Raising, to Enhance Rule Mining in Web Marketing with the Use of an Ontology

Xuan Zhou, James Geller
Department of Computer Science, New Jersey Institute of Technology, Newark, NJ 07104

1. Introduction

Marketing has faced new challenges over the past decade. The days of the mass market are definitely over. Consumers now are exposed to numerous cable channels and satellite channels. Many people do not get their information from TV at all, but use Web sites. The population has also developed. Minorities have grown and asserted their own tastes and needs. A product that is attractive to the average white Anglo-Saxon or Italian citizen might be completely uninteresting to a first generation South American immigrant. Similarly, the market has split up by preferences. Chinese and Indian food have made major inroads and many consumers would like to cook the same food in their homes. In short, the mass market is dead, and marketers today face the problem of advertising to many disjoint niche markets.

With the increase in available, cheap data storage, companies are keeping terabytes of information about their customers. Today it is not outrageous anymore to talk about one-to-one marketing. However, marketers face two problems. They may have information about previous customers, but how could they get personal information about potential customers? Secondly, if information about individuals is truly not accessible, how could they classify such individuals into small categories and then market effectively to these small categories?

To provide a solution for these two problems, the Web Marketing Project (Scherl & Geller, 2002; Geller, Scherl, & Perl, 2002) was created. This project targeted millions of publicly accessible home pages on the Web, on which people freely express their likes and dislikes. These pages are a valuable source of data for marketing purposes. One approach is to use the contact information for direct (email) marketing. For example, if someone expressed his interest as music, then he might be a potential customer of music CDs. Thus the marketing can be directed towards a very narrow niche. If someone lists very detailed interests, such as *The Simpsons*, the Season 8 DVD coming this August could be one of his must-buy products.

A second important use of this data is for finding interesting marketing knowledge. The data may be mined for useful correlations between interests and also between interests and demographic categories. If someone is interested in *The Simpsons*, what is the likelihood that he is interested in another comedy? What age groups are interested in particular types of TV series? The available data can be used for such investigations. The results may again be useful for marketing.

In the Web Marketing Project, we collected people's demographic and interest information from home pages and stored them in a database. There are six modules in this project, which are Web Search, Glossary-Based Extraction, Database, Data Mining, Ontology, and Front End, as described in detail in (Zhou, 2006). In this chapter, we only focus on the Ontology

and Data Mining modules. The ontology consists of two taxonomies, one of which describes different customer classifications, while the other one contains a large hierarchy, based on Yahoo, which contains 31,534 interests. For the customer classification, an intersection ontology (Zhou, Geller, Perl, & Halper, 2006) was developed.

The data mining module uses well-known data mining algorithms to extract association rules from the given data. The WEKA (Witten & Frank, 2000) package was used at the beginning of the project. From the WEKA package, the Apriori algorithm (Agrawal & Srikant, 1994) for data mining was used. The real world data about real people tends to produce rules with unsatisfactory support values. Thus, in this research a method was developed for improving the support values of rules by using the existing ontology. This method is called "*Raising*" and will be discussed in depth in Section 3. Moreover, due to the limitations of WEKA found during the project, the FP-Growth algorithm (Han, Pei, & Yin, 2000; Han, Pei, Yin, & Mao, 2004) was implemented and used in the second stage to correct some errors and improve the results.

Section 2 presents previous literature on ontologies used in rule mining. In Section 3, we introduce the Raising method and show how an ontology can be used to improve the support of mined rules. The effects caused by Raising on derived rules are discussed in Section 4. Future trends and conclusions are presented in Sections 5 and 6, respectively.

2. Background

A concept hierarchy is present in many databases either explicitly or implicitly. Some previous work utilizes a hierarchy for data mining. Han (1995) discusses data mining at multiple concept levels. His approach is to use associations at one level (e.g., milk → bread) to direct the search for associations at a different level (e.g., milk of brand X → bread of brand Y). As most of our data mining involves only one interest, the problem setting in this research is quite different. Han and Fu (1995) introduce a top-down progressive deepening method for mining multiple-level association rules. They utilize the hierarchy to collect large item sets at different concept levels. The approach in this research utilizes an interest ontology to improve support in rule mining by means of concept raising. To the best of our knowledge, the implementation of ontologies with association rule mining for the purpose of finding generalized rules with high support from sparse data has not appeared in the literature before our publication (Chen, Zhou, Scherl, & Geller 2003).

Fortin and Liu (1996) use an object-oriented representation for data mining. Their interest is in deriving multi-level association rules. As only one data item in each tuple is typically used for Raising, the possibility of multi-level rules does not arise in our problem setting. Srikant and Agrawal (1995) present Cumulative and EstMerge algorithms to find associations between items at any level by adding all ancestors of each item to the transaction. In this research, items of different levels are added to candidates during the mining. Psaila and Lanzi (2000) describe a method how to improve association rule mining by using a generalization hierarchy in data warehouses. Páircéir, McClean, and Scotney (2000) also differ from the work of this research in that they are mining multi-level rules that associate items spanning several levels of a concept hierarchy. Data mining has been viewed as an operation with a query language in (Novacek, 1998; Elfeky, Saad, & Fouad, 2000).

Zaki and Hsiao (2002) present a method that greatly reduces the number of redundant rules generated by previous rule miners. They define closed frequent item sets, which are sufficient for rule generation, to replace traditional frequent item sets. They show that this may lead to a reduction of the frequent item sets by two orders of magnitude, for a given support value. The concern is not with the efficiency of generating association rules, but with the total support of the resulting rules. However, *any rule mining algorithm* may be plugged into the Web marketing Project, as mining and raising are performed in a modular way. Thus, the Web Marketing Project would benefit from the improved efficiency of a data mining algorithm such as CHARM (Zaki & Hsiao, 2002). Mannila, Toivonen, & Verkamo (1994) worked on improving algorithms for finding associations rules, by eliminating unnecessary candidate rules.

Berzal, Blanco, Sanchez, and Vila (2001) have worked on a problem that is in some sense the diametrical opposite of the problem in this research. They are trying to eliminate misleading rules which are the result of too high support values. The problem of generating association rules when the available support is too low to derive practically useful rules is being addressed here.

