NARFO* ALGORITHM
Optimizing the Process of Obtaining Non-redundant and Generalized Semantic Association Rules

Rafael Garcia Miani
IBM Brazil Software Laboratory, 04753-080, São Paulo, Brazil

Cristiane Akemi Yaguinuma, Marilde Terezinha Prado Santos and Vinícius Ramos Toledo Ferraz
Department of Computer Science, Federal University of São Carlos, P.O. Box 676, 13565-905, São Carlos, Brazil

Keywords: Data Mining, Generalized Semantic Association Rules, Redundant Rules, Fuzzy Ontology.

Abstract: This paper proposes the NARFO* algorithm, an algorithm for mining non-redundant and generalized association rules based on fuzzy ontologies. The main contribution of this work is to optimize the process of obtaining non-redundant and generalized semantic association rules by introducing the *minGen* (Minimal Generalization) parameter in the latest version of NARFO algorithm. This parameter acts on generalize rules, especially the ones with low minimum support, preserving their semantic and eliminating redundancy, thus reducing considerably the amount of generated rules. Experiments showed that NARFO* produces semantic rules, without redundancy, obtaining 68,75% and 55,54% of reduction in comparison with XSSDM algorithm and NARFO algorithm, respectively.

1 INTRODUCTION

Many approaches on mining association rules are motivated by finding new ways of dealing with different attribute types or increasing computational performance. In the mean time, a crescent number of approaches have been developed concerning semantics of mined data, aiming at improving the quality of obtained knowledge. In this sense, ontologies have been widely employed to represent semantic information defined by knowledge experts, and can also be applied to enhance association rule mining. Some researches, like (Srikant and Agrawal, 1995) and (Hou, et al., 2005), have extended the process of mining association rules in order to mine association rules that represent relation between basic data items, as well as between items at any level of the related taxonomy (is-a hierarchies) or ontology, resulting in the so called *generalized association rules*.

The use of traditional ontologies based on propositional logic is a common point in approaches to generalize association rules. Such restriction becomes inappropriate to represent some concepts and relationships of the real world. Therefore, some researches have used fuzzy ontologies in order to extract semantically richer association rules, such as (Chen, et al., 2003), extended SSDM (Escovar, et al., 2006) and NARFO algorithm (Miani, et al., 2009). However, in cases, by adopting either crisp or fuzzy ontologies in generalized association rule mining, the process of extracting association rules usually brings to users a great amount of rules. Many of these rules represent the same information and are considered redundants. Therefore, it is interesting to avoid mining unnecessary rules that express redundant knowledge, while focusing on the semantic richness provided by fuzzy ontologies.

Considering this context, this paper proposes the NARFO* algorithm to mine non-redundant and generalized association rules based on fuzzy ontologies. The main contribution of this research is the introduction of *minGen* (minimal generalization) parameter. This works on generalizing rules, especially with low minimum support, preserving their semantic and eliminating redundancy.

Garcia Miani R., Akemi Yaguinuma C., Terezinha Prado Santos M. and Ramos Toledo Ferraz V. (2010).
NARFO* ALGORITHM - Optimizing the Process of Obtaining Non-redundant and Generalized Semantic Association Rules.
In *Proceedings of the 12th International Conference on Enterprise Information Systems - Artificial Intelligence and Decision Support Systems*, pages 320-325
DOI: 10.5220/0002968803200325

2 RELATED WORK

Many researches have been employing taxonomies and ontologies as background knowledge in mining association rules in order to enhance the knowledge discovery process. (Hou, et al., 2005) uses domain knowledge to generalize low level rules discovered by traditional rule mining algorithms, in order to get fewer and clearer high level rules. In (Brisson, et al., 2005), domain knowledge is used in pre and post-processing steps. The preprocessing step uses ontology to guide the construction of specific datasets for particular mining tasks. In the post-processing step, mined rules are interpreted and filtered, as terms are generalized based on the ontology.

