Enhanced Frequent Itemsets Based on Topic Modeling in Information Filtering

Than Than Wai, University of Computer Studies, Mandalay, Myanmar
Sint Sint Aung, University of Computer Studies, Mandalay, Myanmar

ABSTRACT

In order to generate user’s information needs from a collection of documents, many term-based and pattern-based approaches have been used in Information Filtering. In these approaches, the documents in the collection are all about one topic. However, user’s interests can be diverse and the documents in the collection often involve multiple topics. Topic modeling is useful for the area of machine learning and text mining. It generates models to discover the hidden multiple topics in a collection of documents and each of these topics are presented by distribution of words. But its effectiveness in information filtering has not been so well explored. Patterns are always thought to be more discriminative than single terms for describing documents. The major challenge found in frequent pattern mining is a large number of result patterns. As the minimum threshold becomes lower, an exponentially large number of patterns are generated. To deal with the above mentioned limitations and problems, in this paper, a novel information filtering model, EFITM (Enhanced Frequent Itemsets based on Topic Model) model is proposed. Experimental results using the CRANFIELD dataset for the task of information filtering show that the proposed model outperforms over state-of-the-art models.

KEYWORDS

Information Filtering, Pattern Mining, Text Documents, Topic Modeling, User Interest Modeling

INTRODUCTION

Information filtering 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. Term based approach, one of the IF model, is efficient in computational performance such as BM25, Racchio, etc (Beil et al., 2002; Robertson et al., 2004). But, term-based document representations suffer from the problems of polysemy and synonymy. To overcome the limitation of term based approach, pattern mining technique is used (Bastide et al., 2000; Cheng et al., 2007). Patterns carry more semantic meaning than term. Pattern mining algorithms depends on developing data mining algorithms to find out interesting, surprising and functional pattern in databases. Pattern mining algorithms can be applied on various types of data such as transactional databases, sequence databases, streams, spatial data, graphs, etc. The goal is to discover all patterns whose frequency in the basis dataset exceeds a user specified threshold. Database model filtering that helps you to create mining models that use subset of data in a mining structure. Pattern based topic filtering used to filter out the irrelevant document and gives relevant document from the collection of documents (Vishnu, 2016). The number of test cases are reduced in order to minimize the time and cost of executing them. Sever techniques can be used to reduce

DOI: 10.4018/IJSI.2017100103

Copyright © 2017, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.
test cases such as information retrieval, data mining and pairwise testing. Data mining approach are used, mainly because of the ability of data mining to extract patterns of test cases that are invisible (Saifan et al., 2016). Some work (Néji et al., 2014) focuses on the problem of Information Retrieval System (IRS) that integrates the human emotion recognition to recognize the degree of satisfaction of the user for the result found through its facial expression, its physiological state, its gestures and its voice. They proposed an algorithm for recognizing the emotional state of a user during a search session in order to issue the relevant documents that the user need and also presented the architecture of the envisaged system and the organizational model. Topic modeling (Blei & Jordan, 2003; Blei & Wang, 2011; Croft & Wei, 2006) is one of the text modeling techniques. It can automatically classify documents into topics and represent every document with multiple topics and their corresponding distribution. Two representative approaches are PLSA (Hofmann, 1999) and LDA (Blei & Jordan, 2003). The topic model contains cluster of words with similar meanings and text, it contains different terms of topic modeling. It also includes model topics with taking into account time based on user interest model and it will cofound the topic discovery. Further it has been mentioned in some of the applications that have been in these methods (Vishnupriya, 2015). The comparison of different topic model features is essential to design a new proposal for information filtering based on user interest model. All of these models considers the time as a most vital factor. Directly applying topic models for IF, two problems are generated. Due to limited number of dimensions to represent documents, the two problems are occurred. First, topic distribution is insufficient. Second, represent documents in word based topic have different semantic content. To overcome these problem patterns enhanced LDA (Gao et al., 2015) is used. It carries more concrete and identifiable meaning than word based representations using LDA (Blei & Jordan, 2003). 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 (Gao et al., 2015) 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. But, the number of patterns in some of the topics can be huge to represent specific topics. The main distinctive features of the proposed model are as follows:

1. User information needs are generated in terms of multiple topics,
2. Each topic is indicated by patterns,
3. Patterns are generated from topics models,
4. The resulting enormous amount of patterns are optimized in order to effect in mining process and to estimate the document relevance to the user.

The paper is organized as follows. We discuss the related work about some state-of-the-art IF models and related techniques. Next, we describe a brief introduction to the background works of LDA and present the details of our proposed model and algorithms. We show data and the experimental results. Finally, we discuss the proposed model and also conclude the whole work.

RELATED WORK

Information filtering system gets user interest or user information needs based on the ‘user profiles’. Information filtering systems expose users to the information that is more relevant to them (Mostafa, 1997). As a classification task or a ranking task, the document filtering can be regarded. Naive Bayes, kNN and SVM are methods that assign binary decisions to documents (relevant or irrelevant) as a special type of classification (Cao et al., 2006). The relevance of a document can be modeled by various approaches that include a term-based model (Beil et al., 2002), a pattern-based model (Wu
Service-Driven Approaches to Software Architecture: Principles and Methodology
Raja Ramanathan (2013). *Service-Driven Approaches to Architecture and Enterprise Integration* (pp. 1-42).
[www.igi-global.com/chapter/service-driven-approaches-software-architecture/77943?camid=4v1a](www.igi-global.com/chapter/service-driven-approaches-software-architecture/77943?camid=4v1a)

Access Control in Service Compositions: Challenges and Solution Architecture
Aurélien Faravelon and Stéphanie Chollet (2013). *Service-Driven Approaches to Architecture and Enterprise Integration* (pp. 165-187).
[www.igi-global.com/chapter/access-control-service-compositions/77949?camid=4v1a](www.igi-global.com/chapter/access-control-service-compositions/77949?camid=4v1a)