Journal of Computer Science

Original Research Paper

A User-Driven Association Rule Mining Based on Templates for Multi-Relational Data

[1]Carlos Roberto Valêncio, [1]Guilherme Henrique Morais,
[2]Márcio Zamboti Fortes, [2]Angelo Cesar Colombini,
[1]Leandro Alves Neves, [3]Mario Luiz Tronco and [1]William Tenório

[1]São Paulo State University (Unesp), Institute of Biosciences, Humanities and Exact Sciences (Ibilce), Campus São José do Rio Preto, São Paulo, Brazil
[2]Fluminense Federal University (UFF), Niterói, Rio de Janeiro, Brazil
[3]São Paulo University (EESC-USP), São Carlos, São Paulo, Brazil

Article history
Received: 10-04-2018
Revised: 18-05-2018
Accepted: 05-11-2018

Corresponding Author:
Carlos Roberto Valêncio
Department of Computer Science and Statistics - DCCE, São Paulo State University (Unesp), Institute of Biosciences, Humanities and Exact Sciences (Ibilce), Campus São José do Rio Preto, São Paulo, Brazil
Email: carlos.valencio@unesp.br

Abstract: Data mining algorithms to find association rules are an important tool to extract knowledge from databases. However, these algorithms produce an enormous amount of rules, many of which could be redundant or irrelevant for a specific decision-making process. Also, the use of previous knowledge and hypothesis are not considered by these algorithms. On the other hand, most existing data mining approaches look for patterns in a single data table, ignoring the relations presented in relational databases. The contribution of this paper is the proposition of a multi-relational data mining algorithm based on association rules, called TBMR-Radix, which considers previous knowledge and hypothesis through the using of the Templates technique. Applying this approach over two real databases, we were able to reduce the number of generated rules, use the existing knowledge about the data and reduce the waste of computational resources while processing. Our experiments show that the developed algorithm was also able to perform in a multi-relational environment, while the MR-Radix, that does not use Templates technique, was not.

Keywords: Data Mining, Templates, Association Rules, Knowledge Discovery in Databases, Multi-relational Data Mining, User-Driven Filter

Introduction

In recent years the speed at which data are generated and collected has increased the volume of stored data. While the resulting datasets has proved to be efficient structures for storing, managing and retrieving these data, the obtainment of knowledge is not a trivial task (Larose and Larose, 2014). Thus, the development and evolution of algorithms and techniques that enhance the performance and quality of this process constitutes an important contribution (Han *et al*., 2011).

Among the different methods available to fulfill this task, it is possible to highlight the algorithms that perform the extraction of association rules (Han *et al*., 2011), such as the traditional Apriori algorithm (Agrawal *et al*., 1993; Liu, 2010), FP-Growth (Wu *et al*., 2007; Zhang *et al*., 2008), PatriciaMine (Pietracaprina and Zandolin, 2003) and MR-Radix (Valêncio *et al*., 2012).

When considering association rules as a tool to extract useful knowledge from databases, there is a challenge related to the quality and relevance of them. Most algorithms produce a high amount of association rules and do not have any measurement of relevance and quality, which difficult the analysis and extraction of useful knowledge (Rameshkumar *et al*., 2013). As a consequence the analyst, which is a person who is responsible for the data analysis process, has to manually analyze the results in order to find rules that contribute to a specific aim.

To beat this problem, several measures have been proposed to evaluate the interestingness of a rule, in addition to the traditional support and confidence measures. As mentioned by Dahbi *et al*. (2016) these interestingness measures fall into two broad groups: User-driven and data-driven. The user-driven are subjective measures based on comparing the discovery rules with the previous knowledge or beliefs of the analyst, while data-driven are objectives measures that give the interestingness in terms of statistics or information theory (Scott *et al*., 2014; Wu *et al*., 2009).

Depending on its purpose, the analyst should choose an appropriate data-driven interestingness measure to

automatically filter the huge amount of rules. However this choice is not easy since the data-driven interestingness measures have many different qualities and some of those properties are incompatible, as Lenca *et al*. (2008) shows. Also, the abundance of measures gave rise to a new problem, namely the heterogeneity of the evaluation results, in which specific rules can be considered relevant with regard to a measure and irrelevant with regard to another.

In this context, several works aim to help the user in the choice of the most adequate measure as surveyed by Dahbi *et al*. (2017). However, is known in the literature that there is no measure better than others in all application domains (Tan *et al*., 2002) and that there are situations in which many measures are correlated with each other.

On the other hand, user-driven measures proved to be efficient when dealing with the aforementioned problem, since it allows the researches to consider the presence of experts in a given analysis, such as done in other steps of the knowledge discovery in databases (Han *et al*., 2011). In this scenario, visualization has been used as a tool to support the analysis of association rules.

Visualization techniques, such as scatter plot (Liu *et al*., 2012) and node-link graph (Leung *et al*., 2008), were used for exploring and explaining the association rules. These approaches focus on the final association rule result.

Hahsler and Chelluboina (2011), Zhao and Liu (2001) and Bruzzese and Davino (2008) provided visualization techniques to explore association rules and their changing behaviors over the time. However, the proposed techniques are either static or only support basic interactions and are not capable to considerer tasks that require user expertise and domain knowledge.

Chen *et al*. (2017) proposed a reinforcement of the conventional association rule mining process by mapping the entire process into a visualization process, reducing the analyst's workload, while providing a wide set of visual explorations tools to support the inspection and manipulation. However, the analyst still has to choose an appropriate interestingness measure, as well as its threshold for finding interesting rules.

In that way, there is still a need for techniques that allow the filtering as a pre-step, before the association rule generation, in order to reduce the computational cost. Also, there is a need for techniques that allow the use of previous knowledge to guide the data mining process.

In addition to the several methods for data mining it is important to note that the algorithms commonly require the data to be stored in a single table: Reducing the one-to-one and many-to-many relationships, 1:N and N:M respectively, to a single table can produce duplications and inconsistencies in the dataset and result in a loss of information (Valêncio *et al*., 2011).

In this context, this work makes use of the concept of Templates as a mechanism to remove uninteresting and excessive amount of rules, as well as to make possible the use of previous knowledge, when considering a multi-relational database. In this sense, the work enables the analysis of data spread among several tables, preserving the existing semantic between then and encourages the use of previously obtained knowledge to guide the process.

Background

This section discusses the essential concepts related to the proposed work.

Multi-Relational Data Mining

As previously mentioned, several approaches to mine useful information from databases require the data to be stored in a single structure, such as a single tables or files. To apply these traditional algorithms, the data must go through a preprocessing to integrate them through joining or aggregation operations.

Although the application of this operation is possible and sufficient for some applications, it can produce unsatisfactory results, since a joining process could generate a table with many registers when the extraction is done from large databases, as well as the appearance of inconsistencies, duplications or loss of information during the data preprocessing. It imposes a limitation when dealing with data presented in a structured form, such as relational databases (Jiménez *et al*., 2012).