"However, in many real-world applications, the taxonomic structure may not be crisp but fuzzy." (Chen, et al., 2000, p.47). For this porpose, (Chen, et al., 2000) developed an algorithm to mine generalized association rules with fuzzy taxonomic structures. The algorithm considers the *R-interest* measure, which is used to eliminate redundants and inconsistent rules. Fuzzy association rule mining, developed by (Farzanyar, et al., 2006), is driven by domain knowledge in order to make the rules more visual, more interesting and understandable. (Escovar, et al., 2006) have proposed another approach that uses fuzzy ontologies to represent the semantic similarity relations among mined data. In this case, the mining algorithm (XSSDM algorithm) considers a new measure, called minimum similarity (*minsim*). If two itens have the similarity degree greater than or equal the *minsim,* fuzzy associations are made and can be expressed in the association rules extracted by the algorithm. For example, if item$_1$ and item$_2$ have the degree of similarity greater than or equal *minsim* in the fuzzy ontology, a fuzzy association is made and a fuzzy itemset is created (represented as item$_1$~item$_2$).

Although generalized association rule mining approaches based on fuzzy ontology express semantically richer information, they may result in a great amount of extracted rules and redundant rules. Then, redundancy treatment has been an interesting research topic. In (Han and Fu, 1999) a multiple level association rule was proposed to reduce the number of generalized association rules. This consists in defining different *minsup* values for each level of a given taxonomy. Higher levels have bigger *minsup* values. Other approaches like (Kunkle, et al., 2008) implement its method to reduce the amount of generalized association rules and redundant rules during the pattern extraction process. (Kunkle, et al., 2008) implemented the MFGI_class algorithm based on maximal frequent itemset theory (Bayardo, 1998). (Chen, et al., 2000), and (Oliveira, et al., 2007) are some approaches that treat the problem after the processing stage. The first one does the generalization process based on the *R-interest* measure. This measure prunes redundant rules, only considering the rules which degree of support or confidence is *R* times the expected degree of support or confidence. (Oliveira, et al., 2007) generalizes only if the descendents of an ancestor generate rules, and the rule of the ancestor has support value *x%* greater than the descendent that generate a rule with the biggest support among its siblings. (Miani, et al, 2009) proposed NARFO algorithm, which decreases the number of redundancy rules by its generalizing and redundancy treatment.

NARFO* and NARFO have the same features, but NARFO* reduces the amount of NARFO's rules by the introduction of *minGen* parameter, with more semantic and no equivocated information.

3 NARFO* ALGORITHM

This section explains the NARFO* algorithm. The introduction of *minGen* parameter is the main contribution of this work. *MinGen* works especially generalizing rules with low minimum support, without semantic lost. If *X%* of descendents is included in rules, the generalization is done, and the items that are not part of the generalization are showed to avoid wrong information. Considering the fuzzy ontology of figure 1, if *minGen* value is 0,6 and rules like *Apple* → *Turkey, Kaki* → *Turkey* are generated, the algorithm generalizes these rules to *Fruit* → *Turkey*, since *Apple* and *Kaki* are more than 60% of *Fruit's* descendents (*minGen* value is 0,6).

The rule are showed as *Fruit(-Tomato)* → *Turkey*, indicating that the item tomato did not compose the generalization.

Besides minGen parameter, the algorithm also eliminate redundant rules, preserving their semantic. If both *Apple~Tomato* → *Chicken* and *Apple* → *Chicken* are extracted, NARFO* only considers the first one, exhibiting with a plus (+) the item more relevant in the fuzzy itemset of the rule A*pple(+)~Tomato* → *Chicken.*

3.1 Data Scanning

This step identifies the items in the database generating itemsets of one size (1-itemset). Considering the fuzzy ontology of figure 1, this

