ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 11, November 2015

Semantic Pattern-Based Topics Filtering for Document Modeling

Pallavy Nath. S , Annie George

PG Student, Dept. of CSE, Sree Buddha College of Engineering for Women, Elavumthitta, Pathanamthitta, Kerala, India

Assistant Professor, Dept. of CSE, Sree Buddha College of Engineering for Women, Elavumthitta, Pathanamthitta, Kerala, India

ABSTRACT: Topic filtering (Such as Latent Dirichlet allocation (LDA) and Maximum matched Pattern-based Topic Model MPBTM) provide a suitable way to analyze large number of unclassified text. Pattern mining is an important research area in data mining and knowledge discovery. The data mining concept is used in the field of information filtering for generating user's information needs from a collection of documents. However, the large amount of discovered patterns hinder them from being effectively and efficiently used in real applications, therefore selection of the most discriminative and representative semantic patterns from the huge amount of discovered patterns becomes crucial. To deal with the above mentioned problems, here proposed NFA based Maximum matched Pattern-based Topic Modeling (MPBTM), Enhanced LDA, Open English Natural language processing (NLP) and Gibbs sampling for topic modeling method. The main features of the proposed model include: (1) each topic is represented by patterns (2) Generate relevant topic document (3) the most discriminative and representative patterns, estimate more information retrieval from the document library according to the user's information.

KEYWORDS: Topic model; information filtering; pattern mining; relevance ranking; user interest model

I. INTRODUCTION

Information filtering (IF) is a system to remove redundant or unwanted information from an information or document stream based on document representations which represent user's interest. Traditional IF models were developed based on a term-based approach, whose advantage is efficient computational performance, as well as mature theories for term weighting [1], [2]. But term based document representation suffers from the problems of polysemy and synonymy. To overcome the limitations of term based approaches, pattern mining based techniques have been used for information filtering and achieved some improvements on effectiveness [3], [4], since patterns carry more semantic meaning than terms. Also, data mining has developed some techniques (i.e., maximal patterns, closed patterns and master patterns) for removing the redundant and noisy patterns [5]–[6].

Topic modeling [7] 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 corresponding distribution. Two representative approaches are Probabilistic Latent Semantic Analysis (PLSA) [8] and LDA [9]. However, there are two problems in directly applying 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 (i.e. a pre-specified number of topics). The second problem is that the word based topic representation (i.e. each topic in a topic model is represented by a set of words) is limited to distinctively represent documents which have different semantic content since many words in the topic representation are frequent general words [10].

In this project propose to overcome the limitation of existing system by using Natural Language Processing Natural language processing (NLP), i.e., the open English NLP 2.0 library used in enhanced LDA algorithm for filtering semantic meanings of patterns from the collections of topics. Here the LDA apply through the Gibbs sampling method and here also proposed Maximum matched Pattern-based Topic Model (MPBTM) for maximum matched pattern

DOI: 10.15680/IJIRCCE.2015. 0311189

representation and document relevance ranking and also it to select the most representative and discriminative patterns, which are to represent topics instead of using frequent patterns.

After installation of this application, it helpful for document searching inefficient and easy way from number of different documents and also available to download and view the document based on user's interested area or patterns. In this system efficiently find out relevant document from collection of document.

II. Related work

Information filtering deals with the delivery of information that the user is likely to find interesting or useful. An information filtering system assists users by filtering the data source and deliver relevant information to the users. When the delivered information comes in the form of suggestions an information filtering system is called a recommender system. Because users have different interests the information filtering system must be personalized to accommodate the individual user's interests. This requires the gathering of feedback from the user in order to make a user profile of the preferences [11]. Two major approaches exist for information filtering: content-based filtering and collaborative filtering system. A content-based filtering system selects items based on the correlation between the content of the items and the user's preferences, while a collaborative filtering system chooses items based on the correlation between people with similar preferences.

