ISSN: 2321-7782 (Online)
Impact Factor: 6.047

Volume 4, Issue 11, November 2016

International Journal of Advance Research in Computer Science and Management Studies

Research Article / Survey Paper / Case Study

Available online at: www.ijarcsms.com

Pattern Based Document Recommendation using Maximum Matched Equivalence Classes

Smita K. Thakare[1]
PG student Department of Computer Engineering,
Late G.N.Sapkal college of Engineering Savitribai Phule,
Pune University – India

Prof. J. V. Shinde[2]
Assistant Professor Department of Computer Engineering
Late G.N.Sapkal college of Engineering, Savitribai Phule
Pune University – India

Abstract: ***Topicmodelling has been widely accepted in the areas of machine learning and text mining, etc. It was proposed to generate statistical models to classify multiple topics in a collection of documents.Existing model I.e. pattern based model, term based model suffered with polysemy and synonymy ,noise generated by this model . All this model only consider that user interested in in only one topic but in situation user are interested in at time many topic in the filled on information filtering Patterns are always thought to be more discriminative than single terms for describing documents. Selection of the most representative and discriminative patterns from the huge amount of discovered patterns becomes essential. To deal with the above mentioned limitations a novel information filtering model is proposed. Proposed model includes user information needs are generated in terms of multiple 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 discriminative and representative patterns are proposed to estimate the document relevance to the user's information needs in order to filter out irrelevant documents. To evaluate the effectiveness of the proposed model TREC data collection and Reuters Corpus Volume 1 are used.***

Keywords: Topic model, information filtering, and pattern based model, term based model, maximum matched pattern.

I. INTRODUCTION

All data mining and text mining techniques assume that the user's interest is only related to a single topic. Actually, this is not necessary in the case. When a user asks for information about a product like "CAR", the user not able to typically mean to find documents which consistently 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 the collection by a no. of topic which represent every document with multiple topics and their corresponding distribution. When we are comparing with pattern-based model and word-based model, pattern-based model generate most meaningful and useful content as per the use requirement. But some time pattern are small in size or large in size and that pattern is not carry the meaning related to the particular topic.so to avoid this

Problem related to pattern we have to find out 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. Hence we proposed the innovate system i.e,A Maximum matched Pattern-based Topic Model which 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. To find out most meaningful pattern we using ranking method and most ranked pattern is most useful

pattern for information filtering and generate user interest document when pattern is achieved that pattern is well structure pattern and used powerfully and excellentlyin this system.

II. Literature Survey

Simple BM25 extension to multipleweighted Field [1] proposed term based approach is efficient computational performance, as well as mature theories for term weighting. But it create problem of polysemy and synonymy. Mining frequent pattern with counting

inference, proposed pattern mining based pattern mining based techniques have been used to develop patterns to represent users interest and have attained some developments in usefulness since patterns transmit more semantic meaning than terms.

Latent Dirichlet allocation,[12] proposedTopic modeling has become one of the most popular probabilistic text modeling techniques and has been quickly accepted by machine learning and text mining communities. It can automatically classify documents in a collection by a number of topics and represents every document with multiple topics and their equivalent distribution .it has ambiguity.

Topical n-grams: Phrase and topic discovery,

with an application to information retrieval, proposed The topical n-Gram (TNG) in is seamlessly integrated into the languagemodeling based IR task, but the improvement this provides is not that significant .

Enriching text representation with frequent pattern mining for probabilistic topicmodeling, proposed frequent patterns are generated from the original documents and then inserted into the original documents as part of the input to a topic modeling model such as LDA. The resulting topic representations contain both individual words and ore generated patterns.

Collaborative topic modeling for recommendingscientific articles, [14] proposed Probabilistic topic modeling can also extract long term user interests by analyzing content andrepresenting 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 users information needs. 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.

Topical n-grams: Phrase and topic discovery,with an application to information retrieval, proposed the topical n-Gram modelproposed in robotically and concurrently determines topics and citations topically relevant slogans. It has been faultlessly combined into

the language modelling based IR task. Compared with word representation, phrases are more discriminative and carry more real semantics. Since phrases are less unclear than words, they have been broadly discovered as text illustration for text retrieval, but few studies

In this area have shown significant progresses in success. Enriching text representation with frequent

Pattern mining for probabilistic topic modeling,[15] proposed in the sense that the topics in the MPBTM model are represented by patterns only.