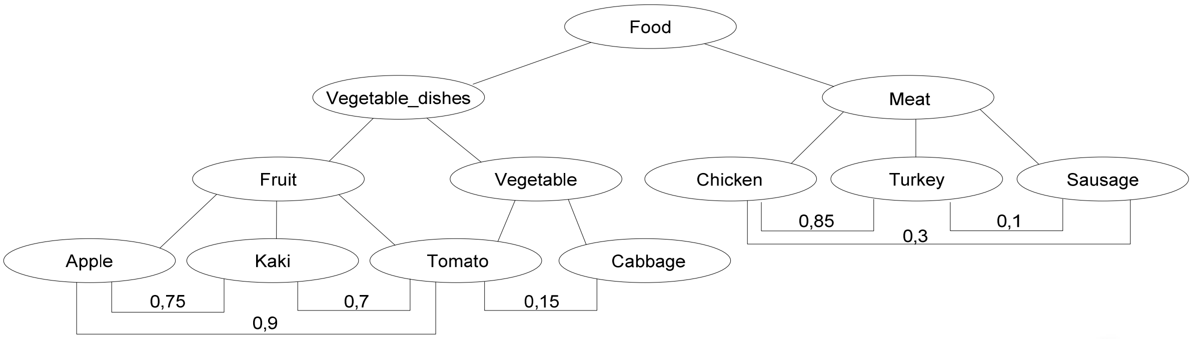

Figure 1: Example of Fuzzy Ontology with fuzzy similarity degrees between items.

step identified the following items: *Apple*, *Kaki*, *Tomato*, *Cabbage*, *Chicken*, *Turkey* and *Sausage.*

3.2 Identifying Similar Items

The similarity degree values between items are supplied by a fuzzy ontology, which specifies the semantics of the database contents. This step navigates through the fuzzy ontology structure to identify semantic similarity between items. If the similarity degree between items is greater than or equal to the *minsim* parameter explained in section 2, a semantic similarity association is found and this association is considered similar enough. A fuzzy association of size 2 is made by these pair of items found and are expressed by fuzzy items, where the symbol ~ indicates the similarity relation between items, for example, *Kaki~Tomato*.

After that, this step verifies the presence of similarity cycles as proposed in (Escovar, et al., 2005). The similarity cycles involve only sibling items. The minimum size of a cycle is 3, and the maximum is the number of descendents that the ancestor of its items has.

3.3 Generating Candidates

The generation of candidates in this algorithm is similar to Apriori algorithm. However, in NARFO* algorithm, besides the items identified in the step described in section 3.1, fuzzy items, which represent fuzzy associations, are added to generated itemsets candidates.

At the end of this step, the algorithm has all itemsets candidates of size *k* that are sent to calculate the weight of itemsets candidates.

3.4 Calculating the Weight of Candidates

If a candidate itemset is fuzzy, the weight of candidates is calculated based on the *Fuzzy weight equation* proposed in (Escovar, et al., 2005). This equation was created due to the presence of fuzzy logic concepts.

The weight reflects the number of occurrences of an itemset in the database. In this step, the database is scanned, and each of its rows is confronted with the set of candidate itemsets, one after one. For each occurrence of a non-fuzzy candidate itemset in a row, its weight is incremented by 1. Otherwise, it is incremented by the fuzzy value, obtained by the *Fuzzy weight equation*.

Finished this step, the itemsets candidates are ready to be used in next step.

3.5 Evaluating Candidates

In this step, the support of each itemset is evaluated. If the item set is fuzzy, the algorithm divides the *fuzzy weight* to the total of transactions. If a candidate itemset is frequent, its support is greater than or equal *minsup*.

However, during the process, besides verifying if each candidate itemset is frequent or not, the algorithm also verifies the non-frequent itemsets, as implemented on NARFO algorithm.

In the end of this step, we have all frequents itemsets of size *k*, where *k* is the size of the itemset. The algorithm return to the step explained in section 3.3 to generate the candidate itemsets of size *k + 1*. If *k* is equal to the number of domains, NARFO* algorithm goes to step 3.6.

3.6 Generating Rules

In this step, the association rules are generated for each itemset of the set of frequent itemsets. All possibilities of antecedents and consequents are generated for each itemset of the set of frequent itemsets. The rules containing the confidence value greater than or equal *minconf* are considered strong.

Then, the algorithm verifies these rules in order to check if a fuzzy item can be generalized. A fuzzy item can be generalized if all descendents of an ancestor are contained in the fuzzy item. If the rule *Tomato~Cabbage* → *Turkey* exists, then the rule *Vegetable* → *Turkey* is generated, since *Vegetable* comprises *Tomato* and *Cabbage* as descendant concepts.

