KATHOLIEKE
UNIVERSITEIT
LEUVEN

DEPARTEMENT TOEGEPASTE ECONOMISCHE WETENSCHAPPEN

RESEARCH REPORT 0020

POST-PROCESSING OF ASSOCIATION RULES

by
B. BAESENS
S. VIAENE
J. VANTHIENEN

D/2000/2376/20

Post-Processing of Association Rules

Bart Baesens, Stijn Viaene & Jan Vanthienen
Leuven Institute for Research in Information Systems (LIRIS)
Department of Applied Economic Sciences
Katholieke Universiteit Leuven
Naamsestraat 69, B-3000 Leuven, Belgium
{Bart.Baesens;Stijn.Viaene;Jan.Vanthienen}@econ.kuleuven.ac.be

Abstract

In this paper, we situate and motivate the need for a post-processing phase to the association rule mining algorithm when plugged into the knowledge discovery in databases process. Major research effort has already been devoted to optimising the initially proposed mining algorithms. When it comes to effectively extrapolating the most interesting knowledge nuggets from the standard output of these algorithms, we are faced with an extreme challenge, since it is not uncommon to be confronted with a vast amount of association rules after running the algorithms. The sheer multitude of generated rules often clouds the perception of the interpreters. Rightful assessment of the usefulness of the generated output introduces the need to effectively deal with different forms of data redundancy and data being plainly uninteresting. In order to do so, we will give a tentative overview of some of the main post-processing tasks, taking into account the efforts that have already been reported in the literature.

Accepted for the workshop Post Processing in Machine Learning and Data Mining: Interpretation, Visualization, Integration, and Related Topics, Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 20 - 23, 2000, Boston, MA, USA

1 Introduction

Nowadays, companies are faced with massive amounts of data stored in large, often corporate data warehouses. However, extracting useful knowledge from this source of potential intelligence, also known as knowledge discovery in databases, is a tricky and difficult challenge. It is commonly defined as the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data [8]. A number of very promising pattern mining techniques have been introduced as core tools to the knowledge discovery process. Amongst them, the technique of mining for association rules has acquired major interest. The technique basically allows to discover intra-transactional patterns between items in a database of transactional records [3]. Over the last couple of years, we have seen a surge in research on improving the algorithmic performance of the initially proposed algorithms, among which the Apriori-algorithm [2] is most known. To date, some successful applications illustrating their usefulness have been reported in the literature [24,18,4].

Definitely a strong element of association rule mining is its ability to discover all associations that exist in the transaction database. Unfortunately, this also turns out to be its main weakness, when trying to interpret the output of the algorithm. Typical to the association rule mining process is that it yields a very large number of rules, making it hard for the user to identify the interesting ones. This is the case particularly for data sets whose attributes are highly correlated.

To overcome the above, post-processing is believed to play a pivotal role when it comes to enhancing the quality of knowledge discovery. In this paper, we provide a basic rationale for post-processing the patterns generated by an association rule mining process, taking into account the main efforts that have already been reported in the literature. The paper is organised as follows. Section 2 will provide an overview of the basic association rule semantics. This will be followed by a concise description of algorithmic essentials in section 3. A critical assessment of the support –confidence rationale will be given in section 4. Section 5 will then discuss some major tasks believed to be important in the post-processing phase of the knowledge discovery process.

2 Basic Association Rule Semantics

The basic association rule semantics are based upon a well-defined data set-up and at the same time rest upon a very limited set of conceptual statistics [3]. Let D be a database of transactions (tuples). Each transaction t_p consists of a transaction identifier and a set of items (attributes) $\{i_1, i_2, \ldots, i_n\}$ selected from the universe I of all possible descriptive items. An association rule is an a-typical implication of the form: $X \rightarrow Y$ where $X \subset I$, $Y \subset I$ and $X \cap Y = \emptyset$. Traditional (crisp) 'if...then' implication semantics are however not applicable. Each association rule is characterised by means of its support and its confidence defined as follows:

$$\mathrm{Sup}(X \rightarrow Y) = \frac{\text{number of transactions supporting } X \cup Y}{\text{total number of transactions}}$$

$$\mathrm{Conf}(X \rightarrow Y) = \frac{\mathrm{Sup}(X \cup Y)}{\mathrm{Sup}(X)} = \frac{\mathrm{p}(X \cup Y)}{\mathrm{p}(X)} = \mathrm{p}(Y \mid X).$$

According to the above definitions, the support measure can be considered as the percentage of database transactions for which ($X \cup Y$) evaluates to true. Equivalently, the confidence measure is understood to be the conditional probability of the consequent given the antecedent. Association rule mining essentially boils down to discovering all association rules having support and confidence above user-specified thresholds minsup and minconf for respectively the support and the confidence of the rules.