This work is similar to (Z. Zhou, Liu, Li, & Chua, 2001) in that it incorporates prior knowledge into the rule mining process. Like Z. Zhou et al., a directed acyclic graph structure is used to present such additional knowledge. However, this research is not using the numeric (probabilistic) dependencies of (Z. Zhou et al., 2001).

Tsujii and Ananiadou (2005) compared the effect on text mining of using a thesaurus versus a logical ontology and argued that a thesaurus is more useful for text mining applications than formal ontologies. On the contrary, Missikoff, Velardi, and Fabriani (2003) did not focus on how to use an ontology for mining but how to build one by mining. *SymOntos*, an ontology management system, and the text mining approach were discussed to support ontology construction or updating.

Mädche and Volz (2001) have combined ontologies with association rules, but in a completely different way than what is done in this research. Their purpose is to semi-automatically construct ontologies. They are using an association rule miner in the service of this activity.

Li and Zhong (2006) have introduced an approach to automatically discover ontologies from data sets in order to build complete concept models for Web user information needs. They proposed a method for capturing evolving patterns to refine discovered ontologies and established a process to assess relevance of patterns in an ontology by the dimensions of *exhaustivity* and *specificity*.

Our approach is similar to (Xodo & Nigro 2005), in that we are interested in potential customers as opposed to previous customers. However, their domain is tourism.

3. Raising

Data mining has become an important research tool for the purpose of marketing. It makes it possible to draw far-reaching conclusions from existing customer databases about connections between different products commonly purchased. If demographic data are available, data

mining also allows the generation of rules that connect them with products. However, companies are not only interested in the behavior of their existing customer. They would also like to find out about potential future customers. Typically, there is no information about potential customers available in a company database, which can be used for data mining. However, it is possible to perform data mining on potential customers when people express their interests freely and explicitly on their home pages and there is a close relationship between specific interests and potential purchases.

Applying well-known data mining algorithms to the data extracted from the Web and stored in the database, association rules representing marketing knowledge are derived in this research for marketing purposes. However, when mining this data for association rules, what is available is often too sparse to produce rules with reasonable support values. Thus, when we initially derived rules by data mining, we found some rules which were interesting but with fairly low support values. In other words, those rules were not representative enough to predict future purchases. Since an interest ontology was created in the project, taking advantage of the ontology hierarchy provided a path to solve this problem.

3.1 Using Raising to Improve Support

The Raising method has been introduced in (Chen et al., 2003). A formal definition of **raising a tuple to its parents** and **raising a tuple to a level *k*** was given in (Geller, Zhou, Prathipati, Kanigiluppai, & Chen, 2005). However, in (Geller et al., 2005), the definition was given based on the implementation using the WEKA package (Witten & Frank, 2000). In order to use WEKA, an ARFF (Attribute-Relation File Format) format input file is required. In an ARFF file, all the attributes together with all their possible values need to be listed before the data. Thus, once an attribute has a large number of values, such as the 31,534 interests in our case, the input file size becomes extremely large. Moreover, the ARFF format does not allow multiple values for an attribute, which happened in our database since people often express more than one interest. Due to the limitations of the ARFF format, every tuple only contains one interest (one value of the attribute *Interest*) in the ontology. Since people normally express more than one interest, such a representation brings about some spurious records after Raising, as described in (Geller et al., 2005). More details about the ARFF limitations and effects can be found in (Zhou, 2006). Realizing this, here we introduce the revised definition of **raising a tuple to a level *k*** to better represent the situation of multiple interests.

For convenience, it is assumed that every interest in the hierarchy is assigned a level L by a revised breadth-first search algorithm, LEVEL-BFS(), as described in (Zhou, 2006, pp. 24-25). Then the level function L(*T*) is defined to return this level as a number. Interests nearer to the root have lower level numbers. The root is by definition at level 0. Lower levels in the diagram correspond to higher level numbers. In a DAG, such as Figure 3.1, the concept *V* is at level 3 and has parents at two different levels, *X* at level 1 and *W* at level 2.

The major difference between the previous definitions in (Geller et al., 2005) and this revised definition is the format of the input and the output. The input of the previous definition only contains one single term (interest) due to the limitations of the ARFF format. In the new definition, the input contains a set of terms, which are used for multiple interests of one person. For the output, the previous definitions returned a set of tuples, which mistakenly added spurious people into the dataset after Raising. However, the new definition only returns one tuple for each person.

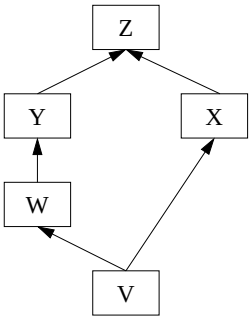

Figure 3.1: Example of parents at different levels.

Definition : An **operation R^k**, called **raising a tuple to the level *k***. Given is a data tuple $T = <N_s,D>$ where N_s is a set $\{N_1,N_2,..,N_n\}$ of interests. D stands for one or several items of demographic information. N_i is derived from a rooted ontology O. In O, each N_i has a uniquely determined, ordered sequence of m_i ($m_i \geq 1$) parents $< A^k_{i\_1}, A^k_{i\_2},.., A^k_{i\_mi} >$, all at level k. If N_i is at a higher level with a number less than k, it does not have an ancestor at level k. In this case, $A^k_{i\_1} = N_i$. Therefore, R^k is defined as the operation that takes T as input and returns the raised tuple

$$T^k = R^k(T) =< A^k_{1\_1}, A^k_{1\_2},.., A^k_{1\_m1}, A^k_{2\_1}, A^k_{2\_2},.., A^k_{2\_m2},.., A^k_{n\_1}, A^k_{n\_2},.., A^k_{n\_mn}, D >$$

as output for every T in O, except for the tuple <Root(O),D>. For the latter $R^k(T)$ is undefined. Moreover, a duplication check is performed during Raising. Thus, every $A^k_{i\_j}$ that appears in T^k is unique and all the duplicates have been removed. Also, as the result for every T in O, all the $A^k_{i\_j}$ in T^k is at a level with a number less than or equal to k.

