

Exploratory Analysis of Retail Sales of Billions of Items

Dunja Mladenić* William F. Eddy† Scott Ziolko‡

Abstract

This paper describes several different approaches to analysis of a data set collected over the past year from a retail grocery chain containing hundreds of stores. Each record in the data set represents an individual item processed by an individual checkout laser scanner at a particular store at a particular time on a particular day. Each record contains additional information such as store department, price, etc. together with identifying information such as the particular checkout scanner and, for some transactions, customer identification. The total data set contains billions of items which can be aggregated into hundreds of millions of transactions for millions of repeat customers. In order to get some insights in the data, we used several different approaches including some statistical analysis, some machine learning, and some data mining methods. Some of these have simply focused on ascertaining the “quality” of the data while others have been more narrowly focused on simple questions like “which pairs of items are most frequently purchased together” or “what is the relationship between basket size and number of baskets”. The sheer size of the data set has forced us to go beyond usual data mining methods and utilize *Meta-Mining*: the post processing of the results of basic analysis methods.

1 Introduction

We have obtained a data set from a retail grocery chain which contains all checkout scanner records for approximately one year. The data are delivered to the corporate headquarters on a weekly basis and are processed into a large corporate data warehouse at that time. We obtained a data feed from the in-house data processing activity. The corporate programs are written in COBOL and run on a large IBM mainframe computer. Thus, the records we

*J.Stefan Institute, Ljubljana, Slovenia and Carnegie Mellon University, Pittsburgh, USA

†Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213, USA

‡Center for Automated Learning and Discovery, Carnegie Mellon University, Pittsburgh, PA 15213, USA

obtain are in EBCDIC (rather than ASCII) and contain fields which are packed decimal and other non-standard formats.

A brief description of our hardware and software may help the reader understand some of the constraints that affected our operations. Our basic data storage facility is an IBM Magstar 3494 robotic tape storage device with a capacity of about four terabytes of data. This device is physically connected to both of our IBM Netfinity 7000M10 dual processor servers with 2GB of RAM each running the Windows NT 4.0 Server operating system. These two servers have Fibrechannel connections to our terabyte RAID hard disk storage device. We also have four dual processor IBM Netfinity 7000M10 servers running RedHat Linux Version 6.2 and we use Samba to mount the NT RAID device onto the Linux servers.

The data arrive weekly on a IBM 3590 cartridge tape which we insert into the Magstar tape library. The files are copied from tape onto the RAID storage and compressed (from 6GB to less than 2GB file size) so that they can be read on the Linux systems (which have a 2GB file size limit). We run our custom conversion program to convert the data from EBCDIC to ASCII and packed decimal to ASCII, etc. This produces one file for each hour of the week. These files are then sorted (by store, time, transaction number, etc.) to put all items in a market basket together.

At this point the data are organized sufficiently for many different subsequent processing steps. Our standard processing generates a new file with one record per basket, listing the items in the basket on that record. We have other specialized projects which perform different processing on the basic sorted files.

2 Data Description

Our data set consists of about a year of data collected over several hundreds of supermarket stores having different sizes and locations. Each record in the data set represents an item that was scanned at one of the checkout stations at a given store, day, and time. For each record we have a number of fields giving details about the conditions under which it was sold, such as price, information about potential price reductions applied (coupon sale, regular sale, . . .), department inside the store, checkout scanner number, and customer number. There are a few million baskets each week and a total of several million customers that are registered as “loyal customers”.

Each item is associated with a Universal Product Code (UPC) and there is additional information about the items themselves given in a 4-level hierarchical taxonomy. The top level includes nearly 100 categories of items, such as *bakery*, *dairy*, *frozen food*, *salads*, *seafood*, *wine*, etc. The next level, giving a finer grouping of items, includes several

hundred groups, such as *bakery mixed*, *bakery needs*, *candy*, *deli*, *fresh bread & cake*, *juice drinks*, *lettuce*, *milk fresh*, *pet food*, etc. The third level includes a couple of thousand subgroups such as *fresh fish*, *frozen fish*, *other seafood*, *cake decor and food color*, *fruit snacks*, *carrots*, *peppers*, *tomatoes*, *other vegetables*, *pasta sauce*, etc. The leaf level contains a couple of hundred thousand items with their UPC codes, such as *cherry angel food cake* (within *cupcakes* within *cakes* within *bakery*), *Hanover whole baby carrots* (within *carrots* within *frozen vegetables* within *frozen*), *48% vegetable oil spread* (within *margarine-bowls* within *margarine and butters* within *dairy*), “*wht zinfandel*” (within *wine-misc* within *wine* within *alcohol beverage*).