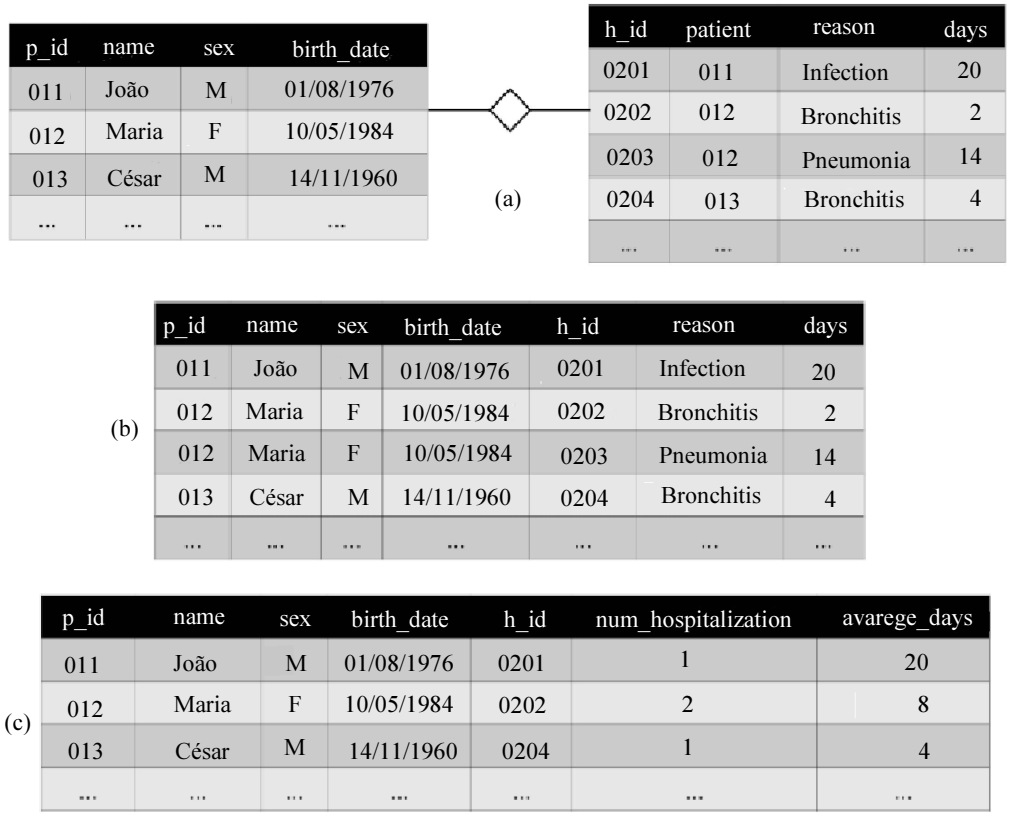
An example, given by Valêncio *et al*. (2012), is shown in Fig. 1. An one-to-many relationship between patients and their hospitalizations is given in Fig. 1a. In Fig. 1b is possible to observe the use of natural joining through foreign keys, which results in the redundancy of data related to patient Maria. Figure 1c presents the use of the aggregate functions sum and average: The first to count the amount of hospitalization per patient and the latter to calculate the average of days the patients remain hospitalized. The use of aggregate functions cause loss of information, such as the reason why the patients were hospitalized, since this attribute is non-aggregatable.

In this context, is possible to find a recent technique in the literature called multi-relational data mining, which aims to minimize the limits imposed by the traditional algorithms. Their mainly feature is to makes possible the obtainment of knowledge directly from the relational tables without the necessity of joining or aggregate operations, as well as preserves the semantics among them (Spyropoulou and De Bie, 2011).

Association Rules and Templates

One of the most used and important technique to extract knowledge from databases is the association rules, which are statements that help uncover relationships between seemingly unrelated data.

As stated by Agrawal *et al*. (1993), let's considerer $I = \{i_1,i_2,...,i_n\}$ a set of n binary attributes called *items* and $T = \{t_1,t_2,...,t_m\}$ a set of transactions called *database*. Each transaction t_i is called *ItemSet* and $t_i \subseteq I$. A rule is defined as an implication of the form $X \rightarrow Y$ where $X,Y \subseteq I$ and $X \cap Y = \emptyset$ and an association rule is a rule that comply with the user-specified minimum support and confidence values.



(a)

p_id	name	sex	birth_date
011	João	M	01/08/1976
012	Maria	F	10/05/1984
013	César	M	14/11/1960
...	...	...	...

h_id	patient	reason	days
0201	011	Infection	20
0202	012	Bronchitis	2
0203	012	Pneumonia	14
0204	013	Bronchitis	4
...	...	...	...

(b)

p_id	name	sex	birth_date	h_id	reason	days
011	João	M	01/08/1976	0201	Infection	20
012	Maria	F	10/05/1984	0202	Bronchitis	2
012	Maria	F	10/05/1984	0203	Pneumonia	14
013	César	M	14/11/1960	0204	Bronchitis	4
...	...	...	...	...	...	...

(c)

p_id	name	sex	birth_date	h_id	num_hospitalization	avarege_days
011	João	M	01/08/1976	0201	1	20
012	Maria	F	10/05/1984	0202	2	8
013	César	M	14/11/1960	0204	1	4
...	...	...	...	...	...	...

Fig. 1: (a) Patients and their hospitalizations (b) Result of natural joining (c) Use of aggregate functions

The support of the rule is the percentage of the transactions in database that contain $X \cup Y$, i.e., the frequency of the rule in the database. The confidence of the rule is the percentage of the transactions containing X which also contain Y, i.e., the reliability of the rule.

In a more practical representation, an association rule is presented as follows:

$$ItemA, ItemB, [...] \rightarrow ItemX, ItemY, [...] \ \%S; \ \%C$$

where, *ItemA*, *ItemB*, *ItemX* and *ItemY* are *items* from the set *I* and the occurrence of *ItemA* and *ItemB* implicates the occurrence of *ItemX* and *ItemY*. *%S* and *%C* are the relevance metrics support and confidence, respectively.

The *items* in a multi-relational environment are called *relational items*. They are composed by a triple that represents (1) the table it belongs to, (2) the attribute it refers to and (3) its value in the data base, according to the pattern:

$$Table.Attribute = "Value"$$

In order to find the rules, sets of two or more relational items are interesting (Chi and Fang, 2011), which constitutes an *ItemSet* t_i. When an *ItemSet* comply with the minimum support and confidence values, it is called *Frequent ItemSet*.

High values of support and confidence restrict the results to rules with high occurrence, but also could result in loss of interesting rules in databases with high variance in the attribute value. On the other hand, low values results in a greater variety of rules, but the amount of rules can difficult the interpretation of the results. Even though the optimum values are found, the resulted set of association rules could present uninteresting items. Despite the several other interestingness measures proposed in the literature, there is still a need for techniques to deal with this problem.

