Labeling Methods for Association Rule Clustering

Veronica Oliveira de Carvalho[1], Daniel Savoia Biondi[1],
Fabiano Fernandes dos Santos[2] and Solange Oliveira Rezende[2]
[1]*Instituto de Geociências e Ciências Exatas, UNESP - Univ Estadual Paulista, São Paulo, Brazil*
[2]*Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Paulo, Brazil*

Keywords: Association Rules, Post-processing, Clustering, Labeling Methods.

Abstract: Although association mining has been highlighted in the last years, the huge number of rules that are generated hamper its use. To overcome this problem, many post-processing approaches were suggested, such as clustering, which organizes the rules in groups that contain, somehow, similar knowledge. Nevertheless, clustering can aid the user only if good descriptors be associated with each group. This is a relevant issue, since the labels will provide to the user a view of the topics to be explored, helping to guide its search. This is interesting, for example, when the user doesn't have, a priori, an idea where to start. Thus, the analysis of different labeling methods for association rule clustering is important. Considering the exposed arguments, this paper analyzes some labeling methods through two measures that are proposed. One of them, Precision, measures how much the methods can find labels that represent as accurately as possible the rules contained in its group and Repetition Frequency determines how the labels are distributed along the clusters. As a result, it was possible to identify the methods and the domain organizations with the best performances that can be applied in clusters of association rules.

1 INTRODUCTION

Association rules are widely used in many distinct domain problems due to their ability to discover the frequent relationships that occur among sets of items stored in databases. Although this characteristic motivates its use, the main weakness of the association technique occurs when it is necessary to analyze the mining results. The huge number of rules that are generated makes the user's exploration a difficult task. Many approaches have been developed to overcome this post-processing problem, such as *Querying*, *Evaluation Measures*, *Pruning*, *Summarizing* and *Grouping* (Zhao et al., 2009; Natarajan and Shekar, 2005; Jorge, 2004). There are other ways to reduce the number of rules before post-processing be done, using, for example, extraction algorithms that are not exhaustive as *Apriori* (Agrawal and Srikant, 1994). However, the focus of this work is the post-processing phase. Thus, it is considered, in this work, that it is better not to eliminate rules (knowledge) during the extraction process, but to work with all of them later.

Grouping is a relevant approach related to the structure of the domain, since it organizes the association rules, previously obtained by algorithms like *Apriori* (Agrawal and Srikant, 1994), in groups that contain, somehow, similar knowledge. These groups can improve the presentation of the mined patterns, providing the user a view of the domain to be explored (Reynolds et al., 2006; Sahar, 2002). The papers that use clustering for post-processing association rules, as seen in (Reynolds et al., 2006; Jorge, 2004; Sahar, 2002; Toivonen et al., 1995), are only concerned with the domain organization. However, it is essential that the organizations be used to aid the user during the exploration process, minimizing its effort. Aiding can be obtained from a structured domain by: (i) highlighting the groups (clusters[1]) that are interesting to be explored; (ii) generating good labels for the groups that allow an easier browsing in the domain.

Regarding (i), (Carvalho et al., 2011), for example, proposed the PAR-COM methodology that, by combining clustering and objective measures, reduces the association rule exploration space by directing the user to what is potentially interesting. Thus, the user only explores a small subset of the groups that contain the potentially interesting knowledge. Regarding (ii), it is essential that groups be represented by

[1]The words groups and clusters are used in this paper as synonymous.

