A Survey on Documentation Modelling in Information Filtering using Pattern Recognition

Ketaki Gadwale[1] Dr. Emmanuel M[2]
[1]PG Student [2]Professor
[1,2]Department of Information Technology
[1,2]PICT, Pune, India

Abstract— In the field of information filtering many full-fledged term-based or pattern-based approaches have been used by assuming the documents in a collection are all about one topic. However, user's interest can be varied and generally multiple topics are involved in the collection of documents. Topic modeling, such as Latent Dirichlet allocation (LDA), has been widely exploited in the fields of machine learning and information retrieval etc. but its effectiveness in information filtering has not been so well explored. A novel information filtering model, Maximum matched Pattern-based Topic Model (MPBTM), is proposed. Following features makes the proposed model peculiar: (1) In terms of multiple topics the user information need are generated and each topic is represented by patterns; (2)the most discriminate and representative patterns, called Maximum Matched Patterns are generated from topic models and are systematized in terms of their statistical and taxonomic features and proposed to estimate the documents relevance in order to filter out irrelevant documents suited to user to user's information needs.

Key words: Topic modeling, Pattern mining, User interest model, Document relevance, Information filtering, Information retrieval

I. INTRODUCTION

Information Filtering is the process of monitoring large amounts of dynamically generated information and pushing to a user the subset of information likely to be of user interest (based on user information needs). Information filtering system assists users by filtering the data source and deliver relevant information to the users. Recent years have witnessed a dramatic increase in web information. Hence, advanced programs and formulas are required to understand what exactly users need and to deliver the best results based on users' information needs. The classical Information filtering models were evolved using term based[1] approach.

Data mining is the process of inspecting data from exclusive point of view to examine colossal pre-present databases with the intention to produce new information. It is the method of analyzing data from extraordinary perspectives and summarizing it into valuable expertise. In a similar way, textual content mining is the method of extracting data from the massive set of data. It additionally refers back to the process of deriving excessive high-quality information from textual content. Polysemy method has become some of the fashionable probabilistic text modeling strategies and has been quickly accepted via machine learning and textual content mining communities. It could possibly routinely classify files in a set through a number of contents and represents every report with more than one meanings and their corresponding distribution.

Statistical topic models such as latent Dirichlet allocation (LDA) have been shown to be strong tools in topic extraction and analysis. These models can seize phrase correlations in a set of textual files with a low-dimensional set of multinomial distributions. Up to date work on this subject has investigated richer structures to also describe inter-topic correlations and resulted in discovering big numbers of more correct, fine-grained issues. The topics learned by LDA capture correlations amongst words, but LDA does now not explicitly represents correlations among themes [2].

Topic models are a set of algorithms that uncover the hidden thematic constitution in record collections. These algorithms aid to develop new approaches for searching, browsing and summarizing gigantic archives of texts.

As our collective knowledge is still digitized and stored—within the form of information, blogs, internet sites, scientific articles, books, photos, sound, video, and social networks—it becomes extra intricate to seek out and detect what we are looking for. We want new computational tools to aid prepare, search, and have an understanding of these huge amounts of knowledge. Right now, we work with on-line information making use of two main tools—search and hyperlinks. We type keywords right into a search engine and find a set of documents regarding them. There are mainly two issues in immediately making use of topic modeling. First issue is restricted quantity of topics that's predefined which is inadequate for record illustration. Second issue is word model always generate widespread word set some word have that means and some usually are not helpful for document representation. The representation by way of single words with probabilistic distributions breaks the relationships between associated phrases. Thus, topic modeling wishes expanded modeling customers' interests in terms of topics' interpretations. In this work, a pattern-based topic modeling is proposed to increase the semantic interpretations of subject matters.

The pattern based topic modeling can be considered as a "post-LDA" model since right here patterns are built from the topic representation of the LDA model. After we are comparing pattern based topic models with the word based topic items we are able to analyze that the pattern-based topic model can be used to symbolize the semantic content of the user's documents more effectively. However, patterns in some subject matters can be colossal and some patterns should not discriminative sufficient to represent particular topics.

On this work, to symbolize topics rather of utilizing typical patterns a model is proposed to prefer probably the most representative and discriminative patterns, which can be called maximum matched Patterns. A new topic model, referred to as maximum matched pattern based topic modeling subject Modeling (MPBTM) is proposed for file

representation and report relevance ranking. The patterns within the MPBTM contained patterns are well structured in order that the maximum matched patterns will also be efficiently selected and used to symbolize and rank documents [4].