3 Algorithmic Essentials

Discovering association rules is essentially a two-step process [3]:

- Step 1: Identification of all (large) itemsets having support above *minsup*, i.e. 'frequent' itemsets;
- Step 2: Discovery of all derived association rules having confidence above *minconf*;

Of the above steps, the first is computationally more intensive and therefore often the focal element of current research. Multiple algorithms have been devised to optimise the efficiency of frequent, large itemset discovery. Most of them are extensions of the Apriori-algorithm as introduced by Agrawal et al. [2]. The basic intuition behind the latter algorithm is that all subsets of a frequent set are frequent as well. This property is also known as the downward-closure or *Apriori* property. Exploiting this key fact allows to effectively discover all frequent itemsets without dramatically increasing the complexity of the algorithm when the number of items in a transactional database increases.

The standard Apriori-algorithm makes k+1 passes over the data, with k being the size of the maximal frequent itemset. Because each pass is very I/O-intensive, several authors have proposed other techniques to minimise the number of database passes. A parallel algorithm is devised by Agrawal et al. [1]. Park et al. [17] suggest a hash based algorithm allowing to minimise the number of candidate 2-itemsets. They also provide instructions to trim the transaction database size in order to reduce the I/O-cost during later iterations. Zaki et al. [25] present an algorithm that scans the database only once and generates the frequent itemsets by means of lattice traversal techniques. Brin et al. [7] propose the Dynamic Itemset Counting (DIC)-algorithm to minimise the number of database passes.

Once the frequent itemsets have been discovered, the 'useful' association-rules can be inferred using the confidence measure. Unfortunately, this measure does not posses a downward – or upward – closure property. This makes the process of discovering

'useful'-association rules less efficient, because no pruning of the hypothesis space (i.e. the itemset lattice) can be performed to reduce the complexity of the discovery process. Silverstein et al. [20] propose the use of what they call 'minimal dependence' which is an upward-closed property in the itemset lattice. This allows to effectively prune away a lot of candidate lattice associations.

4 An initial critical assessment of the support-confidence rationale

The gist of the association rule mining process, specified in the previous section, rests upon the semantics of the support-confidence framework described in section 2. Before embarking on a quest to mine association rules, we should at least include some time to critically reflect on some of the following issues.

- Mining for the presence of items is not enough.

The standard support-confidence framework only looks for association rules containing a specified set of items. However, the absence of certain items may also yield important information with respect to the consequent. Several alternatives have been presented in the literature to accomplish this. First of all, we could include additional columns indicating the absence of all items in a transaction. Running the Apriori-algorithm on this transformed data set would yield the desired association rules. Silverstein et al [20] propose the use of contingency tables and the Chi-squared test statistic to discover "dependence rules". They suggest starting with building a contingency table indicating both the presence and absence of items. The following step is to compute the Chi-squared test statistic and look for its significance. Small p-values indicate a departure from independence and suggest the presence of dependencies between the items in the contingency table.

- Item quantities have to be taken into account in order to quantify profits.

So far, we have discussed boolean association rules which only pay attention to whether an item is present in a transaction or not. Quantitative association rules also take into account the quantities of the items in a transaction. Srikant et al. [22] propose an adjustment of the Apriori-algorithm to mine quantitative association rules. They start with partitioning each quantitative attribute into a set of intervals, which may overlap, and map the problem to a boolean association rules problem. Korn et al. [13] present the use of linear "ratio-rules" based on Principal Component Analysis. By means of the 'key' Principal Components (PC's), they try to explain the variance in the transaction data. These PC's are subsequently used to derive linear "ratio-rules" taking into account the ratio of the quantities of the items present within a particular principal component.

