Interesting Association Rules in Multiple Taxonomies

Jeannette M. de Graaf [a] Walter A. Kosters [a]
Jeroen J.W. Witteman [a]

[a] Leiden Institute of Advanced Computer Science
Universiteit Leiden
P.O. Box 9512, 2300 RA Leiden, The Netherlands
{graaf,kosters,jwittema}@liacs.nl

Abstract

In this paper we study association rules in order to understand customer behaviour. We examine the case where many customers may choose from a long list of products. Suppose that several taxonomies for these products are given: the products are grouped in different ways, e.g., by colour, by price, by brand and so on. Then a rule is called interesting if its support, i.e., the number of customers satisfying the rule, deviates substantially from the predictions that are generated through one or more taxonomies. Such a prediction is found by replacing any product in a rule with its parent in the taxonomy at hand, and then estimating the support of the original rule through the support of the parent rules and the conditional probabilities of the "lifted" products. This notion of interestingness is easy to handle and adheres to the intuition.

1 Introduction

In this paper we study association rules, i.e., rules such as "if a person buys products a and b, then he or she also buys product c". Such a rule has a certain *support* (the number of people satisfying the rule, i.e., buying a, b and c) and *confidence* (the fraction of people buying the products from the "then part" out of those buying the products from the "if part"). In most practical situations an enormous number of these rules, usually containing two or three products, is present. One of the major problems is to decide which of these rules are interesting.

If we only consider the support, there is no emphasis on either "if part" or "then part", and in fact we rather examine the underlying *itemset*, in our example $\{a, b, c\}$. A k-itemset consists of k elements. Such a set is called *frequent* if its support is larger than some threshold, which is given in advance. It is also possible to introduce different support thresholds at different levels.

Now we suppose that a *taxonomy* for the products is given. In this setting association rules may involve categories of products; abstraction from brands gives generalized rules, that are often more informative, intuitive and flexible. Since in this case the number of rules increases enormously, a notion of interestingness, cf. [5, 8], is necessary to describe them. It might for instance be informative to know that people who buy a history book also tend to buy a crime novel; on a more detailed level one might find that people who buy "The Rise and Fall of the Roman Empire" often also buy "The Hound of the Baskervilles". The more detailed rule is only of interest if it deviates substantially from what is expected from the more general one.

A taxonomy is a hierarchy in the form of a tree, where the original products are the leaves, and the root is the "product" *All*. We consider the case where several taxonomies are given; note that in principle it is possible to convert these taxonomies into a single one by introducing extra nodes, but this gives rise to problems to be discussed later on. In this setting, an itemset is allowed to be any set of nodes from different levels from the taxonomies. Often we will restrict an itemset to belong to a single taxonomy. Note that the (internal) nodes of the taxonomies are in fact sets of original products, these being singleton sets; every parent is the union of his or her children. In order to "buy" such a node, it is sufficient for a customer to buy one product out of the set. The root product *All* is the set of all original products, and is the root of all taxonomies at hand.

Let us look at the following simple example. A book shop sells cheap, moderate(ly priced) and expensive books in three genres: crime, history and dictionary. So we might have the following two taxonomies, called **price** and **genre**, assuming that moderately priced history books do not exist:

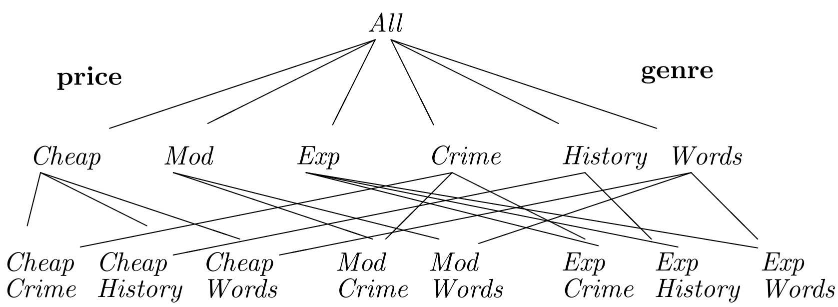

Figure 1: Example — Two taxonomies in a book shop, in one graph.