Oliveira de Carvalho V., Savoia Biondi D., Fernandes dos Santos F. and Oliveira Rezende S..
Labeling Methods for Association Rule Clustering.
DOI: 10.5220/0003970001050111
In *Proceedings of the 14th International Conference on Enterprise Information Systems* (ICEIS-2012), pages 105-111
ISBN: 978-989-8565-10-5

labels that can provide the user a view of the topics contained in the exploration space, helping to guide its search. Finding good labels is a relevant issue in many tasks as in Text Mining (TM) and Information Retrieval (IR) (see some applications in (Manning et al., 2009)). It is necessary, for example, that good descriptors be presented to the user to facilitate exploratory analyses, interesting when the user doesn't have, a priori, an idea where to start. Furthermore, although many methods have been proposed to label document clusters in TM and IR, the papers related to association rule clustering have not explored this issue. Thus, as in other tasks, the analysis of different labeling methods for association rule clustering is also relevant, since it is necessary to identify the methods that present good results. Besides, the integration of good labeling methods with other methodologies can allow association rule clustering to become a powerful post-processing tool. The integration with PARCOM (Carvalho et al., 2011), for example, can enable the identification of the potentially interesting topics in the domain.

Considering the exposed arguments, this paper aims to analyze some labeling methods in order to identify: **(a)** the methods that are more adequate for association rule clustering; **(b)** the domain organizations that provide the best results, since the performance of the methods are affected by them, i.e., by a clustering algorithm combined with a similarity measure; **(c)** a consequence of **(b)**, the domain organizations that best structure the knowledge. Two measures are proposed and used to evaluate the methods. The ideal is that the labels of each cluster represent as accurately as possible the knowledge of its group (Precision (*P*) measure) and be as different as possible of the labels of the other groups (Repetition Frequency (*RF*) measure). It is important to mention that this paper doesn't fit in the post-processing approaches itself. The labeling methods here presented have to be applied to clustering of association rules, i.e., along with a post-processing methodology.

The paper is structured as follows: Section 2 presents some related works; Section 3 and Section 4 the labeling methods that were selected and the measures that were proposed to evaluate the experiments results, respectively; Section 5 the configurations used in experiments; Section 6 the results and discussion; Section 7 the conclusions and future works.

2 RELATED WORKS

Since this paper aims to analyze some labeling methods for association rule clustering, this section presents some papers related to the clustering approach and the labeling methods they use.

In order to structure the extracted knowledge, different clustering strategies have been used for post-processing association rules. In (Reynolds et al., 2006) clustering is demonstrated through partitional (K-means, PAM, CLARANS) and hierarchical (AGNES) algorithms using Jaccard as the similarity measure. In this case, the Jaccard between two rules r and s, expressed by $\text{J-RT(r,s)}=\frac{\#\{t \text{ matched by } r\}\cap \#\{t \text{ matched by } s\}}{\#\{t \text{ matched by } r\}\cup \#\{t \text{ matched by } s\}}$, is calculated considering the common transactions (t) the rules match – we refer to this similarity measure as Jaccard with Rules by Transactions (J-RT). A rule matches a transaction t if all the rule items are contained in t. (Jorge, 2004) demonstrates the use of clustering through hierarchical algorithms (Single Linkage, Complete Linkage, Average Linkage) also using Jaccard as the similarity measure. However, the Jaccard between two rules r and s, expressed by $\text{J-RI(r,s)}=\frac{\#\{\text{items in } r\}\cap \#\{\text{items in } s\}}{\#\{\text{items in } r\}\cup \#\{\text{items in } s\}}$, is calculated considering the items the rules share – we refer to this measure as Jaccard with Rules by Items (J-RI). (Toivonen et al., 1995) proposes a similarity measure based on transactions and uses a density algorithm to carry out the clustering of the rules. (Sahar, 2002) also proposes a similarity measure based on transactions considering (Toivonen et al., 1995)'s work, although using a hierarchical algorithm to carry out the clustering.

All the above papers, related to the structure of the domain, are only concerned with the domain organization. In general, each paper only uses one family of clustering algorithms along with one similarity measure to cluster the association rules and a unique labeling method to present the mined results to the user. (Reynolds et al., 2006) and (Jorge, 2004) select as labels of each group the items that appear in the rule which is more similar to all the other rules in the group (the medoid of the group). (Toivonen et al., 1995) doesn't mention how the labels are found, but provides some traces that the labels represent the more frequent and distinct items in the group. On the other hand, (Sahar, 2002) proposes an approach to summarize each cluster by finding the patterns $a \Rightarrow c$ that cover all the rules in the cluster; a and c are items in the domain and a pattern $a \Rightarrow c$ covers a rule $A \Rightarrow C$ if $a \in A$ and $c \in C$. As observed, although the proposed approach is used to summarize the clusters and not, in fact, to define the cluster's labels, the idea can be used for this purpose.

