Visual Post-Analysis of Association Rules[1]

Dario Bruzzese and Cristina Davino

Department of Mathematics and Statistics
University of Naples Federico II
Via Cintia Monte S. Angelo
I-80126 Naples, Italy
{dbruzzes, cdavino}@unina.it

Abstract. Association Rules (AR) represent a consolidated tool in Data Mining applications as they are able to discover regularities in large data sets. The information mined by the rules is very often difficult to exploit because of the presence of too many associations where to detect the really relevant logical implications. In this framework, by combining methodological and graphical pruning techniques, AR post-analysis tools are proposed. The methodological techniques will ensure the statistical significance of the AR which were not pruned while the graphical ones will provide interactive and powerful visualization tools.

1 Introduction

The post analysis is one of the most crucial steps in the knowledge extraction process, mainly when data mining is performed through Association Rules (AR) [1]. Even if mining rules is a quite simple task, their analysis and interpretation is often difficult due to the huge number of rules that can not be manually inspected.

The main approaches used to face this problem are graphical tools and pruning methods. AR visualization tools proposed in literature [5] [9] [14] allow to obtain a global view of all the discovered rules but they are still lacking because a huge number of rules is displayed and most of them are uninteresting. On the other hand, pruning methods [8] [12] [13] [10] [11] allow to kill the redundant rules but the pruned set still requires graphical tools to be interpreted.

In order to exploit the synergic power of both the visualizing and pruning phase, we propose two different strategies to help the user analyzing and interpreting AR.

The first strategy, "Display and Prune", gives priority to the visualization phase that becomes an interactive tool for pruning unuseful and redundant rules. In the second strategy, "Prune and Display", the capability of Factorial Methods [2] to synthesize and to visualize multidimensional patterns is exploited on those rules survived to a pruning process, which is based on statistical tests.

[1] This research was partially supported by "Data Mining e Analisi Simbolica" COFIN2000 grant (Prof. C. Lauro).

The two strategies will be applied on a real data set provided by RAI regarding the Italian national television channel preferences.

The outline of the reminder of this paper is as follows. In section 2 we introduce Association Rules and some basic notations. Section 3 formalizes the proposed visual post-analysis strategies from the methodological and application point of view. Some concluding remarks and further developments are summarized in section 4.

2 Association Rules: some basic notations

Let $I = i_1, i_2, \ldots, i_m$ be a set of items, called literals, (e.g. all products bought by a group of customers), and $T = t_1, t_2, \ldots, t_n$ be a set of n transactions, where each transaction t_i is a subset of I (e.g. all products in a customer's basket).

An association rule R is an implication of the form: $A \rightarrow C$, where $A \subset I$ is the set of the antecedent items of the rule and $C \subset I$ is the set of the consequent items of the rule such that $A \cap C = \emptyset$. Each rule of the form $A \rightarrow C$ is characterized by two ratios:

- $Support : S_R = \frac{n_R}{n}$ where n_R is the number of transactions in T holding $A \cup C$;
- $Confidence : C_R = \frac{n_R}{n_A}$ where n_A is the number of transactions in T holding the antecedent items A.

The *Support* measures the proportion of transactions in T containing both A and C and it is not related to the possible dependence of C from A. On the other hand, the *Confidence* aims at measuring the strength of the logical implication described by the rule and it refers to the conditional probability of the consequent given the antecedent. The concept of Support can be also referred to a generic item set if the proportion of transactions sharing the item set is considered.

Usually minimum support and minimum confidence values are fixed by the user before mining association rules. These values are generally user-dependent and unproper choices may cause many drawbacks: if they are set very low a huge number of rules (some of which being meaningless) will be found. On the contrary, if they are set very high, trivial rules will be found [13]. Another problem is the strength of the associations being commonly evaluated only by means of the confidence values and no assessment of the statistical significance of these values is made. Furthermore, even if the visualization of the rules has to be framed in a multidimensional context, most of the solutions proposed in literature force their representation in two-dimensional grids without considering the interaction among them.

3 Visual post-analysis of Association Rules

The visual post-analysis approach proposed in this paper aims at exploiting the synergic power of both the visualizing and the pruning phases in order to improve

the profitability of the discovered association rules. The approach is structured into two concurrent strategies where the graphical and the pruning phases have different priorities.