For simplicity we let every leaf correspond with a single product, i.e., a single book title. Let us assume that 50 persons buy $\mathcal{A} = \{$*Cheap Crime*, *Exp History*$\}$, (the same) 50 persons buy $\mathcal{B} = \{$*Crime*, *Exp History*$\}$ (one of the "parents" of $\mathcal{A}$ through the **genre** taxonomy), and 100 persons buy $\mathcal{C} = \{$*Cheap*, *Exp History*$\}$ (one of the "parents" of $\mathcal{A}$ through the **price** taxonomy). Suppose furthermore that the following conditional probabilities are known: the chance that a cheap purchase happens to be a cheap crime novel is 50%, and the chance that a crime purchase happens to be a cheap crime novel is also 50%. Then itemset $\mathcal{A}$ cannot be understood through itemset $\mathcal{B}$, which would predict 25 *Exp History* buyers to buy

$\mathcal{A}$, but itemset $\mathcal{C}$ does explain $\mathcal{A}$. In the **price** taxonomy $\mathcal{A}$ is not interesting, but in the **genre** taxonomy it is. In this (artificial) example it might be the case that people who buy expensive history books are biased towards cheap crime novels, in the sense that if they buy an expensive history novel and some crime novel, the crime novel always is cheap.

In the sequel we shall define a precise notion of interestingness, based on parents of itemsets in the taxonomies. We shall discuss several options, and using both real life data and artificial data we illustrate the relevance of this notion. Our goal is to find a moderate number of association rules describing the system at hand, where uninteresting rules that can be derived from others are discarded. Interestingness of itemsets based on a hierarchy for the items is also discussed in [8]. Several other measures of interestingness not involving taxonomies are mentioned in [2, 3, 6, 7] and references in these papers.

2 Interestingness

An itemset (or rule) should be called *interesting* if it is in some sense "special" with respect to what it is expected to be in the light of its parents. Let us therefore first give some formal definitions concerning the connection between parent itemsets and their children.

A one generation ancestor itemset of a given itemset is created by replacing one or more of its elements by their immediate parents in the taxonomy. For the moment we choose to stay within one taxonomy, but it is also possible to use several taxonomies simultaneously. The only difference in that case is that elements can have more than one parent. The support of an ancestor itemset gives rise to a prediction of the support of the k-itemset $\mathcal{I} = \{a_1, a_2, \ldots, a_k\}$ itself: suppose that the nodes $a_1, a_2, \ldots, a_\ell$ $(1 \leq \ell \leq k)$ are replaced by (*lifted* to) their ancestors $\widehat{a_1}, \widehat{a_2}, \ldots, \widehat{a_\ell}$ (in general not necessarily their parents: an ancestor of a is a node on the path from the root *All* to a, somewhere higher in the taxonomy) giving an itemset $\widehat{\mathcal{I}}$. Then the support of $\mathcal{I}$ is estimated by the support of $\widehat{\mathcal{I}}$ times the confidence of the rule "$\widehat{a_1}, \widehat{a_2}, \ldots, \widehat{a_\ell}$ implies $a_1, a_2, \ldots, a_\ell$":

$$\begin{aligned}
&\mathit{EstimatedSupport}_{\widehat{\mathcal{I}}}(\{a_1, a_2, \ldots, a_\ell, a_{\ell+1}, \ldots, a_k\}) = \\
&\qquad \mathit{RealSupport}(\{\widehat{a_1}, \widehat{a_2}, \ldots, \widehat{a_\ell}, a_{\ell+1} \ldots, a_k\}) \times \frac{\mathit{RealSupport}(\{a_1, a_2, \ldots, a_\ell\})}{\mathit{RealSupport}(\{\widehat{a_1}, \widehat{a_2}, \ldots, \widehat{a_\ell}\})} \quad .
\end{aligned}$$