Text clustering methods can be used to structure large sets of text or hypertext documents. The well-known methods of text clustering, however, do not really address the special problems of text clustering: very high dimensionality of the data, very large size of the databases and understand ability of the cluster description. Here introduced novel approach which uses frequent item (term) sets for text clustering. Such frequent sets can be efficiently discovered using algorithms for association rule mining. To cluster based on frequent term sets; here measure the mutual overlap of frequent sets with respect to the sets of supporting documents [12]. We present two algorithms for frequent term-based text clustering, FTC which creates at clustering and HFTC for hierarchical clustering. An experimental evaluation on classical text documents as well as on web documents demonstrates that the proposed algorithms obtain clustering of comparable quality significantly more efficiently than state-of-the art text clustering algorithms. Furthermore, our methods provide an understandable description of the discovered clusters by their frequent term sets. Text clustering methods can be applied to structure the large result set such that they can be interactively browsed by the user. Effective knowledge management is a major competitive advantage in today's information society. To structure large sets of hypertext available in a company's intranet, again methods of text clustering can be used.

Direct Discriminative Pattern Mining for Effective Classification: direct discriminative pattern mining approach, DDP Mine, to tackle the efficiency issue arising from the two-step approach. DDP Mine performs a branch-and bound search for directly mining discriminative patterns without generating the complete pattern set. Instead of selecting best patterns in a batch, a "feature centred" mining approach that generates discriminative patterns sequentially on a progressively shrinking FP-tree by incrementally eliminating training instances. The instance elimination effectively reduces the problem size iteratively and expedites the mining process. Empirical results show that DDP Mine achieves orders of magnitude speedup without any downgrade of classification accuracy. It outperforms the state-of-the-art associative classification methods in terms of both accuracy and efficiency.

III. Proposed System

Topic 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 corresponding distribution. Two representative approaches are Probabilistic Latent Semantic Analysis (PLSA) and LDA [23] Here proposed a promising way to meaningfully represent topics by patterns rather than single words through combining topic models with pattern mining techniques. Specifically, the patterns are generated from the words in the word-based topic representations of a traditional topic model such as the LDA model. LDA based on sample occurrence and co-occurrence of the words in the documents.

Information filtering model based on pattern enhanced LDA.

DOI: 10.15680/IJIRCCE.2015. 0311189

This phase consist of four stages. They are:

- Pattern Equivalence Class
- Topic-based User Interest Modeling
- Topic-based Document Relevance Ranking
- Algorithms

A. Latent Dirichlet Allocation

Topic modelling algorithms are used to discover a set of hidden topics from collections of documents, where a topic is represented as a distribution over words. Topics models provide interpret able low-dimensional representation of documents (i.e. with a limited and manageable number of topics). LDA [10] is a typical statistical topic modelling technique and the most common topic modelling tool currently in use. It can discover the hidden topics in collections of documents using the words that appear in the documents

B. Pattern Enhanced LDA

Pattern-based representations are considered more meaningful and more accurate to represent topics than word based representations. Moreover, pattern based representations contain structural information which can reveal the association between words. In order to discover semantically meaningful patterns to represent topics and documents, two steps are proposed: 1st construct a new transnational data set from the LDA [11] model results of the document collection secondly, generate pattern-based representations from the transnational data set to represent user needs of the collection. Next Generate Pattern Enhanced Representation, the basic idea of the proposed pattern-based method is to use frequent patterns generated from each transnational data set to represent.

C. Maximum Matched Patterns

Maximum Matched Patterns (MPBTM), the patterns which represent user interests are not only grouped in terms of topics, but also partitioned based on equivalence classes in each topic group. The patterns in different groups or different equivalence classes have different meanings and distinct properties. Thus, user information needs are clearly represented according to various semantic meanings as well as distinct properties of the specific patterns in different topic groups and equivalence classes. Here NFA based Maximum Matched Patterns are used, this help to expression based searching and NLP is used in the enhanced LDA model for Semantic meaning patterns information or topics filtering. The general structure of the proposed IF model is depicted. In the training part, Here first generate user interest models from user profile (documents) by utilizing the proposed two-stage pattern-based topic modelling. The two models are proposed to generate different patterns to represent topics in the user interest models. The first IF model is based on Pattern-Based Topic Model(PBTM), in which user's interests are represented by multiple topics and each topic is represented by using all frequent patterns or closed patterns.