Although many methods have been proposed to label document clusters in tasks of Text Mining (TM)

and Information Retrieval (IR), as in (Moura and Rezende, 2010; Lopes et al., 2007; Kashyap et al., 2005; Fung et al., 2003; Glover et al., 2002; Popescul and Ungar, 2000; Larsen and Aone, 1999; Cutting et al., 1992), the papers related to association rule clustering have not explored this issue. However, as presented in next section, many of these methods used to label document clusters are similar to the ones used to label association rule clusters, i.e, they are, somehow, related. Thus, some methods, apart from the ones presented in the next section, could be adapted from TM and IR for association rule clustering.

3 LABELING METHODS

Aiming to analyze some labeling methods (LM) for association rule clustering regarding their behavior in relation to precision and distinctiveness, four methods were selected and implemented. These methods represent the ideas of many of the methods previously described and cited in Section 2 (both for association rules (AR) as for documents (TM and IR)). In order to understand the methods, consider a clustering composed of three clusters of association rules: C_1={r_1: coffee $\Rightarrow$ butter; r_2: milk $\Rightarrow$ coffee; r_3: milk & butter $\Rightarrow$ coffee}; C_2={r_1: butter $\Rightarrow$ coffee; r_2: milk $\Rightarrow$ butter}; C_3={r_1: butter $\Rightarrow$ milk; r_2: coffee $\Rightarrow$ milk}. The example is merely illustrative. The four methods described below are **LM-M**, **LM-T**, **LM-S** and **LM-PU**.

In **LM-M** (*L*abeling *M*ethod *M*edoid) the labels of each cluster are built by the items that appear in the rule of the group which is more similar to all the other rules in the cluster (the medoid of the group). So, is computed the accumulated similarity (a_s) of each rule considering its similarity with respect to the other rules and the one with the highest value is selected. Considering C_1 of the above example and that r_1 covers {t_1, t_3, t_5, t_7}, r_2 {t_1, t_3, t_5, t_7, t_9}, r_3 {t_3, t_5, t_7}, the similarities s(r_1,r_2)=s(r_2,r_1)=$\frac{4}{5} = 0.8$, s(r_1,r_3)=s(r_3,r_1)=$\frac{3}{4} = 0.75$, s(r_2,r_3)=s(r_3,r_2)=$\frac{3}{5} = 0.6$, considering J-RT (Section 2), are obtained and the following a_s are found: a_s(r_1)=s(r_1,r_2)+s(r_1,r_3)=1.55; a_s(r_2)=s(r_2,r_1)+s(r_2,r_3)=1.40; a_s(r_3)=s(r_3,r_1)+s(r_3,r_2)=1.35. Thus, r_1 is selected and C_1's labels are defined to be {coffee, butter}. These similarities among rules can be obtained through any similarity measure, as the ones presented in Section 2. In this paper we used J-RT as in the most of the literature works. The papers related with this idea are (Reynolds et al., 2006; Jorge, 2004) from AR and (Kashyap et al., 2005; Larsen and Aone, 1999; Cutting et al., 1992) from TM and IR. In this case, the user can also know the existing relationship among the labels through the rule.

In **LM-T** (*L*abeling *M*ethod *T*ransaction) the labels of each cluster are built by the items that appear in the rule of the group that covers the largest number of transactions. A rule covers a transaction t if all the rule items are contained in t. Considering C_1 of the above example and that r_1 covers {t_1, t_3, t_5, t_7}, r_2 {t_3, t_5, t_7}, r_3 {t_1, t_3, t_5, t_7, t_9}, r_3 is selected and C_1 labels are defined to be {milk, butter, coffee}. The paper related to this idea is (Fung et al., 2003) from TM and IR. In this case, the user can also know the existing relationship among the labels through the rule.

