IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

nCORETech 2017

LBS College of Engineering, Kasaragod

Vol. 6, Special Issue 3, March 2017

Maximum Matched Pattern-based Topics for Document Modeling in Information Filtering

Ms. Raveena Sukumaran .M

LBS College of Engineering, Kasargod, Kerala, India

Abstract: In the field of Information Filtering we have many term-based or pattern –based methods for generating user's needed form information from a set of documents .A basic general thinking is that documents in a set of particular collection is related to only a single topic .But in real life user's interest is different and documents in a set or collection includes multiple topics. Most commonly used topic modeling method is Latent Dirichlet Allocation (LDA) which generates a structural model to represent multiple topics in a set of documents. Patterns generally are more descriptive and efficiently used in real time applications. So to select most descriptive and efficient patterns from the discovered set of patterns here a Maximum matched Pattern-based Topic Model is introduced. It helps us to get the relevant document according to user needs by filtering out unwanted documents.

Keywords: Topic Model, Information Filtering, Pattern mining, relevance ranking, user interest model.

INTRODUCTION

Information filtering (IF) is a way of removing unwanted information from a set of information or documents on the basis of user needs. Many traditional models include Term-based way of filtering documents or information, which faces the monosemy and antonym .To overcome this a Pattern-based model was introduced which gave a more effective result. This model took into consideration the user's interest on the particular search..For improving the quality of patterns used in the search data mining techniques like maximal patterns, closed patterns, master patterns was included. Hence it helped to remove the unwanted and noisy patterns in the search.

Topic modeling is commonly used now a day in the field of machine learning and text mining. It classifies the documents in a set by a given number of topics and presents each document as multiple topics and related distribution. Here introduces a efficient way to represent the topics as patterns than single term words, because patterns can describe a user's needs or interest more than a single word. By using patterns we are able to get the documents filter effectively and efficiently upto to an extent than before .But achieving the best pattern from the huge collection is a crucial thing .For this a new topic model called Maximum matched pattern based model (MPBTM) is introduced.

Contributions of MPBTM for information filtering include:

1) Considers user's interest with multiple topics than one topic based on the thinking that user's information needs are diverse.

2) Integrates data mining techniques with statistical topic modeling techniques to get documents and collection of documents.

3) In the structured pattern-based representation of topics, the patterns are divided into groups called equivalence classes on the basis of structural and similar features Each group have words with same frequency and similar meanings

4) A new approach called Ranking method for determining the relevance of new documents is introduced.

EXISTING SYSTEM

In the information filtering technique the aim is to perform mapping from a set of incoming documents to a user relevant document. Let us denote the set of incoming documents as D ,the mapping rank be: D->R, such that rank(d) gives the relevant document .As in the traditional models like term-based have limitations in expressing semantics and also monosemy and antonym. As the number of returned patterns is large selecting reliable patterns is very hard.. Probabilistic topic modeling is another model which helps to extract long term user's needs by verifying content and by representing latent topics which are discovered from user profiles. Lack of explicit discrimination in most of the languages model based approaches and probabilistic topic models .This problems are driven out by labelling topic techniques which considers phrases instead of words for information filtering. In this model n-gram structure is included along with latent topic variables for generating topic relevant phrases. Here it faces the low frequency problem. To overcome all the problems in the topic modeling techniques we introduce the maximum matched pattern based topic model which also have the relevance ranking mechanism which generate an efficient and descriptive patterns from a huge document and also rank the maximum matched patterns

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

nCORETech 2017

LBS College of Engineering, Kasaragod

Vol. 6, Special Issue 3, March 2017

PROPOSED SYSTEM

Generally topic modeling techniques are used for discovering a set of hidden documents from a group of collection ,where topics are a set of words. Latent Dirichlet Allocation is a commonly used topic modeling now a days for information filtering. It can discover the hidden
topics in a set of documents .Let document be taken as D = fd1; d2; . . . ; dM } .The basis goal of LDA is that every documents consist of a number of topics and every topics consist of a number of words. Mainly the LDA process consists of two parts. The document level and collection level .At the document level, each document say is represented by set of topics such as $\theta_{di} = (\vartheta_{di,1}, \vartheta_{di,2}, \ldots\ldots \vartheta_{di,V})$.,V is the number of topics in the document. At collection level these each topics is represented by set of words such as fj , for topic j ,such as $\Phi = \{\Phi_1, \Phi_2, \ldots\ldots \Phi_V\}$.

