

Using an Interest Ontology for Improved Support in Rule Mining

Xiaoming Chen¹, Xuan Zhou¹, Richard Scherl², and James Geller^{1,*}

¹ CS Dept., New Jersey Institute of Technology, Newark, NJ 07102

² Monmouth University, West Long Branch, New Jersey 07764

Abstract. This paper describes the use of a concept hierarchy for improving the results of association rule mining. Given a large set of tuples with demographic information and personal interest information, association rules can be derived, that associate ages and gender with interests. However, there are two problems. Some data sets are too sparse for coming up with rules with high support. Secondly, some data sets with abstract interests do not represent the actual interests well. To overcome these problems, we are preprocessing the data tuples using an ontology of interests. Thus, interests within tuples that are very specific are replaced by more general interests retrieved from the interest ontology. This results in many more tuples at a more general level. Feeding those tuples to an association rule miner results in rules that have better support and that better represent the reality.³

1 Introduction

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 purchased. If demographic data are available, data mining also allows the generation of rules that connect them with products. However, companies are not just interested in the behavior of their existing customers, they would like to find out about potential customers. Typically, there is no information about potential customers available in a company database, that can be used for data mining.

It is possible to perform data mining on potential customers, if one makes the following two adjustments: (1) Instead of looking at products already purchased, we may look at interests of a customer. (2) Many people express their interests freely and explicitly on their Web home pages. The process of mining data of potential customers becomes a process of Web Mining. In this project, we are extracting raw data from home pages on the Web. In the second stage, we raise specific but sparse data to higher levels, to make it denser. In the third stage we apply traditional rule mining algorithms to the data.

When mining real data, what is available is often too sparse to produce rules with reasonable support. In this paper we are describing a method how to

* This research was supported by the NJ Commission for Science and Technology

³ Contact Author: James Geller, geller@njit.edu

improve the support of mined rules by using a large ontology of interests that are related to the extracted raw data.

2 Description of Project, Data and Mining

Our Web Marketing system consists of six modules.

- (1) The Web search module extracts home pages of users from several portal sites. Currently, the following portal sites are used: LiveJournal, ICQ and Yahoo, as well as a few major universities.
- (2) The Object-Relational database stores the cleaned results of this search.
- (3) The data mining module uses the WEKA [13] package for extracting association rules from the table data.
- (4) The ontology is the main knowledge representation of this project [4, 11]. It consists of interest hierarchies based on Yahoo and ICQ.
- (5) The advanced extraction component processes Web pages which do not follow simple structure rules.
- (6) The front end is a user-friendly, Web-based GUI that allows users with no knowledge of SQL to query both the raw data in the tables and the derived rules.

The data that we are using for data mining consists of records of real personal data that contain either demographic data and expressed interest data or two different items of interest data. In most cases, we are using triples of age, gender and one interest as input for data mining. In other cases we are using pairs of interests. Interests are derived from one of sixteen top level interest categories. These interest categories are called interests at level 1. Examples of level 1 interests (according to Yahoo) include RECREATION_SPORTS, HEALTH_WELLNESS, GOVERNMENT_POLITICS, etc. Interests are organized as a DAG (Directed Acyclic Graph) hierarchy.

As a result of the large size of the database, the available data goes well beyond the capacity of the data mining program. Thus, the data sets had to be broken into smaller data sets. A convenient way to do this is to perform data mining on the categories divided at level 1 (top level) or the children of level 1. Thus there are 16 interest categories at level 1, and the interest GOVERNMENT_POLITICS has 20 children, including LAW, MILITARY, ETHICS, TAXES, etc. At the time when we extracted the data, ENTERTAINMENT_ARTS was the largest data file at level 1. It had 176218 data items, which is not too large to be handled by the data mining program.

WEKA generates association rules [1] using the Apriori algorithm first presented by [2]. Since WEKA only works with clean data converted to a fixed format, called .arff format, we have created customized programs to do data selection and data cleaning.

3 Using Raising for Improved Support

A concept hierarchy is present in many databases either explicitly or implicitly. Some previous work utilizes a hierarchy for data mining. Han [5] discusses data mining at multiple concept levels. His approach is to use discovered associations at one level (e.g., milk \rightarrow bread) to direct the search for associations at a different level (e.g., milk of brand X \rightarrow bread of brand Y). As most of our data mining involves only one interest, our problem setting is quite different. Han et al. [6] 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. Our approach utilizes an interest ontology to improve support in rule mining by means of concept raising.