In **LM-S** (*L*abeling *M*ethod *S*ahar due to its reference to (Sahar, 2002)), a simplified version of the process described in (Sahar, 2002) from AR and explained in Section 2, the labels of each cluster are built as follows: (i) considering a set $I = \{i_1, ..., i_m\}$ containing all the distinct cluster items, a set $R = \{r_1, ..., r_n\}$ containing all the possible relationships $a \Rightarrow c$, where $a, c \in I$ – each one of these relationships represents a rule pattern; (ii) the number of rules that each pattern $r_i \in R$ covers is computed (N_c); a pattern $a \Rightarrow c$ covers a rule $A \Rightarrow C$ if $a \in A$ and $c \in C$; (iii) the pattern with the highest cover is selected; in the event of a tie all tied pattern are selected; (iv) all the selected patterns compose a set $P \subseteq R$; (v) at the end, all the distinct items in P compose the labels. Considering C_1 of the above example we have: I={coffee, butter, milk}, R={$r_1 : coffee \Rightarrow butter$, $r_2 : butter \Rightarrow coffee$, $r_3 : coffee \Rightarrow milk$, $r_4 : milk \Rightarrow coffee$, $r_5 : butter \Rightarrow milk$, $r_6 : milk \Rightarrow butter$}, N_c={$r_1 : 1$, $r_2 : 1$, $r_3 : 0$, $r_4 : 2$, $r_5 : 0$, $r_6 : 0$} and P={r_4}. Thus, C_1's labels are defined to be {milk, coffee}. In this case, the user can also know the existing relationship among the labels through the rule(s).

In **LM-PU** (*L*abeling *M*ethod *P*opescul and *U*ngar due to its reference to (Popescul and Ungar, 2000)) the labels of each cluster are built by the N items in the cluster that present the best tradeoff between frequency and predictiveness; formally we have: $f(i_n|C_n) * \frac{f(i_n|C_n)}{f(i_n)}$. The $f(i_n|C_n)$ measure computes the frequency f of each item i_n in its cluster C_n. The $\frac{f(i_n|C_n)}{f(i_n)}$ measure computes the frequency f of each item i_n in its cluster C_n divided by the item frequency in all the clusters. The i_n items are all the distinct items that are present in the rules of the cluster. Each time an item i_n occurs in a rule its frequency is incremented by one. Therefore, the labels are built by the N items that are more frequent in their own cluster and infrequent in the other clusters. Considering C_1 of the above example, its distinct items {coffee,

butter, milk} and $N = 1$ we have: coffee=$3 * \frac{3}{5}$=1.8; butter=$2 * \frac{2}{5}$=0.8; milk=$2 * \frac{2}{5}$=0.8. Thus, C_1's labels are defined to be {coffee}. The papers related to this idea are (Toivonen et al., 1995) from AR and (Lopes et al., 2007; Glover et al., 2002; Popescul and Ungar, 2000) from TM and IR. In this case, the user doesn't know the existing relationship among the labels.

4 EVALUATION METHODOLOGY

In order to evaluate the precision and distinctiveness of the four labeling methods, two measures, presented in Equations 1 and 2, were proposed, where N refers to the number of clusters. Both measures range from 0 to 1. To understand the measures, consider a clustering composed of three clusters of association rules: C_1={coffee $\Rightarrow$ butter; milk $\Rightarrow$ butter} with the labels {butter, coffee, milk}; C_2={butter $\Rightarrow$ coffee; milk $\Rightarrow$ coffee} with the label {milk}; C_3={butter $\Rightarrow$ milk; coffee $\Rightarrow$ milk} with the labels {butter, milk}. The example is merely illustrative.