- Specification of *minsup* and *minconf* is a difficult task.

When mining association rules, both the minimum support, *minsup*, and the minimum confidence, *minconf*, need to be specified in advance by the user. Determining the optimal levels of both thresholds is a difficult task. Setting *minsup* too low will lead to a combinatorial explosion of the number of candidate-itemsets.

Otherwise, setting it too high will miss some important association rules of rare but interesting items. Hence, choosing appropriate levels for both parameters should be done carefully and preferably in co-operation with an expert of the domain under investigation. At this moment, there is a shortage of research devoted to assessing the economical underpinnings of a motivated choice for the critical parameters of the association rule discovery process in its current form.

- Post-processing the results is a necessity.

The support-confidence based association rule mining process often yields a large number of rules, making it hard for the user to select the *interesting* ones. This is particularly the case for data sets where attributes are highly correlated. It is thus not always clear how to act upon the results that are produced by the algorithm. The sheer multitude of generated rules often clouds the perception of the interpreters. Rightful assessment of the usefulness of the generated output introduces the need to effectively deal with different forms of data redundancy and data being plainly uninteresting.

Several authors have tried to alter the support-confidence knowledge discovery framework in order to deal with some of the above elements. Two major lines of thought can be distinguished. There are those that focus on altering the basic support-confidence semantics of the association rule mining algorithm. Direct integration of a richer and more discriminative mining criterion, focusing on more than just the numbers, as is the case for the support-confidence rationale, is believed to directly produce a superior form of knowledge. On the other hand, there are those that rather believe in post-processing of the results generated by the basic semantics introduced in section 2. Thus, giving rise to a process of fine-tuning and filtering of the produced association rules in order to rid them of all kinds of semantically elusive anomalies (e.g. data redundancy, transitivity, etc.). Conceptually, the post-processing phase is made up of an integrated whole of transformations distilling the most interesting knowledge nuggets from the candidate knowledge base produced by the association rule mining algorithm.

Whether integrating a criterion believed more interesting than support and/or confidence within the core association rule mining algorithm or transforming the produced rules as generated from the basic support-confidence rationale in a post-processing phase, all these attempts basically envision the same thing: an increase of the resulting business intelligence. In the next section, we will give a tentative suggestion to the conception of a post-processing phase to the basic support-confidence association rule-mining step, taking into account the main efforts that have already been reported in the literature.

5 Post-Processing of Association Rules

Post-processing is considered to be an important step to help the user discover the useful knowledge nuggets in the huge set of generated association rules. It usually consists of the following steps: pruning, summarising, grouping and visualisation. In the pruning phase, rules are deleted because they are uninteresting or redundant. The

summarising phase tries to summarise the rules into more general concepts. The remaining rules are grouped into rule packets in the grouping phase. Finally the extracted, useful knowledge is depicted in a visualisation phase. Although the distinction and the interaction between these phases is not all that strict, they present different steps that could beneficially be integrated in any post-processing attempt.

In the remainder of this section, we give an overview of the above post-processing tasks conceived in the context of support-confidence based association rule mining, without however in any way claiming to be exhaustive. The overview will be built up according to the main transformations that have been posited in the literature. All of these transformations directly highlight the need for ongoing and intensified integration of economically motivated argumentation when interpreting results from any data mining process. It is our firm belief that this is the only way to realise the form of maturity to generate the potential that is undoubtedly present in the knowledge discovery activity through data mining.

5.1 Pruning of Association Rules

Pruning essentially amounts to the elimination of a number of generated rules because they are manifestations of some kind of unwanted phenomenon, i.e. an anomaly or simply prove to be uninteresting. Some of the most common forms of anomaly are posited and discussed next. It is to be noticed that in the following discussion we make abstraction of any form of generated inconsistencies or other phenomena, a.k.a. anomalies, typically researched in the context of the Verification of (association) rule bases. Anomaly detection as a basis of the Verification process of the association rule generation lies beyond the scope of this discussion, although we fully realise that this may be an artificially constructed boundary imposed upon this discussion.

- Subsumed Rules
  "One rule is subsumed by another if both rules have the same consequent but one contains additional conditions in its antecedent."