Fortin et al. [3] use an object-oriented representation for data mining. Their interest is in deriving multi-level association rules. As we are typically using only one data item in each tuple for raising, the possibility of multi-level rules does not arise in our problem setting. Srikant et al. [12] present Cumulative and EstMerge algorithms to find associations between items at any level by adding all ancestors of each item to the transaction. In our work, items of different levels do not coexist in any step of mining. Psaila et al. [9] describe a method how to improve association rule mining by using a generalization hierarchy. Their hierarchy is extracted from the schema of the database and used together with mining queries [7]. In our approach, we are making use of a large pre-existing concept hierarchy, which contains concepts from the data tuples. Páircéir et al. also differ from our work in that they are mining multi-level rules that associate items spanning several levels of a concept hierarchy [10]. Joshi et al. [8] are interested in situations where rare instances are really the most interesting ones, e.g., in intrusion detection. They present a two-phase data mining method with a good balance of precision and recall. For us, rare instances are not by themselves important, they are only important because they contribute with other rare instances to result in frequently occurring instances for data mining.

There are 11 levels in the Yahoo interest hierarchy. Every extracted interest belongs somewhere in the hierarchy, and is at a certain level. The lower the level value, the higher up it is in the hierarchy. Level 0 is the root. Level 1 is the top level, which includes 16 interests. For example, FAMILY_HOME is an interest at level 1. PARENTING is an interest at level 2. PARENTING is a child of FAMILY_HOME in the hierarchy. If a person expressed an interest in PARENTING, it is common sense that he or she is interested in FAMILY_HOME. Therefore, at level 1, when we count those who are interested in FAMILY_HOME, it is reasonable to count those who are interested in PARENTING. This idea applies in the same way to lower levels.

A big problem in the derivation of association rules is that available data is sometimes very sparse and biased as a result of the interest hierarchy. For example, among over a million of interest records in our database only 11 people expressed an interest in RECREATION_SPORTS, and nobody expressed an interest in SCIENCE. The fact that people did not express interests with more general terms does not mean they are not interested. The data file of

RECREATION_SPORTS has 62734 data items. In other words, 62734 interest expressions of individuals are in the category of RECREATION_SPORTS. Instead of saying “I’m interested in Recreation and Sports,” people prefer saying “I’m interested in basketball and fishing.” They tend to be more specific with their interests. We analyzed the 16 top level categories of the interest hierarchy. We found users expressing interests at the top level only in two categories, MUSIC and RECREATION_SPORTS. When mining data at higher levels, it is important to include data at lower levels, in order to gain data accuracy and higher support.

In the following examples, the first letter stands for an age range. The age range from 10 to 19 is represented by A, 20 to 29 is B, 30 to 39 is C, 40 to 49 is D, etc. The second letter stands for Male or Female. Text after a double slash (//) is not part of the data. It contains explanatory remarks.

Original Data File:

```
B,M,BUSINESS_FINANCE //level=1
D,F,METRICOM_INC //level=7
E,M,BUSINESS_SCHOOLS //level=2
C,F,ALUMNI //level=3
B,M,MAKERS //level=4
B,F,INDUSTRY_ASSOCIATIONS //level=2
C,M,AOL_INSTANT_MESSENGER //level=6
D,M,INTRACOMPANY_GROUPS //level=3
C,M,MORE_ABOUT_ME //wrong data
```

The levels below 7 do not have any data in this example. Raising will process the data level-by-level starting at level 1. It is easiest to see what happens if we look at the processing of level 3. First the result is initialized with the data at level 3 contained in the source file. With our data shown above, that means that the result is initialized with the following two lines.

```
C,F,ALUMNI
D,M,INTRACOMPANY_GROUPS
```

In order to perform the raising we need to find ancestors at level 3 of the interests in our data. Table 1 shows all ancestors of our interests from levels 4, 5, 6, 7, such that the ancestors are at level 3. The following lines are now added to our result.

```
D,F,COMMUNICATIONS_AND_NETWORKING // raised from level=7 (1st
ancestor)
D,F,COMPUTERS // raised from level=7 (2nd ancestor)
B,M,ELECTRONICS // raised from level=4
C,M,COMPUTERS // raised from level=6
```

That means, after raising we have the following occurrence counts at level 3.

