ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 7, July 2016

Similar Document Retrieval using Pattern-Based Topic Modelling for Information Filtering

Nayana M.K. [1], Sabeeha K.P. [2]

P.G. Student, Department of Computer Engineering, M E A Engineering College, Perinthalmanna, Kerala, India[1]

Assistant Professor, Department of Computer Engineering, M E A Engg. College, Perinthalmanna, Kerala, India[2]

ABSTRACT: Topic modelling is widely accepted in the areas of machine learning and text mining, etc. It was proposed to generate statistical models to classify multiple topics in a collection of documents. Where each topic is represented by a distribution of words. The word-based or term-based topic representations may not be able to semantically represent documents. Patterns are always more discriminative than single terms for representing documents. In this work, we propose to combine the statistical topic modelling with pattern mining techniques to generate pattern-based topic models with the purpose of enhancing the semantic representations of the traditional word-based topic models. Utilizing the proposed pattern-based topic model, users' interests can be modelled with multiple topics and each of which is represented with semantically rich patterns. A novel information filtering model is proposed here. In information filtering model user information needs are created in terms of multiple topics where each topic is represented by patterns.

KEYWORDS: Topic Model, User Interest Modelling**,** Pattern Mining, Information Filtering.

I. INTRODUCTION

The user interest modelling is a process to understand the user's information needs based on the most relevant information that can be found and delivered to the user. In order to extract precise user's interests, traditionally, many term-based approaches are used due to their efficient computational performance, as well as mature theories for term weighting, such as Rocchio, BM25, etc. But term-based features suffer from the problems of polysemy and synonymy. Phrase based approaches are more discriminative and should carry certain semantic meaning. However, the performance of using phrases in real applications is discouraging. To overcome the limitations of term-based and phrase-based approaches, pattern mining - based techniques have used patterns to represent users' interest and achieved some improvements on effectiveness, since patterns contain both underlying semantic meaning and high frequency of occurrences in documents.

Topic modelling [1] is one of the most popular probabilistic text modelling techniques, and has been quickly accepted by machine learning and text mining communities. The most inspiring contribution of topic modelling is that it automatically classifies documents in a collection by a number of topics and represents every document with multiple topics and their corresponding distribution. The topic-based representation generated by using topic modelling can conquer the problem of semantic confusion compared with the traditional text mining techniques. The representation by single words with probabilistic distributions breaks the relationships between associated words. Therefore, topic modelling needs improved modelling users' interests in terms of topics' interpretations. In this work, a pattern-based topic model is proposed to enhance the semantic interpretations of topics. This work focuses on how the proposed pattern-based topic model can be used in the area of information filtering (IF) for constructing content-based user interest modelling.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 7, July 2016

The pattern based topic model can be considered as a "post-LDA" model because here patterns are constructed from the topic representation of the LDA (Latent Dirichlet Allocation) [2] model. When we are comparing pattern-based topic models with the word-based topic models we can analyse that the pattern-based topic model can be used to represent the semantic content of the user's documents more accurately. However, patterns in some topics can be huge and some patterns are not discriminative enough to represent specific topics.

Information filtering (IF) is a system to remove redundant or unwanted information from an information or document stream based on document representations which represent users' 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. Once the user's profiles are collected, in this thesis, we focus on modelling user's interests with multiple topics. By utilising the classical topic models, user's interests can be represented by a pre-defined number of topics, each of which is represented by words and their distribution. In this work, 'relevance' of a document refers to the relevance between the user's interests and the document. Suppose the user's interests are well represented with pattern-based topics. Since very often 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, we will propose relevance ranking modelling methods for document representation and relevance estimation. The topical patterns for document modelling should be effectively selected.

In this work, to represent topics instead of using frequent patterns we proposed to select the most representative and discriminative patterns, which are called Maximum matched Patterns. A new topic model, called Maximum matched Pattern Based Topic Modelling (MPBTM) is proposed for document representation and document relevance ranking. The patterns in the MPBTM contained patterns are well structured so that the maximum matched patterns can be efficiently selected and used to represent and rank documents.