Because N_i has m_i ancestors at level k, the result of Raising T, namely $R^k(T)$, is a new tuple with $\sum m_i + n_D$ terms, one for each ancestor at level k. The n_D is added for the items of demographic information D. The sum assumes a case with no duplicates. For the previous example (Figure 3.1),

$$R^2(<V,D>) =< W,D >, \text{ but } R^1(<V,D>) =< X,Y,D >$$

For example, if the given tuple T says that one male (M) in the age range 20-24 is interested in Jennifer Lopez and Richard Gere:

T = < LOPEZ_JENNIFER, GERE_RICHARD, M, (20-24)>

and the interest LOPEZ_JENNIFER has two ancestors at level 3, ACTRESS and SINGER while GERE_RICHARD has only ACTOR as a level 3 ancestor, then the result of Raising to level 3 is:

$R^3(T)$ = <ACTRESS, SINGER, ACTOR, M, 20-24>

meaning that one person of male gender in the age group (20-24) is interested in actress, singer, and actor. One issue arising at this point is whether we accept the generalization that has occurred. This was described in (Geller et al. 2005) and will also be discussed further in Section 3.3.

After having raised the data item $<N_s, D>$, any traditional association rule mining algorithm may be applied to the result of Raising, instead of applying it to the original data item $<N_s,D>$. Thus, Raising replaces a data item by its ancestors before performing rule mining.

3.2 An Example of Raising

Below is an example to illustrate how the Raising method works to improve the quality of derived association rules. The following dataset is the input to a data mining algorithm and is going to be raised to level 3. Each tuple, i.e. each line, stands for one interest instance with demographic information. As before, the values of three attributes (Age, Gender and Interest) are all included in each tuple. Age values are represented as a range while Gender values of male and female are abbreviated as M and F. Text after a double slash (//) is not part of the data. It contains explanatory remarks.

```
(20-29), M, BUSINESS_FINANCE                    //level=1
(40-49), M, METRICOM_INC                        //level=8
(50-59), M, BUSINESS_SCHOOLS                    //level=2
(30-39), F, ALUMNI                              //level=3
(20-29), M, MAKERS                              //level=4
(20-29), F, INDUSTRY_ASSOCIATIONS               //level=2
(30-39), M, AOL_INSTANT_MESSENGER               //level=6
(30-39), F, INTRACOMPANY_GROUPS                 //level=3
```

Once this original dataset is fed into a data mining algorithm, the best association rules, as measured by support value, that can be derived are {(20-29)}→{M} and {(30-39)}→{F}. Both rules have confidence values of 0.67 and support values of 2. This is also an example of sparse data. Every interest only appears once in the dataset. Though rules with a confidence of 1.0 can be derived from the data, such as {(50-59), M}→{BUSINESS_SCHOOLS}, as discussed before, the low support value of 1 does not make the rules useful for marketing purposes.

Table 3.1: Relevant ancestors

Interest Name	Its ancestor(s) at Level 3
METRICOM_INC	COMPUTERS, COMMUNICATIONS_AND_NETWORKING
MAKERS	ELECTRONICS
AOL_INSTANT_MESSENGER	COMPUTERS, INSTANT_MESSAGING

While performing Raising, ancestors are found at level 3 of the interests in the data. Table 3.1 shows all ancestors of the interests from levels below 3 such that the ancestors are at level 3. Table 3.1 is based on the DAG hierarchy in Figure 3.2. As seen in Figure 3.2, among the eight interests in eight tuples, two of them (ALUMNI, INTRACOMPANY_GROUPS) are already at level 3, and three of them (BUSINESS_FINANCE, BUSINESS_SCHOOLS, INDUSTRY_ASSOCIATIONS) are at levels above. Therefore, the Raising process will only function on the other three interests (METRICOM_INC, MAKERS, AOL_INSTANT_MESSENGER) which are at levels 4, 6 and 8 respectively. Table 3.1 lists their ancestors at level 3. The interest ontology is not a tree but a DAG. Some interests have more than one parent and thus more than one ancestor at a certain level. While the interest MAKERS has only one ancestor (parent) at level 3, the other two interests METRICOM_INC and AOL_INSTANT_MESSENGER both have two ancestors at level 3. The Raising to level

3 then replaces all the interests below level 3 in the original dataset by their ancestors at level 3. By doing so, all the interests in the new dataset are at level 3 or above. Thus, the new dataset after being raised to level 3 becomes:

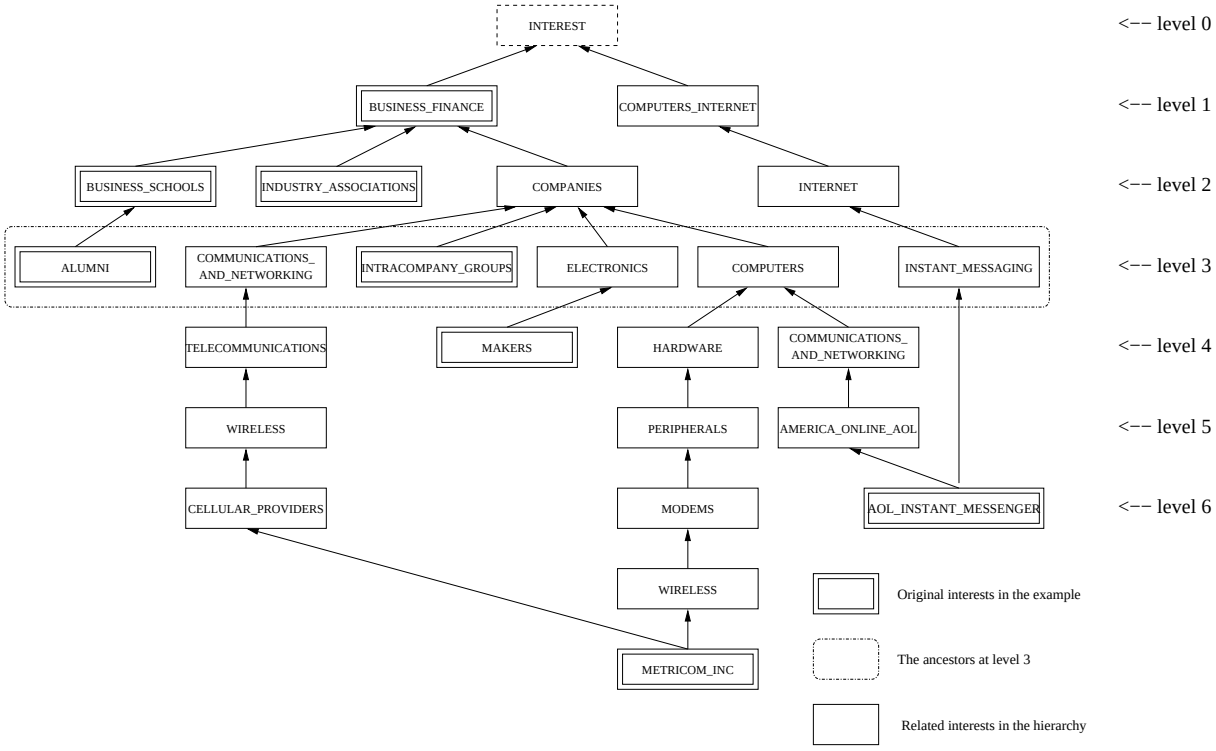

Figure 3.2: An example of Raising to level 3.

(20-29), M, BUSINESS_FINANCE
(40-49), M, COMMUNICATIONS_AND_NETWORKING, COMPUTERS
(50-59), M, BUSINESS_SCHOOLS
(30-39), F, ALUMNI
(20-29), M, ELECTRONICS
(20-29), F, INDUSTRY_ASSOCIATIONS
(30-39), M, INSTANT_MESSAGING, COMPUTERS
(30-39), F, INTRACOMPANY_GROUPS

By feeding the new dataset as input to the data mining algorithm, a new association rule, with a support value greater than 1, is derived other than the two demographic rules derived before:

{M}→{COMPUTERS} Confidence: 0.5 Support: 2

The overall Raising procedure is performed in the following steps (Zhou, 2006, pp. 74-75):

- Data Preparation
- Raising Operation
- Rule Generation
- Result Analysis

3.3 Effects of Generalization

Raising does lose detail and specificity during the process by replacing interests by their ancestors. This is a fact that has positive and negative consequences. A disadvantage would be the missing details due to replacing interests by their ancestors. Those details could have been

used as a direct business act indicators about a product. Thus, a rule which involves deep-level interests can explicitly express a connection between customers and products. Thus, such a rule might connect a specific movie DVD with a demographic group. It is, of course, possible to perform data mining before Raising. Thus, no real loss occurs. However, if a confidence and a support threshold are given, any rule has to *qualify* to appear in the results. That is, a rule must have a greater confidence value and a greater support value than the threshold to qualify. Once a rule *qualifies*, it will appear in the results no matter whether Raising is used later or not. Many rules that are mined before Raising tend to have low support values. Thus these rules would not show up anyway. Thus, no new loss is introduced due to Raising. If a rule is not *qualified*, it does not meet the expectations of a *useful* rule. Therefore, to discard such a rule of *little use* and to lose those details is reasonable.

On the other hand, the generalization has the advantage that it provides better indicators for new product promotions. A new product would never appear in any existing rule, thus no exact match can be found. However, it is not a hard problem to categorize the new product into an existing category, or a higher level interest. For example, the 2005 TV comedy "American Dad" had not been listed in Yahoo at the time of this research, i.e. no rule can be found for it by mining. If the FOX TV network would like to attract a potential audience for the new show, the rules involving the interest "television comedy" would be a nice option to consider. Thus, Raising can help to generalize the information from specific interests such as "The Simpsons" or "Family Guy" (two other TV comedies) to "television comedy," if such a rule is not there before Raising. Even if this rule exists before Raising, the new increased support value after Raising would bring about a better rule quality.

As a conclusion, the Raising operation is not meant to replace the existing mining result rules. Instead, Raising is used to strengthen the derived rules and to provide more possible rules by generalizing detail information.

3.4 Results of Raising

The Raising method has been implemented using the new definition, as described in Section 3.1. This implementation avoids the problems caused by spreading out interests of one person over several lines in the ARFF format, applying the previous definitions, and makes it possible to derive better rules. An FP-Growth mining program was implemented in this research, using JAVA. The input specifies a file which has to be formatted as required for *set-based input mining*, i.e., a person's record is in a single line, including age, gender and all his/her interests. For example, some lines in the input file after Raising at level 6 in the category "BUSINESS & FINANCE" are as follows:

B,FEMALE,1600840341,1602248651,
C,MALE,1600000236,1600001222,1600909559,1600909561,
C,MALE,1600840352,1602216980,1602216981,1602219139,1602236895,
D,MALE,1600000393,1600001278,1600001779,1600193918,

The letters A to F are used to represent age ranges. For example, B stands for an age range from 20 to 29 and D stands for a range of 40-49. A 10-digit number is a unique Yahoo ID for an interest in our ontology. For example, the Yahoo ID 1600840341 stands for the interest TUPPERWARE.

The increments of support values after Raising are shown at the left side in Table 3.2. The increment percentage I_i is computed as the difference between the average support values of

mining results from raised data at level i (S_{ai}) and from unraised data (S_b), at the lowest level appearing in this data, and then divided by S_b. Thus,

$$I_i = \frac{S_{ai} - S_b}{S_b} \times 100\% \quad (3.1)$$

and I_i is the increment rate of the support value at level i relative to the original value before Raising. Since result level 1 only contains one interest and results from levels below level 5 only have sparse data or even do not exist, only level 2 to level 5 appear. The data show the concrete increments of support values from lower levels to higher levels. The right side of Table 3.2 shows the number of newly discovered {Interest}→{Interest} rules, which we could not derive using the WEKA implementation. For example, some rules mined from the input file raised to level 6 in category "BUSINESS & FINANCE" are as follows:

{INTERNET_MARKETING_AND_ADVERTISING}→{INTERNET_BUSINESS}
{HOME_BUSINESS, INTERNET_MARKETING_AND_ADVERTISING}→{INTERNET_BUSINESS}
{STARTUPS}→{INTERNET_BUSINESS}
{NETWORK_MARKETING}→{INTERNET_BUSINESS}
{INTERNET_BUSINESS, HOME_BUSINESS}→{INTERNET_MARKETING_AND_ADVERTISING}
{HOME_BUSINESS, SMALL_BUSINESS}→{INTERNET_BUSINESS}