The first strategy, "Display and Prune", is based on the use of parallel coordinates in order to visualize the discovered rules; each item is a dimension of the graph and it spans according to the utility it provides to the rule. The previous utility is defined by an index called Item Utility (IU) able to take into account the confidence of the rule with or without each item. The user can visually inspect the rules and prune those ones whose items are below a specified IU threshold.

In the second strategy, "Prune and Display", a pruning method [4] based on three statistical tests is followed by the introduction of Factorial Methods in order to synthesize the information stored in the rules and to represent themselves, the items and their interactions on 2-D graphs [3].

The two strategies will be applied on a real dataset provided by RAI (Italian public television) regarding the channel preferences of the official users panel (9965 units) collected by the RAI in order to study television customers behaviors. The preferences refer to 39 different typologies of television programs (variety show, cartoons, musical, sports time, religious time, etc.) and each user is described by the sub-set of genres he has watched for more than five minutes per day during a week (a threshold equal to five minutes allows to avoid users that frequently change channel). In appendix, the list of the considered genres is reported.

The aim of the application is to explain the television customers behaviour through the discovery of the logical associations among the various typologies of television programs in the set of 9965 transactions. Considering only those rules with at most three items in the antecedent and one item in the consequent, fixing very low support (0.01) and confidence values (0.01), a huge number of rules (35783) is obtained.

3.1 The "Display and Prune" strategy

One of the input parameters of association rules mining algorithms is represented by the number of different items in the association, defined as the order of the rule. It can happen that not all the antecedent items give a real contribute to the rule confidence value but if those items are drawn away and a lower order rule is considered, the rule confidence can improve. Viceversa, increasing the order of the rule by adding an item in the antecedent part can significantly enhance the confidence value if the added item is very useful in explaining the consequence. The two previous cases are graphically represented in figure 1 where each rectangle describes an item and its surface is proportional to the item support. From figure 1(a) it results that the presence of item B is relevant because in case of its absence (1(c)) the confidence value reduces while from figure 1(b) it results that by drawing away B from the association, the different interactions among the items cause a rise in the confidence value.

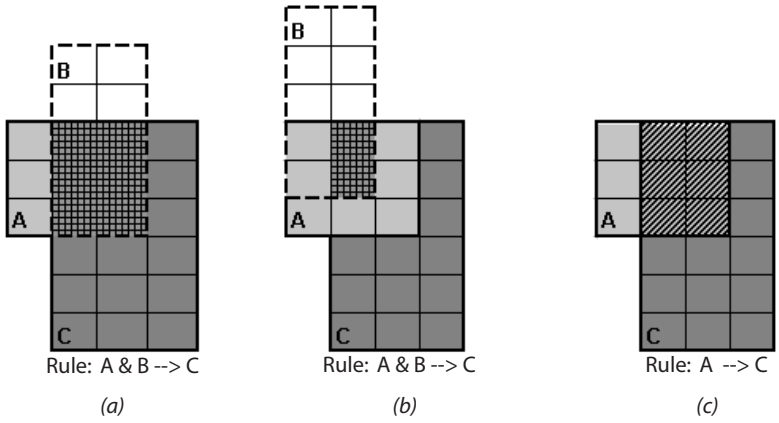

Fig. 1. A graphical representation of the item B utility in the rule $A \ \& \ B \rightarrow C$

In order to measure the real utility of an item i in the premise of a rule R, we introduce an index called *Item Utility* (IU) based on the comparison between the confidence of the rule with or without item i:

$$IU_i = \frac{C_R - C_{R(-i)}}{max(C_R; C_{R(-i)})} \tag{1}$$

If $IU_i \in \ ]-1;0[$, the item is not useful but dangerous and the rule can be pruned as a lower order rule is the real association; the case $IU = 0$ refers to the presence of a redundant item because $A \subseteq B$, the confidences of the rules $A\&B \rightarrow C$ and $A \rightarrow C$ are equal and the real association is described by the rule $A \rightarrow B$; if $IU_i \in \ ]0;1[$, the item is very useful in the rule as it improves the capability of the antecedent to explain the consequence.