Subsumed rules are a special case of information redundancy, one of the most common anomalies in rule bases. Rules that are very specific (i.e. with many conditions in the antecedent) tend to overfit the data and should be removed. Simple and general rules are to be preferred. Consider the following example:

R1: If $A_1 = v_1$ and $A_2 = v_2$ then $A_3 = v_3$
R2: If $A_1 = v_1$ then $A_3 = v_3$.

where Ai = attribute i and v_j = value j. Clearly, whenever **R1** holds, **R2** will also hold. At first sight we can leave out **R1**. However, remember that we are faced with rules that are probabilistic in nature. Therefore, it could be possibly interesting to preserve both rules. Bayardo Jr. et al. [5] propose a measure called the minimum amount of improvement defined as: $conf_{R1} - conf_{R2}$. When this amount exceeds a certain user-defined threshold, it could be interesting to preserve both rules because the extra element in the antecedent significantly augments the confidence of the rule.

- Transitivity of Rules

Consider the following set of rules:

R1: If $A_1 = v_1$ then $A_3 = v_3$
R2: If $A_2 = v_2$ then $A_3 = v_3$
R3: If $A_1 = v_1$ then $A_2 = v_2$.

The presence of 'rule chains' induced by transitivity is a phenomenon that is quite often encountered in rule bases. The above rule set illustrates the anomaly in its basic form. The question arises whether we can delete rule **R1** because it can be deduced by means of the transitivity principle out of rules **R2** and **R3**. In classical rule-based systems this would be no problem. However, the statistical nature of the association rules complicates matters. In order to prune away rule **R1**, the following three conditions should be met [19]:

1. **R3** should have a very high confidence;
2. **R1** and **R2** should have similar strength as measured by their confidence;
3. **R4:** If $A_2 = v_2$ then $A_1 = v_1$ should have a very low confidence.

- Circular Rule Chains

This form of anomaly is induced by the transitivity of rules, making up a rule chain in which the antecedent at the start of the rule chain is in some way incorporated in the tail consequent part of the rule chain.

Another interesting attempt to prune the generated rule base, in line with reducing redundancy within the whole of generated rules, is proposed in [23]. Consider an association rule set Γ. The rule set Γ describes a number of database transaction rows. A 'rule cover' is then defined as a subset Δ of the rule set Γ, essentially describing the same database transaction rows as Γ. Pruning is thereupon performed by reducing Γ to Δ. Toivonen et al. [23] provide an algorithm to efficiently extract a rule cover out of a set of given rules. Brijs et al. [6] further improve this algorithm by means of integer-programming techniques.

- Interestingness Based Pruning

Besides the pruning of (redundant) rules that are in some way deducible from other rules present in the rule set, the generated rules may also be compacted by introducing a pruning paradigm based on some measure of 'interestingness'.

The confidence-measure is a rather poor measure to detect the dependence of the consequent with respect to the antecedent [20,10]. Consider the following situation:

$X \rightarrow Y$ [Sup=25%; Conf=80%] **and** [*minsup*=20%; *minconf*=75%].

At first sight, this rule may look quite interesting. However, further inspection reveals that the a priori-probability of Y equals 85%. In other words, a transaction

satisfying X is less likely to satisfy Y than a transaction about which we have no information. This indicates a clear shortcoming of the confidence-measure. To overcome this problem, different alternatives have been suggested in the literature. The *interestingness*-measure takes into account the probability of the consequent in the following way:

$$\mathrm{I}(X \rightarrow Y) = \frac{\mathrm{Sup}(X \cup Y)}{\mathrm{Sup}(X)\mathrm{Sup}(Y)}.$$

In this way, it measures a kind of departure from independence. A quantity less than 1 indicates a negative dependence (substitution-effect) while a quantity larger than 1 indicates a positive dependence (complementary-effect). Guillaume et al. [10] propose the *Intensity of Implication* measure to avoid the above problem. Unfortunately this measure is less intuitive and harder to understand.

One of the major weaknesses of the *interestingness*-measure is that it only measures co-occurrence and not implication because of its symmetry. Therefore, Brin et al. [7] propose an alternative measure, *conviction,* defined as:

$$\mathrm{Conv}(X \rightarrow Y) = \frac{\mathrm{Sup}(X)\,\mathrm{Sup}(\neg Y)}{\mathrm{Sup}(X, \neg Y)}.$$

This measure takes into account the direction of the rule. Unlike the *interestingness*-measure, it manifests its highest possible value, i.c. infinity, for rules that hold 100% of the time.