ALUMNI: 1
 INTRACOMPANY_GROUPS: 1
 COMMUNICATIONS_AND_NETWORKING: 1
 COMPUTERS: 2
 ELECTRONICS: 1

Before raising, we only had two items at level 3. Now, we have six items at level 3. That means that we now have more data as input for data mining than before raising. Thus, the results of data mining will have better support and will much better reflect the actual interests of people.

Table 1. Relevant Ancestors

| Interest Name | Its Ancestor(s) at Level 3 |
|-----------------------|-------------------------------|
| METRICOM_INC | COMMUNICATIONS_AND_NETWORKING |
| METRICOM_INC | COMPUTERS |
| MAKERS | ELECTRONICS |
| AOL_INSTANT_MESSENGER | COMPUTERS |

Due to the existence of multiple parents and common ancestors, the precise method of raising is very important. There are different ways to raise a data file. One way is to get the data file of the lowest level, and raise interests bottom-up, one level at a time, until we finish at level 1. The data raised from lower levels is combined with the original data from the given level to form the data file at that level. If an interest has multiple parents, we include these different parents in the raised data. However, if those parents have the same ancestor at some higher level, duplicates of data appear at the level of common ancestors.

This problem is solved by adopting a different method: we are raising directly to the target level, without raising to any intermediate level. After raising to a certain level, all data at this level can be deleted and never have to be considered again for lower levels. This method solves the problem of duplicates caused by multiple parents and common ancestors. The data file also becomes smaller when the destination level becomes lower.

In summary, the raising algorithm is implemented as follows: Raise the original data to level 1. Do data mining. Delete all data at level 1 from the original data file. Raise the remaining data file to level 2. Do data mining. Delete all data at level 2 from the data file, etc. Continue until there's no more valid data. The remaining data in the data file are wrong data.

4 Results

The quality of association rules is normally measured by specifying support and confidence. Support may be given in two different ways [13], as absolute support and as relative support. Witten et al. write:

The coverage of an association rule is the number of instances for which it predicts correctly – this is often called its support. . . . It may also be convenient to specify coverage as a percentage of the total number of instances instead. (p. 64)

For our purposes, we are most interested in the total number of tuples that can be used for deriving association rules, thus we will use the absolute number of support only. The data support is substantially improved by means of raising. Following are two rules from RECREATION_SPORTS at level 2 without raising:

age=B interest=AVIATION 70 \Rightarrow gender=M 55 conf:(0.79) (1)

age=C interest=OUTDOORS 370 \Rightarrow gender=M 228 conf:(0.62) (2)

Following are two rules from RECREATION_SPORTS at level 2 with raising.

age=A gender=F 13773 \Rightarrow interest=SPORTS 10834 conf:(0.79) (3)

age=C interest=OUTDOORS 8284 \Rightarrow gender=M 5598 conf:(0.68) (4)

Rule (2) and Rule (4) have the same attributes and rule structure. Without raising, the absolute support is 228, while with raising it becomes 5598. The improvement of the absolute support of this rule is 2355%.

Not all rules for the same category and level have the same attributes and structure. For example, rule (1) appeared in the rules without raising, but not in the rules with raising. Without raising, 70 people are of age category B and choose AVIATION as their interest. Among them, 55 are male. The confidence for this rule is 0.79. After raising, there is no rule about AVIATION, because the support is too small compared with other interests such as SPORTS and OUTDOORS. In other words, one effect of raising is that rules that appear in the result of WEKA before raising might not appear after raising and vice versa.

There is a combination of two factors why rules may disappear after raising. First, this may be a result of how WEKA orders the rules that it finds by confidence and support. WEKA primarily uses confidence for ordering the rules. There is a cut off parameter, so that only the top N rules are returned. Thus, by raising, a rule in the top N might drop below the top N.

There is a second factor that affects the change of order of the mined rules. Although the Yahoo ontology ranks both AVIATION and SPORTS as level-2 interests, the hierarchy structure underneath them is not balanced. According to the hierarchy, AVIATION has 21 descendents, while SPORTS has 2120 descendents, which is about 100 times more. After raising to level 2, all nodes below level 2 are replaced by their ancestors at level 2. As a result, SPORTS becomes an interest with overwhelmingly high support, whereas the improvement rate for AVIATION is so small that it disappeared from the rule set after raising.