The discovered association rules are visualized plotting on parallel coordinates the IU of each item belonging to the antecedent of a rule. Parallel coordinates, introduced by Inselberg in 1981 [6], represent a very useful graphical tool fort the visualization of high dimensional data-sets in a two-dimensional space. They appear as a set of vertical axes where each axis describes a dimension of the domain and each case is represented by a line joining its values on the parallel axes. In the proposed parallel visualization of association rules , each antecedent item is a dimension of the graph and it spans according to the utility provided to each rule.

Some of the interaction tools of parallel coordinates [7] are exploited in order to visualize, interpret and reduce the number of rules. In particular, data analysis can be facilitated by:

- selecting a subgroup of rules with one or more items belove a specified IU threshold in order to remove selected lines from the plot;

- identifying axes (items) with very dense positive values, given a consequent, in order to highlight items with a high explicative power;
- adding two supplementary dimensions corresponding to the support and confidence of the rules in order to remove those rules with values of these parameters below a specified threshold;
- selecting high confidence rules in order to identify sets of items involved in very strong associations;
- changing the order of the dimensions on the basis of IU distributions.

The "Display and Prune" strategy has been applied to the rules discovered on the RAI dataset. In figure 2, a coordinate plot of the utilities of the items in the 1652 rules sharing the same consequent (FLC: series) is provided. The lines corresponding to rules with the item INP (parliamentary news), that has very dense negative IU values, are highlighted in black. This set of 96 rules can be pruned because lower order but more explicative rules will remain.

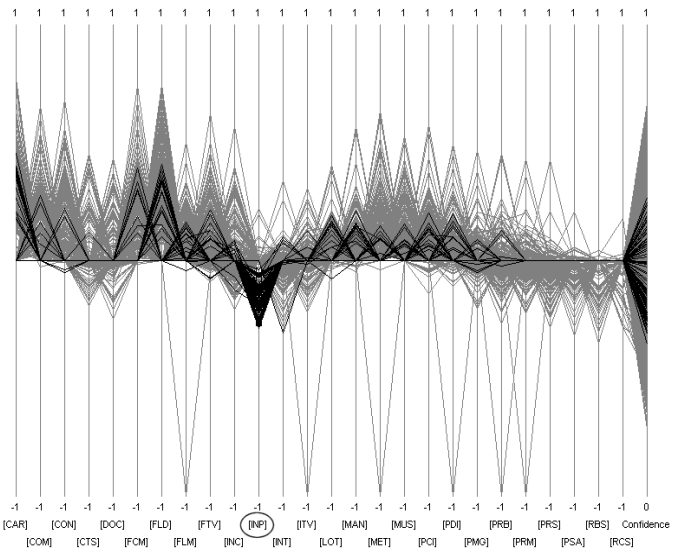

Fig. 2. A parallel coordinate plot of the rules with the consequent equal to FLC.

In figure 3, a subset of 1084 rules with the consequent equal to FLC and with the confidence greater than 0.5 (the choice of this threshold will be justified in Section 4.2) is shown. It is worth of noticing that a huge number of rules with one or more items having negative values for IU still survive thus confirming the guess that high confidence does not ensure the explicative power of each item involved in the rule. A further consideration regards items with high IU values that are still present after the confidence filter and representing very important dimensions in strong associations.

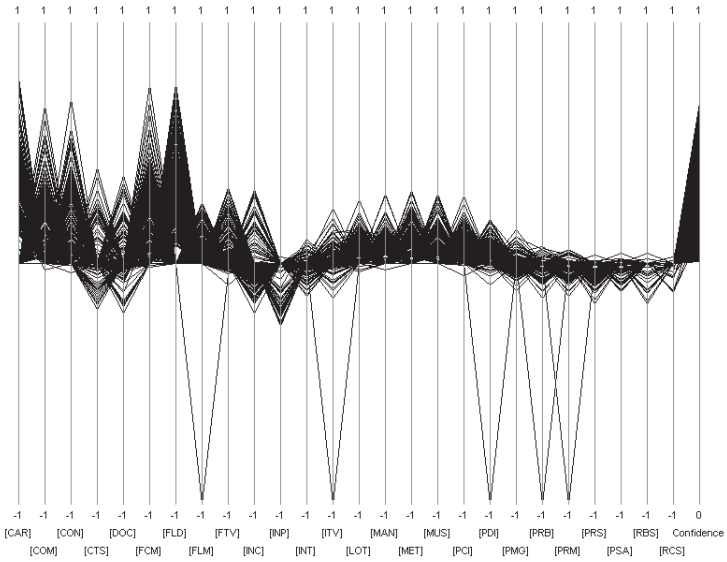

Fig. 3. Plot of the rules with the consequent equal to FLC and $C_R > 0.5$.

