A human-centered approach for mining hybrid-dimensional association rules

Eduardo Corrêa Gonçalves
Instituto Brasileiro de Geografia e Estatística

A Human-Centered Approach for Mining Hybrid-Dimensional Association Rules

Eduardo Corrêa Gonçalves
Instituto de Computação
Universidade Federal Fluminense (UFF)
Niterói, Rio de Janeiro 24210–240
Email: {egoncalves,plastino}@ic.uff.br

***Abstract*—This work presents a new method to mine hybrid-dimensional association rules in databases originated from multiple sources. We adopted an approach in which hybrid associations represent transactional rules that become either exceptionally weaker or exceptionally stronger in some subsets of an integrated database, which satisfy specific conditions over selected attributes. We propose an interest measure to evaluate hybrid-dimensional rules, as well as an algorithm to mine these patterns. This algorithm was applied to a real dataset that keeps information about purchases made by families residing in different Brazilian cities. The obtained results show that the proposed technique provides valuable information for decision making.**

I. Introduction

The goal of the association rule mining task (ARM) is to find hidden and interesting relationships between sets of items or attribute values in large databases [3], [6], [16]. These uncovered relationships are expressed in the form of rules of type $A \Rightarrow B$. The ARM task was originally proposed to solve the market basket analysis problem [1], [9], which consists in the process of analyzing a database of sales records in order to determine what products are likely to be bought together. For instance, considering a hypothetical database that stores the sales transactions of a supermarket over a period of time, an ARM algorithm could be able to discover the following association rule: {*salami*} $\Rightarrow$ {*beer*}. This rule indicates that clients who buy salami are more likely to also buy beer.

The strength of an association rule is often assessed with the use of interest measures, such as the support and confidence factors. The support of an association rule $A \Rightarrow B$, denoted by $Sup(A \Rightarrow B)$, represents the percentage of transactions that contain both A and B, indicating the rule's relevance. The confidence of $A \Rightarrow B$, $Conf(A \Rightarrow B)$, is the probability that a transaction contains B, given that it contains A, indicating the validity of the association rule. The most common framework for mining association rules, introduced in [1], consists in finding all rules that have support and confidence equal to or greater than user-provided minimum support and minimum confidence values, respectively denoted by $MinSup$ and $MinConf$.

Association rules mined from transactional databases[1], such as {*salami*} $\Rightarrow$ {*beer*}, are referred to as transactional association rules. It is worth to note that although a transactional rule might be composed by various items, these always refer to the same concept or dimension (e.g.: the "Product" concept). In other words, transactional rules involve solely a single dimension. As a consequence, they are commonly referred to as single-dimensional association rules [6], [14]. However, it is also possible to mine association rules in data repositories other than transactional databases, such as data warehouses and relational databases. In this case, the association rules can be composed of diverse categorical and numeric attributes, being referred to as multidimensional association rules [5], [6], [7]. To demonstrate this concept, suppose a relational table that stores demographic data and other statistics of cities in a given country. An example of multidimensional rule that could be mined from this table is given by:

(Region = "South") $\wedge$ *(Population* $<$ *10,000)* $\Rightarrow$ *(Main-Activity = "Livestock").*

This hypothetical rule indicates that southern cities with less than 10,000 residents are more likely to have the livestock sector as its main economic activity. Observe that this rule involves three attributes (or dimensions), one of which is numeric (Population) and the other two categorical (Region and Main-Activity).

There is yet a third type of association rule conceptually defined in [6]: the hybrid-dimensional association rule. This kind of rule consists in a mixture of both types, transactional and multidimensional. More specifically, a hybrid-dimensional rule corresponds to an association rule where one of the dimensions can occur multiple times. An example is given by:

(Region = "South") $\wedge$ *(Population* $<$ *10,000)* $\wedge$ *(Product = "salami")* $\Rightarrow$ *(Product = "beer").*

This hypothetical rule indicates that in the southern cities with less than 10,000 residents, the supermarket consumers who buy salami are more likely to also buy beer in their purchases. The above example involves three dimensions (Region, Population and Product), where one of them occurs more than once in the body of the rule (Product). In spite of being very simple, the example is capable of evidencing an appealing property of hybrid-dimensional association rules in the context of information fusion: they are able to represent, at the same time and in an intuitive way, relationships between attributes of different data sources. Thus, hybrid-dimensional rules are capable of potentially revealing useful and actionable knowledge

[1] It is important to clarify that, in data mining, the term "transactional database" corresponds to a database where each record represents a collection of items associated to a identifier. E.g.: market basket data (each transaction corresponds to a collection of products bought by a customer), web usage data (each record contains a list of web pages visited by a user), textual data (each transaction is a set of words that occur in a text document), etc.

to the users of both data mining and information integration systems. Nonetheless, it is surprisingly to notice that just a few proposals of algorithms for mining hybrid-dimensional rules have appeared in the literature [14]. Although [6] defines the concept of hybrid-dimensional rule, it does not propose a specific algorithm to mine such patterns. Moreover and unfortunately, current data mining / information integration tools do not offer hybrid-dimensional mining functionality.

