A Review on Pattern-based Topics for Document Modelling in Information Filtering

Ms. P. S. Ajabe[1,] Dr. P. M. Jawandhiya[2]
P.G.Student of CSE, Pankaj Laddhad Institute of Technology and Management Studies, Buldana,India
Email: poojaajabe@gmail.com[1]
Principal of Pankaj Laddhad Institute of Technology and Management Studies, Buldana,India
Principal_plit@rediffmail.com[2]

Abstract- 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. In the field of information filtering many term-based or pattern-based approaches have been used to generates users information needs from a collection of documents. A fundamental assumption for these approaches is that the document in the collection are all about one topic, but in reality users interest can be diverse and document in collection involve multiple topics. Pattern mining is an important research area in data mining and knowledge discovery. However, large amount of discovered patterns hinder them from being effectively and efficiently used in real applications ,therefore selection of most discriminative and representative semantic patterns from huge amount of discovered patterns becomes crucial .To deal with above mentioned problems, a novel information filtering model, maximum matched Pattern-based Topic Model(MPBTM) is proposed. MPBTM is proposed to estimate the document relevance to the user's information needs in order to filter out irrelevant documents.

Index Terms-Information filtering; Topic Model; Pattern mining; relevance ranking; user interest model

1. 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 modelling[7] has become one of the most popular probabilistic text modelling 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 Latent Dirichlet Allocation (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, propose to overcome the limitation of existing system by using 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 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 in efficient 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

1.1 Topic modelling

A Topic model is a type of statistical model for discovering the abstract "topics" that occur in a collection of documents. Topic models are a suite of algorithms that uncover the hidden thematic structure in document collection. Topic models provide a convenient way to analyze large of unclassified text.

- A topic contains a cluster of words that frequently occur together. A topic model contains a collection of text as input it discovers a set of "topics" recurring themes that are discuss in the collection of documents.
- A topic modelling can connect words with similar meanings and distinguish between uses of words with multiple meanings. So, the idea of topic models is that term which can be working with documents and these documents are mixtures of topics, where a topic is a probability distribution over words.
- In other word, topic model is a generative model for documents. It specifies a simple probabilistic procedure by which documents can be generated. It creates a new document by choosing a distribution over topics. After that, each word in that document could choose a topic at random depends on the distribution.
- Topic modelling has become one of the most popular probabilistic text modelling technique and has been quickly accepted by machine learning and text mining communities.
- A two representative approaches are Probabilistic Latent Semantic Analysis (PLSA) and Latent Dirichlet Allocation (LDA)[13].

1.2 Pattern mining

- Pattern mining is an important research in data mining and knowledge discovery.Patterns can be discovered from text documents in effective manner.
- The patterning provides a reusable architecture which speeds up many computer programs, it offers more characteristics meaning than the single words. Pattern based topic model can be recycled to represent the acceptable content of the user text more truthfully compared with the word based topic models.
- The Pattern is always thought to be more discriminative than single terms for describing documents. The pattern based topic filtering used to filter out the irrelevant document and gives relevant document from the collection of document[15].

1.3 Information filtering

- 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 [1],[16].
- Two major approaches exist for information filtering:

1.3.1 Content-based filtering system

- A content-based filtering system selects items based on the correlation between the content of the items and the user's preferences.

1.3.2 Collaborative filtering system

- A collaborative filtering system chooses items based on the correlation between people with similar preferences

.