Because there are a couple of hundred thousand different items it is useful to group them somehow. We found it extremely difficult to create groups by clustering and other methods because the text descriptions do not provide common unique identification and it is sometimes difficult to group common products together. For example, there are 1909 entries in our database which contain the text string “MILK,” including “MILKY WAY,” BUTTERMILK PANCAKES,” etc. Of these, 291 contain the string “FRESH MILK.” Of those, 49 contain the string “2%.” Five of those contain the string “1/2%.” Thus there are 44 items which correspond to 2% FRESH MILK coming from different suppliers, in different size containers, made of different materials. Because of these difficulties, in our work here we only use the third level of the taxonomy; all of our subsequent analyses are based on the couple of thousand subgroups.

3 A First Look

The main unit we are dealing with is a basket, corresponding to the content of a physical basket the customer presented at the counter. The number of baskets varies over hours and stores and so does the number of items in an individual basket. It is interesting to see how the average basket size varies during the day. Figure 1 shows the distribution of the basket size over different hours of one day for all the stores. As expected the most items are purchased during the daytime (10 a.m. to 8 p.m.), where 25% of the baskets contain more than 10 items with a considerable number of baskets having around 100 items. There are also some outliers with over 150 items, even one basket with about 200 items that was purchased around midnight (Hour 0 in the graph). All these outliers potentially reflect noise or error in the processed data and some simple statistical processing can help in identifying such situations.

Distribution of basket sizes per hour across all stores for one day

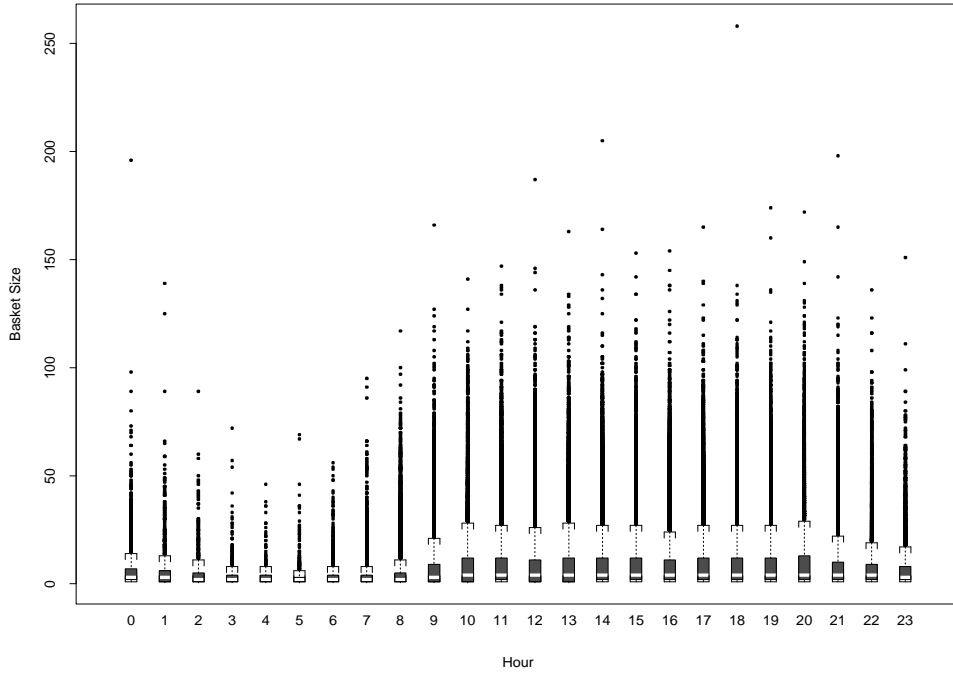
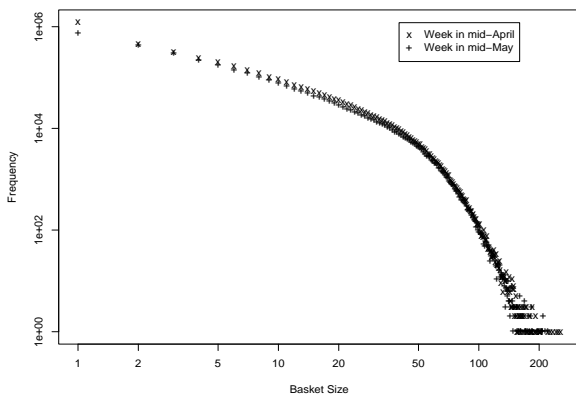


Figure 1: Side-by-side boxplots showing the distributions of basket sizes for each hour of one day across all stores.