III. 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.

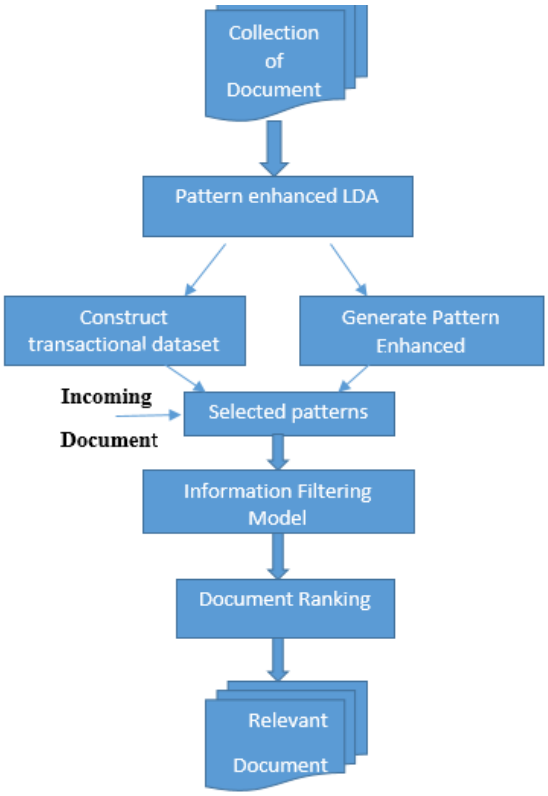

Figure 3.1: Proposed System Architecture (MPBTM System).

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 modelling topics and documents by combining statistical topic modelling techniques and data mining techniques using LDA method.

A. The system architecture is divided into following phases:-

Phase 1: Construct transactional dataset from LDA (User Interest modeling):

1. LDA
2. Word-Topic Assignment .

Phase 2: Information filtering Based on pattern enhanced LDA

1. Frequent pattern
2. Pattern Equivalence class
3. Topic-Based user Interest modelling

Phase 3: Recommendation

1. Matched pattern EC

2. Ranked Document

3. Relevant document

1. LDA

It is the Latent Dirichlet Allocation (LDA) algorithm of Topic modeling algorithms that are used to discover a setof hidden topics from collections of documents, wherea topic is represented as a distribution over words.Topic models provide an interpretable low-dimensionalrepresentation of documents with a limited andmanageable number of topics. Latent Dirichlet Allocation (LDA) is a typical statistical topic modeling technique and the most common topic modeling toolcurrently in use. It can discover the hidden topics incollections of documents using the words that appearing the documents. Let D = {d1, d2,---,d_M} be a collectionof documents. The total number of documents inthe collection is M. The idea behind LDA is that eachdocument is considered to contain multiple topics andeach topic can be defined as a distribution over a fixedvocabulary of words that appear in the documents.

2. Word-Topic Assignment.

Word topic assignment to topic z in the document d. construct a set of word from each word topic assignment instead of using sequence of word.it contain the word which are in document d and assigned to topic z by LDA.

3. Frequent pattern.

Use of frequent pattern generated form the transactional dataset to represent topic it is basic idea which is used in proposed pattern-based model.for a given minimal support threshold σ, an itemset X in the transactional table is frequent if $supp(X) >= \sigma$. Where

Sup(X) is the support of X which is the number of transactions in that contain X.

4. Pattern equivalence class

For a transactional dataset let X be a closed itemset and G'(X) consist of all generators of X,then equivalence class of X in transactional ,denoted as EC(X), is defined as EC(X)=G(X)U{X}.

B. 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 := ϕ

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) > σ_j

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

7: U_E := U_EU {E(Z_j)}

8: **end for**

5. Topic-based user interest modelling

For a collection of document D, the users interests can be represented by the pattern in the topic of Frequent patterns generated from transactional dataset. In the model equivalence classes E(zi) are used to represent user interests which are denoted as U_E={E(Zi)……..,E(Zv)}.

6. Matched pattern EC

Let d be a document ,Z be a topic in the user interest model,EC1,,,,,,,ECn be the the pattern equivalence classes of Z, then a pattern in d is considered a maximum matched pattern to equivalence class. Maximum matched pattern are considered the most significant pattern in d which can represent the topic Z.

7. Ranked document

$$RankE(d) = \sum_{j=1}^{v} \sum_{k=1}^{nj} \left|MC_{jk}^{d}\right| X\ \delta\left(MC_{jk}^{d}, d\right) X\ f_{jk} X \upsilon_{Dj}$$

Where V is the total number of topics ,MC_{jk}^{d} is the maximum matched pattern to equivalence class EC_{jk}, k=1…..n_j and f_{j1}…….,f_{jn} is the corresponding statistical significance of the equivalence classes ,υ_{Dj} is the topic distribution and

δ (X,d)=$\begin{cases}1 \\ 0\end{cases}$if X ε d

Otherwise.

The higher the Rank$E^{(d)}$, the more likely the document is relevant to the user's interest.