Clearly, determining interestingness of a rule is not a simple endeavour. Furthermore, a rule can be interpreted taking into account several extra dimensions for evaluation, which makes it even harder to qualify, let alone quantify, the interestingness of an association rule or association rule packet. On top, a rule can be interesting to one person but not to another.

- Meta-knowledge Guided Pruning

Some authors stress the need to incorporate specific domain knowledge, a.k.a. meta-knowledge or a priori knowledge, to guide the pruning process. Meta-knowledge allows the user to use existing domain knowledge to select the interesting patterns. Again, different approaches have been suggested in the literature. Rule templates [12], also called meta-rules, are general rule skeletons describing the a priori structure of the 'interesting'-rules. Each rule template describes what attributes should occur as antecedents or consequents in a rule. An example of a rule template is the following:

Any $\rightarrow$ Soft Drinks.

This template specifies the desire to look for associations having Soft Drinks as their consequent and any other item as their antecedent. Templates can be either inclusive

or restrictive. To be interesting, a rule has to match an inclusive template. If it matches a restrictive template it is considered to be uninteresting and pruned away.

Liu et al [16] developed a specification language allowing them to specify three levels of domain knowledge: general impressions, reasonably precise concepts and precise knowledge. This meta-knowledge allows them to classify the discovered association rules into 2 categories: conforming and unexpected rules[1]. The former category can be pruned away whilst the latter can be used for further analysis.

5.2 Summarising

Although pruning can significantly reduce the number of discovered associations, this number may still be too large to be tractable. In the summarisation phase, we try to summarise the discovered association rules into more general or abstract concepts that are easier understood by the user. This way, rules can be grouped into several levels of abstraction, all according to the preference of the users. Higher level rules provide a more general overview of the discovered knowledge whilst the lower level rules can be browsed for further details. The abstraction process can be operationalised in a wide variety of ways. Without claiming to be exhaustive, we will concisely highlight some of the more interesting efforts brought forward in literature.

- Taxonomies

A taxonomy specifies a hierarchical, usually tree based, organisation between the items in a transactional database. This taxonomy can then be used to group specific low-level rules into general higher-level rules. This allows for a more compact representation of the generated knowledge. Srikant et al. [21] propose algorithms to mine association rules in the presence of a taxonomy. Integration of the concept of taxonomy in the mining algorithm necessitates the algorithm to be re-run every time the taxonomy changes. Several taxonomical semantics can be devised. Usually taxonomies are conceptualised as a form of 'is a' type of relationship, but this need not be the case.

- Direction Setting Rules

Liu et al [15] introduce the concept of *direction setting* (DS) rules to summarise the essential knowledge generated by the discovered associations. Further details are provided by the non-DS-rules. In this way, the DS-rules provide a summary of the key aspects of the discovered knowledge whilst the non-DS rules provide the user with the relevant details. Consider the following example [15]:

R1: Job=yes → Loan = approved [Sup=40%; Conf=70%]
R2: Own_house=yes → Loan = approved [Sup=30%; Conf=75%]

[1] Notice that Liu et al. [16] further distinguish between unexpected consequent rules, unexpected condition rules and both-side unexpected rules.

By means of Chi-squared analysis it is shown that having a job and owning a house is positively correlated with the grant of a loan. Consequently, the following rule is not very surprising:

R3: Job=yes, Own_house=yes → Loan= approved [Sup=20; Conf=90%]

Rules **R1** and **R2** are DS-rules because they set the direction (positive correlation) for rule **R3**. In this way rules **R1** and **R2** provide a summary of all three rules. Liu et al [15] provide an algorithm for efficiently extracting the DS-rules out of a set of 1-consequent association rules.

5.3 Grouping

So far, we mainly focussed on the statistical properties of the rules in order to prune and/or summarise them. The purpose of the grouping phase is to look at some other characteristics of the rules and group them according to different dimensions. In this section, we highlight the grouping semantics by means of the illustrative dimensions of time and economic relevance, allowing a grouping of the rules into an appropriate amount of rule packets. The economical relevance dimension allows to look at the economical properties of the items in a rule whilst the 'time' dimension focuses on the time distribution of the transactions generating the rules.

