Pattern Based Information Filtering for Author Community Generation

Neena G. Krishnan[1], Neena Joseph[2]

[1]PG Scholar ,Dept of CSE, Mangalam College of Engineering, Mahatma Gandhi University, Ettumanoor, Kottayam, Kerala, India

[2]Assistant Professor of Dept of CSE, Mangalam College of Engineering, Mahatma Gandhi University,
Ettumanoor, Kottayam, Kerala, India

Abstract: ***Pattern mining is one of the important research areas in data mining and knowledge discovery. The data mining concept is used in the field of information filtering. User's interested information is collected using these data mining concepts. Maximum matched pattern based Topic Model provide a suitable way to analyze large number of unclassified text. Large amount of discovered patterns stop them from being effectively and efficiently used in real application, therefore selection of the most separate and representative semantic patterns from the huge amount of discovered patterns become crucial. To deal with the above mentioned problem, propose NFA based Maximum matched Pattern based Topic Modeling .Finally it enhanced to an author community model. Search document within the author community is efficient and easy.***

Keywords: Topic model, information filtering, pattern mining, cluster

1. Introduction

An Information Filtering (IF) system filters the data source and delivers relevant information to the user. It is called as user interest model [10] . Term based approach ,one of the traditional IF model , is efficient in computational performance .But it suffer from problem of polysemy and synonymy .To overcome the limitation of term based approach pattern mining technique is used [1] [2]. Patterns carry more semantic meaning than term .Topic modeling [3],[4],[5] is one of the text modeling technique. It can automatically classify documents into number of topics and represent every document with multiple topics and their corresponding distribution. Two representative approaches are PLSA [6] and LDA [5] . Directly applying topic models for IF two problems are generated .Due to limited number of dimensions to represent documents, topic distribution is insufficient is the first problem and represent document in word based topic have different semantic content is the second problem. To overcome the problem pattern enhanced LDA [8] is used. It carries more concrete and identifiable meaning than word based representations using LDA [5].Number of patterns in some of the topic can be huge and many of the patterns are not distinguishing enough to represent specific topic. To deal with the problem MPBTM Maximum matched Pattern Based Topic Modeling is introduced. MPBTM [8] consists of topic distributions, describing topic preferences of documents or collection of documents and structured pattern based topic representation, representing semantic meaning of the topics in a document. To improve the searching efficiency of document, NFA based MPBTM and Author community are proposed.NFA based MPBTM extract meaningful and expression based document from the author community.

2. Related Work

An information filtering system [7] filters the data source and delivers relevant information to the users. Information filtering consists of two major approaches. They are 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 preference of user, while system of collaborative filtering chooses items based on the correlation between people with similar preferences.

Topic model [3] techniques have been achieved successful retrieval results. The LDA based document models are one of the topic modeling approaches and it achieves good performance compared to other models .Patterns in some of the topic can be huge and many of the patterns are not differentiate enough to represent specific topic. To deal this problem MPBTM [8] is used.

To structure large set of text or hypertext documents text clustering methods can be used. The well-known methods of text clustering do not deal with the special problems of text clustering like very high dimensionality of the data, very large size of the databases and understand ability of the cluster description. Novel approach uses frequent item 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.

Concept of Kmean [9] is set of n data points in real dimensional space Rd, and an integer k, the problem is to find out a set of k points in Rd, called centers. So as to minimize the mean squared distance from each data point to its adjacent center. This measure is often called the squared-error distortion and this type of clustering falls into the general category of variance based clustering.

3. Proposed Architecture

To improve efficiency of searching a document, NFA based MPBTM and Author community are proposed. NFA based MPBTM extract meaningful and expression based document. User can search document based on their input query as well as based on their interested author. That is the selection of document is completely based on the user interest.

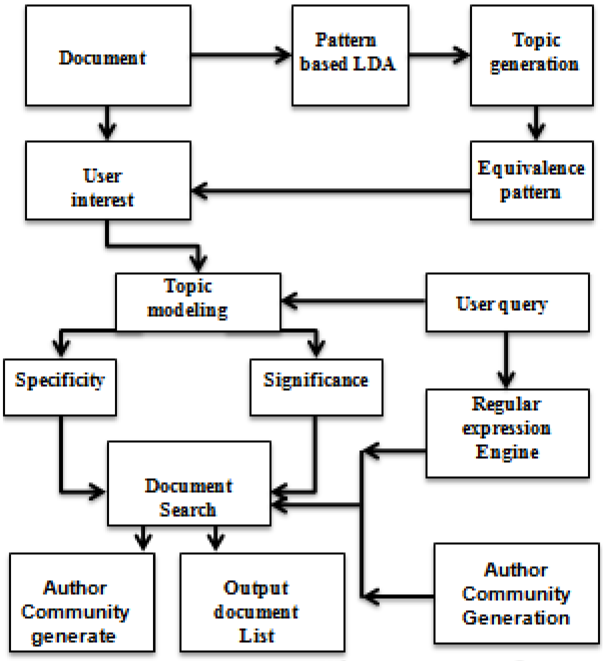

Figure 1: Proposed Architecture

Each block of proposed architecture is important. User can upload document. Uploaded document undergo topic generation, equivalence pattern generation and topic modeling.