2. LITERATURE SURVEY

Sr. No.	Name of Paper	Name of Author	Published Year	Description	Advantages	Disadvantages
1.	Innovative Pattern Mining For Information Filtering System	Vasudevan ,V. Sharmila, Dr. G.Tholkappia Arasu	2012	In this paper, a survey on preprocessing pattern deploying approach, new pattern based information filtering model, revision and mining algorithm, iterative learning algorithm, novel two stage decision model, in addition to some text mining applications are discussed.	The study concludes that the concept-based analysis bridges the gap between Natural Language Processing and Text mining.	The main drawback of pattern based information filtering model are difficulty of occurring long patterns, and low capability of dealing with large discovered patterns.
2	Pattern Enhanced Topic Model	Tincy Chinnu Varghese, Smitha C Thomas	2016	In this paper, LDA Pattern Enhanced LDA, Algorithms in MPBTM is discussed. MPBTM consists of two algorithms: User Profiling Algorithm(generating user interest model) and Document Filtering Algorithm(relevance ranking of incoming document)	MPBTM generates descriptive and semantically rich representations for modelling topics .It is used in the field of content- based extraction of documents ,machine learning etc.	Term based model suffers from problem of polysemy and synonymy. Existing system faces the low frequency problem.
3	Pattern-based Topics for Document Modelling in Information Filtering	T. Devikarani ,Mrs.C. Mohanapriya,M.sc.,M.phil	2016	In this paper, Pattern based topic model is discussed along with algorithm and implementation.	The proposed significantly matched patterns and maximum matched pattern for the StPBTM model.	Many general textual content classification algorithms don't work well for a new person, which typically way no or few training knowledge element
4	A Latent Dirichlet Allocation Algorithm for Pattern-Based Topic Filtering	V.Vishnu Priya ,S.K.Soma Sundaram	2016	In this paper, LDA is used to finding the high values from probability ratio, it gives term weight value and support and confidence based on mining method.	The proposed method is used only for documents eg:notepad files,etc.	It shows the disadvantage of term based approach.
5	Pattern-based topics for Document Modelling in Information Filtering	Yang Gao, Yue Xu,and Yuefeng Li	2015	In this paper, LDA, Pattern Enhanced LDA are discussed which consists of Pattern equivalence class, topic based user interest modelling ,topic-based document relevance ranking. The proposed IF model has two algorithm: User profiling And Document Filtering.	This paper presents an innovative pattern enhanced topic model for information filtering including user interest modelling and document relevance ranking	In this paper, disadvantages of term based approach is discussed

3. PROBLEM DEFINATION

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 the word-based topic representation is limited to distinctively represent documents which have different semantic content since many words in the topic representation are frequent general word. The topic model and the language models are very limited in representing the specificities[1].

4. PROPOSED SYSTEM AND ALGORITHM

4.1 MPBTM (Maximum matched pattern-based topic model)

- In proposed system, a novel information filtering model, Maximum matched Pattern-based Topic Model (MPBTM), is proposed. The patterns are generated from the words in the word-based topic representations of a traditional topic model such as the LDA model. This ensures that the patterns can well represent the topics because these patterns are comprised of the words which are extracted by LDA based on sample occurrence and co-occurrence of the words in the documents[1].
- We propose to model user's interest with multiple topics rather than a single topic under the assumption that user's information interests can be diverse.
- We propose to integrate data mining techniques with statistical topic modelling techniques to generate a pattern-based topic model to represent documents and document collections. The proposed model MPBTM 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.
- We propose a structured pattern-based topic representation in which patterns are organized into groups, called equivalence classes, based on their taxonomic and statistical features. Patterns in each equivalence class have the same frequency and represent similar semantic meaning. With this structured representation, the most representative patterns can be identified which will benefit the filtering of relevant documents.
- We propose 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 users interest. The maximum matched patterns are the most representative and discriminative patterns to determine the relevance of incoming documents.