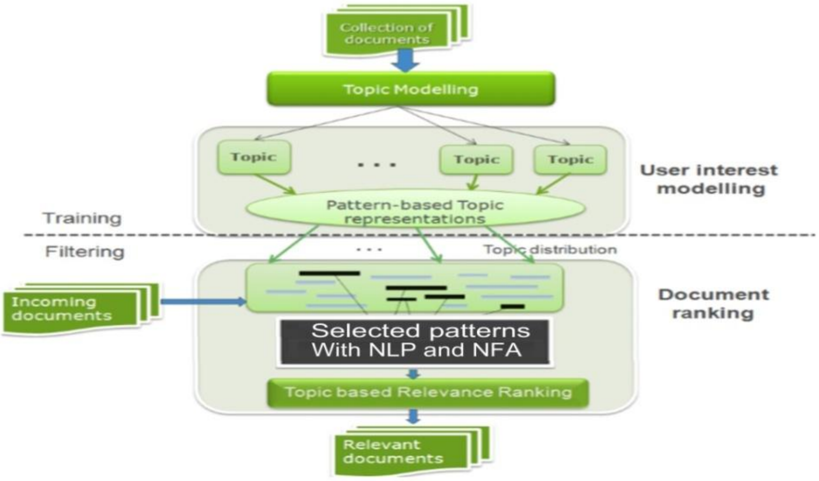

Fig 2: The structure of the proposed IF model for information filtering and retrieval based on NLP and NFA based document search.

The information retrieval or information filtering help users to find the most relevant and reliable information.

The second IF model is based on LDA, in which the patterns in each topic is further organized by different groups of equivalence classes. In the filtering part, for new incoming documents, based on the user interest models, relevant topical patterns are selected and used to calculate the relevance of documents to the user's interest in order to filter out irrelevant documents and provide relevant documents to the user. Further, for the relevance ranking in Pattern-Based Topic Model, corresponding to frequent patterns and closed patterns, there are two ranking models that is Pattern-Based Topic Model FP and Pattern-Based Topic Model FCP. Similarly, there are two ranking models in the Pattern-Based Topic Model. The Pattern-Based Topic Mode uses significant patterns to rank the relevance of documents whereas the Nondeterministic nite automaton (NFA) Pattern-Based Topic Model used for maximum matched patterns and natural language processing (NLP) used for significant patterns and semantic based topics filtering.

The proposed topic model represents topics using patterns with structural characteristics which make it possible to interpret the topics with semantic meanings. As with existing topic models, the proposed model is application independent and can be applied to various domains. Information filtering (IF) is a system to remove redundant or unwanted information from an information or document stream based on document representations which represent user's interests. The input data of IF is usually a collection of documents that a user is interested, which represent the user's long-term interests often called the user's profile. As mentioned before, user's information needs usually involve multiple topics. Hence, the proposed pattern-based topic modelling is applied to extract long-term user's interest through IF. Information retrieval (IR) typically seeks to find documents that are related to a user generated query from a given collection. The input data of IR is a query consisting of a number of terms which represent the user's short-term interest. One significant problem is that the length of queries is usually short and the keywords in a query are very often ambiguous or inconsistent.

For both IF and IR systems, besides user interest modelling, another essential part is document relevance ranking, which estimates the relevance between user's interests and documents. Whether the information retrieval or information filtering, users always to find the most relevant and reliable information. The quality of relevance ranking can potentially impact on user's perception of the specific ranking system's reliability. In this thesis, the relevance ranking models are established based on how to manage topics and extract the most meaningful and useful information from the multiple topic user interest model and the semantic meaning search patterns and NFA extract most meaningful and expression based document information.

Following are the contributions are:

- The important contributions to the domains of topic modelling, information filtering and information retrieval
- For more information filtering and retrieval, here proposed Open English 2.0 NLP and NFA
- To extract efficient semantic meaning pattern based document information by using Open English 2.0 NLP
- To extract more documents based on expression related searching by using regular expression Non-deterministic nite automaton (NFA) method

i. NFA based Maximum Matched Pattern Algorithm

```
Input: Document topics Zj Equivalent class ECj.....n Search expression R
Output: Maximum Matched Pattern MC
                    Process:
                    1. Take search expression R
                    2. Apply NFA expression to get Regular Expression set Rs
                    3. For expression ER on Rs
                    4. Check Rs sum Ec and Rs sum d
                    5. If (valid) Mc ←− Rs
                    6. Update Rank (Mci) +|MCjk|0.5 ∗fci ∗vci
                    7. End for generate Mc
```

