Multirelational Association Mining with Ferda

Martin Ralbovský and Alexander Kuzmin
Department of Information and Knowledge Engineering
University of Economics, Prague
W. Churchill Sq. 4, 130 67 Praha 3, Czech Republic
Email: martin.ralbovsky@gmail.com, alexander.kuzmin@gmail.com

***Abstract*—The abstract goes here.**

I. Introduction

Association rules mining is an important technique widely used in the KDD community [7]. The classical algorithm for association rules mining, *apriori* [1] is restricted to mining over a single relation composed of a set of binary attributes.

There are some approaches that extend the classical algorithm with ability of mining over multiple relations. Dehaspe and De Raedt[2] use techniques from the field of inductive logical programming used in their *WARMR* algorithm. It is suitable for complex structure of relations and had several successful applications. Other approach [8] adapts *support* and *confidence* measures and calculates them on tables without joining them. These approaches are based on the *apriori* algorithm.

There are also approaches to extend expressing power of association rules. By far the most significant work in this field is the ongoing research of the GUHA method [4], [5]. GUHA is the an original method of exploratory data analysis that is nowadays researched mainly in context of generalizing association rules [10], [11]. Attemts to use generalized association rules also for multirelational data were made [9], [6]. The technique uses star schema of the database[1] and introduced *virtual attribute*, attribute that is computed from the detail tables and acts as attribute of master table. There are two types of *virtual attributes* presented: SQL transformation or generalized association rule mined above the detail table[2]. The latter cannot be obtained by any of the methods based on *apriori* algorithm.

Although the theory for multirelational extentions of the GUHA association mining is well developed, until now there was a lack of successful implementation. In our Ferda environment, we implemented the multirelational extentions of GUHA generalized association rules and conducted some very useful experiments to show advisability of these rules. The aim of this paper is to present multirelational association rules based on *virtual attributes* and to demonstrate their implementation in the Ferda system.

Paper is structured as follows: Section II explains in brief the principles of multirelational association mining with *virtual attributes*. Section III describes the demonstrated system. Section IV states the features of the system to be demonstrated and section V concludes the paper.

[1]One master table and several detail tables.
[2]To be explained later

II. Principles of Multirelational Association Mining

A. Generalized Association Rules

Let us first explain the generalized association rules in sense of GUHA method [9], [10], [11][3]. Generalized association rules mined by procedure 4FT [11] extend the "classical" association rules from *apriori* procedure X→Y, where X and Y are sets of items, in two ways.

The first way is to enable *Boolean attributes* for antecedent and consequent. *Boolean attributes* are recursive structures that enable conjunctions, disjunctions and negations of combinations of individual items. Details can be found in [12].

The second way is to enable expressing more general kind of dependency between antecedent and consequent then *confidence* and *support*. We call these dependencies *4ft-quantifiers*. The generalized association rule can be written in form $\varphi \approx \psi$, where φ and ψ are *Boolean attributes* and $\approx$ is a *4ft-quantifier*. The quantifier is computed on the basis of *4ft-table*, as shown in Table I.

$\underline{\mathrm{M}}$	ψ	$\neg\psi$
φ	a	b
$\neg\varphi$	c	d

TABLE I
4FT CONTINGENCY TABLE

A *4ft table* is a quadruple of natural numbers ⟨*a, b, c, d*⟩ so that: a is the number of object from the data matrix satisfying φ and ψ (likewise for other numbers).

The *above average dependence* quantifier is example of such. It is defined by the following condition:

$$\frac{a}{a+b} \geq (1+p)\frac{a+c}{a+b+c+d} \wedge a \leq Base$$

where p and $Base$ are user-defined parameters. It can be verbally interpreted as *Among object satisfying φ, there are at least 100p per cent more objects satisfying ψ then among all observed objects and there are at least Base observed objects satisfying φ and ψ.*

[3]For simplicity, we focus only on *unconditional rules*. More on *conditional rules* can be found in [10].

B. Virtual attributes

We explain our method on following example: there are two tables concerning clients of a bank. The master table contains information about clients' accounts and the detail table contains information about transactions of individual clients[4].

Virtual attributes are attributes from detail data tables, that are created during the process of association rules verification and are treated as normal attributes of master table although not physically stored. The Ferda system allows creation of two types of *virtual attributes*: *aggregation attributes* and *hypotheses attributes*.