This estimate is based on the assumption that given the fact that $\{\widehat{a_1}, \widehat{a_2}, \ldots, \widehat{a_\ell}\}$ is purchased, buying $\{a_1, a_2, \ldots, a_\ell\}$ and $\{a_{\ell+1}, a_{\ell+2}, \ldots, a_k\}$ are independent events. In fact, this is a simple application of conditional probabilities: if

$$\begin{aligned}
&P(\mathcal{I} \,|\, \widehat{a_1}, \widehat{a_2}, \ldots, \widehat{a_\ell}) = \\
&\qquad P(a_1, a_2, \ldots, a_\ell \,|\, \widehat{a_1}, \widehat{a_2}, \ldots, \widehat{a_\ell}) \times P(a_{\ell+1}, \ldots, a_k \,|\, \widehat{a_1}, \widehat{a_2}, \ldots, \widehat{a_\ell}) \quad ,
\end{aligned}$$

then

$$\begin{aligned}
P(\mathcal{I}\,) &= P(\widehat{a_1}, \widehat{a_2}, \ldots, \widehat{a_\ell}) \times P(\mathcal{I} \,|\, \widehat{a_1}, \widehat{a_2}, \ldots, \widehat{a_\ell}) \\
&= P(\widehat{\mathcal{I}}\,) \times P(a_1, a_2, \ldots, a_\ell \,|\, \widehat{a_1}, \widehat{a_2}, \ldots, \widehat{a_\ell}) \quad .
\end{aligned}$$

Now an itemset is called *interesting* if and only if the predicted supports based on all (but one as we shall see soon) of its one generation ancestor itemsets deviate substantially from its real support. If there is at least one parent that predicts the child suitably well, this itemset is not interesting enough. The word "substantially" means that the predicted supports are all larger than the real support, or are all smaller than the real support, by at least some fixed factor. This factor is called the *interestingness threshold*. If all products from an itemset are lifted, estimated support and real support are exactly the same, so it makes sense to omit this prediction. Therefore 1-itemsets are always interesting, in particular the itemset $\{All\}$ (which does not have ancestors): there is no way to predict their support. In order to give a complete description of the "rule database" it is sufficient to describe the interesting rules: the behaviour of the others can then be derived—if one remembers which ancestor itemset provided the best prediction.

The reasons that only one generation ancestor itemsets are used instead of arbitrary ancestors as in [5] (where the number of items in the two itemsets should also be the same) are the following. First, it severely restricts the number of sets that need to be examined. (Note that in a single taxonomy a k-itemset already has $2^k - 2$ one generation ancestor itemsets in principle.) And second, if a set cannot be understood through any of its parents, but some grandparent does predict its support, in our opinion it still deserves attention.

One particular problem is the following. If during the lifting one or more overlapping sets are created, in particular if two or more taxonomies are used, one has to decide which clients satisfy a certain itemset. For instance, if in a book shop the sets *Crime* and *Cheap* have a book *Cheap Crime* in common (see Figure 1), it may or may not be sufficient for a customer to buy *Cheap Crime* in order to satisfy *Crime* and *Cheap* at the same time. In this paper we will stick to the convention that this is indeed sufficient. If we have one taxonomy only we already encounter this problem, but in a somewhat simpler form: it is easy to generate an ancestor such as $\{Cheap, Cheap\}$, which might be called a 2-itemset, needing at least two products. (This occurs if two siblings are lifted to the same parent.) However, we shall consider it to be equivalent to the 1-itemset $\{Cheap\}$, thereby adhering to the convention that a single product may be used to satisfy several different requirements, so for instance *Cheap Crime* satisfies this set. One reason to do so is that otherwise a lot of backtracking would be necessary to perform the checking: if a product is used to fulfill a certain requirement, in a later stage it may be needed for some other requirement. A second reason is that otherwise very many itemsets of the form $\{child, ancestor\}$, such as $\{Cheap\ Crime, Cheap\}$, would enter the already immense list of candidates. Finally, a single purchase may reflect several properties at one time: buying *Cheap Crime* reveals two (or more) clues about the customer behaviour simultaneously.

Besides this type of interestingness we might also consider so-called *right hand side interestingness*. If we choose one particular item a from a k-itemset $\mathcal{I}$, we may lift this to the product *All*. Using the above formula, this leads to a simple estimate for the support of $\mathcal{I}$, or more precisely, of the rule "$\mathcal{I} - \{a\}$ implies $\{a\}$". It gives an extra independent measure for interestingness, which is also easy to compute. More general (cf. [3]), a notion of interestingness can also be based on parents produced by deleting items from itemsets.