This work has the main goal of proposing a new user-driven technique especially targeted for the discovering of interesting hybrid-dimensional association rules in databases originated from multiple sources. In our approach, users can explore a set of transactional rules in order to discover how much the strength of these rules become either unexpectedly weaker or unexpectedly stronger in specific subsets of an integrated database. In order to avoid the generation of uninteresting rules, a hybrid-dimensional rule should be mined only if the change in the strength has been significant. Our approach corresponds to an extension of the technique for mining multidimensional exception rules introduced in [4], [5].

In order to clarify our approach, consider again the association $\{salami\} \Rightarrow \{beer\}$. Suppose this rule has been mined from a transactional database D_1 containing sales records of several stores of a supermarket chain. Also consider that the support and confidence values of this rule are, respectively, equal to 5% and 65% in D_1. Now suppose that D_1 has been combined with a multidimensional dataset D_2 which contains demographic data of the cities where each of the stores is located. Our proposed approach to mine interesting hybrid-dimensional associations allows a user to evaluate if the rule $\{salami\} \Rightarrow \{beer\}$ becomes either exceptionally stronger or exceptionally weaker considering distinct subsets of the combined dataset, which satisfy specific conditions over selected attributes from D_2. As stated before, it is known that when these attributes are not taken into consideration, the support of the rule is 5% and the confidence is 65%. However, if the new attributes incorporated from D_2 pass to be considered, it could be possible to identify, for instance, that $\{salami\} \Rightarrow \{beer\}$ becomes exceptionally stronger in rural cities of the south. Similarly, it could also be possible to identify that $\{salami\} \Rightarrow \{beer\}$ becomes exceptionally weaker (i.e., neither relevant nor valid) in northern cities with large populations.

The rest of this paper is organized as follows. Section II gives an overview of ARM concepts relevant to this paper. Section III is the main section of this work. It first introduces the concepts of negative and positive hybrid-dimensional association rules and then describes the basic framework to mine these patterns. In the same section, a novel interest measure to evaluate hybrid-dimensional rules is proposed. In Section IV we formalize a cooperative algorithm for hybrid association mining. Section V discusses related work. Experimental results are presented and discussed in Section VI. Some concluding remarks are made in Section VII.

II. Background

The concept of association rule was introduced in [1] as follows. Let $\mathcal{I} = \{i_1, i_2, \ldots, i_n\}$ be a set of n distinct items and $\mathcal{D} = \{t_1, t_2, \ldots, t_m\}$ be a set of m transactions (a transactional database) defined over $\mathcal{I}$, where each transaction t_i is a subset of $\mathcal{I}$ ($t \subseteq \mathcal{I}$). An association rule is an implication of the form $A \Rightarrow B$, where $A \subset \mathcal{I}$, $B \subset \mathcal{I}$, $A \neq \emptyset$, $B \neq \emptyset$, and $A \cap B = \emptyset$. A is named antecedent and B is named consequent of the rule. The association $A \Rightarrow B$ holds in database $\mathcal{D}$ with support s and confidence c if, respectively, $s\%$ of the transactions in $\mathcal{D}$ contain $A \cup B$, and $c\%$ of the transactions in $\mathcal{D}$ that contain A also contain B.

The most typical approach for ARM consists in finding all rules that satisfy user-provided minimum support ($MinSup$) and minimum confidence ($MinConf$) thresholds. This conceptual model is known as "support/confidence framework". Most algorithms that follow this approach broke down the original problem into two subproblems:

- Step 1: determine all sets of items present in at least $MinSup\%$ of transactions. These are called frequent itemsets.
- Step 2: for each frequent itemset found in Step 1, generate all association rules with confidence equal to or greater than $MinConf\%$.

Nevertheless, over the two last decades, the data mining literature [4], [5], [6], [8], [9], [10], [12], [15], [16] have evidenced that there is a major drawback associated to the support/confidence framework: the fact that it often leads to the generation of a huge number of association rules, many of which obvious and irrelevant, making it difficult for users to identify those rules that are indeed interesting to them. In other to cope with this problem, some proposals (such as [4], [5], [8], [10], [15]) suggest modify the support/confidence framework by allowing users to guide the mining process into finding only *unexpected rules*, instead of enumerating all possible association rules. An association can be defined as unexpected when it contradicts user beliefs [10]. This is usually the most interesting type of pattern because it allows previously unknown information to be visible to domain specialists.

Motivated by this issue and also by the aforementioned fact that there are a few proposed algorithms especially directed towards the mining of hybrid-dimensional rules (the most useful type of rule in the context of the information fusion field), in the next section we present the main contribution of this paper: a novel approach targeted for the discovering of unexpected hybrid-dimensional association rules in databases originated from multiple sources. To guide the mining process into finding only unexpected rules, the proposed model provides users with opportunities to guide incorporate their prior knowledge right from the start of the search for associations.