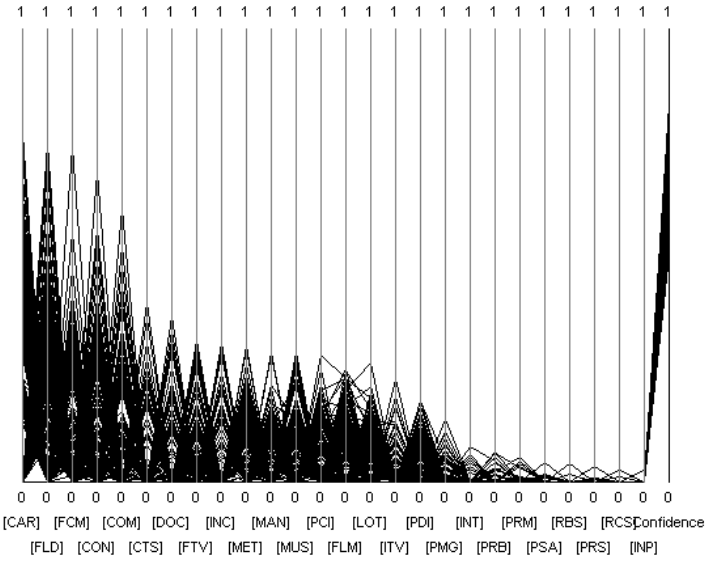

Fig. 4. Plot of the rules survived to the confidence and IU pruning.

Finally, the 463 rules with confidence greater than 0.5 and positive IU for each item in the antecedent are plotted in figure 4. The possibility of reordering the dimensions of the parallel graph on the basis of IU values allows to visually and immediately capture the most relevant items for the analyzed consequent and to face one of the main limits of these graphs which is represented by the arbitrary variables ordering.

3.2 The "Prune and Display" strategy

The "Prune and Display" strategy follows the PEV (**P**runing, **E**xploring, **V**isualizing) approach proposed by the authors in [3] that aims firstly at allowing the user finding automatically a reduced subset of rules, secondly at capturing the most relevant structure inside the pruned set and, finally, at providing graphical tools able to represent the rules, the items and their interactions on 2-D graphs .

AR obtained by a mining process with very low support and confidence values are sequentially pruned using three statistical tests performed on the significance of the consequence, of the antecedent and of the confidence. At each step, the rules are ranked according to the corresponding test statistic and a subset of rules is obtained by pruning the rules out of a suitable threshold. This subset becomes the rules input set for the next step.

The three tests are described in the following.

A test on the significance of the consequence.
Rules whose high confidence is only related to the presence of very frequent items are pruned through a test performed comparing the rule confidence with the support of the consequent part of the rule. This test allows to evaluate how much the antecedent part is able to explain the variability of the consequent part.

A test on the significance of the antecedent.
A chi-2 test is performed for each group of rules given the same antecedent part in order to evaluate how significant the logical implication is and to avoid meaningless associations. The test is based on the comparison between the support of each rule of the group and the theoretical support in case of casual associations with the given antecedent part.