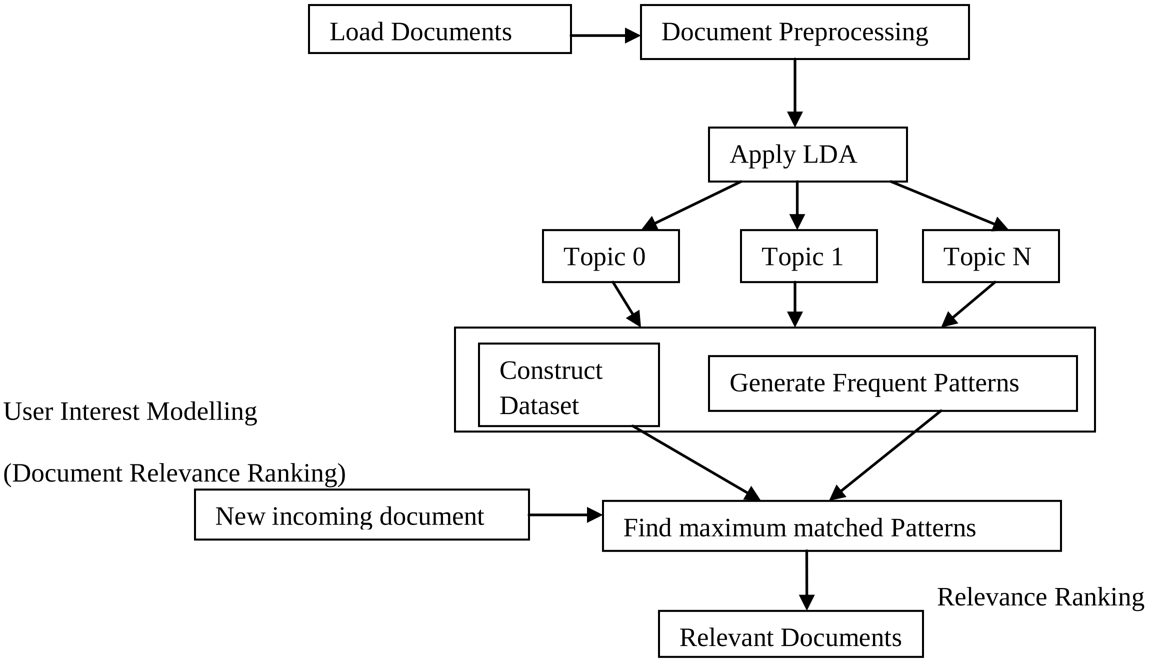

Fig.4.1: Maximum matched Pattern-based Topic Model (MPBTM)

- First of all the user have to load the dataset. Preprocessing includes apply stop words and stemming in the dataset. Find the support value for the words in the file. And sort the data in descending order.
- The splitted data is called as the transactional dataset. The support value is calculated for the splitted data. These splitted data are three types of topics. Pattern-based representations are considered more meaningful and more accurate to represent topics than word-based representations.
- In order to discover semantically meaningful patterns to represent topics and documents, two steps are proposed: firstly, construct a new transactional dataset from the LDA model results of the document collection D; secondly, generate pattern-based representations from the transactional dataset to represent user needs of the collection D. The support value is calculated within the topic type. The probability is calculated for the topics. It is called as the equivalence class.
- User interest modelling is used for the topic distribution.
- Relevance ranking is used for find the topics for the document. The number of patterns in some of the topics can be huge and many of the patterns are not discriminative enough to represent specific topics.
- In this section, one novel IF model, MPBTM, is proposed based on the pattern enhanced topic representations. The proposed model consists of topic distributions describing topic preferences of documents or a document collection and structured pattern-based topic representations representing the semantic meaning of topics in a document. [17].

4.2 Algorithms in MPBTM

- The proposed IF model can be formally described in two algorithms:
  User Profiling (i.e. generating user interest models) **Algorithm** and
  Document Filtering (i.e. relevance ranking of incoming documents) **Algorithm.**
- The former generates pattern-based topic representations to represent the users information needs. The later ranks the incoming documents based on the relevance of the documents to the user's needs.[14]

3.2.1 Algorithm :User Profiling

```
Input: a collection of positive training documents D;
       Minimum support   as threshold for topic Zj;
       number of topics V
Output: = UE ={E(Z1)…E(Zv)}
Generate topic representation ∅ and word-topic assignment  Zd, i by applying LDA to D
UE: = ∅
for each topic Zj ∈ [Z1, Zv] do
    Construct transactional dataset Tj based on ∅ and zd, i
    Construct user interest model Xzj for topic Zj using a pattern mining technique
    Construct equivalence class E (Zj) from Xzj
    UE: = UE ∪ {E(Zj)}
end for
```