- Economic Assessment

As mentioned before, we should take into account the economical identity of the items in a frequent itemset as measured by their price, costs, profit-margin, etc. . Consider e.g. a retail store offering several products with different prices and different profit margins. We may expect that the more expensive products will not be bought as often as the cheaper ones. Nevertheless, associations concerning these products are important to increase profits. One way to deal with this problem is to use different levels of support for the different products. This approach is followed by Liu et al. [14]. An alternative way is to combine all these figures into a new measure called 'the economically adjusted support' of a frequent itemset. This measure could then take into account the economical properties of the items in an itemset and weight them by the support of the itemset. Other measures could make a distinction between the consequent of a rule and its antecedent to maximise e.g. the amount of cross-selling effects.

The question arises whether it would be beneficial to incorporate these measures into the Apriori-algorithm itself. We should look at the downward-closure property of the suggested measures. This insight in combination with a thorough assessment of the time overhead incurred by the need to re-run the association rule algorithm upon a change in the proposed measure threshold will back up or dissuade the decision to do so. Furthermore, setting an appropriate threshold for these measures remains a difficult task.

Several alternative measures are conceivable and could be computed for each rule. The next and logical step is then to cluster the rules according to their scoring on the devised measures. This would enable to group the rules into categories such as high profit-margin rules, medium profit-margin rules and low profit-margin rules. This way, the user gets a more thorough understanding of the economical relevance of the generated patterns.

- Time Based Patterns

A major weakness of association rules is that they do not take into account the distributions of the timestamps of the various transactions generating the association rules. Taking into account his information could allow us to distinguish between time-dependent and time-independent association rules. It could e.g. be that some association rules are generated by transactions occurring during a specific time-frame because of a promotional campaign or a specific season. Associations between ski-boots and ski-pants may be more apparent during the winter than during the summer. By looking at the time distributions of the association rules, we may try to group the rules into packets occurring during the same time frame. This is illustrated in Figure 1.

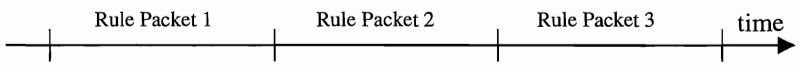

Figure 1 Time clustered rule packets

Both dimensions illustrated above can of course be used separately. They can however also be combined. This is illustrated in Figure 2.

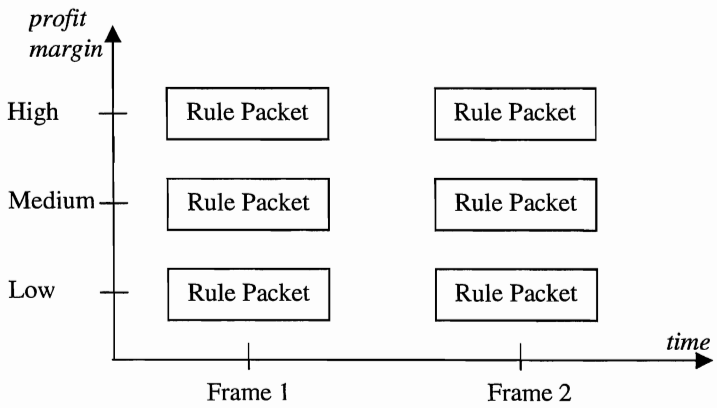

Figure 2 Clustering on different dimensions

Clustering rules according to the dimensions time and profit margin would allow a user to efficiently inspect the rules generating most of the profits during a specific time frame.

The above discussion can be easily extended to incorporate multiple dimensions depending upon the problem domain and the interests of the users.

5.4 Visualisation

Visual exploration of the potentially interesting association rules is to be conceived as an integrated and core element of any exploratory data-mining environment. All of the above transformations can be supported by means of visually enhanced mining tools. Some attempts at visualisation of the mining results are reported in the literature [9,16,11]. The effective and efficient integration with the other elements of the knowledge discovery environment is considered a critical success factor. Other critical success factors that can be distilled from the latter requirement include:

- Support for 'interactive mode' mining

Mining and its subsequent post-processing of results are best characterised as an iterative and intensive process. What is characteristic to many forms of data mining is that the target knowledge is not pre-determined. This definitely is the case for association rule mining [16]. Data miners may want to experiment with different scenarios and parameter settings in order to be able to fine-tune the algorithms. Therefore, many tools provide some form of 'interactive mode' mining to enhance this ad hoc way of mining. Clear presentation of intermediate results may thus come in very handy.

