Pattern Matching And Clustering Of Documents Using MPBTM And LDA

Nisha K. Lagad, Padmapani P. Tribhuvan.

Department of Computer Science and Engineering, Deogiri Institute of Engineering and Management Studies, Aurangabad, INDIA

***Abstract* - Term and pattern related approaches are used in information filtering. These approaches are used for generating user's information needs from a large amount of documents. A prediction for these a technique is the documents in collection are all about same topic. However, in reality users' interests can be diverse and the documents in the collection often involve multiple topics. Topic modelling, such as Latent Dirichlet Allocation is given to generate statistical models to represent multiple topics in a collection of documents.LDA automatically classifies documents into the number of topics and each subject contains a number of words based on their probability. Next, build a new transactional dataset from the LDA result. The resulting transactional data set is the input of the pattern scanning algorithm. In the field of filtering, incoming documents pass through subject modeling, pattern exploration and, finally, the MPBTM selects the maximum matching patterns, Next, compare the incoming document template with the training document template. From this we can find corresponding maximal models and which are used to estimate the relevance of incoming documents. And in proposed system we are getting relevant documents in terms of clusters. Finally we get the relevant information in terms of document clusters.**

Keywords-Topic model, Information Filtering, Pattern Mining, Relevance Ranking, Document Clustering

I. INTRODUCTION

Modeling user interests is a process to understand the user's information needs based on the most relevant information that can be found and delivered to the user. In order to extract the user's specific interests, traditionally, many term-based approaches are used because of their efficient computing performance, as well as mature theories for term weighting, such as Rocchio, BM25, etc. But futures functionality suffers from problems of polysemy and synonymy. The sentence-based approaches are more discriminative and must be semantically meaningful. However, the performance of using phrases in actual applications is discouraging. To overcome the limitations of term- and expression-based approaches, pattern-based techniques have used models to represent user interest and improve efficiency.

II. TOPIC MODELING INFORMATION FILTERING

Topic modeling [1] is one most popular probabilistic text modeling techniques and It was quickly accepted by computer learning and text extraction communities. In this the most inspiring contribution of subject modeling is that it automatically classifies documents into a collection by a number of subjects and represents each document with several subjects and their corresponding distribution. The thematic representation generated by the use of subject modeling can fill the problem of semantic confusion over traditional text extraction techniques. The representation by simple words with probabilistic distributions breaks the relations between the associated words. Therefore, the modeling of subjects requires the improvement of the interests of modelling users in terms of interpretations of subjects. In this work, a model based model is proposed to improve the semantic interpretations of subjects. This work focuses on how the subject model based on the proposed model can be used in the field of information filtering (IF) for constructing content-based user interest modeling.

Topic Modelling is a probabilistic model for collections of discrete data such as text collections. It can automatically divide documents in a collection by a number of topics and represents every document with number topics and their corresponding distribution. Two representative methods are Probabilistic Latent Semantic Analysis PLSA [12] and LDA [11]. However, there are two problems if we directly apply topic models for information filtering. The first problem is that the topic distribution itself is insufficient to represent documents due to its limited number of dimensions. The second problem is that the word based topic representation is limited to distinctively represent documents which have different semantic content since many words in the topic representation are repeated general words.

Information filtering is used for removing repeated and unwanted information from collection of information and from collection of documents which is based on representations of documents that represent the interests of users. Once the user profiles are collected, in this thesis we focus on modeling the interests of the user with multiple subjects. Using classical subject models, the interests of the user can be represented by a predefined number of subjects, each represented by words and their distribution. In this work, the "relevance" of a document refers to the relevance between the interests of the user and the document. Assume that the interests of the user are well represented with models-based subjects. Given that very often the number of models in some of the subjects can be enormous and that many models are not discriminative enough to represent specific subjects, we will propose methods of classification modeling of relevance for the representation of documents and l Of relevance. Topical models for document modeling should be selected. In this work, to represent subjects instead of using frequent models, we proposed to select the most representative and most recurrent models, called corresponding maximum patterns. A new theme model is Topic-Based modeling based on Maximum Compatible Models (MPBTM), is given for document representation and relevance ranking of document. Models in the MPBTM content models are well structured and so that maximum matching models can be effectively selected and used for representing and classifying documents.

III. SYSTEM ARCHITECTURE

Patterns may represent more specific meanings than unique words. By comparing the word based subject model with the model-based subject models, the model-based model can be used to represent the semantic content of user documents more precisely than the word-based document. However, very often, the number of models in some of the subjects can be enormous and many models are not enough to represent specific subjects.