III. Mining Negative and Positive Hybrid-Dimensional Association Rules

A. The CEF database

Throughout this section, in order to facilitate the discussions, we will make use of examples mined from a real database containing observations collected from a household survey called Consumer Expenditure Survey (CEF). This survey has been conducted by a Brazilian institute of research since 1947 to, among other goals, support the analysis of food consumption of Brazilian families. The database studied in this work keeps data about 1540 interviewed families from seven distinct cities. It comprises two tables, one transactional and the other relational, which will be referred to as D_T and D_R, respectively. D_T stores the list of products acquired by each family on their last visit to a supermarket whilst D_R stores demographic data of these families. In the D_R table, each family is characterized by three attributes: monthly income (Income), number of members (Members), and city of residence (City). On its turn, the D_T table involves about 2000 distinct items (i.e., $n \approx 2000$).

B. Negative Hybrid-Dimensional Association Rules

The first contribution of this paper is the definition of the concept of negative hybrid-dimensional association rules. These consist in transactional association rules that become exceptionally weaker in specific subsets of an integrated database. As an example, consider the rule, R_1 :{*milk box*} $\Rightarrow$ {*French bread*}, a real association mined from the D_T table of the CEF database with support and confidence values of 16.88% and 66.84%, respectively. Suppose that a user is interested in discovering if this association becomes weaker (for instance, with lower values of support) considering families that reside in any of the Brazilian cities where CEF was conducted. In other words, the user is interested in knowing if R_1 :{*milk box*} $\Rightarrow$ {*French bread*} becomes weaker on some sub-population of families stratified by the attribute City of the D_R table. In this case, our proposed strategy to mine hybrid-dimensional association rules would be able to extract the following negative pattern:

$$H_1 : \{\textit{milk box}\} \overset{-s}{\Longrightarrow} \{\textit{French bread}\}\ [(\textit{City = “Recife”})].$$

The symbol “$\overset{-s}{\Longrightarrow}$” is employed to indicate that the support value of the transactional association rule {*milk box*} $\Rightarrow$ {*French bread*} is significantly lower than what was expected in the database subset defined by the families who live in Recife. In the adopted notation, the subset at issue is defined by the condition between brackets at the end of the hybrid rule: [(*City = “Recife”*)].

The formal definition of negative hybrid-dimensional association is presented below.

Definition 3.1: (Negative Hybrid-Dimension Rule). Let $\mathcal{D}$ be a database originated from multiple sources. Let $R : A \Rightarrow B$ be a transactional association rule mined from $\mathcal{D}$ with high support value. Let $Z = \{Z_1 = z_1, \ldots, Z_k = z_k\}$ be a set of conditions defined over distinct attributes from $\mathcal{D}$. Z is named *probe set*. A negative hybrid-dimensional association rule related to the rule R is an expression of the form $A \overset{-s}{\Longrightarrow} B\ [Z]$.

A negative hybrid-dimensional association aims at representing how much the presence of a probe set turns weaker a (originally strong) transactional association rule. Negative hybrid-dimensional rules are mined from candidate rules that are generated through the combination of a transactional association rule $A \Rightarrow B$ with a probe set Z. A negative hybrid rule $A \overset{-s}{\Longrightarrow} B\ [Z]$ should be extracted only if it does not achieve an expected support. This expectation is evaluated based on the support of the original rule $A \Rightarrow B$ and the support of the conditions that compose the probe set Z.

Definition 3.2: (Expected Support of a Candidate Rule). Let $C :\ A \Rightarrow B\ [Z]$ be a candidate rule. The actual support of C is givenby $Sup(A \cup B \cup Z)$. The expected support for C, denoted by $ExpSup(C)$ is computed as $ExpSup(C) = Sup(A \cup B) \times Sup(Z)$.

A negative hybrid-dimensional rule $H\ :\ A \overset{-s}{\Longrightarrow} B\ [Z]$ is potentially interesting and should be mined to the user only if the actual support value of the candidate rule $A \Rightarrow B\ [Z]$ is much lower than its expected support value. In order to calculate this deviation, we propose the The IH^- index (*Interest Measure for Negative Hybrid-Dimensional Rules*), defined in Equation 1.

$$IH^-(C) = 1 - \frac{Sup(A \Rightarrow B\ [Z])}{ExpSup(A \Rightarrow B\ [Z])}\ . \tag{1}$$

The IH^- index value grows when the actual support value of the candidate rule is lower and far from the expected support value. The closer the value is from 1 (which is the highest value for this index), the more interesting the negative rule. If $IH^-(C) \approx 0$ the actual support value is closer to the expected, indicating that the negative hybrid rule should not be generated from the candidate rule.