Precision (P), in Equation 1, measures how much the labeling method can generate labels that really represent the rules contained in the clusters. This measure is an adaptation of Recall used in Information Retrieval (see (Manning et al., 2009)); however, in this case, the relevant items to be retrieved are all the rules in a cluster. Considering the above example, the illustrative method has a P of 0.83 ($P(C) = \frac{\frac{2}{2}+\frac{1}{2}+\frac{2}{2}}{3}$), since the labels of C_2 represent only one rule of a total of two. It is considered that a rule is represented (covered) by a set of labels if the rule contains at least one of the labels. Thus, it is expected that a good method must have a high precision. However, it is not enough to be precise if the labels appear repeatedly among the clusters. Therefore, Repetition Frequency (RF), in Equation 2, measures how much the distinct labels that are present in all the clusters don't repeat. Considering the above example, the illustrative method has a RF of 0.33 ($RF(C) = 1 - \frac{2}{3}$): one of the three distinct labels (butter, coffee, milk) that are present in clusters doesn't repeat. The higher the RF value, the better the method, i.e., less repetitions implies in better performance. Observe that RF can be used to compute the repetition frequency if we omit "1-" of Equation 2; however, in this case, the lower the RF value, the better the method. Thereby, the choice of not computing the repetition was to standardize the interpretation of the measures.

$$P(C) = \frac{\sum_{i=1}^{N} P(C_i)}{N}, \text{ where} \quad (1)$$

$$P(C_i) = \frac{\#\{\text{rules covered in } C_i \text{ by } C_i \text{ labels}\}}{\#\{\text{rules in } C_i\}}$$

$$RF(C) = 1 - \frac{\#\{\text{distinct labels that repeat in the clusters}\}}{\#\{\text{distinct labels in the clusters}\}} \quad (2)$$

Considering the labeling methods and the above measures, some experiments were realized, which are next described.

5 EXPERIMENTS

Some experiments were carried out to evaluate the labeling methods regarding precision and distinctiveness through P and RF. The four data sets used in experiments are presented in Table 1. The first three are available in *R Project for Statistical Computing* through "arules" package[2]. The last one was donated by a supermarket located in São Carlos city, Brazil[3]. All the transactions of the Adult and Income contain the same number of items (referred here as standardized-transaction data sets), different from Groceries and Sup (referred here as non-standardized-transaction data sets). Thus, the labeling methods were evaluated on different types of data. The rules were mined using an *Apriori* implementation developed by Christian Borgelt[4] with a maximum number of 5 items per rule and excluding the rules of type $\emptyset \Rightarrow X$, where X is an item contained in data. With the Adult set 6508 rules were generated using a minimum support (min-sup) of 10% and a minimum confidence (min-conf) of 50%; with Income 3714 rules considering a min-sup of 17% and a min-conf of 50%; with Groceries 2050 rules considering a min-sup of 0.5% and a min-conf of 0.5%; with Sup 7588 rules considering a min-sup of 0.7% and a min-conf of 0.5%. These parameter values were chosen experimentally considering the exposed arguments in Section 1 and (Carvalho et al., 2011)'s work.

Table 1: Details of the data sets used in experiments.

Data set	# of transactions	# of distinct items
Adult	48842	115
Income	6876	50
Groceries	9835	169
Sup	1716	1939

Since the papers described in Section 2 only use one family of clustering algorithms and one similar-

[2]http://cran.r-project.org/web/packages/arules/index.html.
[3]http://sites.labic.icmc.usp.br/research/Cjto-Sup.data.
[4]http://www.borgelt.net/apriori.html.

ity measure to cluster the association rules, it was decided to use one algorithm of each family and the two most used similarity measures (J-RI and J-RT (Section 2)). The Partitioning Around Medoids (PAM) was chosen within the partitional family and the Average Linkage within the hierarchical family. PAM was executed with k ranging from 5 to 50 considering a step of 5. The dendrograms generated by Average Linkage were cut in the same ranges (5 to 50 considering a step of 5). All the choices were made considering an analysis of many clustering configurations presented in (Carvalho et al., 2012). Table 2 summarizes the configurations used in the experiments.

Table 2: Configurations used in the experiments.