3.2.2 Algorithm: Document Filtering

```
Input: user interest model UE = {E(Z1)….E(Zv)}, a list of
incoming document Din
Output: rank (d), d ∈ Din
rank(d):=0
for each d ∈ Din do
    for each topic Zj ∈ [Z1, Zv] do
        for each equivalence class EC jk ∈ E(Zj) do
            Scan EC jk and find maximum matched pattern MCjk which exists in d
            update rankE(d) using Equation (3):
            rank(d):=rank(d)+| MCjk|0.5 *f jk*V D,j
        end for
    end for
end for
```

5. CONCLUSION

In this, paper, Pattern based topic model is discussed along with algorithm. Traditional IF models were developed based on a term-based approach, But term based document representation suffers from the problems of polysemy and synonymy. maximum matched pattern based topic model gives an innovative pattern enriched topic model for filtering information from a set of documents including users interest model and relevance ranking. The proposed MPBTM model generates pattern enhanced topic representations to model user's interest 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. It also generates descriptive and semantically rich representations for modelling topics

REFERENCES

[1]Yang Gao,Yue Xu, and Yuefeng Li (JUNE 2015): Pattern-based Topics for Document Modelling in Information Filtering. IEEE TRANSACTIONS OM KNOWLEDGE AND DATA ENGINEERING, VOL.27, NO.6

[2]S.Robertson, H.Zaragoza, and M.Taylor (2004): Simple BM25 extension to multiple weighted fields. InProc.13thACM Int.Conf.Inform.Knowl.Manag.,pp. 42–49.

[3]F.Beil, M.Ester, and X. Xu (2002): Frequent term-based text clustering. in Proc.8th ACM SIGKDD Int. Conf. Knowl.Discov. Data Min., pp. 436–442.

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

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

[6]R.J.BayardoJr (1998): Efficiently mining long patterns from databases. in Proc. ACM Sigmod Record, vol. 27, no. 2, pp. 85–93.

[7]J.Han,H.Cheng,D.X and X. Yan (2007): Frequent pattern mining: Current status and future directions.Data Min.Knowl.Discov.vol.15, no.1, pp 55–86.

[8]M.J.Zaki and C.J.Hsiao (2002):CHARM: An efficient algorithm for closed item set mining. in Proc. SDM, vol. 2, pp. 457–473.

[9] Y.Xu, Y.Li and G. Shaw(2011): Reliable representations for association rules.Data Knowl.Eng., vol.70, no. 6, pp.555–575.[10W.B.Croft [10] X.Wei and W.B.Croft (2006):LDA-based document models for ad-hoc retrieval.in Proc. 29th Annu. Int. ACM SIGIR Conf. Res. Develop. Inform. Retrieval,pp. 178–185.

[11]C.Wang and D.M.Bleim (2011): Collaborative topic modeling for recommending scientific articles.in Proc. 17th ACM SIGKDD Int. Conf. Knowl.Discov. Data Min.,pp. 448–456.

[12]D.M.Blei,A.Y.Ng, and M.I.Jordan (2003): Latent dirichlet allocation.J.Mach. Learn.Res.vol.3, pp.993–1022.

[13] V.Vishnu Priya and S.k.Soma Sundaram (2016): A Latent Dirichlet Allocation Algorithm for Pattern-Based Topic Filtering", 340-345, 2016 ISSN 1990-9233; IDOSI Publications.

[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.

[17]T.Devikarani,Mrs.C.Mohanpriya,M.sc.,M.phil(2016):PATTERN-BASED TOPICS FOR DOCUMENT MODELLING IN INFORMATION FILTERING .International Journal of Research in Computing Science, Technology and Engineering.