In this context, some traditional association rule mining algorithms based on the iterative approach of Apriori algorithm filters the rules at the end of the process, removing those that does not match with an specific interest, such as made by Agrawal and Srikant (1994) and Liu (2010). On the other hand, more effective works aims to filter the rules in the data mining step itself, in order to remove subsets that could not result in the desired rules (Lanfang *et al*., 2009; Li *et al*., 2010).

Thus, the use of pre-defined formats to filter the association rules, which is called of Templates, was presented as an alternative as it allows the use of previously knowledge, while reducing the cost to generate

the rules. A Template can have a restrictive or inclusive type: Restrictive type templates represent the undesired results and the inclusive type templates represent the opposite, i.e., the accepted format of rules (Lanfang *et al*., 2009; Li *et al*., 2010). Even though these works has produced satisfactory results, the existent solutions do not lead with the complexity of relational databases, where the loss of semantic due the joint operation also implies in loss of knowledge among the existing relations (Valêncio *et al*., 2011). Additionally the use of Apriori algorithm in these existent solutions could implies in inefficient performance when processing a large volume of data with templates that allows a high number of results.

The TBMR-Radix Algorithm

The developed algorithm, named Template-based MR-Radix (TBMR-Radix) allows the creation of restrictive and inclusive type templates that are considered while mining the multi-relational databases. Thus, this algorithm allows the analyst to guide the extraction of association rules to a specific focus that meets the aim of the analysis and also contributes to the efficient use of time and computational resources.

It also encourages an iterative analysis of the database, through the use of knowledge previously obtained on other analysis, as shown in Fig. 2. Let's consider a real-world scenario where several data mining techniques are applied to find useful and undiscovered information in a database. Through the analysis the researchers can find an interesting behavior in the data, which requires a deeper investigation. So, TBMR-Radix can be used at this point, to restrict the scope of the analysis.

The use of inclusive and restrictive templates is efficient to determine the useful knowledge. However, even though the results are produced based on the templates, some other values could be still uninteresting for the analysis, such as obvious, blank or missing values. The definition of several restrictive type templates to deal with these particularities would not only result in inefficient performance but can also be an inefficient filtering method, since the uninteresting items would still be present on the frequent item sets used for rule extraction.

In order to solve this problem, an Uninteresting Item List was developed which allows the analyst to individually remove all items that do not contribute for a given analysis. Uninteresting items considered by this approach are filtered out of the frequent item set mining phase, as explained in the next subsection. This solution is very useful when dealing with databases that were not normalized during the preprocessing phase, which can contain several inconsistencies and poorly structured items.

The TBMR-Radix Workflow

The described algorithm is based on the MR-Radix algorithm, proposed by Valêncio *et al*. (2012). The MR-Radix has good results in terms of processing time and memory consumption when compared with other well-known solutions found on the literature, such as FP-Growth, FP-Growth*, Apriori, OportuneProject, CT-Pro e FP-Growth-Tiny, which justify their choice as based algorithm for de development of this work. One of the mainly features that increase their performance is the use of Radix-Tree structure in order to represent the data in the main memory.

The Fig. 3 represents the workflow of TBMR-Radix, where the highlighted boxes represent modifications in the original MR-Radix algorithm.

Initially, after the data selection and definition of support and confidence values, the algorithm receives and stores the restrictive and inclusive type templates defined by the analyst, as well as the uninteresting item list. The process continues as traditionally made through the construction of *ItemList*, *ItemMap* and *Radix-Tree* structures until the step of generation of frequently *ItemSets* ends. During this step the items specified in the inclusive type templates are considered allowed, while the items in the uninteresting list are ignored.

The ItemSets that have interesting items and do not have any ignored items are used to produce the association rules. The format of restrictive and inclusive type templates is considered during this phase in order to only produce rules that have the desired format.

The association rules, which have the format of at least one inclusive type template and have not the format of any restrictive type template are accepted to compose the final results. The algorithm also validates the minimum confidence value of each rule.

Template Syntax

The template syntax used in this study was design to be similar to the most recent object oriented programing languages in order to simplify its use by data analysts and software developers. It was also proposed to be as simple and intuitive as possible in order to allow its use in analysis in areas not related to computational and information technology.

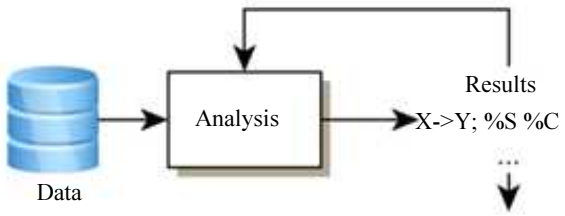


Fig. 2: Iterative analysis using previous knowledge

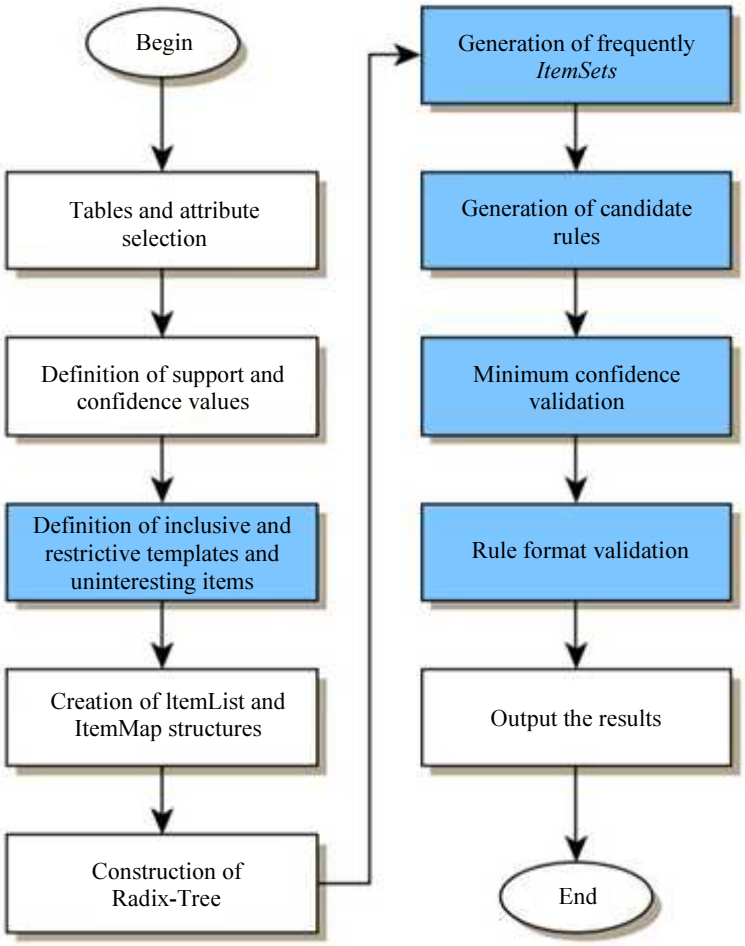


Fig. 3: TBMR-Radix algorithm

The main concept used to define the template syntax is the representation of a database as Relational Items, as previously presented. Four reserved words are used in order to describe the relational items that will compose the template and to provide an abstract description that might be consider relevant. The reserved words are:

- **AnyTable** - used on the *Table* field of the relational item; their use indicates that items from any table could be considered
- **AnyAttrib** - used on the *Attribute* field of the relational item; their use indicates that items from any attribute (in the defined table) could be considered
- **AnyValue** - used on the *Value* field of the relational item; their use indicates that items with any values (from the defined table-attribute) could be considered
- **AnyOthers** - this reserved word does not refer to a relational item. It is a modifier used to indicate that any other relational items could be considered relevant, in addition to the already defined items

To describe a Template, six symbols are also used, as shown in the Table 1.

