GANDIVA: Temporal Pattern Tree for Similarity Profiled Association Mining

Vangipuram Radhakrishna, VNR Vignana Jyothi Institute of Engineering and Technology, Hyderabad, India
Puligadda Veereswara Kumar, Acharya Institute of Technology, Bangalore, India
Vinjamuri Janaki, Vaagdevi College of Engineering, Warangal, India

ABSTRACT

In this research, the authors propose a novel tree structure called GANDIVA which computes true supports of all temporal itemsets by performing a tree-based scan and eliminating the database scan which is required for SPAMINE, G-SPAMINE, MASTER, and Z-SPAMINE approaches. The idea is to construct the tree called GANDIVA which determines support of all time-stamped temporal itemsets from the constructed tree. Another important advantage of the proposed approach is that it does not require the original database to be retained in the memory after a time profiled pattern tree (GANDIVA) is constructed from the original database. The significant advantage of GANDIVA over SPAMINE, G-SPAMINE, Z-SPAMINE, and MASTER is that GANDIVA requires zero database scans after the tree construction. GANDIVA is the pioneering research to propose a novel tree-based framework for seasonal temporal data mining.

KEYWORDS

Distance Bounds, Interest Measure, Membership Value, Pattern Pruning, Support, Temporal Itemset, Temporal Pattern, Time Profiled, Time Stamp Transaction Database

1. INTRODUCTION

Data mining deals with the discovery of useful data patterns from enormous, massive amounts of data (Nong, 2013). Data patterns discovered by applying data mining algorithms can be classified into six major types. These include a) prediction and classification patterns b) data reduction patterns c) cluster patterns d) association patterns e) anomaly patterns f) sequential and temporal patterns. Research on temporal patterns is gaining massive importance and is quickly emerging with various important applications associated with the discovery of temporal patterns. Temporally profiled pattern mining has various applications ranging from financial analysis to healthcare. Determining and unearthing similar temporal patterns in time-stamped temporal databases is a complex and challenging task (Y. C. Chen, 2015) (Y. C. Chen, 2016). The pioneering research (Soung, 2008) (Soung, 2009) (Soung, 2012) addressed retrieval and discovery of similar temporal association patterns by introducing support estimation and pruning approaches to reduce support computations and minimize

DOI: 10.4018/IJITWE.2019100101
computational complexity. However, the approach to overcome the necessity of maintaining the temporal database in the memory is not addressed till today (Shadi, 2018) (Vangipuram, 2018). Also, research by (Song, 2008) (Song, 2009) (Song, 2012) did not address new distance measures for time profiled association mining. Some of the previous studies (Vangipuram, 2015) (Vangipuram, 2016) (Vangipuram, 2017) (Vangipuram, 2018) addressed new distance measures SRIHASS (used in Z-SPAMINE), ASTRA (used in MASTER), and G-SPAMINE (Shadi, 2017) that uses fuzzy distance measure. The major limitation of all the above-mentioned approaches is the inevitable requirement for the database to be present in the main memory. The research embodied in this paper aims to address “…the possibility of eliminating the need for original database retention in the main memory…” Our approach has the major advantage that the original database need not be present in the memory for future computation process. To the best of our knowledge, there is no research till date w.r.t similarity profiled temporal association mining that has proposed a tree-based approach that performed the lattice-based scan. Thus, our approach is the pioneering work in this direction. We propose a new tree structure called “GANDIVA - TIME PROFILED TEMPORAL PATTERN TREE” which eliminates the huge computational overhead involved in finding supports of itemsets.

GANDIVA is inspired and motivated from FP-growth (Han, 2004) for snapshot databases. Traditional FP-tree is applicable for snapshot databases and is not suitable for time-stamped temporal databases. Also, FP-growth deals with support values of items w.r.t snapshot database and hence such support values are single values (1-dimensional). In our case, itemset supports are support sequences over ‘n’ time slots and GANDIVA deals with itemsets of multi-dimensional support time sequences. Further, GANDIVA requires sorting the transaction items based on the similarity of itemsets to a given reference time sequence which is not applicable w.r.t. FP-tree as there is no concept of reference pattern sequence in case of FP-tree based approach. In the next section, we introduce the proposed tree-based structure, GANDIVA and the approach for finding supports of temporal itemsets. GANDIVA shall be the pioneering work and the first tree-based structure w.r.t time stamped temporal databases which substantially reduces the computational complexity and improves time efficiency.

The next section outlines the need for the proposed approach followed by the new tree-based data structure for storing, processing and finding supports of time-stamped temporal itemsets. The subsequent section explains the proposed approach; experiment results and concludes the research.

2. PROPOSED METHOD

In this section, we introduce a new approach for finding supports of time-stamped temporal itemsets that is essentially based on tree construction. Our approach is motivated from FP-tree (Han, 2004) for snapshot databases. The tree constructed from itemset supports of the time-stamped temporal database is named as GANDIVA (also called as time profiled temporal pattern tree). We use the names “GANDIVA” and “time profiled temporal pattern tree” interchangeably throughout this paper.

2.1. Problem Definition

To find true support time sequence values of temporal itemsets using a tree-based approach by constructing time-profiled temporal pattern tree and discovery of time profiled association patterns with reduced space and time complexities.

2.2. Need for Proposed Method

Given a time-stamped temporal database, the task of finding similar temporal itemsets requires knowing support values of temporal itemsets. The number of itemsets possible is $2^n$ for n items. Naturally, finding support values of itemsets is a very important and challenging task for mining of similar itemsets. The overall computation cost for finding all similar itemsets in the database by applying brute-force approach is equal to $\theta(2^n)$. Research studies (Vangipuram, 2015) (Vangipuram,
Related Content

Open Source in Web-Based Applications: A Case Study on Single Sign-On
[www.igi-global.com/article/open-source-web-based-applications/2614?camid=4v1a](www.igi-global.com/article/open-source-web-based-applications/2614?camid=4v1a)

Circular, Smart, and Connected Cities: A Key for Enhancing Sustainability and Resilience of the Cities
[www.igi-global.com/chapter/circular-smart-and-connected-cities/233263?camid=4v1a](www.igi-global.com/chapter/circular-smart-and-connected-cities/233263?camid=4v1a)
A Customized Quality Model for Software Quality Assurance in Agile Environment
[www.igi-global.com/article/a-customized-quality-model-for-software-quality-assurance-in-agile-environment/227688?camid=4v1a](www.igi-global.com/article/a-customized-quality-model-for-software-quality-assurance-in-agile-environment/227688?camid=4v1a)

Development and Testing of New E-Learning Hypermedia System
[www.igi-global.com/article/development-and-testing-of-new-e-learning-hypermedia-system/138292?camid=4v1a](www.igi-global.com/article/development-and-testing-of-new-e-learning-hypermedia-system/138292?camid=4v1a)