3.7 Rules Generalization Treatment

After all rules have been generated by the previous step, they receive generalizing and redundancy treatment. The generalization here is different from NARFO algorithm in resulting of *minGen* introduction. The pseudo-algorithm of these treatments is showed in table 1.

Table 1: Pseudo-algorithm of rules generalization treatment.

```
1  for each rule generated
2    if rule can be generalized considering minGen
3  //generalization on both antecedent and/or consequent
4       generalize rule
5    end if
6    add rule to v // v is a vector of rules
7  end for
8  end for
```

For all rules, the algorithm verifies if the antecedent / consequent of each rule can be generalized (lines 1-8 in table 1) considering *minGen* parameter. For example, if the algorithm generated the rules *Kaki* → *Sausage* and *Apple~Tomato* → *Sausage*, then the algorithm does the generalization to *Fruit* → *Sausage* because all descendants of *Fruit* occurred in rules (including fuzzy items) with the *Sausage* concept as consequent. The generalization process happens not only if all descendents of an ancestor are in a fuzzy item, but also if all descendents of an ancestor are in different rules that have the same correspondent antecedent or consequent. This can happen at the antecedent or consequent of a rule.

Besides that, NARFO* algorithm included the *minGen* parameter on the previous algorithm. *MinGen* performs on generalizing rules, especially the ones with low minimum support. It generalizes rules if a minimal percentage of descendents of an ancestor appears in different rules containing the same antecedents and consequents, except by the descendent. For example, consider *minGen* value 0,6 (60%). Suppose the rule *Apple~Tomato* → *Sausage* was generated. This rule contains two items of the ancestor *Fruit* and, as they are sufficiently similar, they can be considered the same. In this way, 66.67% of *Fruit's* descendents, approximately, are included. So, the rule is generalized to *Fruit (-Kaki)*→ *Sausage*. *Kaki* was showed between parentheses, after a minus symbol, to indicate that this item do not make part of the generalization.

Consequently, the introduction of *minGen* reduces the amount of rules, generating semantically richer rules, without erroneous information.

3.8 Redundancy Treatment

In the redundancy treatment, the algorithm deals with two redundancy issues. The pseudo-algorithm for this treatment is described in table 2. The first kind of redundancy is when a rule is sub-rule of another one (lines from 1 to 8 in table 2). A sub-rule r_1 is a rule that has the same items in antecedent and consequent considering another rule r_2, except that r_1 has at least one item that is descendent of an item in r_2, in the same side of the rule (antecedent or consequent). Then, r_1 is eliminated.

Fuzzy redundancy is also treated by NARFO* (lines 9-15 in table 2). This kind of redundancy occurs when a rule is *fuzzy sub-rule* of another rule that contain a fuzzy item in the same side, in r_2.

Table2: Pseudo-algorithm of redundancy treatment.

```
1  Verify = false / verifies sub-rules
2  for each rule from v // vector of table 3
3    verify = sub-rule(rule, v)
4    //verify if rule is a sub-rule in v
5    if verify == false
6      add rule to v1 // another vector of rules
7    end if
8  end for
9  for each rule from v1
10   verify = fuzzy_sub-rule(rule, v1)
11   //verify if rule is a sub-rule in v1
12   if verify == false
13     add rule to v2 // another vector of rules
14   end if
15 end for
```