A test on the significance of the confidence.
A test on the significance of the confidence is necessary in order to evaluate how strong, from a statistical point of view, the logical implication between A and C is. The test proposed in [4] is based on the comparison between the rule support and the support of the antecedent part of the rule that means evaluating how much the confidence is far from an strong implication. A "soft" version of this test is now introduced by comparing the rule support with the proportion of transactions sharing the antecedent items but not the consequent. This comparison allows to find rules with a confidence not significantly far from a "soft" implication ($C_R = 0.5$).
The proposed "soft" test is based on the following hypothesis:

$$H_0 : S_R = S_A - S_R \qquad H_1 : S_R > S_A - S_R$$

deriving from the assumption that the association between A and C, measured by the support of the rule (S_R), should be at least equal to the association between A and $\bar{C}$ measured by $S_A - S_R$. After simple algebraic manipulations, the test hypothesis becomes as follows:

$$H_0 : C_R = 0.5 \qquad H_1 : C_R > 0.5$$

and, under H_0, C_R is a binomial random variable that can be approximated with a normal distribution:

$$C_R \sim N\left(0.5; \sqrt{\frac{0.5 \cdot (1 - 0.5)}{n_A}}\right)$$

if n_A is sufficiently large.
It follows that the test statistic $V_{ConfSoft}$ is a standardized normal random variable:

$$V_{ConfSoft} = \frac{C_R - 0.5}{\sqrt{\frac{0.5 \cdot (1-0.5)}{n_A}}} \sim N(0; 1).$$

The results of the pruning phase applied to the 35783 rules mined from the RAI data set are summarized in the following table.

Table 1. Information about the pruning process.

	Before Pruning	After Step 1	After Step 2	After Step 3
Number of rules	6901	32872	10649	1562
Minimum Confidence	0.06	0.09	0.09	0.53
Maximum Confidence	1	0.87	0.85	0.85
Minimum Support	0.01	0.01	0.01	0.01
Maximum Support	0.15	0.15	0.15	0.15

After the first step, about 10% of the rules are pruned without influencing significatively the support and confidence ranges. The second step allows to prune a huge number of rules (22232) as it works on groups of rules with the same antecedent items. Using the "soft" test in the third step, 1562 rules survive with a minimum confidence equal to 0.53 compared to the 8536 associations with a confidence greater than 0.53 in the original set of mined rules. This gap is due to the nature of the used pruning approach that allows to kill also rules whose confidence, although high, does not ensure a real logical implication.

The 1562 survived associations still represent a huge number of rules to be manually inspected by the user so that performing a factorial method can allow to synthesize the information and to visualize the associations structure on 2-dimensional graphs. A Multiple Correspondence Analysis (MCA) [2] is applied

to an $n \times (p+2)$ data matrix where n represents the number of rules survived to the pruning steps and p corresponds to the total number of different items both in the antecedent part and in the consequent part of the n rules; support and confidence values are also considered. Each rule is coded by a binary array assuming value 1 if the corresponding column item is present in the rule and value 0 otherwise. Different roles are assigned to the $p+2$ variables in the MCA: the antecedent items are the *active* variables (variables that intervene directly in the analysis and define the factorial planes) while the consequent items and the support and the confidence values are the *supplementary* variables (variables depending by the former and projected later on the defined factorial planes). This choice is related to the logical link that exists between the two components of a rule where the antecedent part represents the logical premise of the consequent part.

Once the MCA is performed, it is possible to represent either the factorial planes allowing to explain at least a user defined threshold of the total variability either a user defined factorial plane or the factorial plane best defined by a user chosen item.

One of the possible views offered by MCA allows to graphically represent the association structures among the antecedent and the consequent items.