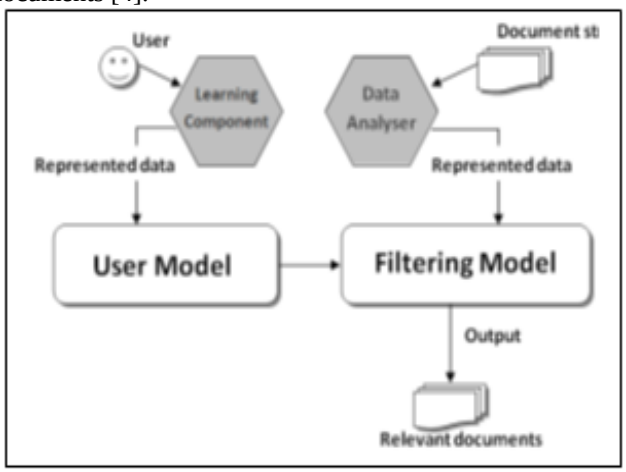

Fig. 1: The general process of Information Filtering system [7]

II. LITERATURE SURVEY

Two technical categories of baseline model include topic modeling methods, pattern mining methods. For the topic modeling category, the baseline models include Latent Semantic Analysis (LSA), Probabilistic Latent Semantic Analysis (PLSA), Latent Dirichlet Allocation (LDA) and Pattern-based Topic Model (PBTM). For pattern mining, the baseline models include Frequent Closed Patterns (FCP), frequent Sequential Closed Patterns (SCP) and phrases (nGram). An important difference between the topic modeling methods and pattern mining methods, the topic modeling methods consider multiple topics in each document in the document collection and use patterns (e.g. PBTM and MPBTM) or words (e.g. LDA) to represent the topics, whereas the pattern mining method assume that the documents within one collection are about one topic and use patterns to represent documents directly. Literature Survey of these baseline models are given below:

A. Topic modeling:

The main aim of probabilistic topic modeling is to discover hidden topics from a large document collection. Topic modeling algorithms are used to analyze the words of the original texts to find out the themes that run through them, how those themes are connected to each other, and how they change over time [4].

SSTM: Ying Liu, Wei Song, Lizhen Liu, Hanshi Wang, 2016, proposed a semantic smoothed topic model to represent document. It takes semantic similarity into consideration for topic of document. This method is useful for capturing the semantic of text to alleviating polysemy and synonyms problem and data sparseness problem[8].

LSA, PLSA: In order to compress large amount of data into useful and manageable form, topic modeling is used. Latent Semantic Analysis (LSA) uses a singular value decomposition of a collection, forming a reduced linear subspace. Another step to this concept is Probabilistic Semantic model, which is a generative data model. Almost all models that can be used are statistical mixture models, in which each word in a document form a mixture model, where the mixture components are multi-national random variables that can be viewed as a representation of topics. Topic modeling techniques can be generally divided into two categories, supervised and unsupervised; where bag of words and sequence words approaches are used respectively. In the field of Information retrieval, Document clustering and Summarization, uses an unsupervised bag of word technique, due to its simplicity. Whereas in the case of supervised models are used in supervised manner, using a pre-assigned labels for training set.

In order to compress large amount of data into useful and manageable form, topic modelling is used. Latent Semantic Analysis (LSA)[1] uses a singular value decomposition of a collection, forming a reduced linear subspace. Another step to this concept is Probabilistic Semantic model, which is a generative data model. Almost all models that can be used are statistical mixture models, in which each word in a document form a mixture model, where the mixture components are multi-national random variables that can be viewed as a representation of topics. Topic modelling techniques can be generally divided into two categories, supervised and unsupervised; where bag of words and sequence words approaches are used respectively. In the field of Information retrieval, Document clustering and Summarization[9],uses an unsupervised bag of word technique [10], due to its simplicity. Whereas in the case of supervised models are used in supervised manner, using a reassigned labels for training set.

LDA: LDA is probabilistic topic model which considers probability distribution functions for assigning words in a document to particular topic. The underlying instinct behind LDA is, documents are mixture of multiple topics. For example document named as computer science, can have topics such as data structure, algorithms, theory of computation, computer network etc means documents are mixture of topics. These topics are distributed over document in equal or unequal proportion. There are mainly two types of variables in LDA as hidden variables and observed variables. Observed variables are words within documents. While hidden variables describes topic structure. More precisely data arises from hidden random variables and these variables form topic structure. The process of inferring hidden structure from document is accomplished by computing posterior distribution. This distribution is conditional distribution of hidden variables in documents. The word 'Dirichlet' in Latent Dirichlet Allocation is a distribution that is used to draw per document topic distribution i.e. it specifies how topics are distributed in particular document. In generative process this output of dirichlet distribution is used to assign words of documents to different topics.