Let us return to the example presented in the beginning of this subsection. The candidate rule C_1 : {*milk box*} $\Rightarrow$ {*French bread*} [(*City = “Recife”*)] was generated by combining the transactional rule R_1 : {*milk box*} $\Rightarrow$ {*French bread*} with the probe set Z_1 = {(*City = “Recife”*)}. The actual support of Z_1 in the CEF database is equal to 13.00% whereas the actual support of the rule R_1 is 16.88%. According to Definition 3.2, $ExpSup(C_1) = Sup(R_1) \times Sup(Z_1) = 16.88\% \times 13.00\% = 2.20\%$. However, the actual support of C_1 in the integrated CEF database is equal to 0.71%. This suggests that the negative hybrid-dimensional rule H_1 : {*milk box*} $\overset{-s}{\Longrightarrow}$ {*French bread*} [(*City = “Recife”*)] might be unexpected and, therefore, potentially interesting to be presented to the user. The interest measure for this negative rule can be computed as $IH^-(HN_1) = 1 - (0.0071 \div 0.0220) = 0.6773$. This result indicates that among the families who live in Recife, the support value of the association between the products milk box and French bread is 67.73% percent smaller than what is expected.

C. Positive Hybrid-Dimensional Association Rules

Positive hybrid-dimensional association rules consist in transactional association rules that become exceptionally stronger in specific subsets of an integrated database. In this section we introduce this kind of pattern, once again making use of a real example extracted from the CEF database. The rule R_2 :{*beer can*} $\Rightarrow$ {*salami*}, with support of 2.08% and confidence of 11.19%, can be regarded as a weak transactional association in the D_T table since it has low values of support and confidence. Suppose a user is interested in discovering if this association becomes stronger on some sub-population of families stratified by the attribute Members of the D_R table. In this case, our strategy to mine hybrid-dimensional association rules would be able to identify the following positive pattern:

$$H_2 : \{\textit{beer can}\} \overset{+s}{\Longrightarrow} \{\textit{salami}\}\ [(\textit{Members = 1})].$$

This positive hybrid-dimensional association rule indicates that in the subset of the D_R table defined by people who live alone (condition Members=1), the support of the transactional rule $\{$*beer can*$\} \Rightarrow \{$*salami*$\}$ is significantly greater than what is expected. The symbol "$\overset{+s}{\Longrightarrow}$" is employed to characterize this situation.

The formal definition of positive hybrid-dimensional association rule is presented below.

Definition 3.3: (Positive Hybrid-Dimension Rule). Let $\mathcal{D}$ be a database originated from multiple sources. Let $R : A \Rightarrow B$ be a transactional association rule mined from $\mathcal{D}$ with low support value. Let $Z = \{Z_1 = z_1, \ldots, Z_k = z_k\}$ be a set of conditions defined over distinct attributes from $\mathcal{D}$ (*probe set*). A positive hybrid-dimensional association rule related to the rule R is an expression of the form $A \overset{+s}{\Longrightarrow} B\ [Z]$.

The goal of a positive hybrid-dimensional rule is to represent how much the presence of a probe set turns stronger a (originally weak) transactional association rule. As with negative hybrid rules, the positive hybrid rules are also mined from candidate rules. However, this time the goal is to identify associations that have actual support value significantly greater the the expected one. In order to compute this deviation, we propose the The IH^+ index (*Interest Measure for Positive Hybrid-Dimensional Rules*), defined in Equation 2.

$$IH^+(C) = 1 - \frac{ExpSup(A \Rightarrow B\ [Z])}{Sup(A \Rightarrow B\ [Z])}\ . \tag{2}$$

In Equation 2, the expected support is placed in the numerator of the equation. Thus, the IH^+ index value is higher when actual support value of the candidate rule is greater and far from the expected support value. The closer the value is from 1 (which is the highest value for this measure), the more interesting the positive hybrid rule. For instance, in the integrated CEF database, the positive hybrid-dimensional association $H_2 : \{$*beer can*$\} \overset{+s}{\Longrightarrow} \{$*salami*$\}$ [(*Members = 1*)] has $IH^+(H_2) = 0.7174$, indicating that the real support of the transactional rule $\{$*beer can*$\} \Rightarrow \{$*salami*$\}$ is exceptionally higher than the expected among those living alone.

IV. Algorithm

In Figure 1 we formalize the HAR algorithm for mining both positive and negative hybrid-dimensional association rules in a database originated from multiple sources. This algorithm requires the following user input parameters.

1) $\mathcal{D}$ - a database originated from multiple sources, containing transactional and multidimensional data.
2) $\mathcal{R}$ - a set of selected transactional association rules involving items in $\mathcal{D}$.
3) $\mathcal{A}$ - a set of selected attributes from $\mathcal{D}$ that will form the probe sets.
4) $MinSup \geq 0$ - minimum value for the support measure.
5) $MinIM \geq 0$ - minimum value for the IH^- and IH^+ measures.