Table 3.2 Support value increments and new rule discovery

Category	Level 2	Level 3	Level 4	Level 5	Interest-Interest Rules
BUSINESS FINANCE	858.79%	371.90%	74.02%	5.60%	115
COMPUTERS INTERNET	946.25%	749.90%	97.66%	3.53%	26
FAMILY HOME	341.41%	146.17%	46.16%	0.15%	6
GOVERNMENT POLITICS	4084.36%	2320.00%	2090.90%	1119.50%	169
RECREATION SPORTS	853.49%	251.86%	64.35%	11.67%	2
SCHOOLS EDUCATION	877.91%	459.82%	249.03%	20.72%	23
SCIENCE	1661.34%	971.87%	894.58%	751.98%	13
Data mined at Confidence≥0.6, Support≥0.02					

This Raising implementation solves the problems caused by the WEKA ARFF implementation. It results in better performance while still improving support values over the previously published Raising method (Chen et al., 2003; Geller et al., 2005). It also eliminates in a natural way the duplication of tuples, which might occur during Raising when an ancestor is reachable by more than one path in a DAG. The application of the FP-Growth mining algorithm results in better efficiency than the previously used Apriori algorithm.

4. Effects on Mining Results by Raising

4.1 Observations about Raising Results

To derive association rules, data are selected from the database and fed into data mining algorithms. In the Web Marketing Project, both the Apriori algorithm (Agrawal, Imielinski, & Swami, 1993, Agrawal & Srikant, 1994) from the WEKA package (Witten et al., 2000) and the FP-Growth algorithm (Han et al., 2000, Han et al., 2004) were used. By providing minimum support and confidence values as input parameters, the data mining algorithms return derived association rules based on the input. However, as we described in (Chen et al., 2003), a problem in the derivation of association rules is that available data is sometimes very sparse and biased. For example, among over a million of interest records in the database of

this research, only 11 people expressed an interest in RECREATION_SPORTS, and nobody expressed an interest in SCIENCE which is counter-intuitive.

Recall the Raising to level 3 example in Section 3.2. The new rule "{M}→{COMPUTERS} Confidence: 0.5 Support: 2" is relatively more attractive for marketing purposes than the results from the original dataset, for the following reasons.

1. The new rule has a better support value, thus it is a rule of higher quality. The occurrence count of an interest at the raised level in the dataset is increased by replacing its descendants in all instances. During the replacement, the demographic information and the interests at levels above are not affected or updated. Thus, while the number of tuples in the dataset is still the same, the support value is improved.

2. The new rule connects demographic information with interest information. The rules derived originally, such as:

   {20-29}→{M} Confidence: 0.67 Support: 2
   {30-39}→{F} Confidence: 0.67 Support: 2
   {F}→{30-39} Confidence: 0.67 Support: 2

   only imply some connections between age and gender. Though those rules are valid, they do not contribute any useful insights for marketing purposes. For marketing usage, an interest should be included in the rules to predict future purchases of potential customers.

3. Last but not least, "brand-new" rules can be derived after Raising. Note that in the original dataset, the interest COMPUTERS did not exist at all. This interest at level 3 is introduced when several interests in the original dataset share it as ancestor. In other words, the newly appeared interest is a generalization of its descendants based on the interest ontology. In the example at the beginning of this section, just because people did not express interests with more general terms does not mean they are not interested. On the contrary, people prefer to express their interests more specifically. In the data file, there are 62,734 data items in the category of RECREATION_SPORTS. These thousands of people prefer saying something like "I'm interested in BASKETBALL and FISHING" instead of saying "I'm interested in RECREATION_SPORTS." By the Raising method, those wide-spread data can be collected and thus new rules can be derived to describe the situation by using high-level interests.

4.2 Effects of Raising on Support and Confidence of Different Rule Types

Since the inputs only include three attributes for each person's record, all the association rules are combinations of those attributes in their antecedents and consequents. For the age and gender attributes, only one single value is allowed for each attribute in every tuple, i.e. one person's record. However, since a person might have expressed more than one interest at the same time, the interest attribute may have multiple values. Here all the possible rule types based on this situation are listed. (The expression {Interest(s)} stands for one or more interests. For example, the rule {Male} → {FISHING, POKER} is categorized by the rule type {Gender} → {Interest(s)}.) Rules with an empty antecedent or consequent are not interesting.

1. {Age} → {Gender}
2. {Age} → {Interest(s)}
3. {Age} → {Gender, Interest(s)}

4. {Age, Gender} → {Interest(s)}
5. {Age, Interest(s)} → {Gender}
6. {Age, Interest(s)} → {Interest(s)}
7. {Age, Interest(s)} → {Gender, Interest(s)}
8. {Age, Gender, Interest(s)} → {Interest(s)}
9. {Gender} → {Age}
10. {Gender} → {Interest(s)}
11. {Gender} → {Age, Interest(s)}
12. {Gender, Interest(s)} → {Age}
13. {Gender, Interest(s)} → {Interest(s)}
14. {Gender, Interest(s)} → {Age, Interest(s)}
15. {Interest(s)} → {Age}
16. {Interest(s)} → {Gender}
17. {Interest(s)} → {Interest(s)}
18. {Interest(s)} → {Age, Gender}
19. {Interest(s)} → {Age, Interest(s)}
20. {Interest(s)} → {Gender, Interest(s)}
21. {Interest(s)} → {Age, Gender, Interest(s)}

These 21 rule types include all the possibilities of derived association rules. By studying these rule types one by one in groups, all the rules which are not proper for marketing purposes are filtered out and the research focuses on the effect of Raising on the remaining rule types.

- **Group (A):** The rule types #1 and #9 only involve Age and Gender. As discussed before, such rules are not useful for marketing purposes and thus are filtered out.

- **Group (B):** The rules which are useful for marketing purposes should be those which connect a certain group of persons to a certain interest, or product. The types #2, #4 and #10 are exactly predicting the relationship between a group of persons and their interests. They tie people and their interests together. Moreover, type #4 is more specific than types #2 and #10. Once such rules with a high confidence and a high support value are found, the group of people described by the antecedent is more likely to make purchases related to the interest.