- Integrated sensitivity analysis

In line with the previous argumentation, is the need for an integrated form of sensitivity analysis, preferably supported in interactive mode. This provides the miner with a means to assess the sensitivity of the mining results to several hot-spot, a priori determined parameters used during the course of the knowledge discovery process e.g. support and confidence thresholds.

- Integrated knowledge navigation

Ideally, a user should be given the opportunity to browse through the mining results in any way he or she sees fit. The provision for several levels of granularity, i.e. supporting the (preferably user imposed) abstraction process, enables the user to start at a high level and dig down for further detail. Support for visualisation at different degrees of granularity is only one of the many ways of navigation through the mined knowledge base. Both horizontal and vertical navigation should be supported. In [11], Liu et al. present DBMiner, a knowledge discovery system that integrates data mining, e.g. association rule mining, with on-line analytical processing (OLAP) further enhanced by means of an integrated set of visualisation tools. The authors effectively make use of data cube technology to allow the visual exploration and navigation of association rules in a flexible and efficient way. Essentially two kinds of associations can be mined in a data cube: inter-dimension association rules and

intra-dimension associations. In this way, a tentative effort to implement 'aspect oriented' navigation can be provided.

With this concise discussion on the topic of visualisation of association rules, all but the last word has been said. The above success factors can be complemented with dozens of other 'nice things to have' as integrated visual mining functionality. Current technological state of the art opens a wide spectrum of possibilities in this respect.

6 Conclusion

In this paper, we situated and motivated the need for a post-processing phase to the association rule mining algorithm when plugged into the knowledge discovery in database process. Association rule mining often yields a huge amount of discovered patterns making it very difficult for the user to focus on the important ones. Rightful assessment of the usefulness of the generated output introduces the need to effectively deal with different forms of data redundancy and data being plainly uninteresting. Hence, post-processing may play a pivotal role in the mining process. In this paper, we covered four post-processing tasks, more specifically, pruning, summarising, grouping and visualisation, taking into account the main efforts that have already been reported in the literature. Although much of the literature is currently focused on the first two transformations, the last two, let alone their integration with the former ones, still remain very unexplored. Furthermore, the need for integration of more than merely statistically inspired argumentation into the post-processing phase is posited as a crucial topic for further research.

Acknowledgement

This work has been realised under the auspices and sponsoring of the KBCInsurance Research Chair that was set up in September 1997 as a pioneering research co-operation between the Leuven Institute for Research in Information Systems (LIRIS) and the KBC bank and insurance group, which is one of the larger super regional bank & insurance groups in the Benelux, Europe with head office in Belgium.

7 References

[1] Agrawal R. & Shafer J., "Parallel Mining of Association Rules", In *IEEE Knowledge & Data Engineering*, 8(6), 1996.

[2] Agrawal R. & Srikant R., "Fast Algorithms for Mining Association rules", In *Proc. of the 20th Int'l Conf. on Very Large Databases*, Santiago, Chile, 1994.

[3] Agrawal R., Imielinski T. & Swami A., "Mining Association Rules between Sets of Items in Massive Databases", In *Proc. of the ACM SIGMOD Int'l Conference on Management of Data*, Washington D.C., USA, 1993.

[4] Ali K., Manganaris S. & Srikant R., "Partial Classification Using Association Rules", In *Proc. of the Third International Conference on Knowledge Discovery and Data Mining (KDD-97)*, Newport Beach, CA, USA, 1997.

[5] Bayardo Jr. R.J. & Agrawal R., "Mining the Most Interesting Rules", In *Proc. of the 5th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining,* San Diego, CA, USA, 1999.

[6] Brijs T., Vanhoof K. & Wets G., "Reducing Redundancy in Characteristic Rule Discovery by Using IP-Techniques", Limburgs Universitair Centrum, ITEO No. 99/03, 1999.