(a)



(b)

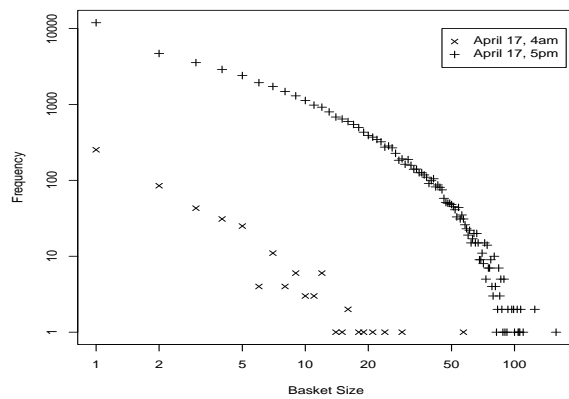


Figure 2: Frequency of different basket sizes. (a) Comparing a week prior to the Easter holiday in April, 2000 and a week in mid May, 2000 we see slightly higher frequencies in April likely due to a larger number of baskets during the week preceding the holiday. (b) Comparing two hours on the Monday prior to Easter, the smallest number of baskets in an hour was for 4A.M. to 5A.M., while the largest number of baskets was observed for 5P.M. to 6P.M.. The pattern observed for the 4A.M. hour appears different than the general pattern found for the relation between the basket size and its frequency.

4 Decision Trees

4.1 Our Approach

Decision trees have often been used in Data Mining tasks such as finding cross-selling opportunities, performing promotion analysis, analyzing credit risk or bankruptcy, and detecting fraud. We use the C5.0 algorithm (an extension of C4.5 proposed by Quinlan[12]) which generates a decision tree for the given dataset by recursive partitioning of the data. The particular implementation of C5.0 we use is part of the commercially available data mining package *Clementine*. In our experiments, we use the Information Gain ratio as the splitting/selection criterion and perform pruning at different levels. Since our goal is to discover patterns in our data and not to classify or predict unseen instances, we derive a rule set from a decision tree by writing a rule for each path in the decision tree from the root to a leaf.

4.2 Experimental Results

We constructed two decision trees for the two one-hour subsets we considered in the previous section. We tried to predict the size of a basket based on several characteristics of the transaction, namely: the basket contents, the particular store number (this is an arbitrary corporate designation for the store), and time. We constructed rules predicting basket size within broad categories (Very Small = 1-3, Small = 4-10, Medium = 11-20, Large = 21-40, Very Large = 41 or more). For clarity we have rewritten the rules generated by the program, in the figures below.

Figure 3 shows all the rules derived from the C5.0 decision tree (with pruning set to severity 75 and at least 10 examples per node) for the tree constructed for the hour with the smallest number of the baskets (4 A.M. to 5 A.M.) on an April Monday. The frequency counts of the five basket size categories are, respectively, (381, 84, 15, 3, 1). Since this hour is in the middle of the night most baskets are very small and the rules reflect this fact. The most significant rule notes that non-loyal customers have very small basket sizes. The remaining rules show that the three subgroups, ICE CREAM & DESSERTS, WHITE SLICED, ENTREE-SIDE DISHES, are important for distinguishing between small and very small baskets in the middle of the night.

The rules for the hour with the highest number of baskets for the same day (5 P.M. to 6 P.M.) are given in Figures 4 and 5. The frequency counts of the five basket size categories are, respectively, (20179, 12855, 6481, 3928, 1195). The number of rules for the same five categories are (5, 32, 93, 114, 33).

The most obvious feature of these is that they contain a considerably larger number of clauses. One rule, which we have omitted from the figure, predicts very small basket

Rules for Very Small Baskets (all rules):

```
if not loyal customer
then Very Small (275, 0.921)

if not (ICE CREAM & DESSERTS, WHITE SLICED,ENTREE-SIDE DISHES)
then Very Small (445, 0.83)
```

Rules for Small Baskets (all rules):

```
if loyal customer
and WHITE SLICED
then Small (11, 0.692)

if loyal customer
and ICE CREAM & DESSERTS
then Small (11, 0.538)

if not ICE CREAM & DESSERTS
and ENTREE-SIDE DISHES
then Small (10, 0.5)
```

Figure 3: All the decision tree rules for the hour from 4 A.M. to 5 A.M.