We use the model based on the improvised maximum patterns (IMPBTM) to select the most representative and recurring models. Maximum matched patterns are referred to as subjects instead of using frequent patterns.

The subject model based on the maximal matching pattern [8] consists of thematic distributions describing the thematic preferences of each document or the collection of documents and the thematic representations based on motifs having the semantic meaning of each subject. Four steps are proposed to generate the subject-based user interest model. First subject modeling algorithm named LDA applying to each document. LDA automatically classifies documents into the number of topics and each subject contains a number of words based on their probability. Next, build a new transactional dataset from the LDA result, which removes duplicate words. The resulting transactional data set is the input of the pattern scanning algorithm. Track frequent mie profiles using an efficient pattern-scanning algorithm. Patterns contain more information than unique words. In the field of filtering, incoming documents pass through subject modeling, pattern exploration and, finally, the MPBTM selects the maximum matching patterns, instead of using all the patterns discovered. Next, compare the incoming document template with the training document template. From this we can find corresponding maximal models and which are used to estimate the relevance of documents. Depending upon output of LDA user gets the most relevant information in terms of document clusters. Figure 1 shows the overall structure of the proposed system.

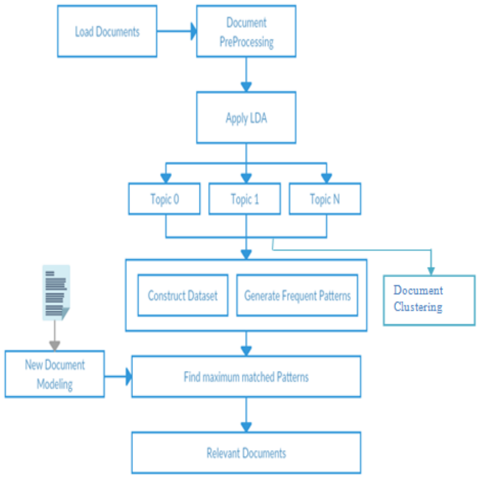

Fig.1 System Architecture

Load Dataset

For this implementation we are using bbc Dataset which we have taken from http://mlg.ucd.ie/datasets/bbc.html this website.

Information Filtering

Information Filtering is used to remove redundant and unwanted information from number of the documents. In this stop-words are get removed from documents**.**

Latent Dirichlet Allocation

LDA is the most commonly used topic modeling algorithm that discovers the hidden topics from collection of documents. Here each discovered topic is represented as distribution over words. LDA discover the hidden topics from the document set by using the word that appears in each document. Let D = {d_1, d_2, ..., d_m} be the collection of documents and the total number of documents in the collection be 'm'. LDA is applied to the whole documents for dividing it into specified number of topics. The main idea behind LDA is under the assumption of each document is considered to contain multiple topics and each topic can be defined as distribution over words.

The LDA model is represented by using two levels, document level and collection level. At document level each document di from the document set is represented by topic distribution θ di = ($v_{di,1}$, $v_{di,2}$, ..., $v_{di,V}$) ,V is the number of topics. At collection level the document set is represented as D. Each document is represented by a probability distribution over words, ϕj for topic j. Overall we have ϕ = {$ϕ_1$, $ϕ_2$, . . . ,$ϕ_v$} for all topics. LDA model also generates the word topic assignment apart from these two levels of representation that is the word occurrence is considered related to the topics.

The topic distribution over the whole document collection D can be calculated from the LDA model, ϕ D = ($v_{D,1}$, $v_{D,2}$,. . . , $v_{D,V}$), where vD, j indicates the importance degree of the topic Zj in the collection D. The most important contribution of LDA model is that the topic representation using word distribution and the document representation using topic representation. The topic representation indicates which words are important to which topic and document representation indicates which topics are important to which document. LDA can learn topics from the collection of documents and decompose the documents according to the topics. Various methods are utilized for new incoming documents to situating the content in terms of trained topics. In this paper we use a pattern based topic model to represent documents and propose an accurate ranking method that determines the relevance of new incoming documents.