[7] Brin S., Motwani R., Ullman J.D. and Tsur S., "Dynamic itemset counting and implication rules for market basket data", In *Proc. of the ACM SIGMOD Conference on Management of Data*, 1997

[8] Fayyad U.M., Piatetsky-Shapiro G., Smyth P. & Uthurusamy, Advances in Knowledge Discovery and Data Mining, AAAI/MIT Press, 1996.

[9] Fukuda T., Morimoto Y., Morishita S. & Tokuyama T., "Data Mining Using Two Dimensional Optimized Association Rules: Scheme, Algorithms and Visualizatio", In *Proc. of the 1996 ACM SIGMOD Conference*, Montreal, Quebec, Canada, 1996.

[10] Guillaume S., Guillet F. & Philippé J.,"Contribution of the integration of intensity of implication into the algorithm proposed by Agrawal ", *In Biennal European Meeting on Cybernetics and Systems Research*, Vienna, 1998.

[11] Han J., "Towards On-Line Analytical Mining in Large Databases", SIGMOD Record, Vol. 27, No. 1, 1998.

[12] Klemettinen M., Mannila H., Ronkainen P., Toivonen H. & Verkamo A.I., "Finding interesting rules from large sets of discovered association rules", In *Proc. of the Third International Conference on Information and Knowledge Management (CIKM'94)*, Gaithersburg, Maryland, USA, 1994.

[13] Korn F., Labrinidis A., Kotidis Y. & Faloutsos C; "Ratio Rules: A New Paradigm for Fast, Quantifiable Data Mining", In *Proc. of the 24th VLDB Conference*, New York, USA, 1998.

[14] Liu B., Hsu W. & Ma Y., "Mining Association Rules with Multiple Minimum Supports", *ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD-99)*, San Diego, CA, USA, 1999.

[15] Liu B., Hsu W. & Ma Y., "Pruning and Summarizing the Discovered Associations", *ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD-99)*, San Diego, CA, USA, 1999.

[16] Liu B., Hsu W., Wang K. & Chen S., "Visually Aided Exploration of Interesting Association Rules", In *Proc. of the Pacific_Asia Conf. on Knowledge Discovery and Data Mining 1999 (PAKDD '99)*, Beijing, China, 1999.

[17] Park J., Chen M. & Yu P., "An effective hash based algorithm for mining association rules". In *SIGMOD Conf.*, 1995.

[18] Pei J., Han J. & Yin Y. , " Mining Access Patterns efficiently from Weg logs", In *Proc. of the Pacific_Asia Conf. on Knowledge Discovery and Data Mining 2000 (PAKDD '00)*, Kyoto, Japan, April 2000.

[19] Shah D., Lakshmanan L.V.S., Sudarshan S. & Ramamritham K., "Interestingness and Pruning of Mined Patterns", In *Proc. of the 1999 ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery (DMKD)*, Philadelphia, Penn, 1999.

[20] Silverstein C., Brin S. & Motwani R., "Beyond Market Baskets: Generalizing Association Rules to Dependence Rules", *Data Mining and Knowledge Discovery*, 2, 1998.

[21] Srikant R. & Agrawal R., "Mining Generalized Association Rules", In *Proc. 1995 Int. Conf. Very Large Data Bases*, Zurich, Switzerland, 1995.

[22] Srikant R. & Agrawal R., "Mining Quantitative Association Rules in Large Relational Tables", In *Proc. of the 1996 ACM SIGMOD Conference*, Montreal, Quebec, Canada, 1996.

[23] Toivonen H., Klemettinen M., Ronkainen P., Hätönen K. & Mannila H., "Pruning and Grouping of Discovered Association Rules", In *Mlnet Workshop on Statistics, Machine Learning, and Discovery in Databases,* Heraklion, Crete, Greece, 1995.

[24] Viveros M.S., Nearhos J.P. & Rothman M.J., "Applying Data Mining Techniques to a Health Insurance Information System", In *Proceedings of the 22nd VLDB Conference,* 1996.

[25] Zaki M., Parthasarathy S., Ogihara M. & Li W., "New Algorithms for fast discovery of association rules", *TR 651*, CS Dept, Univ. of Rochester, 1997.