C. **Algorithm 2:** Document Filtering

Input: user interest model U_E = {E(Z_1),-----, E(Z_V)}, alist 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 $EC_{k,\ j}$ 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**

IV. Result and Analysis

Table 1: 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

Table 2: Performance Measure of F1 Measure using Old base paper system with new proposed system.

No of Topics	F1 MPBTM	F1 CFECA
3	**0.436**	0.746666667
5	**0.457**	0.646153846
10	**0.46**	0.847058824
15	**0.433**	0.72

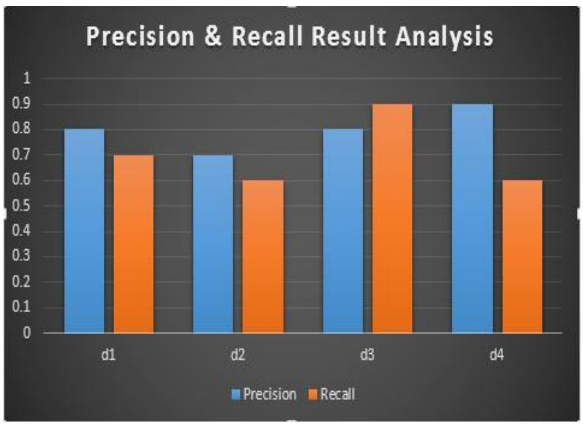

Figure 4.1: Comparison of Precision and recall.

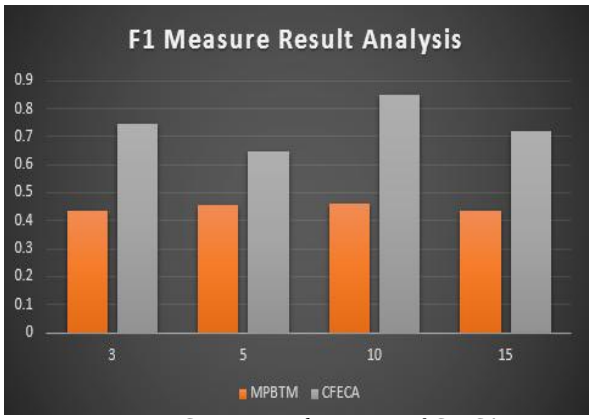

Figure 4.2: Comparison of MPBTM and CFECA

V. Conclusion and Future Scope

This paper presents anew unique MPBTM Architecture for pattern enhancedtopic model for information filtering with user interestmodeling and document relevance ranking. Theproposed MPBTM system produces the pattern enhancedtopic 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 selectsmaximum matched patterns for estimating the relevance of incomingdocuments. The proposed approach incorporates the semantic structure from topic modeling and thespecificity as well as the statistical significance from themost representative patterns. In order to perform the task of information filtering the proposed system hasbeen designed by using the TRECandRCV1 collections systems. In comparison withthe state-of-the-art system, the proposed system showsexcellent results on document modeling with relevance ranking.

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 onKnowledge 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," ACMSIGKDD Explorations Newsletter, vol. 2, no. 2, pp. 66–75, 2000.
4. H. Cheng, X. Yan, J. Han, and C.-W. Hsu, "Discriminative frequentpattern 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. [6].
6. J. Han, H. Cheng, D. Xin, and X. Yan, "Frequent pattern mining:current status and future directions," Data Mining and KnowledgeDiscovery, 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 associationrules," Data & Knowledge Engineering, vol. 70, no. 6, pp. 555–575,2011.
9. X. Wei and W. B. Croft, "LDA-based document models for ad-hocretrieval," in Proceedings of the 29th annual International ACM SIGIRconference on Research and Development in Information Retrieval.ACM, 2006, pp. 178–185.
10. C. Wang and D. M. Blei, "Collaborative topic modeling forrecommending scientific articles," in Proceedings of the 17th ACMSIGKDD International Conference on Knowledge Discovery and DataMining. 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 Proceedingsof the 22nd annual international ACM SIGIR conference onResearch and development in information retrieval. ACM, 1999, pp.50–57.
13. Y. Gao, Y. Xu, Y. Li, and B. Liu, "A two-stage approach forgenerating topic models," in Advances in Knowledge Discovery andData Mining, PADKDD'13. Springer, 2013, pp. 221–232.
14. Y. Gao, Y. Xu, and Y. Li, "Pattern-based topic models for informationfiltering," in Proceedings of International Conference on DataMining Workshop SENTIRE, ICDM'2013. IEEE, 2013.
15. J. Mostafa, S. Mukhopadhyay, M. Palakal, and W. Lam, "Amultilevel approach to intelligent information filtering: model,system, and evaluation," ACM Transactions on Information Systems(TOIS), vol. 15, no. 4, pp. 368–399, 1997.