II. RELATED WORK

Topic modelling algorithms are used to analyze the words of the original texts to discover the themes that run through them, how those themes are connected to each other, and how they change over time. Probabilistic latent semantic analysis (PLSA) [3], is a technique for the analysis of two-mode and co-occurrence data. The probabilistic Latent Semantic Indexing model, which was introduced by Hoffman quickly gained acceptance in a number of text modelling applications. PLSI, also called an aspect model, is a latent variable model for general co-occurrence data which associates an unobserved class (topic) variable with each observation (i.e., with each occurrence of a word). The PLSI model has a problem in that its generative semantics are not well-defined. Therefore there is no natural way to predict a previously unseen document, and the number of parameters of PLSI grows linearly with the number of training documents, which makes the model susceptible to over fitting.

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.

Correlated Topic Model (CTM) [4] is a type of statistical model used in natural language processing and machine learning. They are used to find out the topics that shown in a group of documents. The key for CTM is the logistic normal distribution. It is a novel topic model that extend from LDA which directly model the correlation between topics. Using of logistic normal distribution to create relations among topics. But CTM requires lots of calculation and it having lots of general words inside the topics.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 7, July 2016

Yang Gao, Yue Xu and Yuefung Li, 2015, [5], proposes a two-stage model for modelling the documents in a collections. One of the main discriminative feature of this model is, it combines the data mining techniques to statistical topic modelling to generate discriminative and pattern based representations for modelling 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 modelling method LDA. Another important feature of this representation is, patterns carry more structural and inner relationship within each topic.

H. D. Kim, D. H. Park, Y. Lu, and C. Zhai, proposed Enriching text representation with frequent pattern mining for probabilistic topic modeling [6]. 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 modelling model such as LDA. The resulting topic representations contain both individual words and pre-generated 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.

The work effective pattern discovery for text mining is proposed by N. Zhong, Y. Li, and S.-T. Wu,. 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 [7]. 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.

III. PROPOSED SYSTEM

Pattern carry more meaning than single words. Pattern based representation considered more meaningful and accurate to represent topics than single word based representation. Moreover, pattern based representation gives the structural information, that can reveal the association between words. So pattern-based LDA is an effective approach. A new topic model, called Maximum matched Pattern Based Topic Modelling (MPBTM) is proposed for document relevance ranking.

- Maximum matched Pattern-Based Topic Model (MPBTM)

Maximum Matched Pattern-based Topic Model [8] consists of topic distributions describing topic preferences of each document or the document collection and pattern-based topic representations presenting the semantic meaning of each topic. There are mainly two phases in this model. First one is Document training phase and second one is Document filtering phase. During the document training phase user interest modelling is done. Four steps are proposed to generate the Topic based user interest model. First Topic modelling algorithm named LDA applying to each documents. LDA automatically classifies documents into number of topics and each topics contain number of words based on their probability. Next construct a new transactional dataset from the result of LDA, which removes duplicate words. The resultant transactional dataset is the input to the pattern mining algorithm. Next mine frequent patterns by using efficient pattern mining algorithm. Pattern carry more information than single words. In the filtering stage, incoming documents are passes through topic modelling, pattern mining and finally the MPBTM selects maximum matched patterns, instead of using all discovered patterns. Then compare the incoming document pattern with training documents pattern. From that we can find out Maximum matched patterns and which is used for estimating relevance of incoming documents. Figure 1 represent the overall structure of proposed system.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 7, July 2016

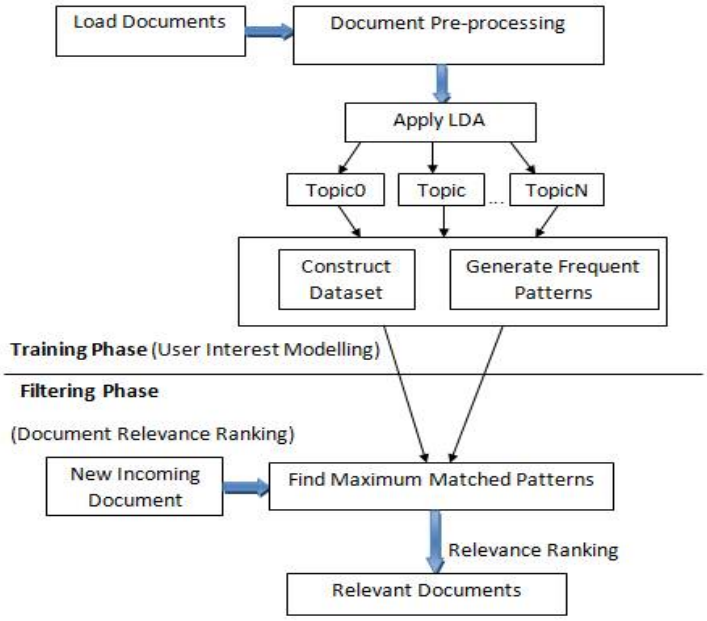

Figure 1: Maximum matched Pattern-based Topic Model (MPBTM)