To obtain search result, document relevance ranking should be calculated. For that calculate specificity and significance from the topic modeling. User input a query for search, regular expression and author community based search is performed in the proposed model. User input query can be a single word or group of word or a regular expression. Input query check based on regular expression, so that output should be relevant. Searching can also be done based on the author. Author cluster is generated for author community. The documents related to a specific area or topic are said to be related documents. The authors of related documents are grouped into a cluster and is given a single cluster id. Therefore user can get related documents of different author, and choose their interested document.

Based on the proposed architecture, an example is shown below. Out of 20 documents uploaded by user, output of pattern based LDA is

Table 1: Example of pattern based LDA

{memory, cpu, register}
{memory,data}
{cpu,instruction}

From the Table 1, select pattern with frequency .6 to generate equivalence class. All patterns in the equivalence class have same frequency. Frequency is the statistical significance.

Table 2: Equivalence class pattern with frequency .6

{memory,cpu}

User input a query ‘memory’; semantically check the input query with document collection. If a document related to the input query is present, then user can download it. Searching can also be done based on the author. Related documents of authors are grouped into a cluster. Same cluster id contains related documents. So the user can select document based on their interested author.

Table 3: Author based document search

Cluster Id	Author	Document
1	Dr.Deepthi Mehrotra	Advanced Computer Architecture

From the above examples, it can be recognized that proposed model provide efficient search result than earlier models. .

3.1.1. LDA

LDA (Latent Dirichlet Allocation) is a technique that automatically discovers topics that are presented in the document.LDA provide topic representation using word distribution and document representation using topic distribution. LDA is widely used in topic modeling tools.

3.1.2. Pattern Enhanced LDA

Pattern based representation is more meaningful and accurate to compare with word based representation. It contains structural information which can reveal the association between words. In order to discover semantically meaningful patterns to represent topics and documents, two steps are needed: 1st construct a new transactional data set from the LDA model results of the document collection and secondly, generate pattern-based representations from the transactional data set to represent user needs of the collection.

3.1.3.Maximum matched Pattern Based Topic Modeling

Maximum Matched Patterns Based Topic Modeling (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. Not only NFA based Maximum Matched Patterns are used for searching document efficiently but also efficient searching can be done within the Author community .First generate user interest models [10] from user profile (documents).

3.1.4. Document Relevance Ranking

Relevance of the documents is estimated based on the user interest model to filter out irrelevant documents. The maximum matched pattern in the equivalent classes are used to estimate the relevance of the new incoming documents to the user interest. Based on the relevance of the documents the new documents will be ranked.

4. Algorithm

Take each pattern from the equivalence class list. Initialize regular expression class for accessing regular expression tool. Internally verify whether the search pattern is present or not. If the search pattern is present, then the corresponding searched pattern is displayed and provides download facility. Not only searching can be done directly but also through Author community. Author Community is generated based on the text clustering algorithm Kmean. Kmean text clustering algorithm is used to find similar document and then list their author into a clusters.

5. Evaluation

Based on the experimental evaluation, out of 20 documents NFA based MPBTM lists more documents as search result than MPBTM .The detailed description is shown in Table 4 and figure 2.

In this evaluation, check the performance of MPBTM and NFA based MPBTM in terms of hit rate. To facilitate, input same query and then compare how much documents is obtained in MPBTM and NFA based MPBTM.

Table 4: Evaluation

Data Set	*MPBTM*	*NFA based MPBTM*
5	1	3
10	4	7
15	9	13
20	11	17

Table 4 shows the performance of MPBTM and NFA based MPBTM in terms of data hit rate.

When 5 data set was uploaded, MPBTM showed 1 and NFA based MPBTM showed 3 documents as search result. Continue this evaluation up to 20 data set .Based on Table 4, Figure 2 is plot.

In Figure 2, X coordinate shows data set that is the number of document uploaded .Y coordinate shows document hit rate that is number of documents retrieved based on the search.

Efficiency of NFA based MPBTM is higher than MPBTM in terms of document hit rate. User can search document based on Author also. That is there are two possible way to search document. One is based on the input query of user other is user interested authors. In both case meaningful and expression based document is retrieved. So that while comparing with previous model, proposed model give efficient searching result.

From the graph, in the case of NFA based MPBTM data hit rate is increasing while increasing the data set. That is data hit rate is directly proportional to data set. That is effective semantic check is done. But in the case of MPBTM we cannot predict that efficient data hit rate obtained for corresponding data set.