The algorithms that deal with subject modeling are mostly used to analyze the words of the base contexts to discover the themes that cross them, how these themes are related to each other and how they evolve over time. Probabilistic latent semantic analysis

(PLSA) [3], is a technique for analyzing data in two modes and co-occurrence. The probabilistic model of latent semantic indexing, which was introduced by Hoffman, was quickly accepted in several text modeling applications. PLSI, called an aspect model, is a latent variable model for general co-occurrence data that associates an unobserved class subject variable with each observation i.e, each occurrence of 'a word. The PLSI model presents a problem because its generative semantics is not well defined. Therefore, there is no natural way to predict a previously invisible document and the number of PLSI parameters increases linearly with the number of training documents, making the model likely to be overvalued. LDA is a probabilistic subject model that considers probability distribution functions to assign words in a document to a particular subject. The underlying instinct behind LDA is, the documents are a mix of multiple subjects. For example, the document named computer, may have subjects such as data structure, algorithms, computational theory, computer network, etc., the documents are a mixture of subjects. These subjects are distributed on a document in equal or unequal proportion. There are mainly two types of variables in LDA as hidden variables and observed variables. The processed variables are usually the words in given documents. While the hidden variables describe the structure of the subject. More precisely, the data come from hidden random variables and these variables form a thematic structure. The process of deducing the hidden structure of the document is performed by calculating the posterior distribution. This distribution is the conditional distribution of the hidden variables in the documents. The word "Dirichlet" in Latent Dirichlet Allocation is a distribution used to draw a distribution by subject of a document, that is, it specifies how subjects are distributed in a particular document. In the generator process, this dirichlet distribution output is used to assign document words to different topics.

The correlated subject model (CTM) [4] is a type of statistical model used in the processing of natural language and machine learning. They are used to find topics that appear in a group of documents. The CTM key is the normal logistic distribution. It is a new subject model that extends from LDA that directly models the correlation between subjects. Use normal logistical distribution to create relationships between subjects. But the CTM requires a lot of calculation and it has many general words in the subjects

Pattern Enhanced LDA

Pattern based representation overcome the limitations of word based representation, which provide an accurate method for represent documents. Moreover in pattern-based representation the structural information is provided by the association among the words. In order to discover semantically meaningful pattern from the document set for representing the topics and documents, two steps are proposed:

(1) Construct a new transactional dataset from the LDA outcomes of the document collection D.

(2) Generate pattern based representations from the transactional dataset to represent user needs.

(3) Obtain Pattern Equivalence Class

1) Construct Transactional Dataset

Let R_{di} and z_j signify the word-topic assignment for topic mentioned in Zj as in the document di. R_{di} Z_j is a sequence of words assigned to topic Zj. For applying LDA the number of topics is specified by the user. The words under each topic occurs in each document is called topical document transaction. Topical document transaction (TDT) is set of words without any duplicates. For all the word-topic assignments $R_{di};Z_j$ to Zj , we can construct a transactional dataset Γj. Let D = {d_1, . . . , d_M} be the original document collection, the transactional dataset Γj for topic Zj is defined as Γj = {I_{1j}; I $_{2j}$; . . . ; I $_{Mj}$}. Where Iij is the set of words which occur in Rdi,Zj. Iij called a topical document transaction. For each of the topics in D, we can construct V transactional datasets (Γ_1, Γ_2 , . . . , Γ_v).

2) Generate Pattern based Representation

In the proposed pattern based method frequent patterns generated from each transactional dataset Γj is used to represent Zj. Patterns is the set of related words. For a given minimal support threshold σ, an itemset X in Γ_j is frequent if and only if supp(X) >= σ, where supp(X) is the support of X which is the number of transactions in Γj that contain X.

Minimal support threshold is specified by the user. The itemset frequency 'X' is defined as the set of all frequent pattern are represented the topic Zj, denoted as Xzi = { X_{i1}, X_{12}, . . . , X_{imi} } , where mi is the total number of patterns in X_{zi} and v is the total number of topics.

3) Pattern Equivalence Class

The number of frequent pattern obtained from the previous stage is considerably large and many of them are not necessarily useful. Several concise patterns have been proposed to represent useful patterns instead of frequent patterns generated from a large dataset such as maximal patterns and closed patterns. For a dataset the number of the concise patterns is significantly smaller than the number of frequent patterns generated.

Let EC1 and EC2 be two different equivalence classes of the same transactional dataset. Then EC1 ∩ EC2 = ϕ which means that the equivalence classes are exclusive of each other. There are two pertaining parts used in the proposed model. In this they have used training part to generate user interest model from the collection of different number of training documents and filtering part determines the relevance of new incoming document.

First subject modeling algorithm named LDA applying to each document. LDA automatically classifies documents into the number of topics and each subject contains a number of words based on their probability. Next, build a new transactional dataset from the LDA result, which removes duplicate words. The resulting transactional data set is the input of the pattern scanning algorithm. Track frequent profiles using an efficient pattern-scanning algorithm. Patterns contain more information than unique words. In the field of filtering, incoming documents pass through subject modeling, pattern exploration and, finally, the MPBTM selects the maximum matching patterns, instead of using all the patterns discovered. Next, compare the incoming documenttemplate with the training document template. From this we can find corresponding maximal models and which are used to estimate the relevance of incoming documents.

