AFARTICA: A Frequent Item-Set Mining Method Using Artificial Cell Division Algorithm

Saubhik Paladhi, University of Kalyani, Kalyani, India
Sankhadeep Chatterjee, University of Calcutta, Kolkata, India
Takaaki Goto, Toyo University, Saitama, Japan
Soumya Sen, University of Calcutta, Kolkata, India

ABSTRACT

Frequent item-set mining has been exhaustively studied in the last decade. Several successful approaches have been made to identify the maximal frequent item-sets from a set of typical item-sets. The present work has introduced a novel pruning mechanism which has proved itself to be significant time efficient. The novel technique is based on the Artificial Cell Division (ACD) algorithm which has been found to be highly successful in solving tasks that involve a multi-way search of the search space. The necessity conditions of the ACD process have been modified accordingly to tackle the pruning procedure. The proposed algorithm has been compared with the apriori algorithm implemented in WEKA. Accurate experimental evaluation has been conducted and the experimental results have proved the superiority of AFARTICA over apriori algorithm. The results have also indicated that the proposed algorithm can lead to better performance when the support threshold value is more for the same set of item-sets.

KEYWORDS
Artificial Cell, Artificial Cell Division, Data Mining, Directed Acyclic Graph, Frequent Item-Sets, Generation Tree

INTRODUCTION

Frequent item-set mining is considered to be one of the fundamental tasks of data mining apart from other frequent problems like the discovery of association rules, correlation mining, multidimensional pattern mining, etc. The essential problem of mining frequent item-sets is based on finding the frequent item-sets from a given set of items that occurs over a given threshold amount of times. The problem is challenging as it is to be performed over a large database in real time.

The target of data mining is to identify patterns in data by using various tools and algorithms (Rajagopalan & Krovi, 2002). Most of the data-mining algorithms designed to solve the problem is built upon the idea of Apriori algorithm (Agrawal et al., 1996). This algorithm uses a breadth first search strategy starting from the bottom of the search space to the top in order to find every possible frequent item-set. An enumeration of all possible subsets of length of the frequent pattern is found

DOI: 10.4018/JDM.2019070104

Copyright © 2019, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.
to be computationally hard in case of dense data. Recent research trends have focused on finding the maximal frequent item-sets to tackle the problems. The present work has been greatly influenced by the methodology of this algorithm. Zaki (Zaki, 2000; Zaki & Hsiao, 1999) carried experiments using both real and synthetic data with new efficient rule mining framework on closed set. Generation of rules only on demand and the method of reduction of redundant rules are found to be interesting.

An efficient pruning methodology has been proposed by Bayardo (Bayardo, 1998) to reduce the search space to a greater extent. It scans the database via looking ahead pruning strategy while searching the search space by employing a breadth first traversal method. A heuristic strategy has been used to reorder the items to prune superset frequency. An iterative approach to find the maximal elements has been proposed (Gunopulos, Mannila & Saluja, 1997). It extends the given pattern till overcoming the failure. In addition, a randomized version of the same has been reported that doesn't assure the return to the maximal pattern. Horizontal data format based Pincer search method (Lin & Kedem, 1998) uses a bottom up approach to tackle the problem by deploying a bi-directional search. It is better in terms of complexity comparable to Apriori which uses only bottom up search strategy. In addition, this effective strategy assists the reduction of the database scan for non-maximal sets. Agarwal et al. (Agarwal, Aggarwal & Prasad, 2001), have reported a lexicographic tree based approach to solve the finding frequent item sets problem. The lexicographic tree of item-sets has been traversed using the depth-first, breadth-first search or both strategies. Successive projection of transactions along with an efficient matrix-counting technique was employed. While, Han et al. (Han, Pei, & Yin, 2000), applied the divide-and-conquer based approach that works efficiently for medium size databases, and considered a new efficient FP-tree based algorithm. A novel one pass algorithm (Jin & Agrawal, 2005) was proposed for frequent item-set mining. It avoided any use of out-of-core structure and found to be very memory efficient along with an outstanding performance on the datasets having large number of distinct items and/or low support levels. Burdick et al. (Burdick, Calimlim & Gehrke, 2001), had reported an algorithm for databases having long item-sets. The algorithm was claimed to combine an efficient depth-first search with a fast pruning of item-set lattice. Endres et al. (Endres & Kießling, 2015) have considered lattice structure and proposed an linear time parallel algorithm to prune and remove unrestricted attribute domain. A relative bitmap compression schema along with a vertical bitmap representation of the database has been used to perform the searching operation. Zaki et al. (Zaki & Hsiao, 2002), had developed Closed Association Rule Mining (CHARM); an efficient closed item-set mining algorithm. The major contribution of the CHARM is employing parallel search strategy to explore both item-set space and transaction space in contrast with the traditional methods’ searching strategy over item-set space only. Along with this it uses a hybrid search process that efficiently prunes the search space. To achieve a more efficient pruning technique, it uses a hash based process to support the hybrid search. Besides, a lot of research work has been done on frequent item-set mining (Zheng, Kohavi & Mason, 2001; Zaki, 1999). A framework, known as XAR-Miner has been introduced by Zhang et al. (Zhang, Liu, Ling, Bruckner & Tjoa, 2006) to mine association rules from XML documents. In this approach, at first the raw data has been converted to XML tree or Multirelational Databases and then a greedy algorithm has been used for generation of itemsets and association rules.

The generation of association rules from frequent (closed) item-sets, has been investigated by Vo et al. (Vo & Le, 2009; Vo & Le, 2011; Vo, Hong & Le, 2013). A frequent item-set Lattice (FIL) has been used to find association rules out of the FIL. It has been reported that mining association rules from frequent item-sets, is more expensive than mining them using FIL. The efficient use of FIL made it possible to perform far better than traditional approaches. From the literature the question of having an efficient pruning strategy to accelerate the process of frequent item-set mining has emerged. Agarwal et al. (Agarwal, Aggarwal & Prasad, 2001), had suggested efficient method of superset pruning. The pruning technique is facilitated by an efficient reordering of the children nodes of the search tree. The all possible combination of the items using apriori resembles to lattice structure. Some lattice optimization techniques have been used in existing literature. Lattice optimization for
Ontology-Based Data Integration
www.igi-global.com/chapter/ontology-based-data-integration/11187?camid=4v1a

Compression Schemes with Data Reordering for Ordered Data
www.igi-global.com/article/compression-schemes-with-data-reordering-for-ordered-data/109930?camid=4v1a