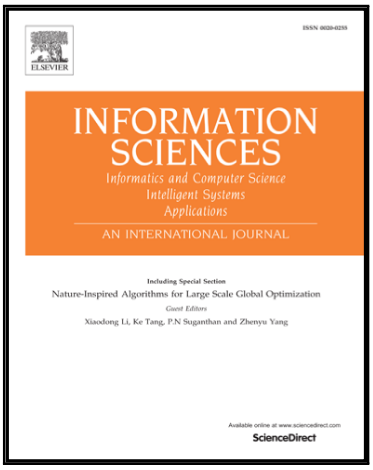

ISSN 0020-0255

ELSEVIER

INFORMATION SCIENCES

Informatics and Computer Science
Intelligent Systems
Applications

AN INTERNATIONAL JOURNAL

Including Special Section

Nature-Inspired Algorithms for Large Scale Global Optimization

Guest Editors

Xiaodong Li, Ke Tang, P.N Suganthan and Zhenyu Yang

Available online at www.sciencedirect.com

ScienceDirect

Mining Maximal Frequent Patterns in Transactional Databases and Dynamic Data Streams: a Spark-based Approach

Md. Rezaul Karim, Michael Cochez, Oya Deniz Beyan,
Chowdhury Farhan Ahmed, Stefan Decker

PII: S0020-0255(17)31126-X
DOI: 10.1016/j.ins.2017.11.064
Reference: INS 13292

To appear in: *Information Sciences*

Received date: 13 February 2016
Revised date: 27 November 2017
Accepted date: 30 November 2017

Please cite this article as: Md. Rezaul Karim, Michael Cochez, Oya Deniz Beyan, Chowdhury Farhan Ahmed, Stefan Decker, Mining Maximal Frequent Patterns in Transactional Databases and Dynamic Data Streams: a Spark-based Approach, *Information Sciences* (2017), doi: 10.1016/j.ins.2017.11.064

Mining Maximal Frequent Patterns in Transactional Databases and Dynamic Data Streams: a Spark-based Approach

Md. Rezaul Karim[a,b], Michael Cochez[a,b,d], Oya Deniz Beyan[b,a], Chowdhury Farhan Ahmed[c], Stefan Decker[a,b]

[a] *Fraunhofer FIT, Schloss Birlinghoven, D-53754 Sankt Augustin, Germany*
[b] *Chair of Computer Science 5 - Information Systems, RWTH Aachen University, Germany*
[c] *Department of Computer Science & Engineering, University of Dhaka, Bangladesh*
[d] *Faculty of Information Technology, University of Jyvaskyla, Finland*

Abstract

Mining maximal frequent patterns (*MFPs*) in transactional databases (*TDBs*) and dynamic data streams (*DDSs*) is substantially important for business intelligence. MFPs, as the smallest set of patterns, help to reveal customers' purchase rules and market basket analysis (*MBA*). Although, numerous studies have been carried out in this area, most of them extend the main-memory based Apriori or FP-growth algorithms. Therefore, these approaches are not only unscalable but also lack parallelism. Consequently, ever increasing big data sources requirements cannot be met. In addition, mining performance in some existing approaches degrade drastically due to the presence of null transactions. We, therefore, proposed an efficient way to mining *MFPs* with *Apache Spark* to overcome these issues. For the faster computation and efficient utilization of memory, we utilized a prime number based data transformation technique, in which values of individual transaction have been preserved. After removing null transactions and infrequent items, the resulting transformed dataset becomes denser compared to the original distributions. We tested our proposed algorithms in both real static *TDBs* and *DDSs*. Experimental results and performance analysis show that our approach is efficient and scalable to large dataset sizes.

Keywords: Big data, transactional databases, dynamic data streams, null transactions, prime number theory, data mining, Apache Spark, maximal frequent patterns.

1. Introduction

The term *data mining* refers to the non-trivial extraction of valid, implicit, potentially useful and ultimately understandable information in large databases with help of the emerging computing technologies. Among them, finding frequent patterns plays a significant role in every data mining task such as association analysis, market basket analysis (*MBA*), clustering, and classification [31, 42, 2, 15, 12]. Consequently, the problem of mining frequent patterns has been discussed widely in data mining research.

Apriori [2] is one of the first algorithms for mining frequent itemsets and association rules from transactional databases. It starts by identifying frequent 1-itemsets in the database and extending them to larger itemsets as long as those itemsets are frequent enough in the database. The frequent itemsets determined by Apriori can be used to determine association rules which highlight general trends in the database for market basket analysis. Apriori uses a breadth-first search strategy to count the support of itemsets and uses a candidate generation function which exploits the *downward closure property* of support. This property states that every subset of a frequent itemset is also frequent.

[*] Corresponding author: rezaul.karim@fit.fraunhofer.de