Hanna M Wallach, 2006,[12] presents a new concept known as Beyond Bag-of-Words.Normally uses bag of words for topic modelling.But it have some disadvantages. The most important among them are it ignores the word order. Generative topic models are generally of two categories,Bigram language models and N-gram topic models.N-gram models does not considers the word order,while bigram models consider pairs of words with the leading word defining a context. Bigram language models uses Hierarchical Dirchlet Language Models whereas N-Gram topic models uses Latent Dirchlet Allocation.It creates

a model which considers both topics and word order.It uses a simple extension of LDA algorithm.The bigram topic model shows improved performance compared to both the bi-gram language model and LDA.It is more feasible to consider word level models when the word order is not ignored.

B. Pattern Mining:

PBTM: Yang Gao, Yue Xu and Yuefung Li, 2015[1], proposes a two-stage model for modeling the documents in a collections. One of the main discriminative features of this model is, it combines the data mining techniques to statistical topic modeling to generate discriminative and pattern based representations for modeling topics in documents. In the first stage, it generates word distributions over topics for documents in the collection whereas in the stage second stage ,it uses the topic representations that are generated in the first stage for representing the topics by using term weighting method and pattern mining methods. The pattern based and discriminative term based representations generated in the second stage are more accurate and efficient than the representations generated by typical statistical topic modeling method LDA. Another important feature of this representation is, patterns carry more structural and inner relationship within each topic.. With a flexible threshold, frequent pattern mining can generate variable number of patterns as needed. Using many patterns can improve the discriminative power of models. However, using too many patterns may decrease the performance of models. Frequent sequential pattern considers `order' of words. Thus frequent sequential pattern mining can give us more discriminative meaningful phrases because it can distinguish phrases whose semantic meanings changes in different order.

FCP:H. D. Kim, D. H. Park, Y. Lu, and C. Zhai[13], proposed Enriching text representation with frequent pattern mining for probabilistic topic modeling. Here frequent patterns are pre-generated from the original documents and it is then added 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 regenerated patterns. To remove redundant patterns, pattern compression methods such as closed pattern and compressed pattern are used. Pattern compression can filter out meaningless patterns and let us use only important ones.

SCP:The work proposed by N. Zhong, Y. Li, and S.T. Wu, shows effective pattern discovery for text mining. Author studied an efficient and effective pattern discovery method which includes the processes of pattern deploying and pattern evolving.. In this work a Pattern Taxonomy Model is considered. There are two main stages in PTM. The first stage describes how to extract useful patterns from text documents, and the second stage is then how to use these discovered patterns to improve the effectiveness of a knowledge discovery system. In PTM, first of all they divide a text document into a set of paragraphs and treat them as an individual transaction, which consists of a set of words (terms). At the subsequent phase, to find frequent patterns from these transactions apply data mining method and generate pattern taxonomies. To get relevant pattern a pruning process is applied next sequential pattern mining algorithm named SPMining is used here[4].

Hong Cheng and Xifeng Yan, 2007,[14] designed discriminative frequent pattern analysis for effective classification. Here the authors give more importance to ”frequent patterns”. Frequent patterns are a set of items, subsequence, sub-graphs etc. Frequent patterns have the capacity of reflecting strong associations between each items and carries the underlying semantics of the data. They are also potentially useful features for classification. But it also have some disadvantages. Due its limited predictive power, the inclusion of infrequent patterns does not increases the accuracy. By building a connection between pattern frequency and discriminative measures like Fischer score,information gain,here a strategy is developed to set a minimum support in frequent pattern mining for generating more useful patterns.A feature selection algorithm was also proposed for building high quality classifiers.The two drawbacks of this approach is the scalability issue and over-fitting issue(Features are not representative).The new feature selection algorithm solves the scalability issue and facilitate the pattern generation.

Roberto J and Bayardo Jr, 1998,[10] proposed a pattern mining algorithm which scales linearly in the number of maximal patterns embedded in a database irrespective of the length of the longest pattern. Apriori like algorithms are used for finding frequent patterns in a database. In apriori algorithms generally uses a bottom up search mechanism and it involves a phase for finding frequent patterns. A frequent itemset is a set of items appearing together in a database records meeting a user specified threshold. There are several disadvantages are there for this apriori like algorithms. Most important among them are it restricts apriori like algorithms to discovering only short patterns. Here they uses a new algorithm called MaxMiner algorithm. Look ahead approach is used instead of bottom up search. and it can also prune all its subset from the consideration. By using the Max-Miner algorithm it can efficiently and effectively extracting only the maximal frequent item sets. The primary task behind every data mining operation is to find the patterns in a database.