In order to exemplify the use of a template, considerer the following hypothetical analysis: *"We want to determine a risk profile of patients that suffered heart attacks"*. In other words we want to identify which patients' features may implicate in the occurrence of heart attack. So the following template could be used:

$$*AnyTable*.*AnyAtrib* = *AnyValue*\#*AnyOthers*$$
$$\Rightarrow TBrecord.symptom = "Heart\ Attack"$$

The left side of the template combines the three main reserved words and represents the general form of the template syntax, meaning the any items could be accepted. It also presents the modifier **AnyOthers**, which means that rules with any amount relational items could be accepted. The right side, after the => symbol, indicates that will be considered relevant the rules that implicates in heart attack symptom only.

Table 1: Symbols considered to define the templates

Symbol	Use
.	Used to separate the table and attribute fields of relational item. [X.Y] means that attribute Y belongs to table X.
=	Used to separate the attribute and value fields of the relational item. [X.Y="Z"] means that attribute Y from table X must have (or have not, in case of restrictive template) a value equals to Z.
,	Used to separate multiple relational items (left side)
=>	Used to separate the left and right side of association rules. [X=>Y] means that the occurrence of relational item X implicates the occurrence of relational item Y.
;	Used to separate multiple templates
#	Used to indicate the end of relational items' definition before the modifier **AnyOthers**