There is another positive effect of raising. Rule (3) above appeared in the rules with raising. After raising, 13773 people are of age category A and gender category F. Among them, 10834 are interested in SPORTS. The confidence is 0.79. These data look good enough to generate a convincing rule. However, there were no rules about SPORTS before raising. Thus, we have uncovered a rule with strong support that also agrees with our intuition. However, without raising, this

rule was not in the result of WEKA. Thus, raising can uncover new rules that agree well with our intuition and that also have better absolute support.

To evaluate our method, we compared the support and confidence of raised and unraised rules. The improvement of support is substantial. Table 2 compares support and confidence for the same rules before and after raising for RECREATION_SPORTS at level 2. There are 58 3-attribute rules without raising, and 55 3-attribute rules with raising. 18 rules are the same in both results. Their support and confidence are compared in the table. The average support is 170 before raising, and 4527 after raising. The average improvement is 2898%. Thus, there is a substantial improvement in absolute support. After raising, the lower average confidence is a result of expanded data. Raising effects not only the data that contributes to a rule, but all other data as well. Thus, confidence was expected to drop. Even though the confidence is lower, the improvement in support by far outstrips this unwanted effect.

Table 2. Support and Confidence Before and After Raising

| Rule (int = interest, gen = gender) | Supp. w/o rais. | Supp. w/ rais. | Improv. of supp. | Conf. w/o rais. | Conf. w/ rais. | Improv. of Conf. |
|--|-----------------|----------------|------------------|-----------------|----------------|------------------|
| age=C int=AUTOMOTIVE \Rightarrow gen=M | 57 | 3183 | 5484% | 80 | 73 | -7% |
| age=B int=AUTOMOTIVE \Rightarrow gen=M | 124 | 4140 | 3238% | 73 | 65 | -8% |
| age=C int=OUTDOORS \Rightarrow gen=M | 228 | 5598 | 2355% | 62 | 68 | 6% |
| age=D int=OUTDOORS \Rightarrow gen=M | 100 | 3274 | 3174% | 58 | 67 | 9% |
| age=B int=OUTDOORS \Rightarrow gen=M | 242 | 5792 | 2293% | 54 | 61 | 7% |
| age=C gen=M \Rightarrow int=OUTDOORS | 228 | 5598 | 2355% | 51 | 23 | -28% |
| gen=M int=AUTOMOTIVE \Rightarrow age=B | 124 | 4140 | 3238% | 47 | 37 | -10% |
| age=D gen=M \Rightarrow int=OUTDOORS | 100 | 3274 | 3174% | 46 | 27 | -19% |
| age=B int=OUTDOORS \Rightarrow gen=F | 205 | 3660 | 1685% | 46 | 39 | -7% |
| age=B gen=M \Rightarrow int=OUTDOORS | 242 | 5792 | 2293% | 44 | 18 | -26% |
| gen=F int=OUTDOORS \Rightarrow age=B | 205 | 3660 | 1685% | 42 | 39 | -3% |
| gen=M int=OUTDOORS \Rightarrow age=B | 242 | 5792 | 2293% | 38 | 34 | -4% |
| int=AUTOMOTIVE \Rightarrow age=B gen=M | 124 | 4140 | 3238% | 35 | 25 | -10% |
| gen=M int=OUTDOORS \Rightarrow age=C | 228 | 5598 | 2355% | 35 | 33 | -2% |
| age=D \Rightarrow gen=M int=OUTDOORS | 100 | 3274 | 3174% | 29 | 19 | -10% |
| gen=M int=AUTOMOTIVE \Rightarrow age=C | 57 | 3183 | 5484% | 22 | 28 | 6% |
| int=OUTDOORS \Rightarrow age=B gen=M | 242 | 5792 | 2293% | 21 | 22 | 1% |
| int=OUTDOORS \Rightarrow age=C gen=M | 228 | 5598 | 2355% | 20 | 21 | 1% |

Table 3 shows the comparison of all rules that are the same before and after raising. The average improvement of support is calculated at level 2, level 3, level 4, level 5 and level 6 for each of the 16 categories. As explained in Sect. 3, few people expressed an interest at level 1, because these interest names are too general. Before raising, there are only 11 level-1 tuples with the interest RECREATION.SPORTS and 278 tuples with the interest MUSIC. In the other

14 categories, there are no tuples at level 1 at all. However, after raising, there are 6,119 to 174,916 tuples at level 1, because each valid interest in the original data can be represented by its ancestor at level 1, no matter how low the interest is in the hierarchy.