Data sets	Adult; Income; Groceries; Sup
Algorithms	PAM; Average Linkage
Similarity measures	J-RI; J-RT
k	5 to 50, step of 5

Table 3: Results for *P* and *RF* considering the ADULT and INCOME data sets.

Labeling method	Mean of *P*	Mean of *RF*
LM-M	0.995310	0.321458
LM-T	0.923752	0.340560
LM-S	0.965381	0.416278*
LM-PU	0.997238*	0.305087
Clustering algorithm	**Mean of *P***	**Mean of *RF***
PAM	0.969465	0.285709
Average	0.971375*	0.405983*
Similarity measure	**Mean of *P***	**Mean of *RF***
J-RI	0.970287	0.269874
J-RT	0.970553*	0.421818*

Table 4: Results for *P* and *RF* considering the GROCERIES and SUP data sets.

Labeling method	Mean of *P*	Mean of *RF*
LM-M	0.924978	0.700539*
LM-T	0.771151	0.696544
LM-S	0.899201	0.641688
LM-PU	0.971076*	0.662681
Clustering algorithm	**Mean of *P***	**Mean of *RF***
PAM	0.873818	0.564347
Average	0.909385*	0.786379*
Similarity measure	**Mean of *P***	**Mean of *RF***
J-RI	0.930973*	0.616215
J-RT	0.852230	0.734511*

Considering the configurations in Table 2, the four labeling methods (LM-M; LM-T; LM-S; LM-PU) were applied in the different domain organizations. In relation to the labeling methods, LM-M and LM-T select only one rule as label, LM-S one or more rules, in case of tie, and LM-PU the 5 items that present the best tradeoff between frequency and predictiveness. Thus, in average, all the labeling methods generate the same amount of labels per cluster. In the end, the performance of each labeling method was evaluated through *RF* and *P*, whose results are presented in the next section. It is important to remember that the aim of the measures is to evaluate, respectively, how much the method can find labels that represent as accurately as possible the knowledge contained in their own groups and how the labels are distributed along the clusters. The ideal is to identify methods that have high values for both measures.

6 RESULTS AND DISCUSSION

As mentioned before, the performance of the labeling methods were evaluated through *P* and *RF*. Thus, in order to identify the methods that are more adequate for association rule clustering and the domain organizations that provide the best results, an analysis based on the mean of each measure was done. Tables 3 and 4 present the results – the best values are marked with "*". Each mean was obtained considering all the results of the experiments[5], which were grouped according to the criteria shown (labeling method, clustering algorithm, similarity measure) and according to the different types of data (standardized-transaction (Table 3) and non-standardized-transaction (Table 4)). It is important to mention that since the results are deterministic no statistical test was done. It can be observed that:

- in the standardized-transaction data sets (Table 3) the method that presents the best result regarding *P* is LM-PU and considering *RF* LM-S. Thereby, the user can choose one of them based on his interests: accurate or distinctiveness. However, it is possible to note that in all the methods *RF* presents low values while *P* presents high values. Thus, it is better to use LM-S when the user wants a tradeoff between *P* and *RF*, since it improves *RF* (difference above 0.1) while maintaining a good *P* (difference of 0.03).
- in the non-standardized-transaction data sets (Table 4) the method that presents the best result regarding *P* is LM-PU and considering *RF* LM-M. Thereby, the user can choose one of them based on his interests: accurate or distinctiveness. On the other hand, it is possible to note that both methods have similar values (difference of 0.05

[5]All the results of the experiments are available in http://veronica1.rc.unesp.br/public/ICEIS-2012-R.pdf.

Table 5: Examples of labels obtained in some of the experiments using Average+J-RT and $k = 5$.