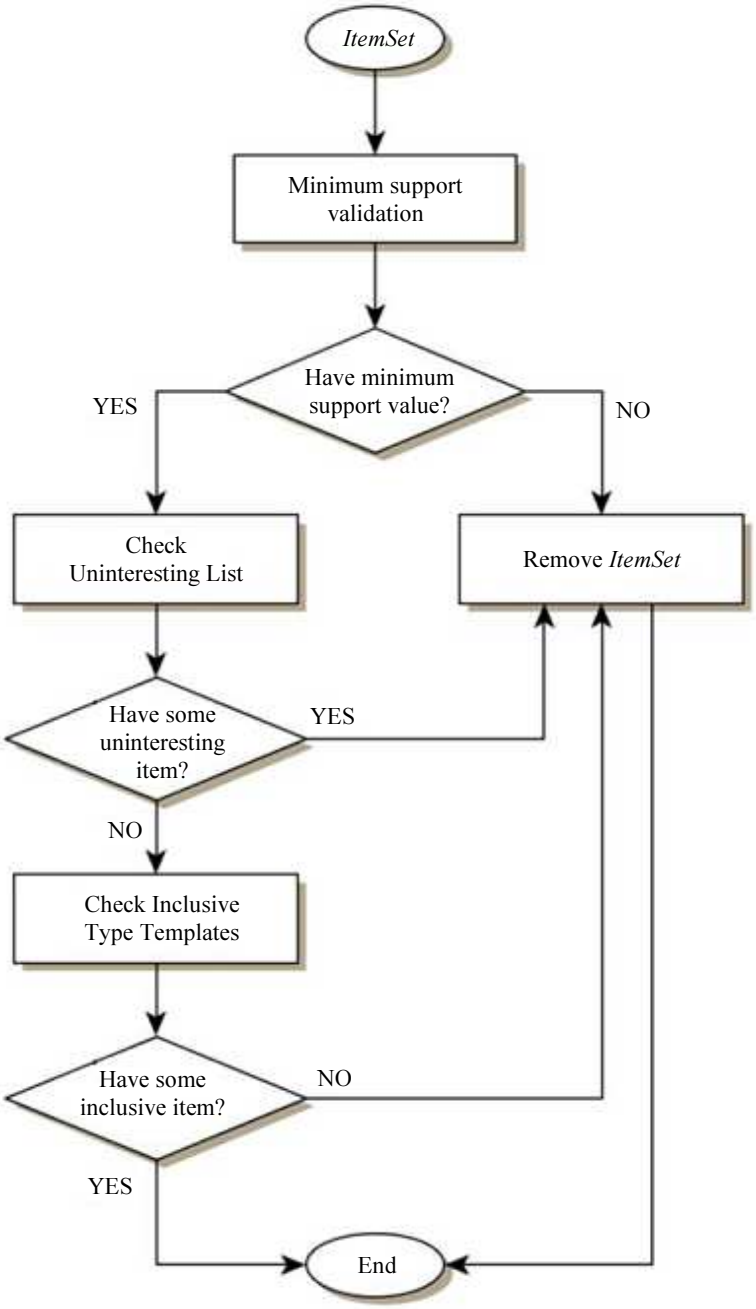


Fig 4: *ItemSets* filtering

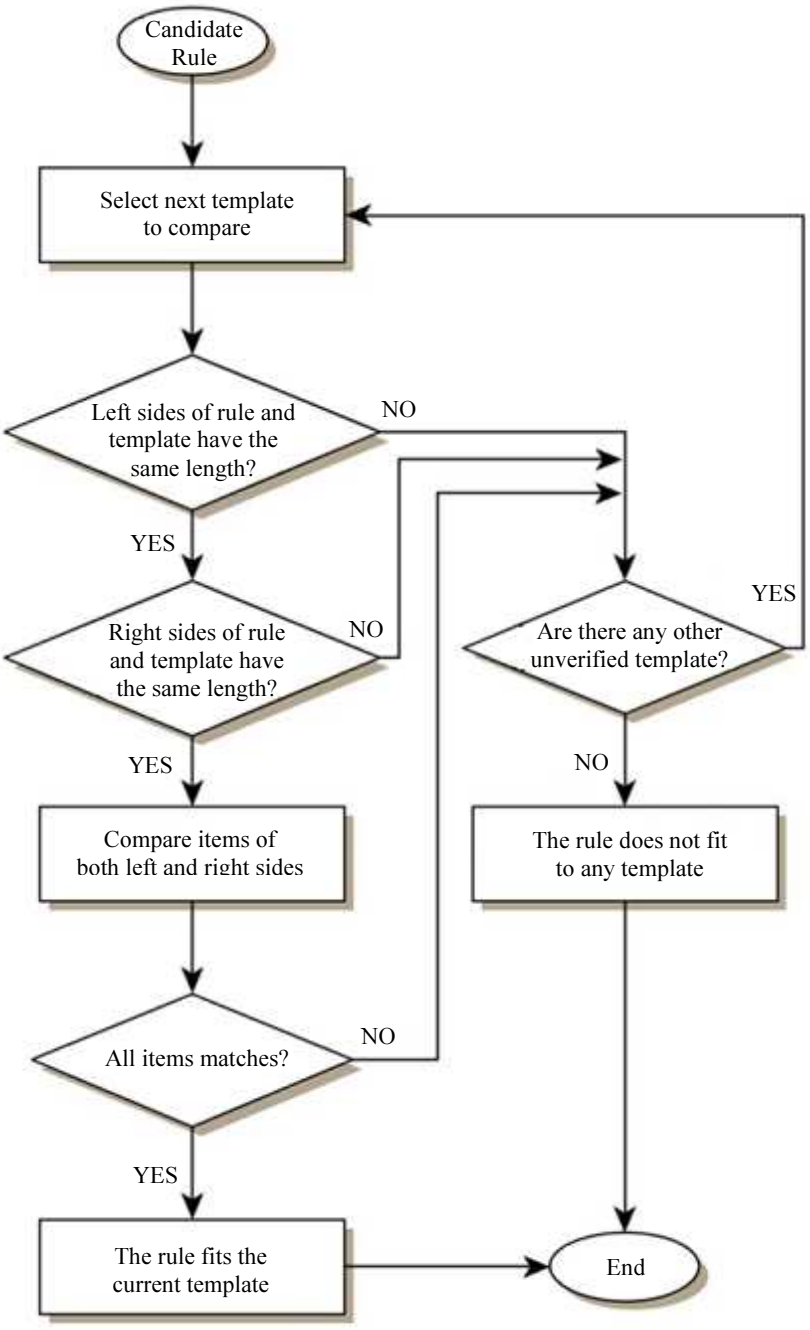


Fig. 5: Association rules

ItemSet Filtering and Generation of Rules

The TBMR-Radix algorithm makes use of three new structures in order to store de templates and the uninteresting items list defined by the analyst: *ItemDB_Temp*, *Template* e *TemplateGroup*. Each one is defined as follows:

- *ItemDB_Temp* is responsible to store the relational item triples produced by the template definition
- *Template* stores the defined templates. It is composed by two linked list of *ItemDB_Temp* elements, each one to store the data triples of a specific side of the rule
- *TemplateGroup*: Groups the Template objects into inclusive and restrictive group

Just as MR-Radix algorithm (Valêncio *et al*., 2012) does, the TBMR-Radix performs a depth-search in the

ItemList and *Radix Tree* structures in order to identify the frequent *ItemSets*. After that, the *ItemSets* are filtered as present in the workflow of Fig. 4.

Once the frequent and relevant *ItemSets* are identified, the combination of all items belonging to each *ItemSet* is used to produce the association rules. After that, the produced association rules are filtered in order to comply with confidence and support values and with the specified form of some inclusive template. The association rules that comply with the specified form of some restrictive template are removed, as shown in Fig. 5.

Materials and Methods

The proposed algorithm called TBMR-Radix is an improvement of the MR-Radix algorithm proposed by Valêncio *et al*. (2012).

There were three experiments conducted in order to produce this paper and to measure the efficiency of the template approach when generating the association rules. For each one, we aimed to compare MR-Radix and TBMR-Radix in terms of relevance of the produced association rules and the memory consumed find them.

The general specifications of the physical computer were an Intel Core i5-2410 M 2.3 GHz, 3 MB Cache (2.9 GHz with Max Turbo) processor, 6 GB Corsair Vengeance DDR3 (1600 MHZ)/(3×2 GBD) RAM and 750 GB-7200 RPM HDD.

The experiments were conducted over the following databases:

- SIVAT Database (Rodrigues *et al*., 2011)
- HEPATITE Database

The SIVAT Database is a real data collection about occupational accidents in the cities of São José do Rio Preto and Ilha Solteira, in São Paulo State, Brazil. It consists of a main table, which have 100.030 records of occupational accidents and three auxiliary tables: The first one to store the external causes, the second one to store the illness, following the International Classification of Disease pattern and the last one to store the body parts affected by the accident. These auxiliary tables have 90.562, 103.497 and 109.998 records, respectively. In addition, the main table has 35 attributes while the others have 3 attributes each. So, this database is relevant to be used, since it stores several records and presents relationship between tables, which characterizes a multi-relational environment. Besides that, this database was already used in real data mining application using the MR-Radix.

The HEPATITE Database is a real data collection of genetic and laboratory information about Hepatitis diseased patients. It consists in a single table with 64 records stored on it and 40 attributes. Also, the attributes have a high variability of values. This database is also relevant since it stores just a few records, but presents a considerable amount of attributes. This configuration proved to be problematic to MR-Radix algorithm in previous real data mining analysis, which motivates the application of TBMR-Radix in this scenario.

The first and second experiments were conducted over the SIVAT database. In these experiments we wish to obtain frequently patterns on the occupational accident records that implicates on medical leave from work. In other words, was aimed to find all frequently item sets that imply in the occurrence of the item *record.medicalLeave ="yes"*.

To do that, the TBMR-Radix considered the following template:

$$*AnyTable*.*AnyAtrib* = *AnyValue*$$
$$\#* AnyOthers* => record.medicalLeave = "yes"$$

The left side of the template consists in the syntax's generic form, meaning that will be considered relevant any amount of any relational items that implicates in the relational item *record.medicalLeave="yes"*.

The third experiment was conducted over the HEPATITE database and aimed to identify which variations in the patients' DNA are related with the Hepatitis disease. So was considered the 23 genetic attributes and 1 laboratorial attribute related to fulminant cases, which can assume the values "fulminant", "acute" or "chronic".

To do that, the TBMR-Radix considered the following template:

$$*AnyTable*.*AnyAtrib* = *AnyValue*$$
$$\#* AnyOthers* => tb\_alvo.fulminant\_cases = *AnyValue*$$

The left side of the template consists in the syntax's generic form of relational items that implicates in some fulminant_cases values, i.e., the relational item *tb_alvo.fulminant_cases=*AnyValue**. Was also inserted an item in the Uninteresting List in order to reject the null and blank values in the fulminant_cases attribute.

In each experiment the support and confidence values were set to 10%, in order to restrict the amount of association rules produced.

Results

The first experiment intended to analyze MR-Radix and TBMR-Radix when considering only ten attributes of the SIVAT database's main table, in order to characterize a traditional data mining environment.

The MR-Radix algorithm produced an amount of 478 association rules from 152 *ItemSets*. It performs in 2604 milliseconds and reached a peak of average memory

consumption of 82.22 MB. An analysis of the 478 produced association rules showed that only 19 of them have the relational item *record.medicalLeave="yes"*, i.e. 96% of produced association rules is uninteresting to the given analysis and will have to be manually removed by the analyst.