The algorithm produces the following output:

1) $NHSet$ - a set of negative hybrid-dimensional association rules in relation to $\mathcal{R}$ and $\mathcal{A}$.
2) $PHSet$ - a set of positive hybrid-dimensional association rules in relation to $\mathcal{R}$ and $\mathcal{A}$.

```
1:  $PSet$ = generate all possible probe sets from $\mathcal{A}$
2:  $CSet = \emptyset$
3:  $CondTree = \emptyset$
4:  for all $rules\ A \Rightarrow B$ in $\mathcal{R}$ do
5:    for all $probesets\ Z$ in $PSet$ do
6:      $CSet = CSet\ \cup\ (A \Rightarrow B\ [Z])$
7:      $X' = \{\{A\}, \{B\}, \{Z\}, \{A, B\}, \{A, Z\}, \{B, Z\},$
            $\{A, B, Z\}\}$
8:      $CondTree = CondTree \cup X'$
9:    end for
10: end for
11: scans $\mathcal{D}$ to compute the support of the sets in
    $CondTree$
12: $NHSet = \emptyset$
13: $PHSet = \emptyset$
14: for all $candidate\,rules\ C : A \Rightarrow B\ [Z]$ in $CSet$ do
15:   if ($Sup(A \cup Z) \geq MinSup$) and ($Sup(B \cup Z) \geq$
      $MinSup$) then
16:     if ($IH^-(C) \geq MinIH$) then
17:       $NHSet = NHset \cup (A \overset{-s}{\Longrightarrow} B\ [Z])$
18:     else if ($IH^+(C) \geq MinIH$) then
19:       $PHSet = PHset \cup (A \overset{+s}{\Longrightarrow} B\ [Z])$
20:     end if
21:   end if
22: end for
```

Fig. 1. The proposed HAR Algorithm for mining positive and negative hybrid-dimensional association rules in databases

The HAR algorithm, shown in Figure 1, can be decomposed into four phases which are explained below.

Phase 1 (line 1) identifies and generates all probe sets that will be used in the composition of the candidate rules. The procedure will generate only probe sets formed by the attributes specified in the user-defined parameter $\mathcal{A}$.

Phase 2 (lines 2-10) generates all candidate rules, combining each transactional association rule in $\mathcal{R}$ with all probe sets in $PSet$ (lines 4-6). These candidate rules are stored in the $CSet$ structure. Other important step in this phase is the creation of the data structure $CondTree$. (lines 7-8). This data structure keeps counters for the support values of all sets of conditions that will be used during the computation of the interesting measures IH^- and IH^+. Remember that in order to compute these indexes for a candidate rule $A \Rightarrow B\ [Z]$, we need to obtain the expected support values for $A \cup Z$, $B \cup Z$ and for the candidate rule $A \Rightarrow B\ [Z]$. To obtain these expected values, it is necessary to count the actual supports for the following sets: $\{A\}$, $\{B\}$, $\{Z\}$, $\{A, B\}$, $\{A, Z\}$, $\{B, Z\}$, and $\{A, B, Z\}$. The data structure $CondTree$ must efficiently stores counters for all these sets. In our implementation, we opted to use the classical hash tree data structure, described in [2].

In Phase 3 (line 11), the entire database is scanned in order to obtain the support values for all elements stored in

$CondTree$. Only a single scan is required since all the sets that must be counted are already stored in the structure.

Finally, Phase 4 (lines 12-22) generates the negative and positive hybrid-dimensional associations. This is simply accomplished by forming the candidate rules and computing the interest measures for each of them. Each candidate rule is analyzed in the following manner. First (line 15), the algorithm evaluates the support of the sets $\{A \cup Z\}$ and $\{B \cup Z\}$ so as to ensure the statistical significance of the hybrid-dimensional rules to be mined. Next (lines 16-20), the values of the measures IH^- and IH^+ are computed for the candidate rule which is currently under evaluation. To perform these calculations, it is only necessary to retrieve the support values stored in $CondTree$ (computed in Phase 3). First, the IM^- value is computed. If the obtained value is equal to or greater than the minimum user-specified value, a negative hybrid-dimensional association rule must be mined and inserted into the results set $NHSet$. Otherwise, the IM^+ value is computed and, according to the obtained value, a positive hybrid-dimensional rule may be mined (in this case it is inserted into the $PHSet$).

The algorithm for mining hybrid-dimensional rules formalized in this section represents a human-centered approach, in the sense that the whole focus of the mining process is specified by the user. The user is given the possibility to define a set of transactional rules whose strength he or she is interested in investigating in different subsets of an integrated database. The user is also responsible for specifying those subsets, by selecting the attributes that will form the probe sets. The algorithm effort is proportional to the user's specifications. The most expensive steps are the definition of the $CondTree$ structure and the support counting of the sets of conditions stored therein. Since only the attributes that compose the candidate rules are relevant, a considerably smaller number of sets need to have involved in support counting operation.