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 11, November 2015

D. GibbsLDA++

Gibbs LDA++ is a C/C++ implementation of Latent Dirichlet Allocation (LDA) using Gibbs Sampling technique for parameter estimation and inference. It is very fast and is designed to analyse hidden/latent topic structures of large-scale data sets including very large collections of text/Web documents. There have been several implementations of this model in C, Java, and Mat-lab. Here decided to release this implementation of LDA in C/C++ using Gibbs Sampling to provide an alternative choice to the topic-model community. Gibbs LDA++ is useful for the following potential application areas:

1. Information Retrieval (analyzing semantic/latent topic/concept structures of large text collection for a more intelligent information search
2. Document Classification/Clustering, Document Summarization, and Text/Web Data Mining community in general
3. Collaborative Filtering
4. Content-based Image Clustering, Object Recognition, and other applications of Computer Vision in general
5. Other potential applications in biological data

IV. Simulation Results

The goal of most Open English Natural language processing NLP systems is to extract meaning from their language input. This meaning might ultimately be expressed as SQL for instance if the interface is for a database but in order to generate target representations NLP systems must first create intermediate representations to capture and rene meanings from their input. This intermediate representation that captures meaning is the semantic representation of the system.

In general, semantic representations need to capture details of objects and their relationships, events and the chains of causality that tie them together. A detailed discussion of semantic representations is beyond the scope of this document but the following section intends to highlight the issues of particular importance to semantic representations used specifically for NLP. Here language processors can implement it two ways: 1) apply the relevant semantic processing each time a new phrase is formed. 2) Apply semantic processing only when a complete and satisfactory parse has been found for an input sentence.

In a non-deterministic finite automaton (NFA), for each state there can be zero, one, two, or more transitions corresponding to a particular symbol. If NFA gets to state with more than one possible transition corresponding to the input symbol that is branches and it gets to a state where there is no valid transition, then that branch dies. An NFA accepts the input string if there exists some choice of transitions that leads to ending in an accept state. Thus, one accepting branch is enough for the overall NFA to accept, but every branch must reject for the overall NFA to reject. This is a model of computation.

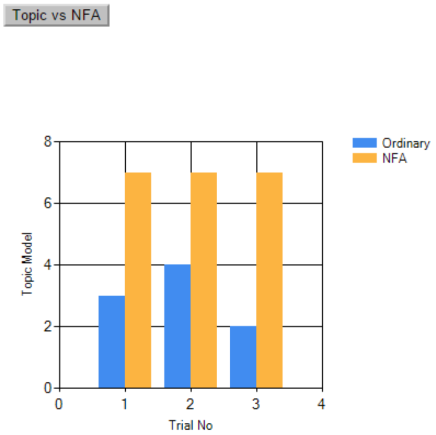

Fig.2: NFA Evaluation, Structure help to find out more information document compare than ordinary search. It is expression based search there for more reliable information document retrieve

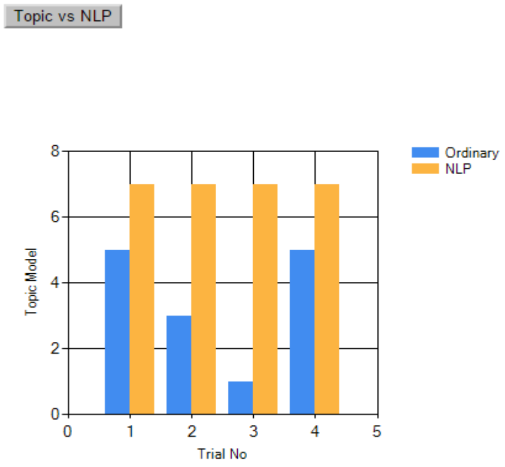

Fig.3: NLP Evaluation, Structure help to find out more semantic based information document model compared with ordinary search. NLP based search help to find meaningful information documents.

NFA accepts any binary string that contains 00 or 11 as a substring. And it accepts all binary strings that end with 101. A non-deterministic nite automaton (NFA) can have zero, one, or multiple transitions corresponding to a particular symbol. It is defined to accept the input if there exists some choice of transitions that cause the machine to end up in an accept state. Non determinism can also be viewed as a tree, or as a "guess-and-verify" concept, where the NFA can change state without consuming an input symbol.