The TBMR-Radix algorithm produced an amount of 19 association rules from 20 *ItemSets*. It performs in 2803 milliseconds and reached a peak of average memory consumption of 86.56 MB. All the 19 association rules produced present the desired structure in the considered problem. So, 100% of the association rules produced is considered relevant.

Table 2 summarizes the results of the first experiment.

In order to analyze the algorithms in a multi-relational environment, the second experiment was conducted over a set of ten attributes of SIVAT's main table and three more attributes from an auxiliary table.

The MR-Radix algorithm produced 1030 association rules from 222 *ItemSets*. It performs in 3863 milliseconds and reached a peak of memory consumption in 85.96 MB. Even though MR-Radix performs well in a multi-relational environment, 95.05% of the produced association rules are irrelevant to the considered analysis.

The TBMR-Radix produced 50 rules from 51 *ItemSets* and performs in 4632 milliseconds with a peak of memory consumption in 84.76%.

Table 3 summarizes the results of the second experiment.

Again, although the processing time of the TBMR-Radix was longer than the MR-Radix, all the 50 association rules produced are relevant, i.e. 100% of them has the relational item *record.medicalLeave="yes"*.

In the third experiment when the HEPATITE database was used, the MR-Radix algorithm was not able to perform, since it exceeds the memory resource of the environment test. Figures 6 and 7 present the processing load and memory consumption of MR-Radix when we were trying to perform this experiment, respectively.

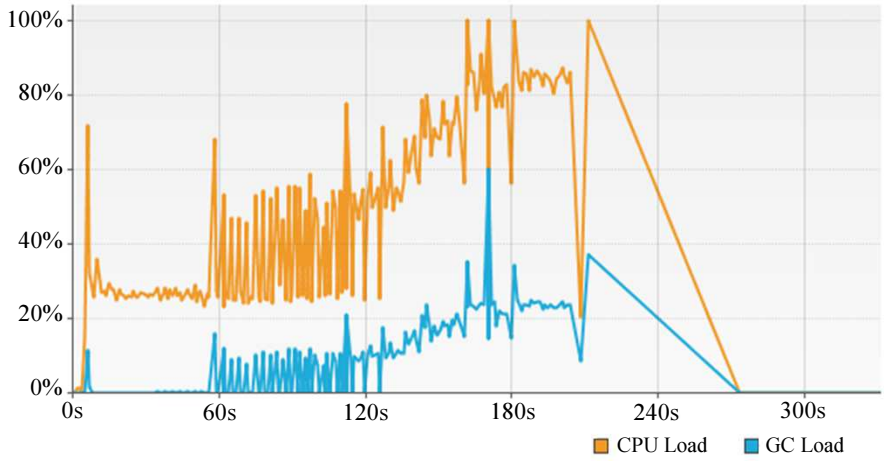


Fig. 6: Processing load of MR-Radix over time

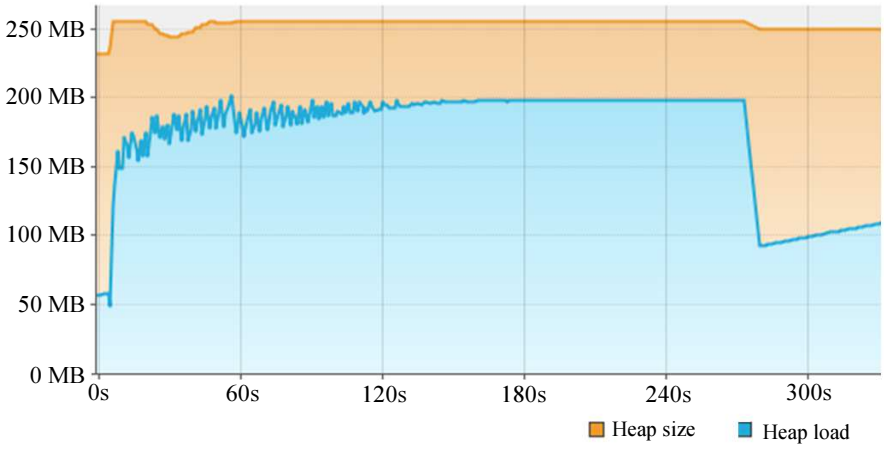


Fig. 7: Memory consumption of MR-Radix over time

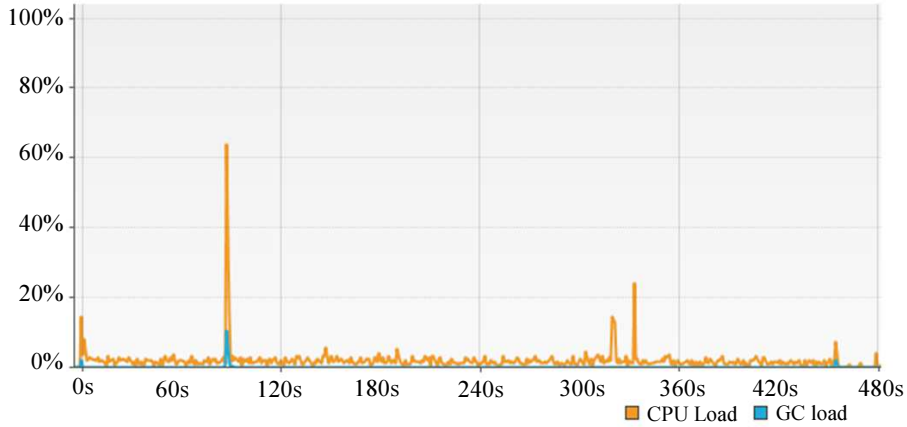


Fig. 8: Processing load of TBMR-Radix over time

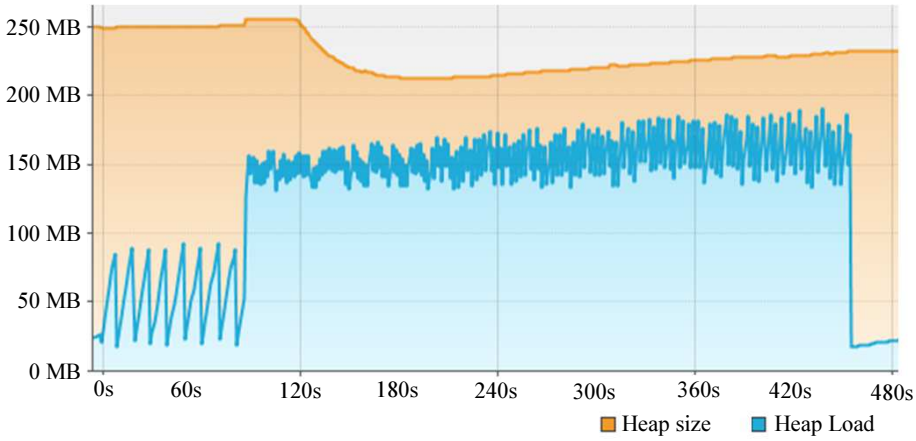


Fig. 9: Memory consumption of TBMR-Radix over time

Table 2: Results of the second experiment

	MR-Radix	TBMR-Radix
Amount of generated rules	1030.00	50.00
Amount of interesting rules	50.00	50.00
Memory consumption (MB)	85.96	84.76
Processing time (ms)	3863.00	4632.00