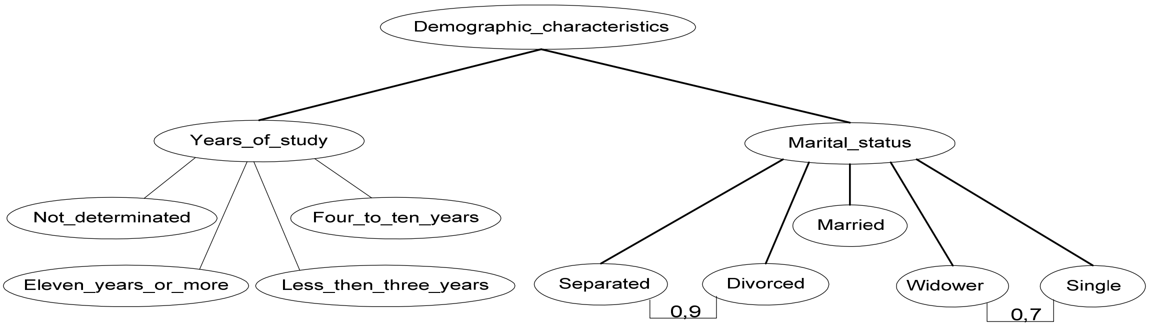

Figure 2: Fuzzy Ontology of IBGE2.

4 EXPERIMENTS

This section shows some results of NARFO* algorithm and compares NARFO* to XSSDM and NARFO algorithms.

In these experiments, data from the Brazilian Demographic Census 2000, provided by IBGE (Brazilian Institute of Geography and Statistics), were used. This database contains information about demographic characteristics of the Brazilian population, *Years of study* and *Marital status* information, called IBGE1.

The tests were done considering the fuzzy ontology IBGE1, which is illustrated in figure 2.

The tests were done with constant *minconf* and *minsim*, whose values are, respectively, 0.2 and 0.6. The *minsup* value varies from 0.05 to 0.5, increasing it by 0.05. *MinGen* value is 0.6 (60%).

Figure 3 presents a comparative between the generated rules extracted by NARFO* and XSSDM algorithms. Figure 3 shows that the number of rules reduced approximately 68,75%. This is the result of rules generalization and redundancy treatment.

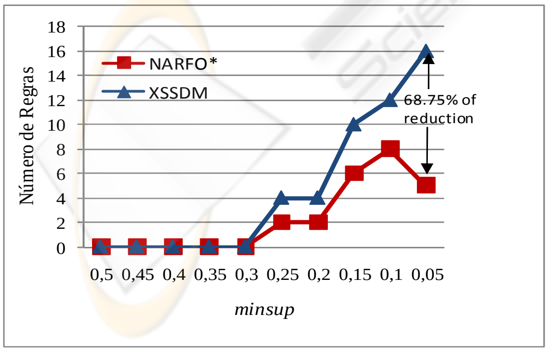

Figure 3: Comparison between NARFO* and XSSDM using IBGE2 database.

If only NARFO* and NARFO algorithm is considered, the reduction is 55,54%. The results are illustrated on figure 4. Note that between 0,1 and 0,05 *minsup* values, NARFO* decreases the amount of rules. This is an effect of the generalization process through *minGen* parameter. NARFO generates rules like *Less_then_three_years* → *Single*, *Four_to_ten_years* → *Single* and *Eleven-years_or_more* → *Single* but does not generalize them. Instead, NARFO* generalizes these rules to *Years_of_study* → *Single*, once the antecedents of those rules are more than 60% of *Years_of_study* descendents. In this way, not only NARFO* reduces the amount of rules but generates more semantic association rules.

Both algorithms generates the same number of rules until *minsup* value 0,15. After that, *minGen* parameter has much effect, generalizing rules that were not generalized before. The first two rules are generalized to *Less_then_three_years* → *Marital_status(-Married)* on NARFO*.

By the results, it could be observed that not only NARFO* reduces the number of

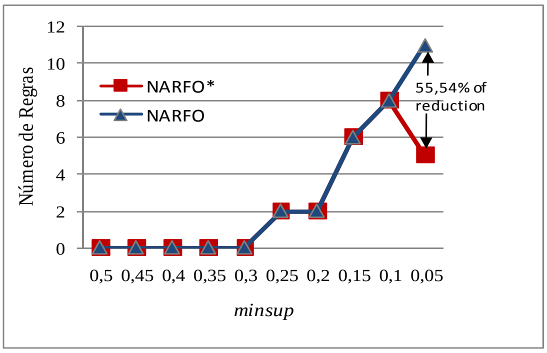

Figure 4: Comparison between NARFO* and NARFO using IBGE1 database.

non-redundant and generalized association rules, but also generates semantically richer rules., without erroneous information.