V. Related Work

The simplest (and, most naive) solution to cope with the problem of mining hybrid-dimensional association rules in databases originated from multiple sources merely corresponds to transform multidimensional information into transactional information. For instance, considering the CEF database, a trivial solution would simply solve the problem by first mapping attribute/value pairs like City="Rio de Janeiro" or Members=1 into distinct products id's and then using a conventional algorithm for mining transactional associations over the transformed dataset. In other words, the distinct values of the multidimensional attributes such as City, Members and Income would be treated as if they were products such as "beer" and "salami". Needless to say that this solution, apart from being inelegant, is also inefficient and not tailored to the needs of discovering interesting (useful and unexpected) patterns to users.

In [14], the authors propose a group of algorithms capable of extracting both multidimensional and hybrid-dimensional rules in data warehouses. The main focus is on the definition of a framework to efficiently transform data stored in fact and dimension tables of a data warehouse into a plain table (which is more suitable to be processed by most algorithms for mining association rules). From this plain table, conventional hybrid-dimensional rules – i.e., rules with support and confidence that match user-provided input thresholds – can be directly mined. Differently for our approach, the proposal of [14] is actually only concerned with efficiency, rather than pursuing the goal of extracting interesting hybrid-rules. Our technique for mining hybrid-dimensional association rules is indeed more related to methods for mining unexpected rules [8], [10], [15], negative patterns [13], rare association rules [12] and exceptions [4], [5], which are discussed in the rest of this section.

As stated in [8], [10], one of main factors that make an association rule subjectively interesting to an user is unexpectedness. An association rule is considered unexpected when it contradicts user beliefs, i.e., when it represents a relationship between items that would not be previously realized or considered as true by this user. According to this principle, some proposals, consider that the models for mining association rules must find a way to represent the expectations of the users and incorporate them into the mining algorithm, so as to allow the mining of unexpected rules. For instance, this idea was applied in the experiment described in [15], where the goal was to find unexpected rules from a real database that keeps data about donations to election campaigns in the USA. A group of domain specialists elaborated a set of beliefs, among them: "residents of Beverly Hills are very wealthy" and "people who are very wealthy tend to donate more than US$ 200". The algorithm for mining unexpected rules proposed in [15] is able to incorporate these knowledge into the mining model and becomes capable of mining the following unexpected rule in relation to the two beliefs: "residents of Beverly Hills tend to donate less than US$ 50.

Differently, the goal of rare association rule mining [11], [12] is to directly search for rules with low support, but high confidence in databases. The motivation lies in the fact that, in many practical situations, rare items may be either more important or more profitable. E.g.: considering the market basket problem, it might be interesting to discover associations involving rare items like "caviar" or "lobster", because they are more expensive, being thus more likely to be associated with other expensive products. The main idea employed by algorithms for mining rare association rules is to adopt a maximum support threshold ($MaxSup$) along with the traditional minimum support ($MinSup$). A rare association rule is typically regarded as interesting if it holds with support value between $MinSup$ and $MaxSup$.

Users may also be interested in finding negative associations or correlations in data. An example of such pattern is: "when costumers buy tea they are less likely to buy coffee". The problem of mining strong negative association rules in transactional databases was introduced in the seminal work of [13]. According to this approach, a negative association rule should be extracted from a database only if it does not achieve an expected support. However, differently from our approach, the expected support in the proposal of [13] is computed based on the existence of a taxonomy, which hierarchically classify items that belong to the application knowledge domain. It is expected, for example, that items falling under the same class, such as *coke* and *pepsi*, have similar associations with other items. An example presented in [13] illustrates that the negative association rule $\{ruffles\} \not\Rightarrow \{pepsi\}$ can be obtained if the support value of the positive rule $\{ruffles\} \Rightarrow \{pepsi\}$ is significantly lower than the support value of the extracted positive rule $\{ruffles\} \Rightarrow \{coke\}$ (regarding the ratio between the support values of $coke$ and $pepsi$). Opposed to our approach that finds both exceptionally positive and exceptionally negative hybrid-rules, the method introduced in [13] only extracts negative transactional rules from databases.

The method for mining hybrid-dimensional associations proposed in this work represents an extension of the human-centered method for mining exceptions in multidimensional databases introduced in [4], [5]. This proposal defines an exception as an unexpected negative multidimensional association rule, which should be mined if it does not achieve an expected support in a specific subset of the database. However, we modified the algorithm proposed in [4], [5] in two key aspects. First, the original method was extended to not only perform the mining of unexpected negative exceptions, but also to allow the searching for unexpected positive exceptions (which may be as useful as negative patterns). Second, the method was adapted to solve the problem of mining hybrid-dimensional rules in integrated databases, instead of mining the more conventional multidimensional and transactional rules. Since hybrid-dimensional rules are capable of discovering patterns from different sources they are naturally much more useful in the context of information fusion.

VI. Experimental Results

The HAR algorithm specified in Section IV was implemented and evaluated on the CEF dataset. The main goal of the evaluation was to perform a subjective analysis over the obtained results, with the major concern of verifying if the mined hybrid rules would indeed represent valid and useful information. The input parameters were configured as follows.