Table 3: Results of the first experiment

	MR-Radix	TBMR-Radix
Amount of generated rules	478.00	19.00
Amount of interesting rules	19.00	19.00
Memory consumption (MB)	82.22	86.56
Processing time (ms)	2604.00	2803.00

On the other hand, the TBMR-Radix was able to perform without errors and produced 1552 association rules from 4419 *ItemSets* in 471257 milliseconds (~7 minutes and 50 seconds) and reached a peak of memory consumption of 182 MB. The Figure 8 and 9 present the processing load and memory consumption of TBMR-Radix in this experiment, respectively.

Discussion

In both first and second experiments, the MR-Radix algorithm generated more than 95% of irrelevant rules, when considering the desired relational item. Although this result is expected from most association rule algorithms, the execution must be followed by an extensive analysis and selection of relevant rules, which can be tiring and may result in loss of useful knowledge due to human error. Thus, this experiment reaffirms the need for studies about techniques that can analyze the relevant of the produced association rules.

Besides that, all the association rules produced by TBMR-Radix were exactly the same as those produced by MR-Radix that contains the desired relational item. Thus, is possible to observe that there is no loss of interesting results through the use of template technique. It is also possible to observe that this technique does not affect the accuracy of the resulting association rules produced by the algorithm. The accuracy remains the same from MR-Radix.

A growth in memory consumption and processing time when executing TBMR-Radix can also be observed, which is related to the extra processing load need to validate the candidate rules. However, this growth seems to be irrelevant and do not prejudice the use of template technique, as it allows a reduction in the number of uninteresting rules.

In the third experiment, the failure of the MR-Radix algorithm is due to the characteristics of HEPATITE Database since it has a greater amount of attributes with high variance, which turns the manipulation of the *ItemSets* more computationally expensive. The Figures 6 and 7 show that there was a constant processing load value and a progressively growing use of memory in the first minute; it refers to the step of frequently *ItemSets* analysis. After that, in the step of processing the *ItemSets* in order to obtain the relevant rules, the processing load grows progressively until the test environment was no longer able to continue the analysis.

On the other hand, even though the processing time of TMBR-Radix is longer than the MR-Radix, the Figures 8 and 9 show that the processing load was considerable better and seemed to be stabilized. This feature is due to the *ItemSets* Selection phase, which reduces the number of *ItemSets* to be analyzed, since the *ItemSets* that does not present any item with the desired relational item are removed.

Also, the TBMR-Radix was able to perform over a multi-relational environment. It is an important feature since the most other association rules mining approaches assume that the data resides in a single table and require preprocessing to integrate data from multiple tables, through joining or aggregation, into a single table before they can be applied, which can cause loss of meaning or information.

Real-World Application

In order to clarify the importance and potential use of the proposed algorithm, we present a real-world application (Valêncio *et al*., 2016) conducted by the Database Research Group - GBD (http://grupogbd.com/PortalGBD/about), from São Paulo State University in Brazil.

Several data mining techniques were applied to a data set composed by more than 38.000 medical records from a psychiatric hospital. During the analysis of a huge amount of obtained results, the analyst was possible to identify a high occurrence of treatment abandon through absconding, when the patient runs away from the hospital. Knowing that this behavior can be linked to serious consequences to the patient, the analyst applied the TBMR-Radix algorithm to find other associated behaviors.

The new findings make possible the creation of a risk profile of “runaway patients”, for example: Patients in social and economic disadvantage, assisted by the Brazil's publicly funded health care system and who were hospitalized due acute intoxication by the use of psychoactive substances, are considered to be in high risk of absconding. It allowed the medical team to take preventive measures in order to avoid this behavior.

Conclusion and Future Works

The application of association rules algorithms is an important tool to extract useful knowledge from databases. However, most of the algorithms produce a high amount of association rules, do not have any measurement of relevance and quality and require the data to be stored in a single table, which difficult the analysis. Besides that, the data mining algorithms commonly requires the data to be stored in a single table, which can produce duplications and inconsistencies, as well as loss of information.

Thus, this work proposed the use of a user-driven technique for mining association rules based on templates in order to reduce the amount of association rules when considering a multi-relational environment. The experiments proved that TBMR-Radix can reduce the number of uninteresting rules while maintaining the computational cost to produce the rules. Besides that, this work contributes to a more effective and efficient analysis, since the analyst can also focus at different scenarios separately and select the most useful result for each one. In a conventional analysis all scenarios are simultaneously considered, which implies on the task of manually differentiate and evaluate them.

An important direction for future work is to considerer ontologies as a way to potentiate the use of previous knowledge obtained on previous analysis.

Acknowledgement

We thank the São Paulo Research Foundation (FAPESP) and Dr. Adolfo Bezerra de Menezes Hospital, from Brazil, for the financial support and the authors and reviewers for their relevant contributions.

Author’s Contributions

In order to elaborate this paper, a bibliographical survey of the related areas was carried out by the authors, where different approaches were found. Based on discussed suggestions for improvements, the authors defined basic objectives for the work, described the algorithm developed and each stage of its operation was detailed.

Ethics

This article is original and contains unpublished material. The authors confirm that are no conflict of interest involved.

References

Agrawal, R., T. Imieliński and A. Swami, 1993. Mining association rules between sets of items in large databases. ACM SIGMOD Record, 22: 207-216. DOI: 10.1145/170035.170072

Agrawal, R. and R. Srikant, 1994. Fast algorithms for mining association rules in large databases. Proceedings of the 20th International Conference on Very Large Data Bases, Sep. 12-15, Morgan Kaufmann Publishers Inc., Chile, pp: 487-499. https://dl.acm.org/citation.cfm?id=672836

Bruzzese, D. and C. Davino, 2008. Visual mining of association rules. Visual Data Min., 4404: 103-122. DOI: 10.1007/978-3-540-71080-6_8

Chen, W., C. Xie, P. Shang and Q. Peng, 2017. Visual analysis of user-driven association rule mining. J. Visual Lang. Comput., 42: 76-85. DOI: 10.1016/j.jvlc.2017.08.007

Chi, X. and Z.W. Fang, 2011. Review of association rule mining algorithm in data mining. Proceedings of 3rd International Conference on Communication Software and Networks, May 27-29, Xian, China, pp: 512-516. DOI: 10.1109/ICCSN.2011.6014622

Dahbi, A., M. Mouhir, Y. Balouki and T. Gadi, 2016. Classification of association rules based on K-means algorithm. Proceedings of the 4th IEEE International Colloquium on Information Science and Technology, Oct. 24-26, IEEE Xplore Press, Tangier, Morocco, pp: 300-305. DOI: 10.1109/CIST.2016.7805061

Dahbi, A., S. Jabri, Y. Ballouki and T. Gadi, 2017. A new method to select the interesting association rules with multiple criteria. Int. J. Intell. Eng. Syst., 10: 191-200. DOI: 10.22266/ijies2017.1031.21

Hahsler, M. and S. Chelluboina, 2011. Visualizing association rules: Introduction to the R-extension package arulesViz. R Project Module. http://cran.ma.imperial.ac.uk/web/packages/arulesViz/vignettes/arulesViz.pdf

Han, J., M. Kamber and J. Pei, 2011. Data Mining: Concepts and Techniques. 3rd Ed. Elsevier, ISBN-13: 9780123814807, pp: 744.