Take an example of a set of documents D ={ d1; d2; d3; d4} be a set of some documents with 12 words in each documents. Let us divide the documents in D into three topics such as .The table below shows the topic and word distributions in the set of documents. The topic representation using word distribution and the document representation using topic distribution are the most important contributions provided by the LDA model. The topic representation indicates which words are important to which topic and the document representation indicates which topics are important for a particular document. Given a collection of documents, the LDA can learn topics and decompose the documents according to the topics. Furthermore, for a new incoming document, various methods can be utilized to situate its content in terms of the trained topics. However, single word based topic representations contain ambiguous semantics. Thus, TNG improves the LDA model by expanding word-based topic representation to phrase-based, which enhances the explicit semantics of topics. However, TNG suffers from the low occurrence problem and fails to significantly improve the LDA model.

In this paper, we propose a new approach for generating a pattern-based topic model to represent documents and also a new ranking method to determine relevant docu-ments based on the topic model.

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: firstly, construct a new transactional dataset from the LDA model results

GENERATE PATTERN ENCHANCED REPRESENTATION

Here the frequent patterns that are generated in each transaction dataset are taken. For a given minimal support threshold σ, an itemset X in T1 is frequent if that the supp(X) supp(X) >= σ, , where supp(X) is the support of X that is the number of transactions in T1 that contain X. The frequency of item set X is defined supp(X)/|Tj| for topic zj. . However, 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.

As a result, documents cannot be accurately represented by these topic representations. That means, these pattern-based topic representations which represent user interests may not be sufficient or accurate enough to be directly used to determine the rele-vance of new documents to the user interests. In this sec-tion, one novel IF model, MPBTM, is proposed based on the pattern enhanced topic representations. The proposed the relevance of incoming documents based on Maximum Matched Pat-terns, which are the most distinctive and representative patterns, as proposed in this paper. The details are described in the following subsections. model consists of topic distributions describing topic preferences of documents or a document collection and structured pattern-based topic representations represent-ing the semantic meaning of topics in a document. More-over, the proposed model estimates

PATTERN EQUIVALANCE CLASS

Normally, the number of frequent patterns is considerably large and many of them are not necessarily useful. Several concise patterns l have been proposed to represent useful patterns generated from a large dataset instead of frequent patterns such as maximal patterns and closed patterns The number of these concise patterns is significantly smaller than the number of frequent patterns for a dataset .In particular the closed patterns drawn great attention due to attractive features.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

nCORETech 2017

LBS College of Engineering, Kasaragod

Vol. 6, Special Issue 3, March 2017

Topic	Z_1		Z_2		Z_3	
Document	$\theta_{d,1}$	words	$\theta_{d,2}$	words	$\theta_{d,3}$	words
d_1	0.6	w_1, w_2, w_3, w_2, w_1	0.2	w_1, w_9, w_8	0.2	w_7, w_{10}, w_{10}
d_2	0.2	w_2, w_4, w_4	0.5	w_7, w_8, w_1, w_8, w_8	0.3	w_1, w_{11}, w_{12}
d_3	0.3	w_2, w_1, w_7, w_3	0.5	w_7, w_3, w_3, w_2	0.4	w_6, w_7, w_{10}, w_{11}
d_4	0.3	w_2, w_7, w_8	0.4	w_9, w_8, w_1	0.3	w_1, w_{11}, w_{10}

transaction	topic document transaction
1	$\{w_1, w_8, w_9\}$
2	$\{w_1, w_7, w_8\}$
3	$\{w_2, w_3, w_7\}$
4	$\{w_1, w_8, w_9\}$

Patterns	*supp*
$\{w_1\}, \{w_8\}, \{w_1, w_8\}$	3
$\{w_9\}, \{w_7\}\{w_8, w_9\}, \{w_1, w_9\}, \{w_1, w_8, w_9\}$	2

PROPOSED SYSTEM METHODOLOGY

A predominant assumption for these methods is that the records within the collection are all about one topic. Nevertheless, definitely customers' pursuits will also be numerous and the records within the assortment most commonly contain more than one topic. Topic modeling, reminiscent of Latent Dirichlet Allocation (LDA), was proposed to generate statistical units to symbolize more than one topic matters in a collection of files, and this has been commonly utilized in the fields of desktop learning and information retrieval, and so forth. But its effectiveness in expertise filtering has not been so well explored. Patterns are perpetually idea to be more discriminative than single phrases for describing documents.