- $\mathcal{R}$ was formed by a set of rules involving eight popular items: beer can, black beans, cereal flakes, garlic, milky meal, milk bag, milk box, and salami.
- The attributes Income, Members and City were used to form the probe sets. The income is given in number of minimum wage salaries.
- The $MinSup$ value was specified as 0.30.
- The $MinIM$ was specified as 0.30.

The obtained results will be now presented and commented. First, Table I presents some interesting hybrid-dimensional rules containing the Members attribute in the probe set. In this table (and the other shown in this section), the first column is used to present the hybrid rule. The second column shows the $IM+$ value if the hybrid rule is positive. Similarly, the third column shows the $IM-$ value if the hybrid rule is negative. The first hybrid rule in Table I, HP_1, has been previously presented in Section III. It reveals that the association between the sales of {*beer can*} and {*salami*} is much stronger among people who live alone. The rules HN_1 and HP_2 provide interesting information regarding the transactional association rule {*cereal flakes*} ⇒ {*milky meal*}. Observe that the support of this rule is significantly lower that what was expected in the subset of the database defined by families with two members. This situation is easily explainable, since the majority of families with two member does not contain a child. On the other hand, the rule HP_2 demonstrate that the association between the same two products is exceptionally higher than what is expected among families with five members (which are probably composed by one or more children). Although these rules might be not actually surprising, they are able to demonstrate that our approach for mining hybrid-dimensional rules was able to extract valid patterns from the database.

TABLE I. Results - hybrid associations involving probe sets based on "Members" attribute

id	Hybrid Rule	IH^+	IH^-
HP_1	$\{\textit{beer can}\} \overset{+s}{\Longrightarrow} \{\textit{salami}\}$ [(Members=1)]	0.7174	-
HN_1	$\{\textit{cereal flakes}\} \overset{-s}{\Longrightarrow} \{\textit{milky meal}\}$ [(Members=2)]	-	0.6540
HP_2	$\{\textit{cereal flakes}\} \overset{+s}{\Longrightarrow} \{\textit{milky meal}\}$ [(Members=5)]	0.3919	-

TABLE II. Results - hybrid associations involving the products "milk box", "milk bag", and "French bread"

id	Hybrid Rule	IH^+	IH^-
HP_3	$\{\textit{milk box}\} \overset{+s}{\Longrightarrow} \{\textit{French bread}\}$ [(City="Porto Alegre"),(Income="18-25")]	0.6425	-
HP_4	$\{\textit{milk box}\} \overset{+s}{\Longrightarrow} \{\textit{French bread}\}$ [(City="Florianópolis")]	0.5307	-
HN_2	$\{\textit{milk box}\} \overset{-s}{\Longrightarrow} \{\textit{French bread}\}$ [(City="Recife")]	-	0.6773
HN_3	$\{\textit{milk box}\} \overset{-s}{\Longrightarrow} \{\textit{French bread}\}$ [(Members=4),(City="Fortaleza")]	-	0,6988
HN_4	$\{\textit{milk box}\} \overset{-s}{\Longrightarrow} \{\textit{French bread}\}$ [(Income="2-5")]	-	0,4716
HP_5	$\{\textit{milk bag}\} \overset{-s}{\Longrightarrow} \{\textit{French bread}\}$ [(Members=5),(Income="2-5")]	0.5894	-

Table II presents some hybrid rules related to the products {*milk box*}, {*milk bag*}, and {*French bread*}. The mined rules evidence that the association between the sales of the products {*milk box*} and {*French bread*} is stronger in the cities located in the south region of Brazil (Porto Alegre and Florianópolis) and among the families with higher income. On the contrary, this association becomes much weaker in the cities of the Northwest (Recife and Fortaleza) and among low-income families. These correspond to examples of rules that are potentially unexpected to domain specialists.

In closing this section, the set of results in III show how much the strength of an association rules can deviate from the average considering families from different cities. Observe that in Belém and Florianópolis the sales of {*garlic*} e {*black beans*} become stronger. However, this association becomes weaker in the city of Belo Horizonte. This is another example of unexpected rule, that can be explained by the fact that residents of Belo Horizonte probably prefer other kinds of beans, such as the brown and white ones.

In summary, the hybrid rules mined in this experiment could reveal valid and, in some cases, unexpected (thus interesting) information regarding the social and cultural differences from the distinct groups of families interviewed by CEF.