- **Group (C):** The types #15, #16 and #18 are the opposite of #2, #10 and #4. The attribute interest is in the antecedent while demographic attributes are in the consequent. The interpretation for such rule types is “If somebody is interested in A, this person is likely to be in a certain demographic group B.” These rules describe the distribution of person groups within all those who are interested in an interest A. The types #5 and #12 can also be categorized in this group since there is only the demographic attribute in the consequents. These rule types are less useful for promotion purposes which this project is focused on.

- **Group (D):** The types #3 and #11 have only demographic attributes in the antecedents. In the consequents are the combinations of a demographic attribute and interest attributes. For example, #3 can be interpreted as “If a person is in the age group B, there is a certain confidence that this person has a gender C and will be interested in the interest A.” A more specific example is “If a person is a teenager, there is a confidence of 0.8 that it's a girl who is interested in SOFTBALL.” By connecting the Age attribute and Gender

attribute in the rules, the interpretation of these two types of rules is confusing and they appear not suitable for marketing.

- **Group (E):** The types #6, #8, #13 and #17 have only the interest attribute in the consequents. The rule type #17 implies a connection between two or more interests. The types #6, #8 and #13 are more specific formats of type #17 by including demographic groups. These kinds of rules are attractive for marketing purposes. When a retailer is going to promote a product, which is categorized by interest X, he might prefer association rules which can lead to persons grouped by age, gender, etc. However, there might not be a rule (due to insufficient support or confidence values during rule mining) of rule type #4 from Group (B). Therefore, rules in this group would greatly support his search as a complement of Group (B).

- **Group (F):** The types #7, #14, #19, #20 and #21 also contain a combination of demographic attribute and interest attribute in consequents, like the rule types in Group (D). The difference to Group (D) is that interest attributes appear in the antecedent. Type #7 and #14 are more specific formats of type #20 and #19 respectively. The usefulness of these rules for marketing is doubtful. For example, a rule could be "If a man is interested in FOOTBALL, there is a certain confidence that his age is 30 to 39 and he is also interested in BEER." These rule types try to connect interest attributes to demographic attributes and are hybrids of Group (C) and Group (E). However, for marketing purposes, these rule types are weaker than those in Group (E).

All the 21 rule types have been categorized into six groups. The effects of Raising can be analyzed group by group. Before Raising, a rule has a support value of $S_0 = Occ_{ante\ \&\ con}$ and the confidence value is calculated by $C_0 = \frac{S_0}{Occ_{ante}}$.

- **Group (A):** After Raising, since the instances of age and gender have not been changed and no interests occur, the confidence and support values are not affected.

$$S_{new} = S_0 \tag{4.1}$$
$$C_{new} = C_0 \tag{4.2}$$

- **Group (B):** After Raising, the occurrences of demographic information in the antecedents are not changed. However, the occurrences of interests in the consequent might be increased by replacing descendants by multiple ancestors. If there is no replacement needed, the occurrences stay unchanged. Thus, the support values always stay unchanged or are increased and the confidence values also stay unchanged or are increased accordingly. Suppose the increment of occurrence is *Inc* ($Inc \geq 0$), then

$$S_{new} = S_0 + Inc \geq S_0 \tag{4.3}$$
$$C_{new} = \frac{S_0 + Inc}{Occ_{ante}} \geq C_0 \tag{4.4}$$

- **Group (C):** After Raising, the occurrences of interests in the antecedent are increased. Suppose the increment of occurrences of the antecedent is Inc_{ante} ($Inc_{ante} \geq 0$). However, among all the tuples updated with these interests, the demographic information might not match those in the rule, thus the increment of occurrence of both antecedent and consequent will be a different variable, $Inc_{ante\ \&\ con}$ ($Inc_{ante} \geq Inc_{ante\ \&\ con} \geq 0$). Therefore,

$$S_{new} = S_0 + Inc_{ante\ \&\ con} \geq S_0 \tag{4.5}$$

$$C_{new} = \frac{S_0 + Inc_{ante\ \&\ con}}{Occ_{ante} + Inc_{ante}} \tag{4.6}$$

The comparison of C_{new} to C_0 depends on the two increments. If $Inc_{ante\&con}$ is much smaller than Inc_{ante}, C_{new} could be less than C_0. Otherwise, $C_{new} \geq C_0$. More specifically,

$$\begin{aligned} C_{new} - C_0 &= \frac{S_0 + Inc_{ante\ \&\ con}}{Occ_{ante} + Inc_{ante}} - \frac{S_0}{Occ_{ante}} \\ &= \frac{(S_0 + Inc_{ante\ \&\ con})Occ_{ante} - S_0(Occ_{ante} + Inc_{ante})}{Occ_{ante}(Occ_{ante} + Inc_{ante})} \\ &= \frac{Inc_{ante\ \&\ con}Occ_{ante} - S_0 Inc_{ante}}{Occ_{ante}(Occ_{ante} + Inc_{ante})} \\ &= \frac{\frac{Inc_{ante\ \&\ con}}{Inc_{ante}} - \frac{S_0}{Occ_{ante}}}{\frac{Occ_{ante}}{Inc_{ante}} + 1} \end{aligned} \tag{4.7}$$

Since all the values in the equation are non-negative, the value of the numerator in the Formula 4.1 shows which is greater, C_{new} or C_0. If the value is non-negative, C_{new} is greater than or equal to C_0. Otherwise, if the value is negative, C_{new} is less than C_0. Thus,

$$\frac{Inc_{ante\ \&\ con}}{Inc_{ante}} \geq \frac{S_0}{Occ_{ante}} \Rightarrow C_{new} \geq C_0 \tag{4.8}$$

$$\frac{Inc_{ante\ \&\ con}}{Inc_{ante}} < \frac{S_0}{Occ_{ante}} \Rightarrow C_{new} < C_0 \tag{4.9}$$

Notice that $C_0 = \frac{S_0}{Occ_{ante}}$. Let's take a look at the $\frac{Inc_{ante\ \&\ con}}{Inc_{ante}}$. The numerator is the increment of the records which contain all the terms in both antecedent and consequent. The denominator is the increment of the records which contain all the terms in the antecedent. Thus, for $C_{Inc} = \frac{Inc_{ante\ \&\ con}}{Inc_{ante}}$, C_{Inc} is the confidence value of the rule mined from the sub-dataset which contains all the updated records, *i.e.* records which have interests being replaced by their ancestors, during the Raising. In other words, the changes of confidence values before and after Raising are based on the confidence value from the sub-dataset which contains all the Raising-affected records.