3 Algorithms

The algorithms that find all interesting rules are straightforward. The well-known Apriori algorithm from [1], or any of its refinements, provides a list of all association rules. The algorithms can be easily adapted to generate all rules including nodes from the taxonomy, where special care has to be taken to avoid parent-child problems (see [8]). In fact, if one augments the list of original products with all non-leaves from the taxonomy, the Apriori algorithm can do the job. Once the list of all rules is known, it is easy to generate the interesting ones by just comparing supports for the appropriate rules.

For every frequent itemset $\mathcal{I}$ all its one generation ancestor itemsets $\widehat{\mathcal{I}}$ are generated, and expected and real support are compared; we define the *support deviation* of $\mathcal{I}$ to be the smallest *interestingness ratio*

$$\mathit{RealSupport}(\mathcal{I}) \ / \ \mathit{EstimatedSupport}_{\widehat{\mathcal{I}}}(\mathcal{I})$$

that occurs. In the example in Section 1 we would get $50/(100 \times 0.5) = 1.0$ through parent $\mathcal{C}$ and $50/(50 \times 0.5) = 2.0$ through parent $\mathcal{B}$. If this support deviation is higher than the interestingness threshold, the itemset is called interesting. Note that the ancestors are automatically frequent, unless—as in [5]—different support thresholds are specified at different tree levels. The frequent itemsets can be ordered with respect to support deviation: the higher this ratio, the more interconnection occurs between the products involved. In fact, the assumption concerning the independence between lifted and non-lifted products clearly does not hold in that case, and an interesting connection is revealed.

It is also possible to look at the average interestingness ratio over the ancestors, but we feel that the minimum approach is more appropriate. Of course it is also a possibility to look at the maximum, and then try to find underestimated supports; however, in most practical applications the minimum approach gives better results. If necessary, the confidence might be used in order to turn the list of interesting itemsets into a list of interesting rules, and in order to further decrease the number of interesting rules.

The run time of the algorithms may—as usual—be long when the number of transactions is large and the threshold minimum support is low. In order to get also information on the product level, and not only on aggregate levels, this minimum support should be small enough. A run time of several hours was quite normal, most of it devoted to the computation of the frequent itemsets using the Apriori algorithm. Once the rules/itemsets are computed, it is however easy to deal with different interestingness thresholds. This is an advantage over methods to detect interestingness during the computation of the frequent itemsets (cf. [3], where no taxonomies are used). In [4] the differences between algorithms with and without the precomputation of itemsets are studied in a somewhat different context.

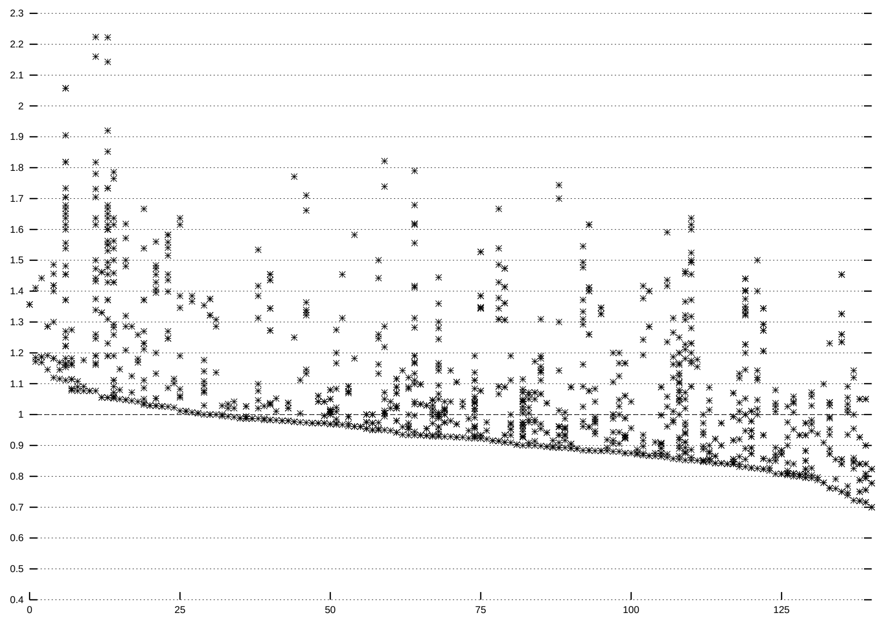

Figure 2: Interestingness ratios of itemsets with at least 20% support in the artificial book shop database.