Aggregation attributes can be created as SQL aggregations over a detail table. Attribute stating the average amount of money transfered by a client is an example of *aggregation attribute*. We will not focus on *aggregation attributes*, because they can be also imported into the master data table by corresponding SQL transformations.

Hypotheses attributes are of more interest. They represent a generalized association rule (as explained above), which is true or false for each key item of the master table. Example of such attribute can be *client that often pays by credit card*, which can be formally written as

$$ClientID \approx Payment(CreditCard)$$

with a suitable *4ft-quantifier* $\approx$. It is obvious, that this rule may be very useful to use in the master table concerning clients' accounts. We name the *virtual attribute* $ClientPayingByCreditCard$. Then one can examine status of a client based on client's payments and address. Example of such generalized association rule is

$$District(SouthEast)\&ClientPayingByCreditCard$$

$$\approx Status(good)$$

Note that these types of rules cannot be obtained by any other method mentioned in section I.

Issues concerning *hypotheses attributes* by far exceed the scope of this paper. As well as WARMR, construction of *hypotheses attributes* suffer from explosion of hypotheses space. [6] gives more information about this problem. Interpretation of *hypotheses attributes* poses another issue. Vast majority of association rules created from the detail data table are not interpretable within the master data table. Yet there are some constrains according to which interpretable rules can be created. The matter is at present under research.

III. The Ferda DataMiner

Ferda DataMiner[5] (or Ferda) is the newest system implementing the GUHA method[4], [5]. It envolved from the older *LISp–Miner* system[6].

[4]The example was greatly inspired by data describing clients of virtual bank Barbora. The dataset was examined e.g. during the PKDD 1999 Discovery challenge, see http://lisp.vse.cz/challenge.

[5]http://sourceforge.net/projects/ferda

[6]http://lispminer.vse.cz

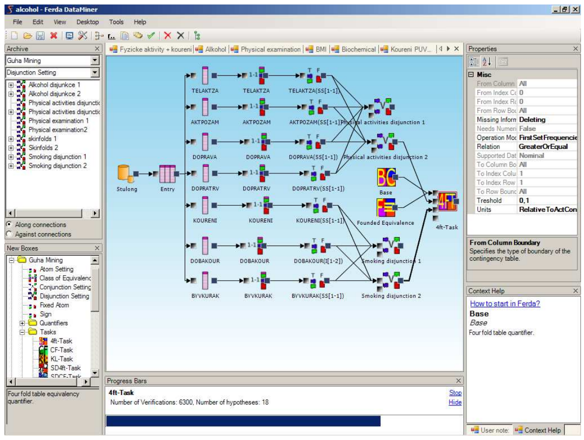

Fig. 1. The Ferda environment

Apart from implementing 4FT, procedure for mining generalized association rules (as described in section II-A) and its multirelational version, Ferda implements five other relational and one multirelational procedures for finding interesting patterns in the data, based on the GUHA method.

One of main features of the Ferda system is visualization of the data mining task setup. Figure 1 shows the working environment. User constructs the task by connecting and setting visual elements called *boxes*. Unlike in other system, *box* does not represent part of the mining process. It represents a function, that has input parameters and computes its output. At present, language of boxes' functions is not recursive, however we are working on a fully recursive language with standard functional constructs like arrays and λ function.

Ferda has also some implementation features worth mentioning. The program is written under GPL license and runs on .NET Framework and Mono. Ferda is a highly modular environment, each *box* can run on different computer over the network and can be written in one of several programming languages. This is achieved by using the Ice middleware[7]. More details on *boxes* and implementation of Ferda can be found in [3].

IV. Demonstrated features

We would like to present the Ferda system, focusing on usage of multirelational association mining, mainly *hypotheses attributes*. After short introduction of the Ferda environment, we will present various multirelational association mining tasks. Figure 2 shows one of such.

The *hypotheses attribute* task setting is located in the upper part of the figure. The setting corresponds to example of *hypotheses attribute* from section II-B. It examines relations between clients and types of operations (deposit, withdrawal or credit card payment). *Above average dependence* quantifier is used.