Thus, if $C_{Inc} \geq C_0$ then $C_{new} \geq C_0$ and if $C_{Inc} < C_0$ then $C_{new} < C_0$.

- **Group (D):** As in Group (B), the occurrences of demographic information in the antecedents are not changed after Raising. The increment of occurrences of both antecedent and consequent $Inc_{ante\ \&\ con}$ ($Inc_{ante\ \&\ con} \geq 0$) depends on the increment of occurrence of interests.

$$S_{new} = S_0 + Inc_{ante\ \&\ con} \geq S_0 \tag{4.10}$$

$$C_{new} = \frac{S_0 + Inc_{ante\ \&\ con}}{Occ_{ante}} \geq C_0 \tag{4.11}$$

Therefore always: $S_{new} \geq S_0$ and $C_{new} \geq C_0$.

- **Group (E) and Group (F):** These two groups can be put together since the interest attributes appear in both antecedent and consequent. After Raising, the increment of occurrence of both antecedent and consequent is $Inc_{ante\&con}$ ($Inc_{ante\&con} \geq 0$). However, there is also an increment of occurrence of interests in the antecedent Inc_{ante} ($Inc_{ante} \geq Inc_{ante\&con}$). Unfortunately, these two increments $Inc_{ante\&con}$ and Inc_{ante} do not have any relationship to each other. Thus,

$$S_{new} = S_0 + Inc_{ante\ \&\ con} \geq S_0 \tag{4.12}$$

$$C_{new} = \frac{S_0 + Inc_{ante\ \&\ con}}{Occ_{ante} + Inc_{ante}} \tag{4.13}$$

  Note the formulas are exactly the same as Formulas 4.5 and 4.6 in **Group (C)**. Therefore, the same Formula 4.7 can also be applied to the relationship between the confidence values before and after Raising for **Group (E)** and **Group (F)**. Thus the changes are also based on the Raising-affected data.

  If $C_{Inc} \geq C_0$ then $C_{new} \geq C_0$ and if $C_{Inc} < C_0$ then $C_{new} < C_0$.

According to the case analysis above, after Raising, the *support values* of all the association rules are *never decreased*, thus Raising guarantees higher or equal quality rules. For *confidence values*, the most important rule types for marketing purposes, Group (B), always have higher confidence values. This ensures high quality association rules with better support values and also better confidence values. The rule types in Groups (A), (C) and (D) are not proper for marketing purposes. Those rules are filtered out from the data mining results in a postprocessing step. The rule types in Groups (F) and (E) have increasing support values but undetermined changes of confidence values. However, as discussed before, those rule types are only used as complements for Group (B).

5. Future Work

In the processing of Raising, there is a duplication check performed while replacing the interests by their corresponding ancestors at a specific level. Such a check eliminates one interest if it would appear in somebody's interest list more than once. However, though it is the case that people will not express the same interest twice in an interest list, it is still possible that two siblings are expressed at the same time. When Raising is performed to the level where the two siblings' parent is located, the new interest list for this level only contains the parent once. One concern arises here whether this parent should be counted twice in the list since two of its children were originally expressed. In other words, should the interests in the list be assigned a weight in such situations? In that case, when somebody expressed those two siblings, one might want to stress his multiple interests in the same category. Should this stress also be considered in the new list after Raising?

To expand this idea to the supermarket shopping cart example, a weight could be assigned according to the values of products. Thus, should a product for which customers paid a lot for be assigned a higher weight and function as more important in the mining than other products which cost less? A mining method with weighted items might be a solution. This kind of work has been done before. In 1998, Cai, Fu, Cheng, and Kwong introduced the MINWAL algorithm, which used a metric called *k-support bound* in the mining process. Wang, Yang, & Yu (2000) extended the traditional association rule problem by allowing a weight to be associated with each item in a transaction and then introducing the weight during the rule generation. Lu, Hu, & Li (2001) also presented an algorithm called MWAR to handle the problem of mining mixed weighted association rules. However, since all these approaches were based on the Apriori algorithm, we would like to include a weighted mining mechanism based on FP-Growth in future work.

The Raising method discussed here was applied to the domain of Web marketing. However, it is not limited to this single domain. The Raising method targets any input datasets, as long as the data items in the datasets use terms from an existing, well-structured ontology. The originality or the domain of the data are not important. By adapting the input formats, Raising can be applied to different domains with a corresponding ontology.

6. Conclusions

The Raising method uses an ontology to perform a preprocessing step on the input datasets before data mining. The preprocessed dataset can be used as input to any data mining algorithm to derive more and better association rules, with higher support values and higher confidence values. Again, we are not inventing a novel data mining algorithm but a preprocess to be applied to the input datasets of ANY data mining algorithm, once the input format is made compatible.

The Raising method takes advantage of the hierarchy structure of an ontology and collects instances at the lower levels of the hierarchy to enrich the derivation of the association rules which involves the ancestors of these instances. In our experiments, the support values of rule sets were greatly increased, up to 40 times. The effects of Raising on the confidence values were also analyzed based on each type of the possible derived rules. Though not all the confidence values for every rule type were increased during Raising, we found that the rule group (B), which is the most useful for marketing purposes, did have their confidence values increased by Raising. Thus Raising resulted in better rules with higher support and confidence values.

References

Agrawal, R., Imielinski, T., & Swami, A. N. (1993). Mining association rules between sets of items in large databases. *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data* (pp. 207-216). New York, NY: ACM Press.

Agrawal, R. & Srikant, R. (1994). Fast algorithms for mining association rules in large databases. *Proceedings of the 20th International Conference on Very Large Data Bases* (pp. 487-499). San Francisco, CA: Morgan Kaufmann.

Berzal, F., Blanco, I., Sanchez, D., & Vila, M.-A. (2001). Measuring the accuracy and interest of association rules: A new framework. *Intelligent Data Analysis*, 6(3), 221-235.

