

# Discovering and Processing Sequential Patterns in Databases

Marek Wojciechowski

Poznan University of Technology, Institute of Computing Science,  
ul. Piotrowo 3a, 60-965 Poznan, Poland  
Marek.Wojciechowski@cs.put.poznan.pl

## 1. Introduction

Data mining, also referred to as knowledge discovery in databases, is a relatively new research area that aims at extracting previously unknown and potentially useful knowledge from large sets of data. One of the most important data mining problems is discovery of frequently occurring patterns in sequential data. Application areas for this problem include analysis of telecommunication systems, discovering frequent buying patterns, analysis of patients' medical records, etc.

The problem of mining frequent patterns in a set of data sequences together with a few mining algorithms was first introduced in [1]. The class of patterns considered there, called *sequential patterns*, had a form of sequences of sets of items. The statistical significance of a pattern (called support) was measured as a percentage of data sequences containing the pattern. In [2], the problem was generalized by adding taxonomy (is-a hierarchy) on items and time constraints such as minimum and maximum gap between adjacent elements of a pattern.

Another formulation of the problem was given in [3], where discovered patterns (called *episodes*) could have different type of ordering: full (serial episodes), none (parallel episodes) or partial and had to appear within a user-defined time window. The episodes were mined over a single event sequence and their statistical significance was measured as a percentage of windows containing the episode (frequency) or as a number of occurrences. Efficient algorithms were presented for serial and parallel episodes. In [4], the model was extended to handle events described by a set of attributes. Episodes mined in sequences of such events were build of a set of unary and binary predicates on event attributes. To make discovery of such complex episodes feasible, it was assumed that a user has to specify a class of interesting patterns by providing a template. In [5], a language capable of specifying episodes of interest based on logical predicates was presented and a few further extensions to the model were added.

The approaches presented above have several drawbacks. First of all, they either do not allow a user to specify any item constraints to define the interesting class of patterns [1][2][3] or they assume that a user has enough knowledge to provide a very specific template [4][5]. Moreover, they do not take into account the fact that the source data is in most cases likely to be stored in databases. The main goal of this thesis is to provide mechanisms leading to the implementation of a knowledge discovery system closely coupled with database systems. Such a system will allow users to specify all the mining tasks concerning discovery of frequent patterns in event sequences via a uniform interface. Issues of integrating user-defined constraints on sequential patterns into mining algorithms and storing discovered knowledge in the database will also be addressed.

Discovery of sequential patterns is the most popular data mining technique applied to event sequences. But in many cases a user might want to perform classification or clustering on sets of objects described by event sequences associated with them. An interesting approach to classification using discovered sequential patterns as features describing objects was presented in [9]. The second most important objective of the thesis is introducing algorithms

for clustering event sequences making use of sequential patterns. There is a need for new clustering algorithms because traditional methods either cannot be applied to event sequences at all or are not able to take into account sequential relationships existing in the data.

## **2. Integration of sequential patterns discovery with databases**

We believe that data mining is an interactive and iterative process. A user formulates a data mining task as a KDD query in a high-level language [6]. The query is sent to the Knowledge Discovery Management System which retrieves the data from the database, chooses the right data mining algorithm, and returns results in a form of frequent patterns to the user. The system should provide mechanisms for storing discovered knowledge in a database for further selective analyses. So far we proposed an SQL-like language for specifying all tasks concerning discovery of frequent patterns in databases [14]. The language is an extension of MineSQL [8], which is an extension of SQL proposed to handle association rules queries. This approach seems to be reasonable because association rules and sequential patterns are very often mined in the same datasets. MineSQL is designed as a query language for advanced users but it can also serve as an Application Programming Interface (API) for building business applications dealing with knowledge discovery [7]. MineSQL provides mechanisms for storing patterns in relational tables by offering new complex data types.

MineSQL allows a user to specify various constraints defining the requested class of patterns (statistical constraints, time constraints, and item constraints). Current algorithms do not handle item constraints at all or require too detailed information on the structure of patterns. In the thesis an algorithm using item constraints in the mining process will be presented. A special emphasis will be laid on the fact that the source data is likely to be stored in relational tables.

## **3. Application of discovered sequential patterns to clustering of event sequences**

Clustering is a discovery process that groups a set of data such that the intra-cluster similarity is maximized and the inter-cluster similarity is minimized [10]. Traditional clustering algorithms employ some measure of distance between data points in n-dimensional space and have problems with categorical attributes and complex objects (e.g. sequences). Usually they work well when the number of dimensions is relatively small. Many new clustering algorithms addressing drawbacks listed above have been proposed recently [11][12][13]. Unfortunately, none of the approaches is suitable for sequences of events described by categorical attributes (names of products, car manufacturers, types of failures, etc.).