4 Experiments

In order to illustrate the relevance of the proposed selection of interesting itemsets, we first present some results for a small artificial example of a book shop database, with two taxonomies as in Figure 1. We created 42 customers, where we deliberately made the itemset {*Exp Words*, *Cheap*} interesting. (By the way, this is not that easy.) Its support deviation was 1.12, making it the sixth most interesting itemset. The itemset {*History*, *Cheap*} turned out to be the best one, with a support deviation of 1.36; this happens to be an itemset with nodes from both taxonomies. In Figure 2 we plot the different interestingness ratios of all 141 frequent itemsets, sorted with respect to their support deviation; the support threshold used was 20%. An interestingness threshold of 1.07 would give 12 interesting itemsets, including one 4-itemset having 21 one-generation ancestors.

Furthermore, we performed experiments on a real database with about 45,000 transactions ("baskets"), and about 1,800 possible products, using a single taxonomy containing about 100 non-leaves. We also generated an artificial random database with the same number of transactions and the same distribution of 1-itemsets. We then took a support threshold of 0.2%, and only considered those itemsets that contained a rule with at least 75% confidence (in order to keep the number of itemsets reasonable). In Figure 3 and Figure 4 we plot the different interestingness ratios of all frequent 3-itemsets, about 7,500 for the real database and about 11,500 for the artificial database. For every itemset (in fact rule) we plot the interestingness ratios for its at most $2^3 - 2 = 6$ parents (remember we agreed to omit the situation where all products were lifted). For the real database the number of rules with a support deviation larger than the interestingness threshold 1.3 is 194, including some unexpected ones, but for the artificial database it is only 50. Also note the more compact nature of the data cloud for the artificial database. These observations suggest that interesting itemsets are indeed interesting.

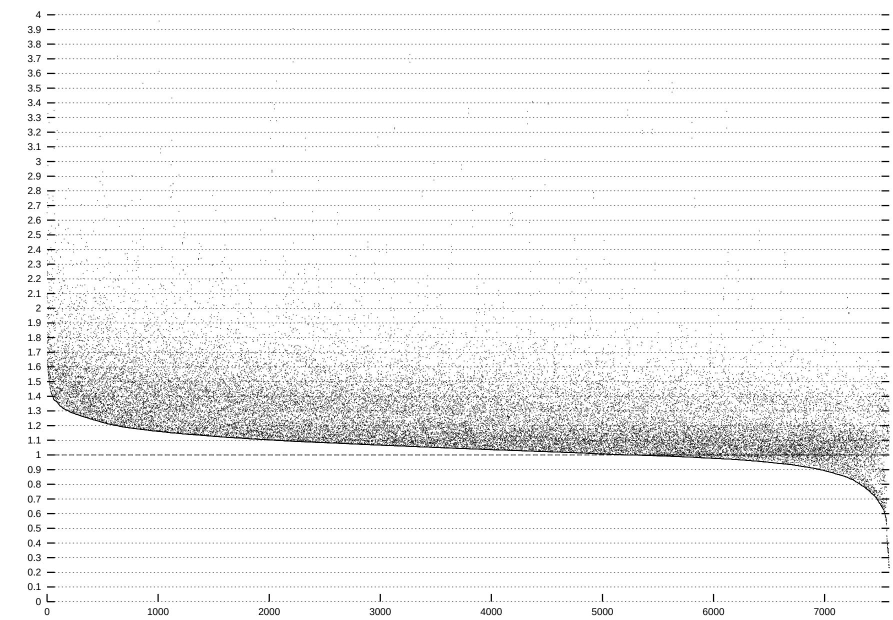

Figure 3: Interestingness ratios of 3-itemsets with at least 0.2% support and 75% confidence in a real database.

