CloFAST: Closed Sequence Mining based on Sparse Id-List

Fabio Fumarola
Dipartimento di Informatica
Universita' degli Studi di Bari
via Orabona 4 Bari, Italy
ffumarola@di.uniba.it

Eliana Salvemini
Dipartimento di Informatica
Universita' degli Studi di Bari
via Orabona 4 Bari, Italy
esalvemini@di.uniba.it

Tim Weninger
University of Illinois at
Urbana-Champaign
Urbana IL USA
weninge1@illinois.edu

Donato Malerba
Dipartimento di Informatica
Universita' degli Studi di Bari
via Orabona 4 Bari, Italy
malerba@di.uniba.it

Jiawei Han
University of Illinois at
Urbana-Champaign
Urbana IL USA
hanj@illinois.edu

ABSTRACT

The mining of closed sequential patterns has attracted researchers because of its capability to use compact results to preserve the same expressive power as traditional mining, and because of its efficiency. In this paper we propose CloFAST a novel algorithm for mining closed frequent sequences. CloFAST combines a new data representation of the dataset (*sparse id-list* and *vertical id-list*) with a new two-step strategy for fast support counting and space pruning. Although CloFAST is based on a *candidate maintenance-and-test approach*, it still outperforms the BIDE algorithm [**?**] by two orders of magnitude. CloFAST is also able to mine long closed sequences by reducing the effort required for support counting, search space pruning, and candidates generation. Experimental evaluation shows that the proposed approach is two orders of magnitude faster than BIDE with a modest increase in memory cost.

Categories and Subject Descriptors

A.4 [**Frequent sets and patterns**]: Frequent Closed Sequential Patterns

General Terms

Algorithm, Sequential Patterns, Closed Sequences

Keywords

Data Mining, Descriptive Models, Sparse id-list, Vertical id-list

1. INTRODUCTION

Since its introduction [**?**], sequential pattern mining has become a fundamental data mining task with large spectrum of applications, including web mining [**?**], classification [**?**], finding copy-paste and related bugs in large-scale software code [**?**] and mining motifs from biological sequences [**?**]. A sequential pattern mining algorithm mines a sequence database looking for repeating patterns that can be used to find associations among the different items or events in their data. The algorithms presented in literature have good performance in databases comprised of short sequences [**?**, **?**, **?**, **?**, **?**]. Unfortunately, when these algorithms are used to mine long sequences they generate an exponential number of sequences, especially for lower support thresholds, which make analysis difficult. Moreover the performance of such algorithms often degrades dramatically in both time and space as the support threshold is lowered.

KDD-2011 USA