Maximum Matched Patterns

Using MPBTM algorithm we are finding maximum matched patterns from testing documents. And depending upon maximum matched patterns we estimate relevant documents.

Document Clustering

The K-means clustering is one of the method of vector quantization and it is given from signal processing, K-means clustering is popular method for cluster analysis in data mining. K-means clustering aims to partition n observations into the k clusters in which each and every observation belngs to the cluster with the nearest mean. This is the results in a partitioning of the data space. On the output of LDA We can apply the k-means algorithm and after applying that user gets the most relevant information in terms of Document Clusters.

IV. PERFORMANCE EVALUATION

Dataset Details

For this implementation we are using bbc Dataset which we have taken from http://mlg.ucd.ie/datasets/bbc.html this website. From this bbc dataset on 5 different domains we are giving experimental results. These domains are such as Politics, Business, Sports, Technical and Entertainment. From each Domain we have taken 50 documents for calculating the results. Results are calculated on the basis of parameters such as Precision, Recall, F-Measure, and Accuracy.

TABLE 1 Result Analysis of Different Datasets

Domain	Precision	Recall	F-Measure	Accuracy
Politics	0.9263	0.8649	0.8946	0.9046
Business	0.9448	0.8969	0.9202	0.9235
Sports	0.9427	0.8933	0.9174	0.9235
Tech	0.9294	0.8701	0.8988	0.9046
Entertainment	0.9306	0.8721	0.9364	0.9046

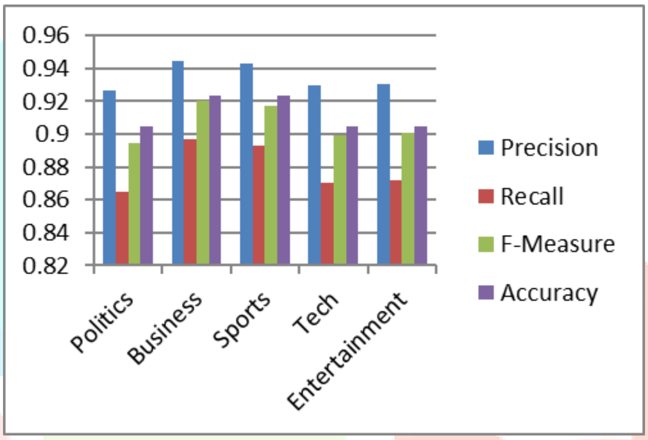

Fig 2: Graphical representation of result

V. CONCLUSION AND FUTURE WORK

This system is used to form Document Clustering. This system gives average results. Average result of precision is 0.9461, Recall is 0.8993, F-Measure is 0.9221, and Accuracy is 0.9282. The model has been evaluated by using bbc dataset for the task of Document Clustering. In comparison with the state-of-the-art models, the proposed model demonstrates excellent strength on document modelling and relevance ranking and document clustering. In Future, we can do Annotation of clustering. Means we can tell which cluster is from which topic.

REFERENCES

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] Y. Cao, J. Xu, T.-Y. Liu, H. Li, Y. Huang, and H.-W. Hon, “Adapting ranking svm to document retrieval,” in Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 2006, pp. 186–193.

[7] S. Robertson, H. Zaragoza, and M. Taylor, “Simple BM25 extension to multiple weighted fields,” in Proceedings of the thirteenth ACM International Conference on Information and Knowledge Management. ACM, 2004, pp. 42–49.

[8] 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.

[9] 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.

[10] S.-T. Wu, Y. Li, and Y. Xu, “Deploying approaches for pattern refinement in text mining,” in 6th International Conference on Data Mining, ICDM’06. IEEE, 2006, pp. 1157–1161.

[11] 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.

[12] 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.

[13] R. J. Bayardo Jr, “Efficiently mining long patterns from databases,” in ACM Sigmod Record, vol. 27, no. 2. ACM, 1998, pp. 85–93.

[14] Tincy Chinnu Varghese,Smitha C Thomas (March 2016): Pattern Enhanced Topic Model. International Journal Of Computer Science And Information Technology Research ISSN2348-120X,vol.4,Issue 1.

[15] Vasudevan, V.Sharmila, Dr.G.Tholkappia Arasu (October 2012): Innovative Pattern Mining for Information Filtering System.ISSN: 2277-3754, vol 2, Issue 4.

[16]PallavyNath.S,AnnieGeorge(November2015):Semantic Pattern-Based Topics Filtering for Document Modelling. International Journal of Innovative Research in Computer and Communication Engineering,vol.3,Issue 11.