We introduced a new algorithm [15] which uses discovered sequential patterns to form the initial sets of clusters. The approach is based on the observation that each frequent pattern defines a cluster containing data sequences sharing a common subsequence. The algorithm iteratively merges clusters until the requested number of clusters is reached. In each iteration the two most similar clusters are merged to form a new larger cluster. The similarity measure we applied is based on the co-occurrence of frequent patterns supported by sequences forming clusters. The advantage of the proposed method is that the source database is read only once and after that initial scan the algorithm works on a compact transformed database where for each frequent pattern a list of sequences supporting it is remembered. The algorithm performs partial clustering since the resulting clusters may overlap. This feature may not be desired in some applications, which is a motivation for further research.

#### 4. Concluding remarks

The thesis has two main objectives. The first is providing support for mining sequential patterns within a Knowledge Discovery Management System. The second is introducing new clustering algorithms suitable for event sequences. The results achieved so far include a new high-level language for specifying mining tasks concerning sequential patterns, extensions to database systems offering data types for transparent storing of discovered knowledge in the database, and a new clustering algorithm grouping event sequences into a set of meaningful overlapping clusters. The ongoing research concentrates on constraint-based algorithms for mining sequential patterns in databases and efficient methods of clustering event sequences into a set of disjoint clusters.

#### Bibliography

1. Agrawal R., Srikant R.: Mining Sequential Patterns. Proc. of the 11th Int'l Conference on Data Engineering (1995).
2. Srikant R., Agrawal R.: Mining Sequential Patterns: Generalizations and Performance Improvements. Proc. of the 5th Int'l Conf. on Extending Database Technology (1996).
3. Mannila H., Toivonen H., Verkamo A.I.: Discovering frequent episodes in sequences. Proc. of the 1st Int'l Conference on Knowledge Discovery and Data Mining (1995).
4. Mannila H., Toivonen H.: Discovering generalized episodes using minimal occurrences. Proc. of the 2nd Int'l Conference on Knowledge Discovery and Data Mining (1996).
5. Guralnik V., Wijesekera D., Srivastava J.: Pattern Directed Mining of Sequence Data. Proc. of the 4th Int'l Conference on Knowledge Discovery and Data Mining (1998).
6. Imielinski T., Mannila H.: A Database Perspective on Knowledge Discovery. Communications of the ACM, Vol. 39, No. 11 (1996).
7. Imielinski T., Virmani A., Abdulghani A.: Datamine: Application programming interface and query language for data mining. Proc. of the 2nd Int'l Conference on Knowledge Discovery and Data Mining (1996).
8. Morzy T., Zakrzewicz M.: SQL-like Language for Database Mining. ADBIS'97 Symposium (1997).
9. Lesh N., Zaki M.J., Ogihara M.: Mining Features for Sequence Classification. Proc. of the 5th Int'l Conference on Knowledge Discovery and Data Mining (1999).
10. Hartigan J., "Clustering Algorithms", John Wiley, 1975
11. Gibson D., Kleinberg J.M., Raghavan P.: Clustering Categorical Data: An Approach Based on Dynamical Systems. Proc. of the 24th Int'l Conference on Very Large Data Bases (VLDB), New York City, New York (1998)
12. Han E., Karypis G., Kumar V., Mobasher B.: Hypergraph Based Clustering in High-Dimensional Data Sets: A summary of Results. Bulletin of the IEEE Computer Society Technical Committee on Data Engineering, Vol.21 No. 1 (1998)
13. Ketterlin A.: Clustering Sequences of Complex Objects. Proc. Of the 3<sup>rd</sup> Int'l Conference on Knowledge Discovery and Data Mining (KDD), Newport Beach, California, USA (1997)
14. Wojciechowski M.: Mining Various Patterns in Sequential Data in an SQL-like Manner, Proc. of short papers of the 3rd East European Conference on Advances in Databases and Information Systems (ADBIS'99), Maribor, Slovenia (1999)
15. Morzy T., Wojciechowski M., Zakrzewicz M.: Pattern-Oriented Hierarchical Clustering, Proc. of the 3rd East European Conference on Advances in Databases and Information Systems (ADBIS'99), Maribor, Slovenia, LNCS 1691, Springer-Verlag (1999)