All the 16 categories have data down to level 6. However, COMPUTERS_INTERNET, FAMILY_HOME and HEALTH_WELLNESS have no data at level 7. In general, data below level 6 is very sparse and does not contribute a great deal to the results. Therefore, we present the comparison of rules from level 2 through level 5 only. Some rules generated by WEKA are the same with and without raising. Some are different. In some cases, there is not a single rule in common between the rule sets with and without raising. The comparison is therefore not applicable. Those conditions are denoted by “N/A” in the table.

Table 3. Support Improvement Rate of Common Rules

| Category | Level2 | Level3 | Level4 | Level5 |
|-----------------------|--------|--------|--------|--------|
| BUSINESS_FINANCE | 122% | 284% | 0% | 409% |
| COMPUTERS_INTERNET | 363% | 121% | 11% | 0% |
| CULTURES_COMMUNITY | N/A | 439% | N/A | 435% |
| ENTERTAINMENT_ARTS | N/A | N/A | N/A | N/A |
| FAMILY_HOME | 148% | 33% | 0% | 0% |
| GAMES | 488% | N/A | 108% | 0% |
| GOVERNMENT_POLITICS | 333% | 586% | 0% | N/A |
| HEALTH_WELLNESS | 472% | 275% | 100% | 277% |
| HOBBIES_CRAFTS | N/A | 0% | 0% | 0% |
| MUSIC | N/A | 2852% | N/A | 0% |
| RECREATION_SPORTS | 2898% | N/A | 76% | N/A |
| REGIONAL | 6196% | 123% | N/A | 0% |
| RELIGION_BELIEFS | 270% | 88% | 634% | 0% |
| ROMANCE_RELATIONSHIPS | 224% | 246% | N/A | 17% |
| SCHOOLS_EDUCATION | 295% | 578% | N/A | 297% |
| SCIENCE | 1231% | 0% | 111% | 284% |
| Average Improvement | 1086% | 432% | 104% | 132% |

Table 4 shows the average improvement of support of all rules after raising to level 2, level 3, level 4 and level 5 within the 16 interest categories. This is computed as follows. We sum the support values for all rules before raising and divide them by the number of rules, i.e., we compute the average support before raising, S_b . Similarly, we compute the average support of all the rules after raising. Then the improvement rate R is computed as:

$$R = \frac{S_a - S_b}{S_b} * 100 \text{ [percent]} \quad (1)$$

The average improvement rate for level 2 through level 5 is, respectively, 279%, 152%, 68% and 20%. WEKA ranks the rules according to the confidence, and discards rules with lower confidence even though the support may be higher.

In Tab. 4 there are three values where the improvement rate R is negative. This may happen if the total average relative support becomes lower after raising. That in turn can happen, because, as mentioned before, the rules before and after raising may be *different* rules. The choice of rules by WEKA is primarily made based on relative support and confidence values.

Table 4. Support Improvement Rate of All Rules

| Category | Level2 | Level3 | Level4 | Level5 |
|-----------------------|--------|--------|--------|--------|
| BUSINESS_FINANCE | 231% | 574% | -26% | 228% |
| COMPUTERS_INTERNET | 361% | 195% | 74% | -59% |
| CULTURES_COMMUNITY | 1751% | 444% | 254% | 798% |
| ENTERTAINMENT_ARTS | 4471% | 2438% | 1101% | 332% |
| FAMILY_HOME | 77% | 26% | 56% | 57% |
| GAMES | 551% | 1057% | 188% | 208% |
| GOVERNMENT_POLITICS | 622% | 495% | 167% | 1400% |
| HEALTH_WELLNESS | 526% | 383% | 515% | 229% |
| HOBBIES_CRAFTS | 13266% | 2% | 7% | 60% |
| MUSIC | 13576% | 3514% | 97% | 62% |
| RECREATION_SPORTS | 6717% | 314% | 85% | 222% |
| REGIONAL | 7484% | 170% | 242% | -50% |
| RELIGION_BELIEFS | 285% | 86% | 627% | 383% |
| ROMANCE_RELATIONSHIPS | 173% | 145% | 2861% | 87% |
| SCHOOLS_EDUCATION | 225% | 550% | 1925% | 156% |
| SCIENCE | 890% | 925% | 302% | 317% |
| Average Improvement | 279% | 152% | 68% | 20% |

