An Approach to Improve Generation of Association Rules in Order to Be Used in Recommenders

Hodjat Hamidi, Department of Information Technology, Faculty of Industrial Engineering, K. N. Toosi University of Technology, Tehran, Iran
Elnaz Hashemzadeh, Department of Information Technology, Faculty of Industrial Engineering, K. N. Toosi University of Technology, Tehran, Iran

ABSTRACT

This paper introduces a method for improving the association rules that will also help improve the performance of recommender systems. Combining sampling and parallelism in the process, the proposed method, in addition to help perform the process more quickly, better quality rules will be generated. The proposed method and the dataset will be segmented into a number of smaller sections based on clustering items, and each section is separately sampled. Frequent itemsets and association rules in any section of the selected samples will be found and by gathering and analyzing the results, the quality and number of rules will be evaluated. One of the innovations of this article is the method of sampling of the database. Instead of random sampling, information about the best users and items was isolated. After determining the precise amount of support, the authors extracted the frequent and favorable rules from the selected sample.

KEYWORDS
Accuracy of the Rules, Frequent Itemsets, Identification, Sampling

1. INTRODUCTION

One important problem in data mining is the discovery of frequent itemsets in a transactional database (Leung, Chan, & Chung, 2006). Frequent itemset mining is a traditional and important problem in data mining. An itemset is frequent if its support is not less than a minimum support specified by users. Traditional frequent itemset mining approaches have mainly considered the problem of mining static transaction databases. In these methods, transactions are stored in secondary storage so that multiple scans over the data can be performed. Frequent patterns, such as frequent itemsets, substructures, sequences term-sets, phrase sets, and sub graphs, generally exist in real-world databases. Identifying frequent itemsets is one of the most important issues faced by the knowledge discovery and data mining community. Frequent itemset mining plays an important role in several data mining fields as association rules (Chandak, Girase, & Mukhopadhyay, 2015; Mousavi et al., 2017; (Bimonte et al., 2017; Hamidi et al., 2016), warehousing (Daraei et al., 2016; Hamidi, 2011, 2012, 2009, 2010, 2011, 2017), correlations, clustering of high-dimensional biological data, and classification (Lin et al., 2002; Hashemzadeh et al., 2016; Mohammadi et al., 2005).

DOI: 10.4018/IJDWM.2017100101

Copyright © 2017, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.
Three kinds of frequent itemset mining approaches over static databases have been proposed: reading-based (Kotsiantis et al., 2006; Moradi et al., 2017; Kumar et al., 2016), writing-based (Esposito et al., 2016; Bimonte et al., 2017), and pointer-based (Ashrafi, 2004; 2005; 2007). Tjoe, 2005, presented a comprehensive survey of frequent itemset mining and discussed research directions. Frequent Itemsets are huge when the given threshold is low; consequently, the condensed representations of frequent itemsets including closed frequent itemsets (Daraei et al., 2017), free frequent itemsets (Tantar et al., 2008), approximate frequent k-sets (Ashrafi et al., 2004), weighted frequent itemsets (Vafaei et al., 2009; 2010; 2012), and non-derivable frequent itemsets (Chen et al., 2002; Esposito et al., 2016; Hamidi et al., 2005; 2006) were proposed; in addition, (Brönnimann, et al., 2003; Hamidi, 2016; Chevers et al., 2016) focused on discovering a minimal set of unexpected itemsets. The use of sampling in identifying frequent itemsets has been popular since inception of association rule mining paradigm. A representative list of the initial work in this direction can be seen in (Chen et al., 2002; Monadjemi et al., 2012; Liu et al., 2016; Shadloo et al., 2016; Wu et al., 2016). Their main focus was on experimentally studying the effectiveness of sampling at different sample sizes. A major limitation, however, was that they considered only small databases, whereas the true power and benefits of sampling come to the fore with very large databases. Implementation of traditional association rule mining algorithms requires several passages of the database, which may take hours or even days, and this will be even more intense in the future. Seeking to resolve this issue, many studies have been performed to improve association rule mining. The major finding in these studies is presenting ways to enhance the efficiency of finding frequent patterns from large datasets, using limits to determine better models, and the elimination of irrelevant association rules using some criteria (Umaran et al., 2009; Ye et al., 2016; Safdar et al., 2016). One of the innovations of this article is the method of sampling of the database. Instead of random sampling, information about the best users and items was isolated. After determining the precise amount of support, we extracted the frequent and favorable rules from the selected sample.

The remaining part of the paper is organized as follows. The literature review of the subject is presented in the second section. An explanation of the frequent itemsets and the proposed method will be presented in the third and fourth section. In the fifth and sixth sections of the article, the implementation is discussed and the results will be evaluated. Finally, the conclusion will be provided

2. RELATED WORK

Existing methods for constructing test databases generally focus on generating data by combining the database schema definition with conflict resolution, disclosure assessment and data perturbation methods (Chen et al., 2002; Johnson et al., 2016; Kakar et al., 2016) and randomized functions for populating testing environments (Zaki, Parthasarathy, Li, & Ogihara, 1997). In (Jannach et al., 2011), the authors focus on techniques to optimize the generation of test databases by using sequential and parallel algorithms given the statistical distribution of the underlying data. Other generic tools generate synthetic data using the schema only (Choi, Yoo, Kim, & Suh, 2012).

Shen et al., 2010, have introduced FAST (Finding Associations from Sampled Transactions), a two-phase sampling-based algorithm for discovering association rules in large databases. In Phase I, a large initial sample of transactions was collected and used to quickly and accurately estimate the support of each individual item in the database. In Phase II these estimated supports were used to either trim “outlier” transactions or select “representative” transactions from the initial sample, thereby forming a small final sample that more accurately reflects the statistical characteristics of the entire database. In an empirical study, FAST was able to achieve 90-95% accuracy using a final sample having a size of only 15-33% of that of a comparable random sample. The sampling technique can be used in conjunction with almost any standard association-rule algorithm, and can potentially render scalable other algorithms that mine “count” data.
Related Content

Data Mining and the Project Management Environment
Emanuel Camilleri (2010). Data Mining in Public and Private Sectors: Organizational and Government Applications (pp. 337-357).
www.igi-global.com/chapter/data-mining-project-management-environment/44296?camid=4v1a

SQL-Based Fuzzy Query Mechanism Over Encrypted Database
www.igi-global.com/article/sql-based-fuzzy-query-mechanism-over-encrypted-database/117159?camid=4v1a
Information Privacy and Security for e-CRM
Jerry Fjermestad and Nicholas C. Romano (2009). Social Implications of Data Mining and Information Privacy: Interdisciplinary Frameworks and Solutions (pp. 48-70).
www.igi-global.com/chapter/information-privacy-security-crm/29144?camid=4v1a

PSSRC: A Web Service Registration Cloud Based on Structured P2P and Semantics
Qian He, Baokang Zhao, Liang Chang, Jinshu Su and Ilsun You (2016). International Journal of Data Warehousing and Mining (pp. 21-38).
www.igi-global.com/article/pssrc/146851?camid=4v1a