Maximum Matched Patterns (MPBTM), the patterns which represent user interests are not only grouped in terms of topics, but also partitioned based on equivalence classes in each topic group. The patterns in different groups or different equivalence classes have different meanings and distinct properties. Thus, user information needs are clearly represented according to various semantic meanings as well as distinct properties of the specific patterns in different topic groups and equivalence classes. Here I used ten text documents for the filtering information and evaluation process. This evaluation proved Natural language processing (NLP) and non-deterministic nite automaton (NFA) is better than natural search. Because NFA and NLP give more relevant information document based on our topic search better than regular search. NFA based Maximum Matched Patterns are used to help expression based searching and NLP is used in the enhanced LDA model for Semantic meaning patterns information or topics filtering.

A non-deterministic finite automaton (NFA)is better than Natural language processing (NLP) Because NFA filtered more document information based on regular expression but in the NLP filter less document information because it filter the document based on semantic meaning. It only focuses on meaningful topics for filtering information that's why it gives less document information.

V. CONCLUSION AND FUTURE WORK

Here presents an innovative pattern enhanced topic model for information filtering including user interest modeling and document relevance ranking. The proposed MPBTM model generates pattern enhanced topic representations to model user's interest’s 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. The proposed approach incorporates the semantic structure from topic modeling and the statistical relevant method from the most representative patterns. The proposed model has been evaluated by using the Open English NLP 2.0 on the enhanced LDA models for retrieving semantic meaning of patterns and NFA based Maximum matched patterns based on regular expression for

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 11, November 2015

the task of more information filtering. In the proposed model demonstrates excellent strength on document modeling and relevance ranking.

The proposed model automatically generates discriminative and semantic rich representations for modeling topics and documents by combining statistical relevant topic modeling techniques and data mining techniques. The technique not only can be used for information filtering , but also can be applied to many content-based feature extraction and modeling tasks, such as information retrieval and recommendations metrics in future with some modifications in design considerations the performance of the proposed algorithm can be compared with other energy efficient algorithm. We have used very small network of 5 nodes, as number of nodes increases the complexity will increase. We can increase the number of nodes and analyze the performance.

REFERENCES

1. Hofmann, T., Unsupervised learning by" probabilistic latent semantic analysis", In IEEE transaction on Machine Learning , vol.42 , pp.177- 196, May 2010.
2. Y. Bastide, R. Taouil, N. Pasquier, G. Stumme, and L. Lakhal, "œMining frequent patterns with counting inference",€ " IEEE transaction on International Conference", vol. 2, pp. 66-75,2007.
3. H. Cheng, X. Yan, J. Han, and C.-W. Hsu, "Discriminative frequent pattern analysis for effective classification," IEEE transaction on. Data Eng., vol.7, pp. 716-725, 2007 .
4. R. J. Bayardo Jr, "Efficiently mining long patterns from data bases" IEEE transaction on "Comparing Passwords, Tokens, and Biometrics for User Authentication", vol. 91, pp. 2019-2020, Dec. 2003.
5. X. Wei and W. B. Croft, "LDA-based document models for ad-hoc retrieval," IEEE transaction on Develop. Information Retrieval", IEEE, vol. 185, pp. 178. June.2006.
6. http://www.wikipeda.com.
7. C. Wang and D. M. Blei,,œ"Collaborative topic modeling for recommending scientific articles", IEEE transaction on Knowledge. Discover Data Min., vol.5, pp. 448-" 456, 2011.
8. D. M. Blei, A. Y. Ng, and M. I. Jordan, □ Latent dirichlet allocation,"œProbabilistic latent semantic indexing, "IEEE transaction on Develop. Inform. Retrieval, vol. 3, pp. 1022, 2003,2008
9. http://www.wikipeda.com.
10. B. Liu, W. Hsu, and Y. Ma, "œIntegrating classification and association rule mining," IEEE transaction on Data Mining, vol.9, pp. 80-" 86, 2005.

BIOGRAPHY

Pallavy Nath. S is a Post Graduate student in Department of Computer Science & Engineering, Sree Buddha College of Engineering for Women, Mahatma Gandhi University. She received Bachelor of Technology (B.Tech) degree in 2013 from Mahatma Gandhi University, Kottayam, Kerala, India. Her research interests are .NET, Java etc.