VII. Conclusions

In this paper we addressed the problem of mining both positive and negative hybrid-dimensional association rules in

TABLE III. RESULTS - HYBRID ASSOCIATIONS INVOLVING THE PRODUCTS "BLACK BEANS" AND "GARLIC"

id	Hybrid Rule	IH^+	IH^-
HP_6	$\{$*black beans*$\} \overset{+s}{\Longrightarrow} \{$*garlic*$\}$ [(City="Belém")]	0.6412	-
HP_7	$\{$*black beans*$\} \overset{+s}{\Longrightarrow} \{$*garlic*$\}$ [(City="Florianópolis")]	0.5237	-
HN_5	$\{$*black beans*$\} \overset{-s}{\Longrightarrow} \{$*garlic*$\}$ [(City="Belo Horizonte")]	-	0.8624

databases composed by multiple sources of information, by proposing a human-centered approach to mine such patterns. In the proposed technique, users can explore a set of transactional association rules in order to examine if these rules become either much weaker (with lower support value) or much stronger (with lower support value) in some specific subsets of the integrated database, which satisfy specific conditions over selected attributes. The negative hybrid-dimensional associations are mined from candidates that do not achieve an estimated expected support. On the contrary, positive hybrid-dimensional associations are mined from candidates that have actual support value significantly greater the the expected support value. We proposed an interest measure to evaluate both positive and negative hybrid rules.

We defined the HAR algorithm which is especially targeted at mining hybrid-dimensional rules. This algorithm was implemented and evaluated over a real dataset containing survey data. The obtained results evidence that the approach is capable of generating valid and useful patterns to users of data mining and information integration tools.

As future work we first intend to perform detailed analysis on the sensitivity of the results to the input parameters $MinSup$ and $MinIM$. In order to accomplish this task, we intend to evaluate the algorithm in more datasets. We also intend to develop a fully automatic version of the HAR algorithm, i.e., a method capable of generating unexpected exceptions without human intervention.

ACKNOWLEDGMENT

This work was supported by CAPES research grant BEX 1642/14-6 (Eduardo Corrêa Gonçalves).

REFERENCES

[1] R. Agrawal, T. Imielinski and R. Srikant, Mining Association Rules between Sets of Items in Large Databases, *Proc. of the ACM SIGMOD Intl. Conf. on Management of Data*, Washington D.C., USA, 1993, 207–216.

[2] R. Agrawal and R. Srikant, Fast Algorithms for Mining Association Rules, *Proc. of the 20th Very Large DataBase Conference (VLDB'94)* Santiago, Chile, 1994, 487–499.

[3] A. Giacometti, D. H. Li, P. Marcel, A. Soulet, 20 Years of Pattern Mining: a Bibliometric Survey, *ACM SIGKDD Explorations*, Vol 15(1), 2013, 41–50.

[4] E. C. Gonçalves, I. M. B. Mendes and A. Plastino, Mining Exceptions in Databases, *Proc. of the 17th Australian Joint Conf. on Artificial Intelligence (LNAI 3339)*, Cairns, Australia, 2004, 1081-1086.

[5] E. C. Gonçalves and A. Plastino, Mining Strong Associations and Exceptions in the STULONG Data Set, *Proc. of the 6th ECML/PKDD Discovery Challenge*, Pisa, Italy, 2004, 44–55.

[6] J. Han, M. Kamber and J. Pei, *Data Mining: Concepts and Techiniques*, 3rd ed, Morgan Kaufmann, 2011.

[7] M. Kamber, J. Han and J. Y Chiang, Metarule-Guided Mining of multidimensional Association Rules Using Data Cubes, *Proc. of the* 3^{rd} *SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining (KDD'97)*, California, USA, 1997.

[8] K. N. Kontonasios, E. Spyropoulou and T. De Bi, Knowledge discovery interestingness measures based on unexpectedness, *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, Vol 2(5), 2012, 386–399.

[9] G. S. Linoff and M. J. A. Berry, *Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management*, 3rd ed, Wiley, 2011.

[10] B. Padmanabhan and A. Tuzhilin, Unexpectedness as a Measure of Interestingness in Knwoledge Discovery. *Decision Support Systems*, Vol 27, 1999, 303–318.

[11] C. Romero, J.R. Romero, J.M. Luna and S. Ventura, Mining Rare Association Rules from e-Learning Data, *Proceedings of the 3rd International Conference on Educational Data Mining (EDM'10)* Pittsburgh, USA, 2010, 171–180.

[12] L. Szathmary, A. Napoli and P. Valtchev, Towards Rare Itemset Mining, *Proceedings od the 19th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'07)*, Patras, Greece, 2007, 305–312.

[13] A. Savasere , E. Omiecinski and S. Navatge, Mining for Strong Negative Associations in a Large Database of Costumer Transactions. *Proceedings of the 14th International Conference on Data Engineering*, Orlando, Florida, USA, 1998,494–502.

[14] H. C. Tjioe and D. Taniar, Mining Association Rules in Data Warehouses, *International Journal of Data Warehousing and Mining*, Vol. 1(3), 2005, 28–62.

[15] K. Wang, Y. Jiang, L. V. S. Lakshmanan, Mining Unexpected Rules by Pushing User Dynamics. *Proc of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'03)* Washington D.C., USA, 2003, 246–255.

[16] S. Zhang and X. Wu., Fundamentals of association rules in data mining and knowledge discovery, *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* Vol 1(2), 2011, 97–116.