Fast incremental mining of web sequential patterns with PLWAP tree

C. I. Ezeife · Yi Liu

Received: 1 September 2007 / Accepted: 13 May 2009 / Published online: 11 June 2009
Springer Science+Business Media, LLC 2009

Abstract Point and click at web pages generate continuous data sequences, which flow into the web log data, causing the need to update previously mined web sequential patterns. Algorithms for mining web sequential patterns from scratch include WAP, PLWAP and Apriori-based GSP. Reusing old patterns with only recent additional data sequences in an incremental fashion, when updating patterns, would achieve fast response time with reasonable memory space usage. This paper proposes two algorithms, RePL4UP (Revised PLWAP For UPdate), and PL4UP (PLWAP For UPdate), which use the PLWAP tree structure to incrementally update web sequential patterns efficiently without scanning the whole database even when previous small items become frequent. The RePL4UP concisely stores the position codes of small items in the database sequences in its metadata during tree construction. During mining, RePL4UP scans only the new additional database sequences, revises the old PLWAP tree to restore information on previous small items that have become frequent, while it deletes previous frequent items that have become small using the small item position codes. PL4UP initially builds a bigger PLWAP tree that includes all sequences in the database using a tolerance support, $t$, that is lower than the regular minimum support, $s$. The position code features of the PLWAP tree are used to efficiently mine these trees to extract current frequent patterns when the database is updated. These
approaches more quickly update old frequent patterns without the need to re-scan the entire updated database.

**Keywords** Incremental mining · Sequential mining · Frequent patterns · Data streams · PLWAP tree · Scalability

1 Introduction

The goal of incremental mining of web sequential patterns is to generate current frequent patterns for the updated database (consisting of both old and incremental data) using mostly only the incremental (or newly added) data and previously mined frequent patterns. When data (like web access patterns) are inserted into a database (like web log), some previous frequent patterns may no longer be interesting, while some new interesting patterns could appear in the updated database. Incremental mining of web sequential patterns is beneficial because it may:

1. Scale web sequential mining to large datasets by speeding up processing time.
2. Be effective for mining fast changing and highly dynamic data environment like in stream processing environment requiring fast and real time responses.
3. Be an approach for more efficient utilization of I/O, memory and CPU resources that may be scarce in some applications.
4. Be more effective for detecting similarities and differences in versions of data and their patterns for purposes of predicting and detecting such phenomena as intrusions.

A web access sequential database is a special case of a general sequential database, where every event $e_i$ in a web access sequence is a single event (or item) and not a set of events as is the case in a general sequential database. Events in a web access sequential database could represent, for example, web pages accessed by users stored in web log data, products accessed in an E-Commerce web site. In such web applications, since each click is on a page, single element sets form members of each sequence. While a general sequence looks like $\langle \{a\}, \{a, b, c\}, \{e, f\} \rangle$, a web access sequence contains only sequences like $\langle a, a, b, c, e, f \rangle$. Prominent and good generic sequence mining algorithms include AprioriAll (Agrawal and Srikant 1995), GSP (Srikant and Agrawal 1995), Suffix Tree (Wang 1997), SPADE (Zaki 2000), FreeSpan and Prefix-Span (Pei et al. 2001). The few algorithms designed specifically for single-element set sequences suitable for web navigational sequences include: WAP-tree (Pei et al. 2000), PLWAP-tree (Ezeife and Lu 2005; Ezeife et al. 2005; Lu and Ezeife 2003), and FS-Miner (El-Sayed et al. 2004). Web sequence mining requires an incremental algorithm. The PLWAP-tree (Ezeife and Lu 2005) sequential miner is a good candidate for incremental web sequential mining because with the use of its position code features, it stores relevant parts of the original database in a comparatively compressed tree structure, and it does not require multiple scans over the entire original data. The PLWAP’s position code labels for its nodes can also be used to maximize the reuse of already mined information for incremental maintenance purposes.