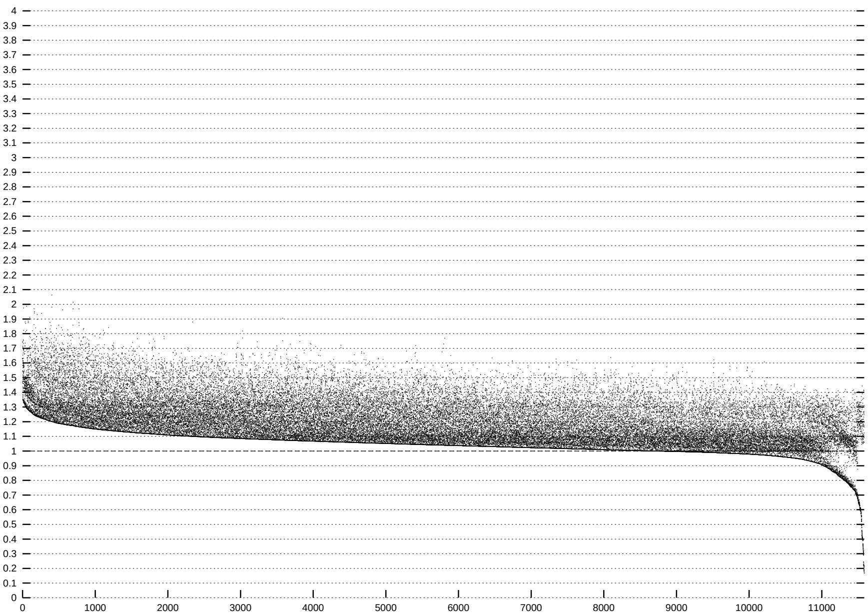

Figure 4: Interestingness ratios of 3-itemsets with at least 0.2% support and 75% confidence in an artificial database.

5 Conclusions and Further Research

We may conclude that interesting rules can be found with a reasonable amount of work. They give a good overview of diverging customer behaviour.

We would like to examine the connection with more traditional statistical methods. We are also interested in time series, where the set of interesting rules/itemsets might be time dependent. It is easy to imagine that interesting behaviour changes throughout the year, and it should be possible to describe this effect in some natural way. It would also be nice to examine the interpretation where every item requires a unique product, so in order to satisfy {*Cheap*, *Crime*}, {*Cheap Crime*, *Cheap*} or even {*Cheap*, *Cheap*}, (at least) two purchases would be necessary. As mentioned before, this would have a negative influence on the run time. Finally, the introduction of different thresholds for different levels deserves further study.

References

[1] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, and A.I. Verkamo. Fast discovery of association rules. In U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors, *Advances in Knowledge Discovery and Data Mining*, pages 307–328. AAAI/MIT Press, 1996.

[2] R.J. Bayardo Jr. and R. Agrawal. Mining the most interesting rules. In S. Chaudhuri and D. Madigan, editors, *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 145–154. ACM Press, 1999.

[3] R.J. Bayardo Jr., R. Agrawal, and D. Gunopulos. Constraint-based rule mining in large, dense databases. *Data Mining and Knowledge Discovery*, 4:217–240, 2000.

[4] B. Goethals and J. Van den Bussche. A priori versus a posteriori filtering of association rules. Presented at the 1999 ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery, 1999.

[5] J. Han and Y. Fu. Mining multiple-level association rules in large databases. *IEEE Transactions on Knowledge and Data Engineering*, 11:798–804, 1999.

[6] M. Klemettinen, H. Mannila, P. Ronkainen, H. Toivonen, and A.I. Verkamo. Finding interesting rules from large sets of discovered association rules. In *Proceedings of the Third International Conference on Information and Knowledge Management (CIKM'94)*, pages 401–407. ACM Press, 1994.

[7] G. Piatetsky-Shapiro. Discovery, analysis, and presentation of strong rules. In G. Piatetsky-Shapiro and W.J. Frawley, editors, *Knowledge Discovery and Data Mining*, pages 229–248. MIT Press, 1991.

[8] R. Srikant and R. Agrawal. Mining generalized association rules. In U. Dayal, P.M.D. Gray, and S. Nishio, editors, *Proceedings of the 21st VLDB Conference*, pages 407–419. Morgan Kaufmann, 1995.