INTERACTIVE MINING ON HIERARCHICAL DATA

Petr Chmelar[1] **& Lukas Stryka**[1]
PhD Degree Programme (1), FIT BUT
E-mail: {chmelarp,stryka}@fit.vutbr.cz

Supervised by: Jaroslav Zendulka
E-mail: zendulka@fit.vutbr.cz

ABSTRACT

In this paper, we propose a framework for interactive, iterative, and intuitive mining of multilevel association, characterization and classification rules on data organized in multilevel conceptual hierarchies. This framework is called OLAM SE (Self Explaining On-Line Analytical Mining) and it is proposed as an extension of OLAP or as an alternative to Han's OLAM. OLAM processes data stored in data cubes structure of which is based on a given conceptual hierarchy. OLAM SE determines minimum support value from user defined cover value of data with usage of entropy coding principle. It also automatically determines the maximum threshold to avoid explaining knowledge that is obvious and so potentially uninteresting. Major part of data is thus described by frequent patterns. The presentation of results is inspired by UML diagram notation. It contains a graph nodes of which are frequent data sets represented as packages including sub packages – data classes or items. Edges represent relations or patterns between packages. This representation could be applicable for characterization and non-naïve Bayesian classification process as well. Patterns can be interactively explored by the user, who gets a detailed view of attractive ones. She can intuitively drive the more detailed knowledge obtaining process.

1. INTRODUCTION

There are huge amounts of data stored in databases. Thus, it is very difficult to make decisions based on this data. Decision support problems have been motivating for a development of sophisticated tools which provide a new view on data for better data understanding. These tools are used for business analysis, medical and scientific research and many other areas. These tools can be based on data mining techniques, OLAP (On-Line Analytical Processing), data warehouses, etc.

There are many algorithms and methods for data mining on transactional and relational data [4, 9]. But following the requirements of science or commercial sphere the expansion of storing structured or semi-structured data has been coming up. Thus it’s necessary to developed new methods or techniques for data mining on this kind of data [5, 8]. The data mining is mostly very time-consuming, so there are some approaches to save the computing time by a reduction of scanned state space [11].

Our approach uses user interactivity and principles from theory of information to determine data sets that are potentially interesting for analysis. The main asset is to provide easy, fast and interactive system for data mining on highly structured data.

2. OLAP + DATA MINING = OLAM

OLAM (On-Line Analytical Mining) [1, 10] integrates OLAP with data mining in multidimensional databases. OLAM server performs mining in a similar manner as OLAP server performs OLAP analysis – both accept user on-line commands via GUI API and analyze the data cube via cube API. OLAM provides on-line selection of the integrated data mining methods on different subsets of data and at different levels of abstraction in sense of the given metadata – a conceptual hierarchy for each dimension by OLAP basic functions (drilling, filtering, slicing and dicing) on data cube.

3. MULTILEVEL ASSOCIATION ANALYSIS

For many mining tasks, it is difficult to find strong patterns in data at low or primi-tive levels of abstraction. Strong associations discovered at high levels of abstraction may represent common sense knowledge. Sometimes common sense knowledge for one user may be novel for another. Therefore, methods providing capabilities for mining association rules at multiple levels of abstraction with sufficient flexibility for easy traversal among different abstraction spaces has been developed.

These methods use a conceptual hierarchy defining a sequence of mappings from a set of low-level concepts (also called classes) to higher-lever, more general concepts. Conceptual hierarchy is represented by rooted tree, where nodes are general item sets, leafs are data items and the root represents most generalized abstraction of all data items. So these data can be generalized by replacing low-level concepts by their higher-level concepts in a concept hierarchy.

4. OLAM SE CONCEPTS

We propose the OLAM SE system (Self Explaining On-Line Analytical Mining) that is similar to the Han's OLAM [1] in the idea of interactive data mining. The main contribution is to simplify on-line analytical data mining to experts who understand their data but want more significant, interesting and useful information. It processes data and its multilevel concept hierarchy.

Above all, it is done by shielding internal concepts (association, classification, characterization) and thresholds (support, confidence) from user. Firstly it derives important thresholds (different for each concept layer) that can be after that modified using drag-and-drop and the litter bin. Secondly the user selects concepts interesting for her. Then the system interactively creates frequent patterns and after some time it iteratively shows association analysis, concept classification or characteristics which are intuitively offered to the user. The generated hypotheses might be then interactively tested and modified by her.

Also an intuitive graphical interface that suggests most relevant items is proposed in this paper.

4.1. SIMPLIFIED THRESHOLDS

Standard data mining techniques are too complicated for beginners – wrong threshold causes long computational time and irrelevant results. In case of multilevel conceptual hierarchy even experts can guess the proper values only.