Cai, C. H., Fu, A. W., Cheng, C. H., and Kwong, W. W. (1998). Mining association rules with weighted items. *Proceedings of 1998 Internatinal Database Engineering and Applications Symposium* (pp. 68-77).

Chen, X., Zhou, X., Scherl, R., & Geller, J. (2003). Using an interest ontology for improved support in rule mining. *Proceedings of the 5th International Conference on Data Warehousing and Knowledge Discovery* ser. *Lecture Notes in Computer Science*, vol. 2738 (pp. 320-329). New York, NY: Springer Verlag.

Elfeky, M. G., Saad, A. A., & Fouad, S. A. (2001). ODMQL: Object Data Mining Query Language. *Proceedings of the 2000 International Symposium on Objects and Databases* (pp. 128-140). New York, NY: Springer Verlag.

Fortin, S. & Liu, L. An object-oriented approach to multi-level association rule mining. *Proceedings of the 5th International Conference on Information and Knowledge Management* (pp. 65-72). New York, NY: ACM Press.

Geller, J., Scherl, R., & Perl, Y. (2002). Mining the Web for target marketing information. *Proceedings of the Collaborative Electronic Commerce Technology and Research (CollECTeR) Workshop*. Toulouse, France.

Geller, J., Zhou, X., Prathipati, K., Kanigiluppai, S., & Chen, X. (2005). Raising data for improved support in rule mining: How to raise and how far to raise. Intelligent Data Analysis, 9(4), 397-415.

Han, J., Pei, J., & Yin, Y. (2000). Mining frequent patterns without candidate generation. *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data* (pp. 1-12). New York, NY: ACM Press.

Han, J. (1995). Mining knowledge at multiple concept levels. *Proceedings of the 4th International Conference on Information and Knowledge Management* (pp. 19-24). New York, NY: ACM Press.

Han, J. & Fu, Y. (1995). Discovery of multiple-level association rules from large databases. *Proceedings of the 21st International Conference on Very Large Data Bases* (pp. 420-431). Zurich, Swizerland.

Han, J., Pei, J., Yin, Y., & Mao, R. (2004). Mining frequent patterns without candidate generation: A frequent-pattern tree approach. *Data Mining and Knowledge Discovery*, 8(1), 53-87.

Li, Y. & Zhong, N. (2006) Mining Ontology for Automatically Acquiring Web User Information Needs. *IEEE Transactions on Knowledge and Data Engineering*, 18(4), 554-568.

Lu, S., Hu, H., & Li, F. (2001). Mining weighted association rules. *Intelligent Data Analysis*, 5(3), 211-225.

Mädche, A. & Volz, R. (2001). The ontology extraction and maintenance framework Text-To-Onto. *Proceedings of the ICDM'01 Workshop on Integrating Data Mining and Mnowledge Management*. San Jose, CA.

Mannila, H., Toivonen, H., & Verkamo, A. (1994). Improved methods for finding association rules. *Proceedings of the AAAI Workshop on Knowledge Discovery* (pp.181-192). Finland.

Missikoff, M., Velardi, P. & Fabriani, P. (2003). Text Mining Techniques to Automatically Enrich a Domain Ontology, *Applied Intelligence*, 18(3), 323-340.

Novacek, V. (1998). Data mining query language for object-oriented database. *Proceedings of the 2nd East European Symposium on Advances in Databases and Information Systems* (pp. 278-283). New York, NY: Springer Verlag.

Páircéir, R., McClean, S., & Scotney, B. (2000). Discovery of multi-level rules and exceptions from a distributed database. *Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 523-532). New York, NY: ACM Press.

Portscher, E., Geller, J., & Scherl, R. (2003). Using internet glossaries to determine interests from home pages. *Proceedings of the 4th International Conference on Electronic Commerce and Web Technologies* (pp. 248-258). Berlin: Springer Verlag.

Psaila, G. & Lanzi, P. L. (2000). Hierarchy-based mining of association rules in data warehouses. *Proceedings of the 2000 ACM Symposium on Applied Computing* (pp. 307-312). New York, NY: ACM Press.

Scherl, R. & Geller, J. (2002). Global communities, marketing and Web mining. *Journal of Doing Business Across Borders*, 1(2), 141-150.

Srikant, R. & Agrawal, R. (1995). Mining generalized association rules. *Proceedings of the 21st International Conference on Very Large Data Base* (pp. 407-419). Zurich, Swizerland.

Tsujii, J. & Ananiadou, S. (2005). Thesaurus or Logical Ontology, Which One Do We Need for Text Mining?, *Language Resources and Evaluation*, 39(1), 77-90.

Wang, W., Yang, J., & Yu, P. S. (2000). Efficient Mining of Weighted Association Rules (WAR). *Proceedings of the sixth ACM SIGKDD International Conference* (pp. 270-274). Boston, MA.

Web Marketing Project Web site. (2005). (http://web.njit.edu/challeng/webmarketing.html)

Witten, I. H. & Frank, E. (2000) *Data Mining, Practical Machine Learning Tools and Techniques with Java Implementations*. San Francisco, CA: Morgan Kaufmann.

Xodo, D. & Nigro, H. O. (2005) Knowledge Management in Tourism. In L. C. Rivero, J. H. Doorn, & V. E. Ferraggine (Eds.), *Encyclopedia of Database Technologies and Applications* (pp. 319-329). Hershey, PA: Idea Group Reference.

Zaki, M. J., & Hsiao, C. J. (2002). CHARM: An efficient algorithm for closed itemset mining.

In *Proceedings of the 2nd SIAM International Conference on Data Mining*. SIAM.

Zhou, X., Geller, J., Perl, Y., & Halper, M. (2006). An application intersection marketing ontology. *Theoretical computer science: Essays in memory of Shimon Even*. ser. *Lecture Notes in Computer Science,* vol. 3895 (pp. 143-153). Berlin: Springer-Verlag.

Zhou, X. (2006). *Enhancing Web Marketing by Using an Ontology*. Doctoral dissertation, New Jersey Institute of Technology, Newark, NJ.

Zhou, Z., Liu, H., Li, S. Z., & Chua, C. S. (2001). Rule mining with prior knowledge - a belief networks approach. *Intelligent Data Analysis*, 5(2), 95-110.