Exploiting Generalized Association Rules*

Marcos Aurélio Domingues[1] and Solange Oliveira Rezende[2]

[1] LIAAD-INESC Porto L.A. – University of Porto
[2] ICMC – University of Sao Paulo
marcos@liaad.up.pt, solange@icmc.usp.br

Abstract. Association Rules is a Data Mining technique frequently used for decision support in market basket analysis, marketing, retail, and so forth. However, the mining of association rules may generate large quantities of patterns, complicating the patterns analysis. An approach that can help the analysis of the association rules is the use of taxonomies. In this paper we propose a system that uses taxonomies to generalize association rules and to analyze the generalized rules.

1 Introduction

The problem of mining association rules was introduced in [2]. Given a set of transactions, where each transaction is a set of literals, called items, an association rule is represented like an expression $LHS \Rightarrow RHS$. The LHS and RHS are, respectively, the *Left Hand Side* and the *Right Hand Side* of the rule, defined by distinct sets of items. The intuitive meaning of such a rule is that transactions in the database which contain the items in LHS tend to also contain the items in RHS. So, the association rules are used to find out the tendency that allows the user to understand and exploit the behavior patterns of the data. An example of such a rule might be that 80% of the customers who purchase the Q product also buy the W product. Here 80% is called the confidence of the rule.

The association rules technique has caught the attention of companies and research centers. Several researches have been carried out with this technique and the results have been used by companies to improve their businesses (marketing, insurance policy, demographics). However the use of association rules may generate large quantities of patterns, complicating the patterns analysis.

An approach to solve the problem of large quantities of patterns, extracted by the association rules technique, is the use of taxonomies [1]. The taxonomies can be used to prune uninteresting and/or redundant rules (patterns). In this paper we propose a system that uses taxonomies to generalize association rules and to analyze the generalized rules.

2 A System for Exploiting Generalized Association Rules

The problem of mining association rules is to find all rules that satisfy a user-specified minimum support and minimum confidence [2]. In most cases, tax-

* Thanks to FCT, Portugal (SFRH/BD/22516/2005) and CNPq, Brazil.

onomies (*is-a* hierarchies) over the items of association rules are available. An example of a taxonomy is presented in Fig. 1. This taxonomy says that t-shirt *is-a* light cloth, short *is-a* light cloth, light cloth *is-a* sport cloth, etc. Generalized association rules generate rules that span different levels of the taxonomy. For example, we can infer a rule that people who buy light cloth tend to buy tennis (light cloth ⇒ tennis) from the fact that people bought t-shirt with tennis (t-shirt ⇒ tennis) and short with tennis (short ⇒ tennis).

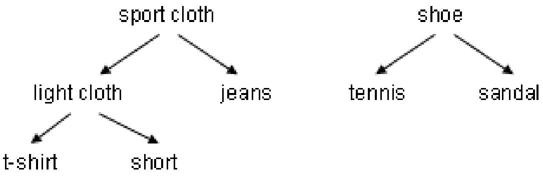


Fig. 1. Example of a taxonomy.

To make use of taxonomies to generalize association rules and to analyze the generalized rules, we propose the system $\mathcal{ENGAR}$ (Environment for Generalization and Analysis of Association Rules). This system is a desktop version of the Web module for generalized association rules, *RulEE-GAR*, proposed in [3]. It was developed due to the low performance of the Web module in processing big data sets. The system (Fig. 2) is composed for 3 modules: Data Entry, Generalization and Analysis.

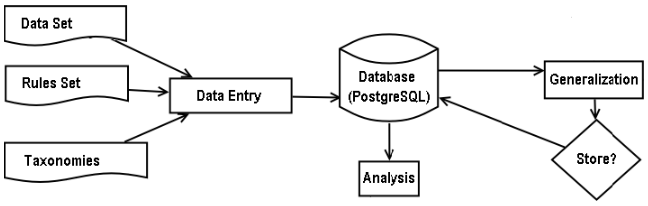
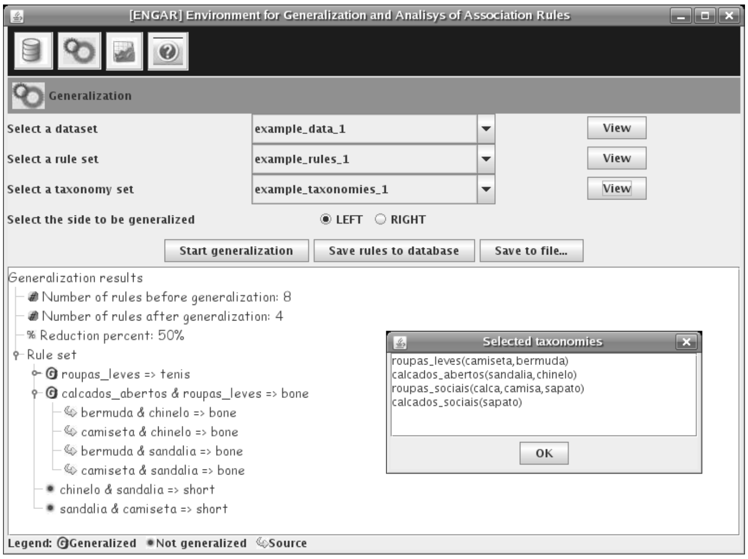


Fig. 2. Architecture of the system $\mathcal{ENGAR}$.