So the first problem of intuitive data mining is to find the significant factors as simple as possible. The very first idea of the simplification was based on Paretto analysis. It shows that usually 80% of consequences stem from 20% of the cause [12].

This has led to the parameter we have called *cover*. It means a percentage cover of examined data - $cover(A, B) = P(A \cup B)$ where *A, B* are subsets of investigated data. For instance A is a set of transactions containing Cycling goods and B Football equipment in the figure 1. We have used ideas of information theory – our algorithm works similarly (but reversely) to the Huffman coding technique while it is building the coding tree. The Inverse Huffman algorithm is merging the most frequent (probable) items on each level of concept hierarchy until it covers the required amount of data (eg. 80%). The *support* is then automatically set to be between the least frequent merged item and the most frequent unmerged item.

The second problem of useful data mining is mining unnecessary, obvious information (eg. Sports Shop sells Sporting Goods). That's the reason we employed the obviosity metric. A high-level concept is supposed to be *obvious* when it (self) has higher support than the minimum cover (eg. 80%).

An example of the algorithm and its results is provided in the figure 1.

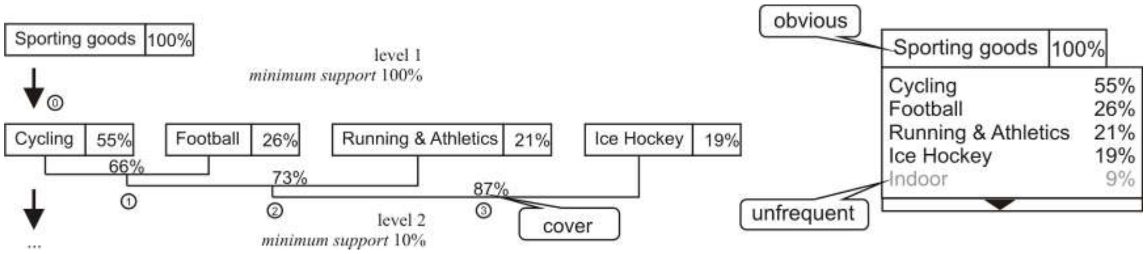

Fig. 1. Illustration of Inverse Huffman algorithm determining frequent (Cycling, Football, …) and obvious (Sporting goods) concepts.

4.2. Presentation of Mined Results

Next important topic of intuitive knowledge discovery is the presentation. We have been inspired by UML structure diagrams [7]. Nodes are aggregated data concepts presented as packages. Edges represent relations of frequent (item)sets (undirected), association rules (directed) and aggregations (terminated by a diamond). Relations appear automatically using drag-and-drop of concepts and (expanding) its sub-concepts (drill-down). Support values related to (sub)concepts and its' relations are displayed as shown in the Figure 2.

