ldots I_{Mj}\}$. For the topics in D, we can construct V transactional datasets (T1,T2, . . . TV). An example of a transactional dataset is given below Table II taken from the Table I.

ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 4, Issue 1, pp: (189-194), Month: January - March 2016, Available at: www.researchpublish.com

Generate Pattern Enhanced Representation:

Here the frequent patterns that are generated in each transaction dataset are taken. For a given minimal support threshold σ, an itemset X in $\boldsymbol{T_j}$ is frequent if supp(X) >= σ, where supp(X) is the support of X that is the number of transactions in $\boldsymbol{T_j}$ that contain X. The frequency of the itemset X is defined as $\frac{\boldsymbol{supp(X)}}{|\boldsymbol{T_j}|}$ Topic $\boldsymbol{Z_j}$ can be represented by a set of all frequent patterns, denoted as $X_{Z_i} = \{X_{i1}, X_{i2}, \ldots . X_{im_i}\}$ where m_i is the total number of patterns in $X_{Z_i}$and V is the total number of topics. In the below Table III shows the frequent generated patterns with a minimum support of 2.

TABLE II Transactional Document Transaction (TDT)

transaction	topic document transaction
1	$\{w_1, w_8, w_9\}$
2	$\{w_1, w_7, w_8\}$
3	$\{w_2, w_3, w_7\}$
4	$\{w_1, w_8, w_9\}$

Γ_2

TABLE III Frequent patterns with min supp, σ = 2

Patterns	*supp*
$\{w_1\}, \{w_8\}, \{w_1, w_8\}$	3
$\{w_9\}, \{w_7\}\{w_8, w_9\}, \{w_1, w_9\}, \{w_1, w_8, w_9\}$	2

Pattern Equivalence Class:

For a transactional dataset T, let *X* be a closed itemset and *G(X)* consist of all generators of *X*, then the equivalence class of X in T, 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. The frequency of the patterns in an equivalence class is used to represent the statistical significance of the equivalence class. The three equivalence classes within the patterns for topic Z_2 in Table 3, where f indicates the statistical significance of each class is given in the below Table IV.

TABLE IV Equivalence Classes in Z_2

$\boldsymbol{EC_{21}(f_{21} = 0.75)}$	$\boldsymbol{EC_{22}(f_{22} = 0.5)}$	$\boldsymbol{EC_{23}(f_{23} = 0.5)}$
$\{\boldsymbol{w_1, w_8}\}$	$\{\boldsymbol{w_1, w_8, w_9}\}$	$\{\boldsymbol{w_7}\}$
$\{\boldsymbol{w_1}\}$	$\{\boldsymbol{w_1, w_9}\}$	
$\{, \boldsymbol{w_8}\}$	$\{\boldsymbol{w_8, w_9}\}$	
	$\{\boldsymbol{w_9}\}$	

Maximum ,matched pattern based topic model has mainly two stages. A training stage which is used to generate user information needs from a set of collection and a filtering stage which is used to determine the importance of incoming documents based on user's interest.

User Interest Modelling:

For a document collection *D* and *V* pre-specified latent topics, from the results of LDA to *D*, generate *V* transactional datasets $T_{1,\ldots\ldots}T_V$. Generates user interest model which is given as $U = \{X_{Z_1}, X_{Z_2}, \ldots . X_{Z_V}\}$ where each $X_{Z_i} = \{X_{i_1}, X_{i_2}, \ldots . X_{im_i}\}$ is a set of frequent patterns which are generated from transactional dataset $\theta_D = (\vartheta_{D,1}, \vartheta_{D,2}, \ldots . \vartheta_{D,V})$,$\vartheta_{D,j} = \frac{1}{M}\sum_{i=1}^{M} \theta_{d_{i,j}}$ represents topic distribution of *D*, which is used to represent the user's topic interest distribution. Let

ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 4, Issue 1, pp: (189-194), Month: January - March 2016, Available at: www.researchpublish.com

E(Z_i) denote the set of equivalence classes for topic Z_i, i.e.$E(Z_i) = \{EC_{i1}, \ldots\ldots EC_{in_i}\}$.In the model MPBTM, the equivalence classes E(Z_i) are used to represent user interests which are denoted as $U_E = \{E(Z_1, \ldots..E(Z_v))\}$.

Document Relevance Ranking:

At filtering stage, document relevance is to filter out irrelevant documents based on the user's information needs. Identify maximum patterns in d which match some patterns in the topic-based user interest model. Then estimate the relevance of d based on the user's topic interest distributions and the significance of the matched patterns. For topic significance, let d be a document,Z_j be a topic in the user interest model .let PA_{jk}^dbe a set of matched patterns in document d for topic Z_j.Then the corresponding topic significance of Z_j can be defined as

$$\boldsymbol{sig(Z_d, d) = \sum_{k=1}^{n_j} spec(PA_{jk}^d) \times f_{jk} = \sum_{k=1}^{n_j} a\,|PA_{jk}^d|^m f_{jk}} \quad (1)$$

For the incoming documents d ,we propose to estimate the relevance of d to the user interest based on the topic significance and topic distribution.The equation is as follows

$$\textbf{Rank(d)} = \boldsymbol{\sum_{j=1}^{V} sig(Z_j, d) \times \vartheta_{D,J}} \quad (2)$$

By equating both these equations we get a equation which is denoted as $Rank_E(d)$ is given below

$$\boldsymbol{Rank_E(d) = \sum_{j=1}^{V} \sum_{k=1}^{n_j} |MC_{jk}^d|^{0.5} \times \delta(MC_{jk}^d, d) \times f_{jk} \times \vartheta_{D,j}} \quad (3)$$

The higher the $Rank_E(d)$,more relevant is the document for the users.