Experiment	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Income+ LM-S	age=14-34 dual_incomes=not_married language_in_home=english	age=35+ language_in_home=english	language_in_home=english number_in_household=1	language_in_home=english occupation=professional/managerial	language_in_home=english sex=female years_in_bay_area=10+
SUP+ LM-M	agua_tonica_antartica coca_cola	coca_cola gatorade	deterglimpol oleo_girassol_salada_bunge	fartrigo_renata gelatina_royal leite_moca	coca_cola leite_salute

in P and of 0.04 in RF). Thus, both of them could be used when the user wants a tradeoff between P and RF. However, it seems more adequate to use LM-M in spite of LM-PU since LM-M (i) can be more easily computed with partitional algorithms, (ii) can allow the user to know the existing relationship among the labels and (iii) presents a better value for RF (above 0.7) while maintaining a good P (above 0.9). Finally, it is possible to note that these types of data sets present better RF values in relation to the RF values in Table 3.

- the algorithm that presents the best performance in all the tests is Average (Tables 3 and 4).
- the similarity measure that presents the best performance in almost all the tests is J-RT (Tables 3 and 4). The only exception is P in Table 4, where J-RI presents a better performance.

Considering the exposed arguments, it can be observed that: (i) for standardized-transaction data sets the method that seems to be more adequate for association rule clustering is LM-S; (ii) for non-standardized-transaction data sets the method that seems to be more adequate for association rule clustering is LM-M; (iii) the methods present better results when the clustering is obtained through Average; (iv) J-RT seems to be a good similarity measure to be used along with Average; (v) as a consequence of (iii), it is possible to verify that Average represents the domain organization which best separates the domain knowledge, independently of the similarity measure used – it can be inferred that a domain is well separated if a domain organization, along with an adequate labeling method, provides good labels. These conclusions cover the three objectives stated in Section 1 (letters **(a)** to **(c)**). Besides, these results can be used with other methodologies, as the methodology described in (Carvalho et al., 2011), to make the association rule clustering a powerful post-processing tool.

Finally, Table 5 presents examples of labels obtained in some of the experiments using Average+J-RT and $k = 5$. One data set of each type of data (standardized or non-standardized) is shown along with its labeling method, according to the results above discussed, that had the best performance. The items that occur more than once are underlined. It can be observed that: (i) the labels of Income describe, with good precision and distinctiveness ($P = 0.835$; $RF = 0.875$), some specificities well defined of the domain – cluster 2, for example, is related to people above 35 years and cluster 5 to people who are female and live for more than 10 years in the San Francisco Bay area; (ii) on the other hand; the labels of SUP describe, also with good precision and distinctiveness ($P = 0.788$; $RF = 0.889$), some types of beverages that can be purchased, as clusters 1, 2 and 5, which are related with distinct shop styles: cluster 1 with water, cluster 2 with soft drink and cluster 5 with milk; (iii) the items that occur in many clusters labels are very frequent in their data sets (language_in_home=english: 91%; coca_cola: 22%), which means that they can be used as complementary information of the clusters. Thus, as observed, it is essential that good labels be found, since they can aid the users in exploratory analyses by guiding their search.

7 CONCLUSIONS

Due to the huge amount of association rules that are obtained, considering the exposed arguments in Section 1, many approaches were suggested, as clustering. However, for clustering to be useful to users it is essential that good descriptors be associated with each cluster to help, for example, in guiding their search. Thus, the analysis of different labeling methods for association rule clustering is a relevant issue. Considering the exposed arguments, this paper analyzed some labeling methods. Two measures were proposed and used to evaluate the methods. Precision, P, measures how much the methods can find labels that represent as accurately as possible the rules contained in their own groups. Repetition Frequency, RF, measures how the labels are distributed along the clusters. As a result, it was possible to identify the methods and the domain organizations with the best performances that can be applied in clusters of association rules.

As future work we will explore some approaches that aim to improve the labels through a generalization process. We want to explore the impact of generic labels on *P* and *RF* to analyze if the results of the labeling methods can be improved. From this generalization process we intend to discover a topic for each cluster considering the context given by the user through ontology. Given, for example, the labels "rice", "bean" and "salad", the topic could be food or lunch, depending on the knowledge codified in the ontology.