Ning Zhong, Yuefeng Li and Sheng-Tang Wu,2012,[12]proposed a technique for effective pattern discovery for data mining. Almost all data mining techniques have been used for finding or extracting useful patterns in a text document, and most of these techniques are adopted from the concept of term based approach, they all suffer from the problems of polysemy and synonymy. Pattern based approaches can well perform than term-based approaches. The most prominent reasons for not using phrases are, they have inferior statistical properties compared with respect to terms, low frequency of occurrence and there exist large amount of redundant and noisy phrases among them. In this effective pattern discovery technique, first calculates the specificity of the discovered patterns and then calculates term weights according to the distribution of terms in the discovered patterns and thus solving the misinterpretation problem. This approach refine the discovered patterns by means of two methods called pattern deploying and pattern evolving. And this method shows an improvement in accuracy of evaluating term weights because discovered patterns are much more specific than the whole documents.

Topic Models	Characteristics	Limitations
LSA	LSA can get synonym words from the topics if there are any. the topic words if	It is hard to obtain and to determine the number of topics.
PLSA	Handles polysemy	At the level of documents PLSA cannot do probabilistic model.
LDA	Need to manually remove stop words.	It is unable to find the relationships among topics.
SVM	Efficient computational performance.	Model suffers from the problems of polysemy and synonymy
Pattern based model	Represent the semantic content of the user‘s documents more accurately.	Many times the patterns are not discriminative enough to represent specific topics.

Table: 1 Comparison of Different Topic Model Methods

III. CONCLUSION

An innovative pattern enhanced topic model for information filtering including user interest modeling and document relevance ranking. The proposed polysemy model generates pattern enhanced topic representations to model user‘s interests across multiple contents in the topics .The polysemy based semantic approach selects maximum matched patterns, instead of using all discovered patterns, for estimating the relevance of incoming documents. The proposed approach incorporates the semantic structure from content based modeling and the specificity as well as the statistical significance from the most representative patterns. The proposed model automatically generates discriminative and semantic rich representations for modeling the documents.

REFERENCES

[1] Y. Gao, Y. Xu and Y. Li, "Pattern-based Topics for Document Modelling in Information Filtering", IEEE Transactions on Knowledge and Data Engineering, vol. 27, no. 6, pp. 1629-1642, 2015.
[2] D. Sundar, "content based mining and extraction from documents using topic modeling", International Journal of Science Technology & Engineering, vol. 2, no. 10, pp. 1-4, 2016. Probabilistic topic models", Communications of the ACM, vol. 55, no. 4, p. 77, 2012.
[3] D. Blei, "Probabilistic topic models", Communications of the ACM, vol. 55, no. 4, p. 77, 2012.
[4] N. M.K. and S. K.P., "Similar Document Retrieval using Pattern- Based Topic Modelling for Information Filtering", International Journal of Innovative Research in Science, Engineering and Technology, vol. 5, no. 7, pp. 1-5, 2016.
[5] P. Nath. S and A. George, "Semantic Pattern-Based Topics Filtering for Document Modeling", International Journal of Innovative Research in Computer and Communication Engineering, vol. 3, no. 11, pp. 1-3, 2015.
[6] S. Thakare and M. Bhandare, "Maximum Matched Pattern-based Topic Model In Information Filtering", IJCTA, vol. 6, pp. 3-5, 2015.
[7] B. Jadhav, D. Bhosale and D. Jadhav, "Pattern Enhanced Topic Model for Information Filtering", International Journal of Advanced Research in Computer Science and Software Engineering, vol. 6, no. 5, pp. 1-3, 2016.
[8] Ying Liu, Wei Song, Lizhen Liu and Hanshi Wang, “Document Representation based on Semantic Smoothed Topic Model”, IEEE, pp. 1-4,2016.
[9] H. Cheng, X. Yan, J. Han, and C.-W. Hsu, Discriminative frequent pattern analysis for effective classication, in Proc. IEEE 23rd Int. Conf. Data Eng., 2007, pp. 71672
[10] R. J. Bayardo Jr, Efciently mining long patterns from databases, in Proc. ACM Sigmod Record, 1998, vol. 27, no. 2, pp. 8593.
[11] Hoffman, Matthew, Francis R. Bach, and David M. Blei. ”Online learning for latent dirichlet allocation.” advances in neural information processing systems. 2010.
[12] 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.
[13] H. D. Kim, D. H. Park, Y. Lu, and C. Zhai, “Enriching text representation with frequent pattern mining for probabilistic topic modeling,” in Proc. Am. Soc. Inform. Sci. Technol., 2012, vol. 49, no. 1, pp. 1–10.
[14] 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.