Architectural Design:

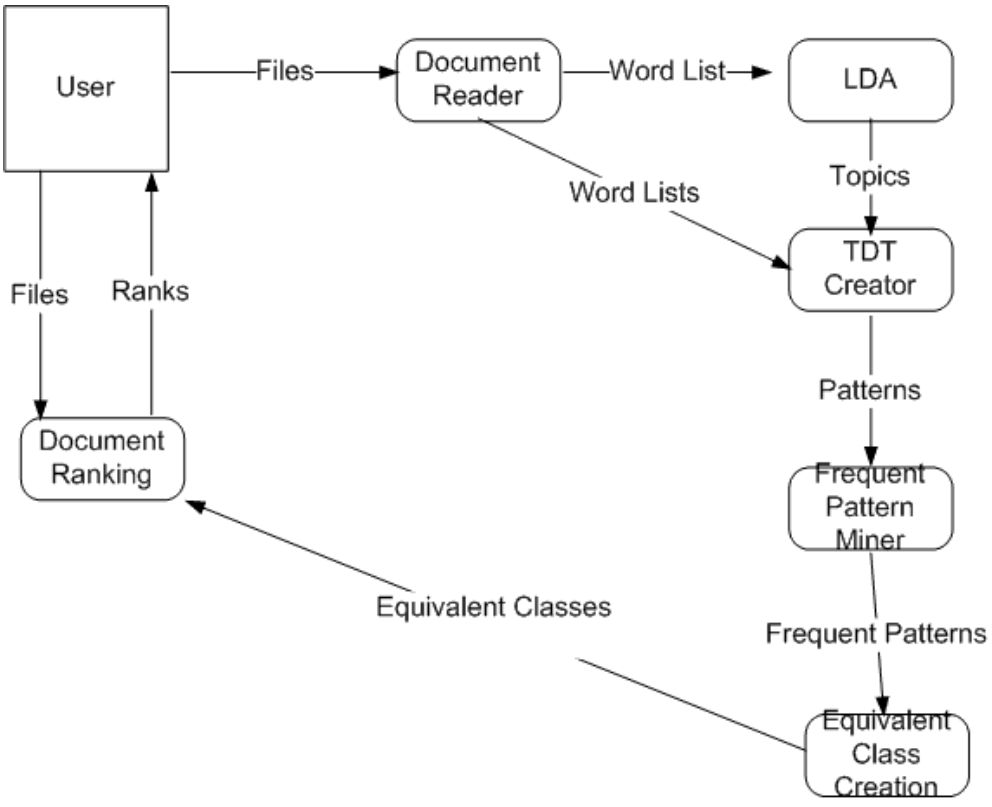

Fig.1: Architectural Design

Algorithms In MPBTM:

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.

ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 4, Issue 1, pp: (189-194), Month: January - March 2016, Available at: www.researchpublish.com

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: $\mathrm{U}_E = \{\mathrm{E}(Z_1), \ldots \ldots \mathrm{E}(Z_V)\}$

1: Generate topic representation $\emptyset$ and word-topic assignment

$Z_{d,i}$ by applying LDA to *D*

2: $\mathrm{U}_E := \emptyset$;

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

4: Construct transactional dataset $\mathbf{T}_j$ based on $\emptyset$ and $Z_{d,i}$

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

6: Construct equivalence class $\mathrm{E}(Z_j)$ from $\mathbf{X}_{Z_j}$

7: $\mathrm{U}_E \coloneqq \mathrm{U}_E \cup \{\mathrm{E}(Z_j)\}$

8: **end for**

Algorithm 2. *Document Filtering*

Input: user interest model $\mathrm{U}_E = \{\mathrm{E}(Z_1), \ldots \ldots \mathrm{E}(Z_V)\}$, a list of

incoming document D_{in}

Output: $rank_E(d), d \in D_{in}$

1: *rank(d)*:=0;

2: **for** each d ∈ 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 $rank_E(d)$ using Equation (3):

7: $rank(d) \coloneqq rank(d) + |MC_{jk}^d|^{0.5} \times f_{jk} \times \vartheta_{D,j}$

8: **end for**

9: **end for**

IV. CONCLUSION

This maximum matched pattern based topic model gives an innovative pattern enriched topic model for filtering information from a set of documents including user's interest model and relevance ranking. The proposed MPBTM model generates pattern enhanced topic representations to model user's interests 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 .It also generates descriptive and semantically rich representations for modeling topics .It is used in the field of content-based extraction of documents ,machine learning etc.

ACKNOWLEDGEMENT

I would like to extend my thankfulness to the reference authors, as well as reviewer of my paper.

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] 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.

[3] 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.

[4] 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.

[5] 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. New York, NY, USA: Springer, 2013, pp 221–232.

[6] H. S. Christopher, D. Manning, and P. Raghavan, An Introduction to Information Retrieval. Cambridge, United Kingdom.: Cambridge Univ. Press, 2009.

[7] C. Zhai, "Statistical language models for information retrieval,"Synthesis Lectures Human Lang. Technol., vol. 1, no. 1, pp. 1–141, 2008