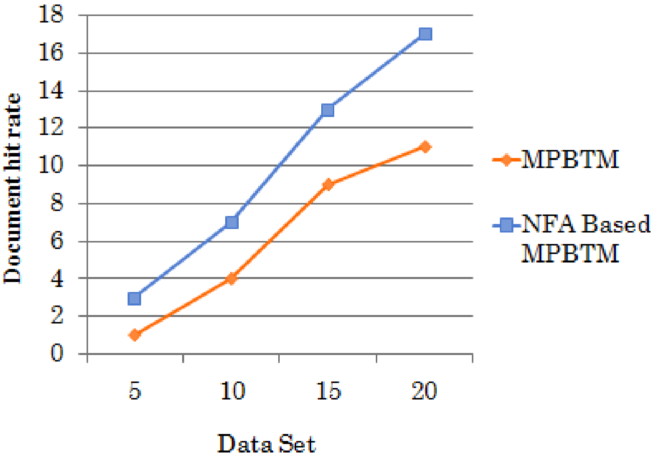

Figure 2: Comparison of NFA based MPBTM and MPBTM in terms of hit rate

From the evaluation, it can be stated that performance of NFA based MPBTM and author community is efficient than MPBTM. Search results are also relevant.

6. Conclusion

Comparing with previous model MPBTM, NFA based MPBTM and author community can give relevant search result that is searching document in terms of the author as well as the semantic meaning of user input.

MPBTM consists of topic distributions, describing topic preferences of documents or collection of documents and structured pattern based topic representation , representing semantic meaning of the topics in a document .To improve the searching efficiency of document , NFA based MPBTM and Author community are proposed. From the evaluation, it is clear that proposed model overcome the limitation of MPBTM.So declare that efficiency of document search of proposed model higher than that of early model.

NFA based MPBTM does not support case sensitivity and it takes more processing time. In future methods for identifying case sensitive topics as well as methods to reduce the processing time for generating the model can be introduced. Then it will be the one of the efficient model in information filtering.

References

[1] Y. Bastide, R. Taouil, N. Pasquier, G. Stumme, and L. Lakhal,“Mining frequent patterns with counting inference,” ACM SIGKDD Explorations Newslett., vol. 2, no. 2, pp. 66–75, 2000

[2] H. Cheng, X. Yan, J. Han, and C.-W. Hsu, “Discriminative frequent pattern analysis for effective classification,” in Proc. IEEE 23rd Int. Conf. Data Eng., 2007, pp. 716–725

[3] X. Wei and W. B. Croft, “LDA-based document models for ad-hocretrieval,” in Proc. 29th Annu. Int. ACM SIGIR Conf. Res. Develop.Inform. Retrieval, 2006, pp. 178–185.

[4] C.Wang and D.M.Beli,”Collaborative topic modelling for recommending scientific articles,” in Proc.17th

ACM SIGKDD Int.Conf.knowl.Discov.Data Min.,2011,pp.448-456

[5] D.M. Beli,A. D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent dirichlet allocation,” J. Mach. Learn. Res., vol. 3, pp. 993–1022, 2003.

[6] T. Hofmann, “Probabilistic latent semantic indexing,” in Proc22nd Annu. Int. ACM SIGIR Conf. on Res. Develop. Inform. Retrieval 1999, pp. 50–57.

[7] Y. Gao, Y. Xu, and Y. Li, “Pattern-based topic models for information filtering,” in Proc. Int. Conf. Data Min. Workshop SENTIRE, 2013, pp. 921–928.

[8] Yang Gao ,Yue Xu ,and Yuefeng Li, “Pattern- based Topics for Document Modelling in Information Filtering,” IEEE Transaction on Knowledge And Data Engineering, June 2015.

[9] Tapas Kananga, David M. Mount, Nathan S. Netanyahu, Christine D. Piatko, Ruth Silverman, and Angela Y. Wu,” An Efficient k-Means Clustering Algorithm: Analysis and Implementation,” IEEE Transaction on Pattern Analysis and Machine Intelligence, VOL. 24, NO. 7, JULY 2002

[10] https://en.wikipedia.org/wiki/User_modeling

Author Profile

Neena G. Krishnan is pursuing M. Tech in Computer Science and Engineering from Mangalam College of Engineering, Mahatma Gandhi University. She received B.Tech degree in 2013 from Mangalam College of Engineering, Mahatma Gandhi University, Kottayam, Kerala, India. Her research interests are .NET, Java etc.

Neena Joseph received the B.tech Degree in Computer Science & Engineering from Mangalam College of Engineering in 2008 and M. Tech Degree from Achariya College of Business & technology in 2012.She is presently working as Assistant Professor at Mangalam College of Engineering. Her Research areas of interest are Relational Database and Compilers.