Jiménez, A., F. Berzal and J.C. Cubero, 2012. Using trees to mine multirelational databases. Data Min. Knowl. Discovery, 24: 1-39. DOI: 10.1007/s10618-011-0218-x

Lanfang, L., P. Qingxian and C. Zhiyu, 2009. An efficiency filtering algorithm for mining association rules. Proceedings of the 4th International Conference on Computer Science and Education, Jul. 25-28, Nanning, China, pp: 1847-1850. DOI: 10.1109/ICCSE.2009.5228261

Larose, D.T. and C.D. Larose, 2014. Discovering Knowledge in Data: An Introduction to Data Mining. 2nd Edn., John Wiley and Sons, New Jersey, ISBN-13: 9781118873571, pp: 336.

Lenca, P., P. Meyer, B. Vaillant and S. Lallich, 2008. On selecting interestingness measures for association rules: User oriented description and multiple criteria decision aid. Eur. J. Operat. Res., 184: 610-626. DOI: 10.1016/j.ejor.2006.10.059

Leung, C.K.S., P.P. Irani and C.L. Carmichael, 2008. WiFIsViz: Effective visualization of frequent itemsets. Proceedings of the 8th International Conference on Data Mining, Dec 15-19, Pisa, Italy, pp: 875-880. DOI: 10.1109/ICDM.2008.93

Li, N., Z. Li and X. Li, 2010. An improved template-based method for mining association rules from defect repositories. Proceedings of the 5th International Conference on Computer Science and Education, Aug. 24-27, China, pp: 1796-1799. DOI: 10.1109/ICCSE.2010.5593798.

Liu, Y., 2010. Study on application of apriori algorithm in data mining. Proceedings of the 2nd International Conference on Computer Modeling and Simulation, Jan. 22-24, Sanya, China, pp: 111-114. DOI: 10.1109/ICCMS.2010.398

Liu, G., A. Suchitra, H. Zhang, M. Feng and S.K. Ng *et al*., 2012. AssocExplorer: An association rule visualization system for exploratory data analysis. Proceedings of the 18th International Conference on Knowledge Discovery and Data Mining, Aug. 12-16, ACM, Beijing, China, pp: 1536-1539. DOI: 10.1145/2339530.2339774

Pietracaprina, A. and D. Zandolin, 2003. Mining frequent itemsets using Patricia tries. Proceedings of ICDM Workshop on Frequent Itemset Mining Implementations, Dec. 19-19, Melbourne, USA. http://ceur-ws.org/Vol-90/pietracaprina.pdf

Rameshkumar, K., M. Sambath and S. Ravi, 2013. Relevant association rule mining from medical dataset using new irrelevant rule elimination technique. Proceedings of the International Conference on Information Communication and Embedded Systems, Feb. 21-22, IEEE Xplore Press, India, pp: 300-304. DOI: 10.1109/ICICES.2013.6508351

Rodrigues, H.P., A.P. Flores, A.F. Silva, C.R. Valêncio and D.C.M. Segura *et al*., 2011. Sistema computacional para análise de notificações de acidentes de trabalho por meio de recursos georeferenciados. Revista Ciência em Extensão, 7: 54-54.

Scott, M., R.P. Boardman, P.A. Reed, T. Austin and S.J. Johnston *et al*., 2014. A framework for user driven data management. Inform. Syst., 42: 36-58. DOI: 10.1016/j.is.2013.11.004

Spyropoulou, E. and T. De Bie, 2011. Interesting multi-relational patterns. Proceedings of the 11th International Conference on Data Mining, Dec. 11-14, Vancouver, Canada, pp: 675-684. DOI: 10.1109/ICDM.2011.82

Tan, P.N., V. Kumar and J. Srivastava, 2002. Selecting the right interestingness measure for association patterns. Proceedings of the 8th International Conference on Knowledge Discovery and Data Mining, July 23-26, ACM, Edmonton, Canada, pp: 32-41. DOI: 10.1145/775047.775053

Valêncio, C.R., F.T. Oyama, P.S. Neto and R.C.G. Souza, 2011. Comparative study of algorithms for mining association rules: Traditional approach versus multi-relational approach. Proceedings of the 12th International Conference on Parallel and Distributed Computing, Applications and Technologies, Oct. 20-22, IEEE Xplore Press, South Korea, pp: 275-280. DOI: 10.1109/PDCAT.2011.29

Valêncio, C.R., F.T. Oyama, P.S. Neto, A.C. Colombini and A.M. Cansian *et al*., 2012. MR-Radix: A multi-relational data mining algorithm. Human-Centric Comput. Inform. Sci., 2: 1-17. DOI: 10.1186/2192-1962-2-4

Valêncio, C.R., G.T. Saturno, G.P. Daniel, V.H.P. Martins and W. Tenório *et al*., 2016. Extração de Conhecimento em Banco de Dados do Hospital Filantrópico Dr. Adolfo Bezerra de Menezes: Doença atual e Evolução. UNESP/IBILCE, São José do Rio Preto, ISSN: 978-85-8224-141-7.

Wu, F., S.W. Chiang and J.R. Lin, 2007. A new approach to mine frequent patterns using item-transformation methods. Inform. Syst., 32: 1056-1072. DOI: 10.1016/j.is.2007.01.001

Wu, J., K.L. Tan and Y. Zhou, 2009. Data-driven memory management for stream joins. Inform. Syst., 34: 454-467. DOI: 10.1016/j.is.2009.02.001

Zhang, W., H. Liao and N. Zhao, 2008. Research on the FP growth algorithm about association rule mining. Proceedings of the International Seminar on Business and Information Management, Dec. 19-19, Wuhan, China, pp: 315-318. DOI: 10.1109/ISBIM.2008.177

Zhao, K. and B. Liu, 2001. Visual analysis of the behavior of discovered rules. Proceedings of the seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Aug. 26-29, ACM, San Francisco, USA, pp: 59-64.