To care for the above acknowledged barriers and problems, in this paper, a novel information filtering mannequin, maximum matched pattern-centered topic model (MPBTM), is proposed. The most important uncommon aspects of the proposed model incorporate: (1) user expertise wishes are generated in phrases of multiple topic; (2) each and every topic is represented via patterns; (3) patterns are generated from topic models and are prepared in phrases of their statistical and taxonomic points; and (four) essentially the most discriminative and consultant patterns, known as maximum Matched Patterns, are proposed to estimate the report relevance to the user's knowledge desires with a view to clear out beside the point files. So as to alleviate the paradox of the topic representations in LDA, in [13], we proposed a promising solution to meaningfully symbolize topics through patterns alternatively than single words via combining topic units with sample mining methods. Above all, the patterns are generated from the phrases within the word-established topic representations of a typical topic mannequin such because the LDA model. This ensures that the patterns can good characterize the issues because these patterns are constituted of the phrases that are extracted by LDA founded on pattern occurrence and co-incidence of the phrases in the files. The pattern based topic mannequin, which has been utilized in IF [14], may also be regarded as a "post-LDA" model in the sense that the patterns are generated from the topic representations of the LDA model.

On account that patterns can characterize extra distinctive meanings than single phrases, the sample-situated matter models can be used to symbolize the semantic content of the person's documents more effectively when compared with the word-founded topic models. However, very most likely the quantity of patterns in some of the topic matters can also be huge and many of the patterns are usually not discriminative enough to represent distinct topics.
In this paper, we advocate to prefer essentially the most consultant and discriminative patterns, which are called maximum matched Patterns, to represent topic matters alternatively of making use of customary patterns. A new topic

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

nCORETech 2017

LBS College of Engineering, Kasaragod

Vol. 6, Special Issue 3, March 2017

model, called MPBTM is proposed for report illustration and record relevance rating. The patterns in the MPBTM are well structured in order that the maximum matched patterns will also be successfully and quite simply selected and used to symbolize and rank records. The usual contributions of the proposed MPBTM to the discipline of IF will also be described as follows:

1) We propose to model users' interest with multiple topics rather than a single topic under the assumption that users' information interests can be diverse.

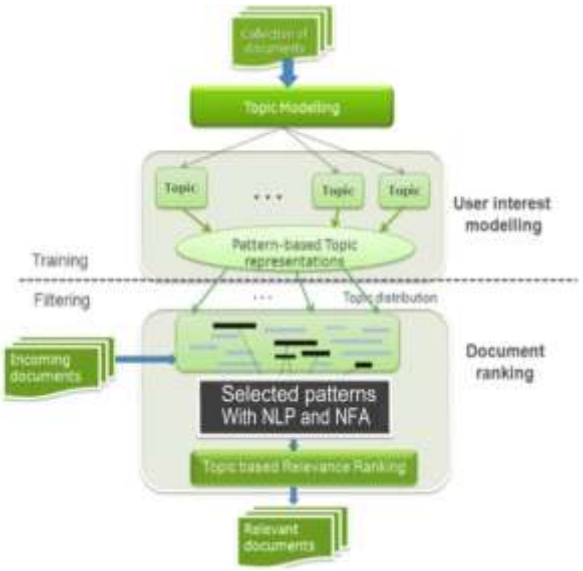
2) We propose to integrate data mining techniques with statistical topic modeling techniques to generate a pattern-based topic model to represent documents and document collections. The proposed model MPBTM consists of topic distributions describing topic preferences of each document or the document collection and pattern-based topic representations representing the semantic meaning of each topic.

3) 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 user's interest. The maximum matched patterns are the most representative and discriminative patterns to determine the relevance of incoming documents

PATTERN BASED TOPIC MODEL

A two-stage approach is proposed to combine the statistical topic modeling process with the classical knowledge mining approaches, with the hope of improving the accuracy of topic modeling in significant record collections. In stage one, the most recognized topic modeling process, Latent Dirichlet Allocation (LDA), is used to generate preliminary topic models. In stage two, the most popularly used term weighting procedure and the established pattern mining system are used to derive extra discriminative phrases and patterns to symbolize topics of the collections. Moreover, the common patterns reveal structural expertise concerning the associations between phrases that make issues more comprehensible, semantically central and canopy broader meanings.



USER INTEREST MODELING

For a collection of documents D, the user's interests can be represented by the patterns in the topics of D. represents the topic distribution of D and can be used to represent the user's topic interest distribution, , and V is the number of topics. In this paper, the topic distribution in the collection D is defined as the average of the topic distributions documents in D, The probability Distribution of topics in u_D represents the degree of interest that the user has in these topics.