IV. EXPERIMENTAL RESULTS

Below graph shows the performance of existing system and the proposed system based on the parameters precision, recall and F1-Score. From this graphical representation it is clear that proposed system perform better than other system.

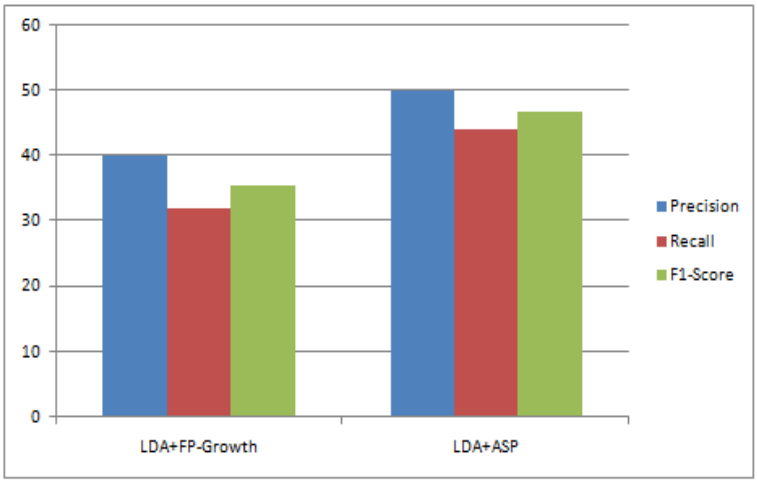

Figure 2: Comparison

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 7, July 2016

V. CONCLUSION

Topic modelling is one of the most popular probabilistic text modelling techniques. The topics that are discovered during topic modelling are represented by the distribution of words. Its importance is quickly accepted by machine learning and text mining communities. They are very useful in information filtering, document ranking, content-based feature extraction and modelling tasks, such as information retrieval and recommendations. A basic assumption for these approaches is that the documents in the collection are all about single topic. But in reality this is not necessarily the case. Patterns can carry more semantic meaning than single words. Here topic modelling in the field of information filtering is mainly considered. The Maximum matched Pattern Based Topic Modelling (MPBTM) can consider multiple topics within a document. The patterns in the MPBTM are well structured so that the maximum matched patterns can be efficiently selected and used to represent and rank documents.

REFERENCES

[1] M. Steyvers and T. Griths, “Probabilistic topic models," Handbook of latent semantic analysis, p. 424-440, January 2007.
[2] Blei, D. M. and M. I. Jordan, “Latent Dirichlet Allocation," the Journal of machine Learning research, p. 993-1022, 2013.
[3] T. Hofmann, “Probabilistic latent semantic indexing," p. 50-57, 1999.
[4] D. M. Blei and J. D. Lafferty, “A correlated topic model of science," The Annals of Applied Statistics, vol. 1, p. 17-35, 2007.
[5] B. Xu, Yue, “A two-stage approach for generating topic models," Advances in Knowledge Discovery and Data Mining, 2012.
[6] Hyun Duk Kim, Dae Hoon Park, “Enriching text representation with frequent pattern mining for probabilistic topic modeling," October 2012.
[7] S. S. Rohini Y. Thombare, “Effective pattern deploying approach in pattern taxonomy model for text mining," International Journal of Computer Applications, p. 0975 -8887, 2013.
[8] Yang Gao, Yue Xu, Yuefeng Li,“Pattern-based Topic Models for Information Filtering,” In Cambria, Erik & Chen, Ping (Eds.) Sentiment Elicitation from Natural Text for Information Retrieval and Extraction (SENTIRE), 7 December 2013.