ACKNOWLEDGEMENTS

We wish to thank Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) (processes numbers: 2010/07879-0 and 2011/19850-9) and Fundação para o Desenvolvimento da Unesp (FUNDUNESP) for the financial support.

REFERENCES

Agrawal, R. and Srikant, R. (1994). Fast algorithms for mining association rules. In Bocca, J. B., Jarke, M., and Zaniolo, C., editors, *Proceedings of the 20th International Conference on Very Large Data Bases, VLDB 1994*, pages 487–499.

Carvalho, V. O., Santos, F. F., and Rezende, S. O. (2011). Post-processing association rules with clustering and objective measures. In *Proceedings of the 13th International Conference on Enterprise Information Systems*, volume 1, pages 54–63.

Carvalho, V. O., Santos, F. F., Rezende, S. O., and Padua, R. (2012). PAR-COM: A new methodology for post-processing association rules. *Lecture Notes in Business Information Processing*, 102. In press. Available due May 19.

Cutting, D. R., Karger, D. R., Pedersen, J. O., and Tukey, J. W. (1992). Scatter/gather: a cluster-based approach to browsing large document collections. In *Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 318–329.

Fung, B. C. M., Wang, K., and Ester, M. (2003). Hierarchical document clustering using frequent itemsets. In *Proceedings of the 3rd SIAM International Conference on Data Mining*, pages 59–70.

Glover, E. J., Pennock, D. M., Lawrence, S., and Krovetz, R. (2002). Inferring hierarchical descriptions. In *Proceedings of the 11th International Conference on Information and Knowledge Management*, pages 507–514.

Jorge, A. (2004). Hierarchical clustering for thematic browsing and summarization of large sets of association rules. In *Proceedings of the 4th SIAM International Conference on Data Mining*. 10p.

Kashyap, V., Ramakrishnan, C., Thomas, C., and Sheth, A. (2005). Taxaminer: An experimentation framework for automated taxonomy bootstrapping. *International Journal of Web and Grid Services*, 1(2):240–266.

Larsen, B. and Aone, C. (1999). Fast and effective text mining using linear-time document clustering. In *Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 16–22.

Lopes, A. A., Pinho, R., Paulovich, F. V., and Minghim, R. (2007). Visual text mining using association rules. *Computers & Graphics*, 31(3):316–326.

Manning, C. D., Raghavan, P., and Schütze, H. (2009). *An Introduction to Information Retrieval*. Cambridge University Press. 544p.

Moura, M. F. and Rezende, S. O. (2010). A simple method for labeling hierarchical document clusters. In *Proceedings of the 10th IASTED International Conference on Artificial Intelligence and Applications*, pages 336–371.

Natarajan, R. and Shekar, B. (2005). Interestingness of association rules in data mining: Issues relevant to e-commerce. *SADHANA – Academy Proceedings in Engineering Sciences (The Indian Academy of Sciences)*, 30(Parts 2&3):291–310.

Popescul, A. and Ungar, L. (2000). Automatic labeling of document clusters. Unpublished manuscript. http://www.cis.upenn.edu/∼popescul/Publications/popescul00labeling.pdf.

Reynolds, A. P., Richards, G., de la Iglesia, B., and Rayward-Smith, V. J. (2006). Clustering rules: A comparison of partitioning and hierarchical clustering algorithms. *Journal of Mathematical Modelling and Algorithms*, 5(4):475–504.

Sahar, S. (2002). Exploring interestingness through clustering: A framework. In *Proceedings of the IEEE International Conference on Data Mining*, pages 677–680.

Toivonen, H., Klemettinen, M., Ronkainen, P., Hätönen, K., and Mannila, H. (1995). Pruning and grouping discovered association rules. In *Workshop Notes of the ECML Workshop on Statistics, Machine Learning, and Knowledge Discovery in Databases*, pages 47–52.

Zhao, Y., Zhang, C., and Cao, L. (2009). *Post-mining of association rules: Techniques for effective knowledge extraction*. Information Science Reference. 372p.