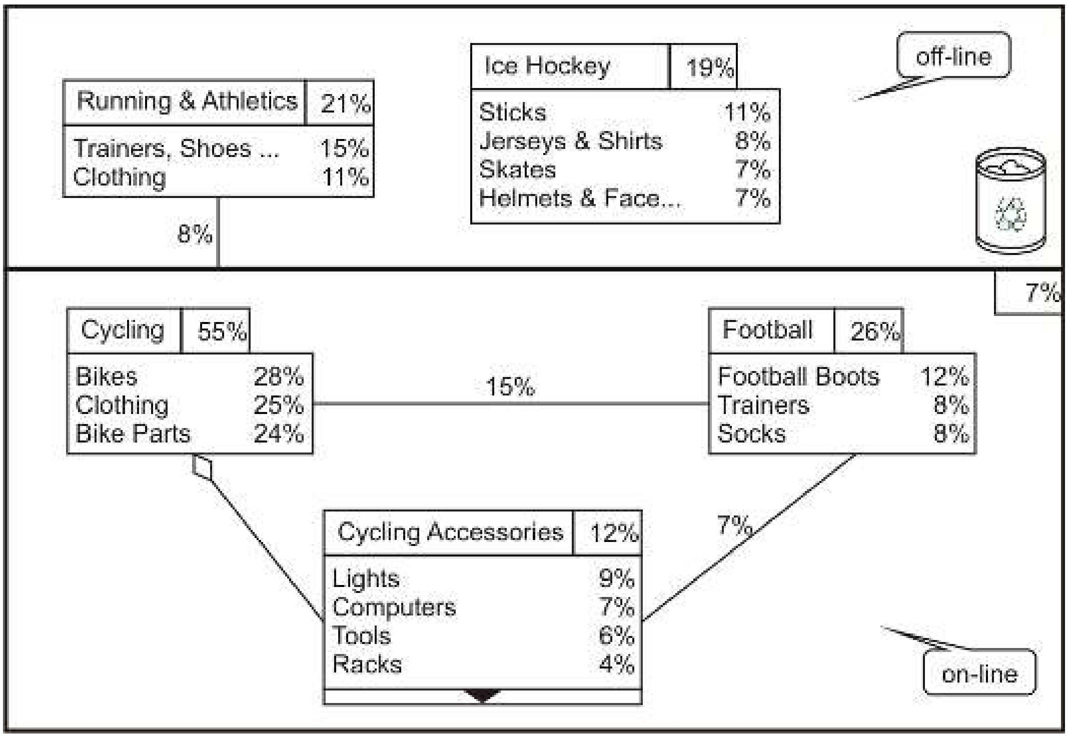

Fig. 2. A screenshot with one on-line (OLAP) and one off-line (mining) workspace.

The application window consists of (at least) two workspaces – off-line and on-line, necessary for slice/dice OLAP operations. Concepts in the on-line (hypotheses testing) workspace are supposed to be interesting for user and interactively examined using the Apriori algorithm [2]. Association rules are progressively derived from frequent data sets.

We propose, the knowledge that cannot be discovered interactively is mined off-line – as a background processes with lower priority. In fact it is only an iterative sorting of off-line concepts according to the relevance (maximum support) to the on-line concepts. We call it progressive hypothesis suggestion – user should wait some while for iterative results that improve its quality without any user interaction.

Suppose that a user moves some products to the on-line workspace and customer characteristics to the off-line workspace (other concepts are e.g. in the litter bin). The system then automatically characterizes customers buying specified products.

In the same way the system provides not-naïve (attributes in on-line workspace are interrelated) Bayesian classification using maximum a posteriori (MAP) [3] by maximizing the relevance of classes (in off-line workspace) to the specified features.

5. CONCLUSIONS

The data mining tasks are very useful in selective marketing, decision analysis, business management and many other areas. We have focused on multi-level frequent pattern analysis, which provides us an information about interesting relationships between data sets on the same or different levels of given conceptual hierarchy.

Our OLAM SE system provides user simplification of processing and understanding huge amounts of data. We use interactivity principles of OLAP to user driven processing of input data. We have established two parameters - cover and obviosity. The cover parameter is based on Paretto analysis and entropy coding to determine interesting patterns. It's leading to the lossy compression of data sets on each level of conceptual hierarchy. The second parameter is the obviosity. On the basis of this parameter the frequent pattern with low information gain are moved to the litter bin.

Our method works in two modes. In online mode it is processed interactive and fast mining of knowledge. In the offline mode the data is processed with data mining algorithms with high computation complexity but iteratively – depending on how much time the user has.

The main goal is that the user that know the data doesn't have to know OLAP or data mining techniques – it is either characterization, not-naive Bayessian classification, frequent pattern analysis, association and correlation rules mining nor the appropriate thresholds and parameters.

REFERENCES

[1] Han J.: Towards on-line analytical mining in large databases. SIGMOD Record (ACM Special Interest Group on Management of Data), (1998).

[2] Han, J., Kamber, M.: Data Mining: Concepts and Techniques. Morgan Kaufmann Publish-ers, Boston, 2nd edition (2006).

[3] Chmelar, P.: Bayesian concepts for human tracking and behavior discovery. Student EEICT 2006, Brno, CZ, Volume 4 (2006) 360-364.

[4] Warner, L.: Data mining techniques. [online] http://www.statsoft.com/textbook/ (2006).

[5] Han, J., Fu, Y.: Discovery of multiple-level association rules from large databases. In Proc. of 1995 Int'l Conf. on Very Large Data Bases, (1995) 420-431.

[6] Fortin S., Liu L., Goebel R.: Multilevel Association Rule Mining An Object Oriented Ap-proach based on Dynamic Hierarchies. University of Alberta (1996).

[7] UML. Unified modeling language. [online] http://www.uml.org (2006).

[8] Zhu, H.: On-line analytical mining of association rules. Master's thesis, Burnaby Univer-sity, Burnaby, British Columbia V5A 1S6, Canada (1998).

[9] Stryka L.: Association rules mining modul. Master's thesis, BUT Brno, Brno (2003).

[10] Han, J. et al.: DBMiner: A system for mining knowledge in large relational databases. In Proc. 1996 Int'l Conf. on Data Mining and Knowledge Discovery, (1996) 250-255.

[11] Messaoud, R. B. et al. Enhanced Mining of Association Rules from Data Cubes. DOLAP'06, (2006).

[12] Juran, M. J.: Juran's quality handbook. McGraw-Hill, New York, 5th ed. (1999).