Lower part of figure 2 shows task setting of the master table. Information about client's district in conjunction with

[7]http://www.zeroc.com

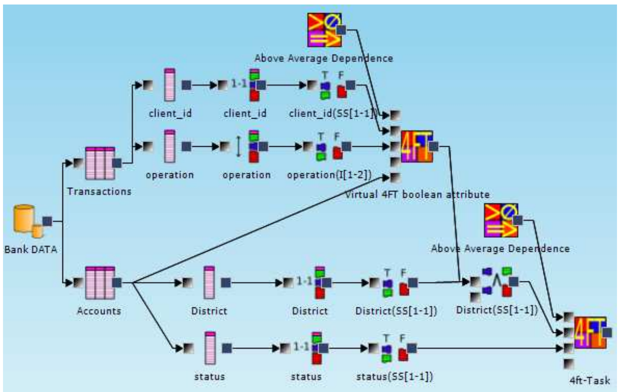

Fig. 2. Example of multirelational task setting

hypotheses attribute are examined against status of the client. The setting again corresponds to exemplary association rule from section II-B. Note similar appearance of the master table task *box* and also the *hypotheses attribute box* – they represent modifications of the 4FT procedure. Also note how inputs of the master table task box represents setting of generalized association rules, $\varphi \approx \psi$.

We would also like to present, how *hypotheses attribute* setting can affect the hypotheses space and thus the running times. So far, our experiments have been based on examining the Barbora dataset. This dataset contains only generated data, so observed rules are not relevant in the real world. We will try to find another source of data and mine interesting multirelational rules. After the implementation, we also encountered problems with interpreting multirelational rules and with setting of intepretable tasks. The issue is a matter of research and by the time of the conference, we should bring findings.

V. Conclusion

The conclusion goes here.

Acknowledgment

This work was supported by the project MSM6138439910 of the Ministry of Education of the Czech Republic, project IG407056 of University of Economics, Prague and by the project 201/05/0325 of the Czech Science Foundation.

We would like to thanks our research colleagues Jan Rauch, Tomáš Kuchař and Michal Kováč for help, valuable comments and reviews.

References

[1] Agrawal R., Mannila H., Srikant R., Toivonen H., Verkamo A.: *Fast discovery of association rules*. Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R., eds.: Advances in Knowledge Discovery and Data Mining. AAAI Press, Menlo Park (1996) p. 307 – 328

[2] Dehaspe L., De Raedt L.: *Mining Association Rules in Multiple Relations*. In Proceedings of the 7th International Workshop on Inductive Logical Programming, Volume 1297, LNAI, pp. 125–132, Springer-Verlag, 1997

[3] Kováč M., Kuchař T., Kuzmin A., Ralbovský M.: *Ferda, New Visual Environment for Data Mining*. Znalosti 2006, Conference on Data Mining, Hradec Králové 2006, p. 118 – 129 (in Czech)

[4] Hájek P., Havel I., Chytil M.: *The GUHA method of automatic hypotheses determination*. Computing 1, 1966, p. 293 – 308

[5] Hájek P., Havránek, T.: *Mechanising Hypothesis Formation - Mathematical Foundations for a General Theory*. Springer-Verlag: Berlin - Heidelberg - New York, 1978.

[6] Karban T.: *Relational Data Mining and GUHA*. in Richta K., Snášel V., Pokorný J.(eds.): Proceedings of the 5th annual workshop DATESO 2005(Databases, Texts, Specifications and Objects), ISBN:80-01-03204-3, pp.103–112

[7] KDNuggets Polls, *Data mining/analytic techniques you use frequently*. www.kdnuggets.com/polls/2005/data_mining_techniques.htm

[8] Pizzi, L.C., Ribeiro, M.X., Vieira, M.T.P.: *Analysis of Hepatitis Dataset using Multirelational Association Rules*. ECML/PKDD Discovery Challenge, 2005.

[9] Rauch J.: *Interesting Association Rules and Multi-relational Association Rules*. Communications of Institute of Information and Computing Machinery, Taiwan. Vol. 5, No 2, May 2002. pp. 77–82

[10] Rauch J.: *Logic of Association Rules*. In: Applied Inteligence, Vol. 22, Issue 1, p. 9 – 28

[11] Rauch J., Šimůnek, M.: *An Alternative Approach to Mining Association Rules* Lin T Y, Ohsuga S, Liau C J, and Tsumoto S (eds): Foundations of Data Mining and Knowledge Discovery, Springer-Verlag, 2005 p. 219 – 239

[12] Ralbovský M., Kuchař T.: *Using Disjunctions in Association Mining*. In: Perner P.: Advances in Data Mining, Springer-Verlag 2007, to appear