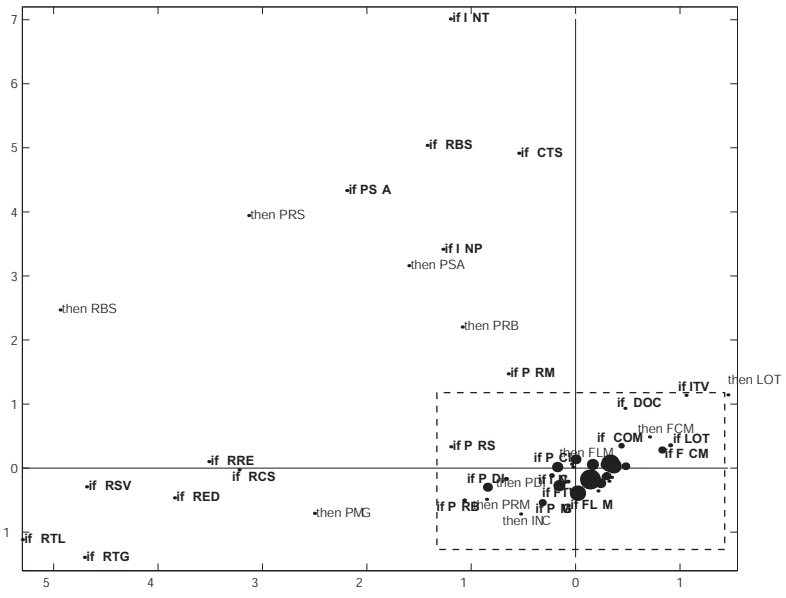

Fig. 5. The Item Representation.

In figure 5 the antecedent and the consequent items are plotted on the first factorial plane with a dimension proportional to the number of rules sharing them and in figure 6 a zoom of the selected area is shown. The support and

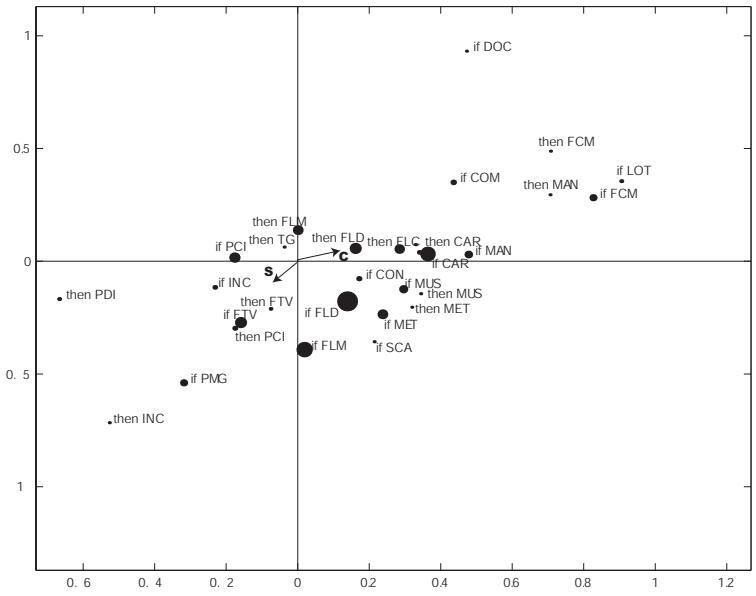

Fig. 6. A zoom of the Item Representation.

the confidence measures can be represented by oriented segments linking the origin of the axes to their projection on the plane as their coordinates are the correlation coefficients with the axes. The previous expedient allows to identify privileged regions in the plane with high supports and confidences: in figure 5 the first quadrant contains items associated to high confidence but low support rules. The proximity between two antecedent items shows the presence of a set of rules sharing them while the proximity between two consequent items is related to a common causal structure. Finally, the closeness between antecedent items and consequent items highlights the presence of a set of rules with a common dependence structure. For example, the consequent item FLC results very close to antecedent items such as *Cartoons* (CAR), *Comic programme* (COM), *Musical* (MUS), etc thus identifying genres that contribute to explain the same consequent. This result confirms the interpretation of figure 4 where the aforementioned items showed very dense IU positive values.

Once the common structure has been grasped in the item visualization, it is possible to come back to the associations by plotting, on the factorial plane, the rules with a dimension proportional to their confidence. In this representation, the proximity among two or more rules gives evidence to the presence of a common structure of the antecedent items associated to different consequences and it allows to change the set of close rules into a higher order macro-rule by linking the common antecedent items to the logical disjunction of the different consequent items.

4 Concluding remarks

In this paper the Association Rules post-processing problem is faced in order to both guarantee statistical significance of mined implications and to make them usable and interpretable through interactive graphical tools. The priority given to the methodological pruning or to the graphical representation depends on the data analyst requirements and it leads to the choice between the "Display and Prune" and the "Prune and Display" strategies.

The modular feature of the proposed approaches gives the possibility to differently combine them, e.g. to plot on parallel coordinates the rules sub-set survived to the statistical pruning or to synthesize the rules with high IU items by MCA.

The calculation of the measures involved in the two post-analysis tools (IU values and test statistics values) has very low computational costs as it is based on the use of information already computed during the mining process.