5 Conclusions and Future Work

In this paper, we showed that the combination of an ontology of the mined concepts with a standard rule mining algorithm can be used to generate data sets with orders of magnitude more tuples at higher levels. Generating rules from these tuples results in much larger (absolute) support values. In addition, raising often produces rules that, according to our intuition, better represent the domain than rules found without raising. Formalizing this intuition is a subject of future work.

According to our extensive experiments with tuples derived from Yahoo interest data, data mining with raising can improve absolute support for rules up to over 6000% (averaged over all common rules in one interest category). Improvements in support may be even larger for individual rules. When averaging

over all support improvements for all 16 top level categories and levels 2 to 5, we get a value of 438%.

Future work includes using other data mining algorithms, and integrating the raising process directly into the rule mining algorithm. Besides mining for association rules, we can also perform classification and clustering at different levels of the raised data. The rule mining algorithm itself needs adaptation to our domain. For instance, there are over 31,000 interests in our version of the interest hierarchy. Yahoo has meanwhile added many more interests. Finding interest – interest associations becomes difficult using WEKA, as interests of persons appear as sets, which are hard to map onto the .arff format.

References

1. R. Agrawal, T. Imielinski, and A. N. Swami. Mining association rules between sets of items in large databases. In Peter Buneman and Sushil Jajodia, editors, *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, pages 207–216, Washington, D.C., 1993.
2. R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In Jorge B. Bocca, Matthias Jarke, and Carlo Zaniolo, editors, *Proc. 20th Int. Conf. Very Large Data Bases, VLDB*, pages 487–499. Morgan Kaufmann, 1994.
3. S. Fortin and L. Liu. An object-oriented approach to multi-level association rule mining. In *Proceedings of the fifth international conference on Information and knowledge management*, pages 65–72. ACM Press, 1996.
4. J. Geller, R. Scherl, and Y. Perl. Mining the web for target marketing information. In *Proceedings of COLLECTeR*, Toulouse, France, 2002.
5. J. Han. Mining knowledge at multiple concept levels. In *CIKM*, pages 19–24, 1995.
6. J. Han and Y. Fu. Discovery of multiple-level association rules from large databases. In *Proc. of 1995 Int'l Conf. on Very Large Data Bases (VLDB'95), Zürich, Switzerland, September 1995*, pages 420–431, 1995.
7. J. Han, Y. Fu, W. Wang, K. Koperski, and O. Zaiane. DMQL: A data mining query language for relational databases, 1996.
8. M. V. Joshi, R. C. Agarwal, and V. Kumar. Mining needle in a haystack: classifying rare classes via two-phase rule induction. *SIGMOD Record (ACM Special Interest Group on Management of Data)*, 30(2):91–102, 2001.
9. G. P. and P. L. Lanzi. Hierarchy-based mining of association rules in data warehouses. In *Proceedings of the 2000 ACM symposium on Applied computing 2000*, pages 307–312. ACM Press, 2000.
10. R. Páircéir, S. McClean, and B. Scotney. Discovery of multi-level rules and exceptions from a distributed database. In *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 523–532. ACM Press, 2000.
11. R. Scherl and J. Geller. Global communities, marketing and web mining. *Journal of Doing Business Across Borders*, 1(2):141–150, 2002. [http://www.newcastle.edu.au/journal/dbab/images/dbab_1\(2\).pdf](http://www.newcastle.edu.au/journal/dbab/images/dbab_1(2).pdf).
12. R. Srikant and R. Agrawal. Mining generalized association rules. In *Proc. of 1995 Int'l Conf. on Very Large Data Bases (VLDB'95), Zürich, Switzerland, September 1995*, pages 407–419, 1995.
13. I. H. Witten and E. Frank. *Data Mining*. Morgan Kaufmann Publishers, San Francisco, 2000.