The Data Entry module is responsible for loading the text files containing the data set, rules set and taxonomies, to be used by the other modules. The Generalization module is the part of the system responsible for executing the algorithms of generalization of association rules. The current release of the system uses the $\mathcal{GART}$ algorithm proposed in [3]. Finally, the third module contemplates the functionality of analysis of the system, where a set of rules, previously generalized and stored in the system, can be evaluated through some kinds of analysis, for example, exploring methods [3] (to visualize the generalized rules expanded, the regular rules that were generalized, and so forth) and quality measures [4] (confidence, support, correlation, lift, laplace, etc). In Fig. 3 we show a screen of the system $\mathcal{ENGAR}$, where we can generalize and exploit the association rules.



[ENGAR] Environment for Generalization and Analisys of Association Rules

Select a dataset: example_data_1

Select a rule set: example_rules_1

Select a taxonomy set: example_taxonomies_1

Select the side to be generalized ● LEFT ○ RIGHT

```
Generalization results
├ # Number of rules before generalization: 8
├ # Number of rules after generalization: 4
├ % Reduction percent: 50%
└ Rule set
   ├ G roupas_leves => tenis
   ├ G calcados_abertos & roupas_leves => bone
   │  ├ bermuda & chinelo => bone
   │  ├ camiseta & chinelo => bone
   │  ├ bermuda & sandalia => bone
   │  └ camiseta & sandalia => bone
   ├ ● chinelo & sandalia => short
   └ ● sandalia & camiseta => short
```

Selected taxonomies

```
roupas_leves(camiseta,bermuda)
calcados_abertos(sandalia,chinelo)
roupas_sociais(calca,camisa,sapato)
calcados_sociais(sapato)
```

Legend: G Generalized ● Not generalized Source

Fig. 3. A screen of the system $\mathcal{ENGAR}$.

3 Case Study

We applied our system in a sales data set of a small supermarket. The database contained sales data of 3 month. We made 5 partitions of the data set to carry out this analysis. The partitions contained sales data along of 1 day (32668 rules generated), 7 days (19166 rules generated), 14 days (16053 rules generated), 1 month (21505 rules generated) and 3 months (19936 rules generated). The rules sets were generated using the *Apriori* algorithm with minimum support value equal 0.5, minimum confidence equal 0.5 and a maximum number of 5 items by rule. We also asked to an expert to make 18 different sets of taxonomies.

We ran the $\mathcal{GART}$ algorithm combining each set of taxonomies with each set of rules. In Fig. 4, a chart shows the reduction rates of the 5 rules sets after running the $\mathcal{GART}$ algorithm, using the 18 sets of taxonomies, to generalize each rules set. In Fig. 4, the sets of taxonomies are called "T" followed by an identification number, as for example: T01. The reduction rates go from 14,61% to 50,11%.

Now the generalized rules can be analyzed using the system $\mathcal{ENGAR}$ and the results, for example, can be used to change the layout of the supermarket. For the sake of confidentiality, we can not show other results.

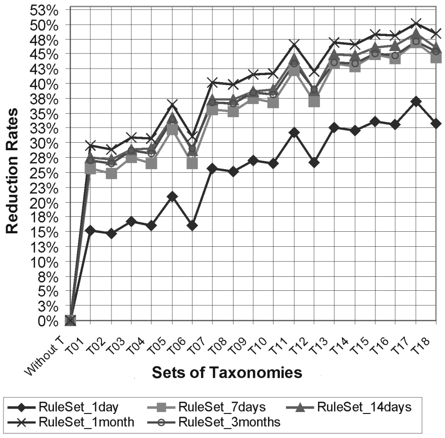


Fig. 4. Reduction rates using taxonomies to generalize association rules.

4 Conclusions and Future Work

In this paper we proposed a system that uses taxonomies to generalize association rules and to analyze the generalized rules. The taxonomies can be used to prune uninteresting and/or redundant rules, facilitating the analysis of large rules sets.

In [5], Li Yang uses parallel coordinates to visualize regular and generalized association rules. Our proposal is not so sophisticated as the one proposed in [5]. However our system provides other functionalities to exploit and to analyze the quality of generalized rules, making it one more option to analyze such rules.

As future work, we plan to make the system $\mathcal{ENGAR}$ an extensible plug-in for the Weka Data Mining Software [3]. We also plan to validate the system carrying out other experiments using artificial and real data sets.

References

1. Jean-Marc Adamo. *Data Mining for Association Rules and Sequential Patterns.* Springer-Verlag, New York, NY, 2001.
2. Rakesh Agrawal and Ramakrishnan Srikant. Fast algorithms for mining association rules. In *Proceedings 20th International Conference on Very Large Data Bases*, pages 487–499, 1994.
3. Marcos Aurélio Domingues and Solange Oliveira Rezende. Post-processing of association rules using taxonomies. In *Proceedings of the 12th Portuguese Conference on Artificial Intelligence (EPIA 2005)*, pages 192–197, Covilha, Portugal, 2005.
4. Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. *Introduction to Data Mining, (First Edition).* Addison-Wesley Longman Publishing Co., Inc., USA, 2005.
5. Li Yang. Pruning and visualizing generalized association rules in parallel coordinates. *IEEE Transactions on Knowledge and Data Engineering*, 17(1):60–70, 2005.

[3] http://www.cs.waikato.ac.nz/ml/weka.