Further developments will regard the statistical evaluation of the IU measure to introduce an objective threshold in order to identify the items characterized by the worse or best IU values and the analysis of the validation and visualization issues in case of Classification Rules.

References

1. Agrawal, R., Imielinski, T. & Swami, A.: Mining Association Rules between Sets of Items in Large Databases, Proceedings of the 1993 ACM SIGMOD Conference, May, Washington DC, USA, (1993) 207–216.
2. Benzècri, J.-P.: L'Analyse des Donnèes, Dunod, Paris (1973).
3. Bruzzese, D. & Davino, C.: Pruning, Exploring and Visualizing Association Rules, Statistica Applicata, vol. 12, n. 4, (2000) 461–472 .
4. Bruzzese, D. & Davino, C.: Statistical Pruning of Discovered Association Rules, Computational Statistics, vol. 16 (2001) 387–398.
5. Hofmann, H. & Wilhelm, A.: Validation of Association Rules by Interactive Mosaic Plots. In Bethlehem, J.G., van der Heijden, P.G.M. (eds.): Compstat 2000 - Proceedings in Computational Statistics, Physica-Verlag, Heidelberg (2000) 499–504.
6. Inselberg, A.: N-dimensional Graphics, part I - Lines and Hyperplanes, in IBM LASC Tech. Rep. G320-2711, 140 pages. IBM LA Scientific Center (1981).
7. Inselberg, A.: Visual Data Mining with Parallel Coordinates, Computational Statistics, vol. 13, n.1, (1998) 47–64.
8. Klemettinen, M., Mannila, H., Ronkainen, P., Toivonen, H., & Verkamo, A.I.: Finding interesting rules from large sets of discovered association rules, Proceedings of the Third International Conference on Information and Knowledge Management CIKM-94, (1994) 401–407.
9. Liu, B., Hsu, W. & Ma, Y.: Pruning and Summarizing the Discovered Associations, Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-99), August 15-18, San Diego, CA, USA (1999).
10. Liu, B., Hsu, W., Wang, K. & Chen, S.: Visually Aided Exploration Interesting Association Rules. Proceedings of the Pacific Asia Conference on Knowledge Discovery and Data Mining (PAKDD-99). Springer Eds., April 26-28, Beijing (1999).

11. Shah, D., Lakshmanan, L.V.S., Ramamritham, K. & Sudarshan S.: Interestingness and Pruning of Mined Patterns, Workshop Notes of the 999 ACM SIGMOD Research Issues in Data Mining and Knowledge Discovery (1999).
12. Toivonen, H., Klemettinen, M., Ronkainen, P. , Hatonen, K. & Mannila, H.: Pruning and grouping of discovered association rules. Workshop Notes of the ECML-95 Workshop on Statistics, Machine Learning, and Knowledge Discovery in Databases, Heraklion, Greece, April 1995 (1995) 47–52.
13. Weber, I.: On Pruning Strategies for Discovery of Generalized and Quantitative Association Rules. Proceedings of Knowledge Discovery and Data Mining Workshop, Singapore (1998).
14. Wong, P.C., Whitney, P., Thomas, J.: Visualizing Association Rules for Text Mining. In Wills, G., Keim, D. (eds.): Proceedings of IEEE Information Visualization '99, IEEE CS Press, Los Alamitos, CA (1999).

Appendix: Television genres legend

Table 2. TV genres.

CAR: cartoons	COM: comic programme	CON: concert
CTS: customs and society	DOC: documentary	FCM: short
FLC: series	FLD: film with discussion	FLM: film
FTV: TV film	INC: survey	INP: parliamentary news
INT: suspension	ITV: break	LOT: lottery
MAN: demonstrations	MET: weather-forecast	MUS: musical
PCI: showing	PDI: discussion	PMG: montage
PRB: children programme	PRM: advertising	PRS: coming shortly
PSA: drama	RBS: sport programme	RCS: school programme
RED: editorial	RRE: religious magazine	RSV: documentary programme
RTG: tv news programme	RTL: current affairs	RUB: programme
SCA: Science	SCN: serial	TG: news
TLD: teaching programme	TLF: TV film	TQZ: quiz show