5 CONCLUSIONS AND FUTURE WORK

This paper proposed NARFO* algorithm, an algorithm for mining non-redundant and semantic generalized association rules based on fuzzy ontologies. The main contribution of this work is the introduction of *minGen* parameter in the latest version of NARFO algorithm.

The experiments proved that NARFO* generates semantically richer rules, with no erroneous information, reducing the amount of rules in 67,05% and 55,54% in comparison to XSSDM and NARFO algorithms, respectively.

It is being developed a method to automatize *minsup* and *minconf* values, without the need of final user know what these parameters means. A fuzzy membership degree, which considers how much an ancestor belongs to one or more ancestors, is also being added to NARFO*, so the generalization process can be improved.

ACKNOWLEDGEMENTS

IBM, UFSCar, CAPES, FAPESP and CNPq.

REFERENCES

Bayardo, J. R. J. (1998). Efficiently mining long patterns from databases. In *Proceedings of the 1998 Annual Conference on Management of Data*. Seattle, USA. 2-4 June 1998.

Brisson, L., Collard, M., Pasquier, N. (2005). Improving Knowledge Discovery Process Using Ontologies. *In International Workshop on Mining Complex Data*. Houston, USA. 27-30 November 2005.

Chen, G., Wei, Q., Kerre, E. E. (2000). Fuzzy Data Mining: Discovery of Fuzzy Generalized Association Rules. In: G. Bordogna and G. Pasi, eds. *Recent Issues on Fuzzy Databases*. Wurzburg: Physica-Verlag. 45–66.

Chen, X., Zhou, X., Scherl, R. B., Geller, J. (2003). Using an Interesting Ontology for Improved Support in Rule Mining. *In 5th InternationalConference on Data Warehousing and Knowledge Discovery*. Prague, Czech Republic. 3-5 September 2003.

Escovar, E. L. G., Biajiz, M., Vieira, M. T. P. (2005). SSDM: A Semantically Similar Data Mining Algorithm. *In 20th Brazilian Symposium of Databases*. Uberlândia, Brazil. 3-7 October 2005.

Escovar, E. L. G., Yaguinuma, C. A., Biajiz, M. Using Fuzzy Ontologies to Extend Semantically Similar Data Mining. *In 21th Brazilian Symposium of Databases*. Florianópolis, Brazil. 16-20 October 2006.

Farzanyar, Z., Kangavari, M., Hashemi, S. (2006). A New Algorithm for Mining Fuzzy Association Rules in the Large Databases Based on Ontology. *In Sixth IEEE International Conference on Data Mining – Workshops*. Hong Kong, China. 18-22 December 2006.

Han, J., Fu, Y. (1999). Mining Multiple-Level Association Rules in Large Databases. *IEEE Transactions on Knowledge and Data Engeneering*, 11(5), 798-805.

Hou, X., Gu, J., Shen, X., Yan, W. (2005). Application of Data Mining in Fault Diagnosis Based on Ontology. *In 3th Third International Conference on Information Technology and Applications*. Sydney, Australia. 4-7 July 2005.

Kunkle, D. Zhang, D. H., Cooperman, G. (2008). Mining Generalized Frequent Itemsets and Generalized Association Rules Without Redundancy. *Journal of Computer Science and Technology*, 23(1), 77-102.

Miani, R. G., Yaguinuma, C. A., Santos, M. T. P., Biajiz, M. (2009). NARFO Algorithm: Mining Non-redundant and Generalized Association Rules Based on Fuzzy Ontologies. *In 11 International Conference on Enterprise Information Systems*. Milan, Italy. 6-10 May 2009.

Oliveira, V. C., Rezende, S. O., Castro, M. (2007). Evaluating Generalized Association Rules Through Objective Measures. *In Proceedings of 25th International Multi-Conference on Artificial Intelligence and Applications* . Innsbruck, Austria. 12-14 February 2007.

Srikant, R., Agrawal, R., (1995). Mining Generalized Association Rules. *In Proceedings of the International Conference of Very Large Data Bases*. Zurich, Suíça. 11-15 September 1995.