By using the methods described in Section 4, for a document collection D and V pre-specified latent topics, from the results of LDA to D, V transactional datasets, G_1; . . . ; G_V can be generated from which the pattern-based topic

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

nCORETech 2017

LBS College of Engineering, Kasaragod

Vol. 6, Special Issue 3, March 2017

representations for the collection, U ¼ fX$_{Z1}$; X$_{Z2}$; . . . ; X$_{ZV}$ g, can be generated, where each X$_{Zi}$ ¼ fX$_{i1}$; X$_{i2}$; . . . ; X$_{imi}$ g is a set of frequent patterns generated from transactional dataset G$_i$. U is considered the user interest model, the patterns in each X$_{Zi}$ represent what the user is interested in terms of topic Z$_i$.

As mentioned before, normally, the number of fre-quent patterns generated from a dataset can be huge and many of them may be not useful. A closed pattern reveals the largest range of the associated terms. It covers all the information that its subsets describe. Closed patterns are more effective and efficient to represent topics than fre-quent patterns. However, only using closed patterns to represent topics may impact the effectiveness of docu-ment filtering since closed patterns often may not exist in new incoming documents. On the other hand, frequent patterns can be well organized into groups based on their statistics and coverage. equivalence class is a useful structure which collects the frequent patterns with the same frequency into one group. The statistical significance of the patterns in one equivalence class is the same. This distinctive feature of equivalence classes can make the patterns more effectively used in document filtering. In this paper, we propose to use equivalence classes to represent topics instead of using frequent patterns or closed patterns.

Assume that there are n$_i$ frequent closed patterns in X$_{Zi}$, which are c$_{i1}$; . . . ; c$_{ini}$, and that X$_{Zi}$ can be partitioned into n equivalence classes,. For simplicity, the equivalence classes are denoted as or simply for topic denote the set of equivalence classes for topic In the model MPBTM,

ALGORITHM

Algorithm 1. *User Profiling*

Input: a collection of positive training documents D;
minimum support σ_j as threshold for topic Z_j;
number of topics V

Output: $U_E = \{\mathbb{E}(Z_1), \ldots, \mathbb{E}(Z_V)\}$

1: Generate topic representation ϕ and word-topic assignment $z_{d,i}$ by applying LDA to D
2: $U_E := \emptyset$
3: **for** each topic $Z_j \in [Z_1, Z_V]$ **do**
4: Construct transactional dataset Γ_j based on ϕ and $z_{d,i}$
5: Construct user interest model $\mathbf{X}_{Z_j}$ for topic Z_j using a pattern mining technique so that for each pattern X in $\mathbf{X}_{Z_j}$, $supp(X) > \sigma_j$
6: Construct equivalence class $\mathbb{E}(Z_j)$ from $\mathbf{X}_{Z_j}$
7: $U_E := U_E \cup \{\mathbb{E}(Z_j)\}$
8: **end for**

Algorithm 4 *Document Filtering_F*

Input: a list of incoming document D_{in}
Output: $rank_F(d), d \in D_{in}$

1: Call *User Profiling* to construct $U_F := \{\mathbf{X}_{Z_1}, \mathbf{X}_{Z_2}, \cdots, \mathbf{X}_{Z_V}\}$
2: $rank'(d) := 0$
3: **for** each $d \in D_{in}$ **do**
4: **for** each topic $Z_j \in [Z_1, Z_V]$ **do**
5: Scan $\mathbf{X}_{Z_j}$ and find frequent pattern X^d_{jk} which exists in d
6: update $rank_F(d)$ using Equation 4.3:
7: $rank_F(d) := rank'(d) + |X^d_{jk}|^{0.5} \times f_{jk} \times \vartheta_{D,j}$
8: $rank'(d) := rank_F(d)$
9: **end for**
10: **end for**

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319-5940
International Journal of Advanced Research in Computer and Communication Engineering
nCORETech 2017
LBS College of Engineering, Kasaragod
Vol. 6, Special Issue 3, March 2017

Topic-based Document Relevance Ranking
In terms of the statistical significance, all the patterns in one equivalence class are the same. The differences among them are their size. If a longer pattern and a shorter pattern from the same equivalence class appear in a document simultaneously, the shorter one becomes insignificant since it is covered by the longer one and it has the same statistical significance as the longer one.

In the filtering stage, document relevance is estimated to filter out irrelevant documents based on the user's information needs. In this paper, for a new incoming document d, the basic way to determine the relevance of d to the user interests is firstly to identify maximum patterns in d which match some patterns in the topic-based user interest model and then estimate the relevance of d based on the user's topic interest distributions and the significance of the matched patterns.

RANKING TASK METHODS
1).NAIVE BAYES METHODS
2).kNN
3)SVM

PATTERN MINING
Sequential pattern mining is a topic of data mining concerned with finding statistically relevant patterns between data examples where the values are delivered in a sequence. It is usually presumed that the values are discrete and thus time series mining is closely related, but usually considered a different activity . sequential pattern mining is a special case of structured data mining

MINING TECHNIQUES
1)A priori algorithm
2) prefix span
3)FP-Tree

Comparisons with Pattern-based Models
The comparison results among the proposed model and pattern-based baseline models are in the middle part of Table 6. We can see that all the three pattern-based topic modeling models, i.e. MPBTM, PBTM_FCP and PBTM_FP, outperform the three pattern-based baseline models, i.e. SCP, n-Gram, and FCP, which clearly shows the strength obtained by combining topic modeling with pattern-based models. Among the three baseline models, the SCP outperforms the other two models for b=p, MAP and F_1, while the FCP model performs the best for top20. The bottom line of the pattern-based part in the table provides the percentage of improvement achieved by the MPBTM against the SCP for b/p, MAP and F_1, and against the FCP model for top20. The MPBTM achieves excellent performance in improvement percentage with a maximum of 32.3 percent and a minimum of 17.9 percent.

Comparisons with Term-based Models
From the bottom section of Table 6, we can see that the SVM achieved better performance than the BM25, while the MPBTM and the PBTM_FCP and the PBTM_FP consist outperform the SVM. The maximum and minimum improvement achieved by the MPBTM against the SVM is 23.5 and 9.3 percent, respectively.

We also conducted the T-test to compare the MPBTM with all other PBTM models and baseline models. The results are listed in Table 7. The statistical results indicate that the proposed MPBTM significantly outperforms all the other models (all values in Table 7 are less than 0.05) and the improvements are consistent on all four measures. Therefore, we conclude that the MPBTM is an exciting achievement in discovering high-quality features in text documents mainly because it represents the text documents not only using the topic distributions at a general level but also using hierarchical pattern representations at a detailed specific level, both of which contribute to the accurate docu-ment relevance ranking.

EXPERIMENTS AND FUTURE WORK

I introducing a new method for relevance ranking to include a naive bayes algorithm for ranking relevant and non relevant algorithm .here we will analyses the user interest models and search the entire search database to know users interest by talking a probability of currently available information. The algorithm is best for text clustering methods so it is very use full one for the relevance ranking method. We are introducing a technique for pattern mining it is apriori algorithm which is very efficient because the last set is selected through this algorithm is very simple and it is effective for identify the document and for selecting topic.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

nCORETech 2017

LBS College of Engineering, Kasaragod

Vol. 6, Special Issue 3, March 2017

CONCLUSION

This thesis begins by using combating a protracted-standing task in internet know-how, which is expertise overload. Working out customers' actual information desires can support us distinguish most valuable know-how from enormous quantities of non-imperative know-how. We for that reason gave main emphasis to searching for superior models to appropriately mannequin underlying structure for customers' pursuits. And utilizing the optimized user curiosity modeling to extract the relevance of records and ranking essentially the most important documents at high via constructing relevance ranking process.

The user interest modeling in IF combines the statistical items with semantic function representations, which outlines the user's pursuits with distribution of topic matters at a basic level as well as interpretable features at element degree. On relevance rating system, established patterns and closed patterns for the PBTM model, the proposed significantly matched patterns and maximum matched patterns for the StPBTM model, are selected to represent the relevance of documents.

REFERENCES

[1] N.J. Belkin, W. Bruce Croft, "Information filtering and information retrieval: Two sides of the same coin?" Special issue on information filtering. ACM transcation, vol-35, issue-12, pp: 29-38 (1992).
[2] S.Deerwaster, S. Dumas, G.Furnas, T. Landauer, R. Harsman, "Indexing by Latent Semantic analysis". Journal of the American Society of Information Science, vol. 41, pp. 391-407 (1990).
[3] M. W. Berry, S. T. Dumais, G. W.O" Brein, "Using linear algebra for intelligent information retrieval". SIAM Review, vol. 37, pp. 573-595 (1995).
[4] T.Kolda, D. O"Leary, "A semi-discrete matrix decomposition for latent semantic indexing in information retrieval". ACM Trans.Inform. Systems, vol. 16, pp. 322- 346 (1998).
[5] B. T. Bartell, G.W. Cottrell, R.K. Belew, "Latent Semantic Indexing is an Optimal Special Case of Multidimensional Scaling". SIGIR, pp. 161-167 (1992).
[6] C.H.Q. Ding, "A Similarity-based Probability Model for Latent Semantic Indexing". SIGIR, pp.58-65 (1999).