size for loyal customers if the basket contains EGGS and does not contain any of 39 other specific subgroups. We looked in more detail at the very small baskets and found 11,931 baskets with one item, 4,691 with two items, and 3,557 with three. The most frequent items in the one-item baskets (in descending frequency) were SNACK BAR, HOT PREPARED FOOD, EGGS, REGULAR COLAS, BREADS, 2% MILK, BABY FORMULA-LIQUID, etc. The most frequent combination pair of subgroups in the two-item baskets were HOT PREPARED FOODS (twice), FILM COUPON and FRONT RACK ITEM, and MILK (twice) (Film is usually displayed in the front racks). The most frequent triple of subgroups was BABY FORMULA-LIQUID, BABY FOOD-CEREALS, BABY FOOD-JUICES which only occurred 12 times. It is interesting to note that some of the rules use the (arbitrary) store number to predict basket size and one rule uses the time of day. Several of the rules apply only to loyal customers. It seems difficult to generalize further from these rules and the others we have omitted were no more helpful.

Rules for Small (a subset of 32 rules):

```
if storeNo <= MediumBound
and loyal customer
and (MAGAZINES, FRONT END MAGAZINES)
and not (MEATS DELI, MAINSTREAM COOKIES, YOGURT, DRY PACKAGED DINNERS,
        BUNS, BANANAS, VEGETABLES, NUTS CANNED, BABY FOOD-CEREALS-COOKIES,
        POTATOES&ONIONS, SKIM, REF ORANGE JUICE, BACON)
then Small (24, 0.962)

if loyal customer
and (HOT PREPARED FOODS > 1)
and not (MEATS DELI, ICE CREAM & DESSERTS, BUNS, EGGS, APPLES:LOOSE,
        BERRIES, VEGETABLES, CONVENIENCE FOODS, POTATOES & ONIONS,
        DISH DETERGENT, FACIAL TISSUES, 2%, SKIM, LUNCHMEAT)
and not (CAT FOOD-WET > 2, BABY FOOD-2ND FOODS > 2)
then Small (45, 0.851)
```

Rules for Medium (a subset of 93 rules):

```
if loyal customer
and (ENTREE-SIDE DISHES, ORANGE JUICE)
and not (YOGURT, BANANAS, VEGETABLES, POTATO CHIPS, POTATOES & ONIONS,
        FACIAL TISSUES, BABY FOOD-CEREALS-COOKIES, BACON, LUNCHMEAT)
then Medium (35, 0.784)

if LowerBound < StoreNo <= UpperBound
and loyal customer
and LUNCHMEAT
and not (MEATS DELI, PASTA, EGGS, POTATO CHIPS, BABY FOOD-CEREALS-COOKIES,
        POTATOES & ONIONS)
and CAT FOOD-WET > 3
then Medium (21, 0.826)

if time <= 17:20:00
and loyal customer
and (SHREDDED CHEESE, POTATOES & ONIONS)
and not (MEATS DELI, PASTA, YOGURT, EGGS, BACON, UNK,
        BABY FOOD-CEREALS-COOKIES, CONVENIENCE FOODS)
then Medium (21, 0.826)
```

Figure 4: Some of the decision tree rules for the hour from 5 P.M. to 6 P.M. for Small and Medium basket sizes.

Rules for Large (a subset of 114 rules):

```
if (REGULAR COLAS, EGGS, 2%, SOUR CREAM)
and not ( CONDENSED CANNED SOUPS, BACON)
then Large (31, 0.667)
```

```
if loyal customer
and (MEATS DELI, EGGS, POTATOES & ONIONS, BACON)
and not (CANNED CHUNK LITE TUNA, PASTA, POLY BAG POTATOES, PAPER TOWELS)
then Large (61, 0.635)
```

Rules for Very Large (a subset of 32 rules):

```
if (YOGURT, EGGS, POTATOES & ONIONS, UNK > 3)
then VeryLarge (48, 0.9)
```

```
if (REGULAR COLAS, EGGS, BACON, LUNCHMEAT)
then VeryLarge (77, 0.722)
```

```
if loyal customer
and (MEATS DELI, BANANAS, VEGETABLES > 1, TOMATOES)
then VeryLarge (104, 0.604)
```

Figure 5: Some of the decision tree rules for the hour from 5 P.M. to 6 P.M. for Large and Very large basket sizes.

5 Association Rules

5.1 Description of Our Approach

In order to find associations between the items in the data, we used the third level of taxonomy having a couple of thousand of subgroups and replacing each item by the associated subgroup. As usual in the market basket analysis, each example in our experiments corresponds to a single basket of items that the customer has purchased. Each example is thus represented as a Boolean vector giving information about presence of items in the basket. Using the data we generated association rules by applying the Apriori algorithm [1] using the publicly available implementation [3], a version of which is incorporated in the commercially available data mining package "Clementine" [4].

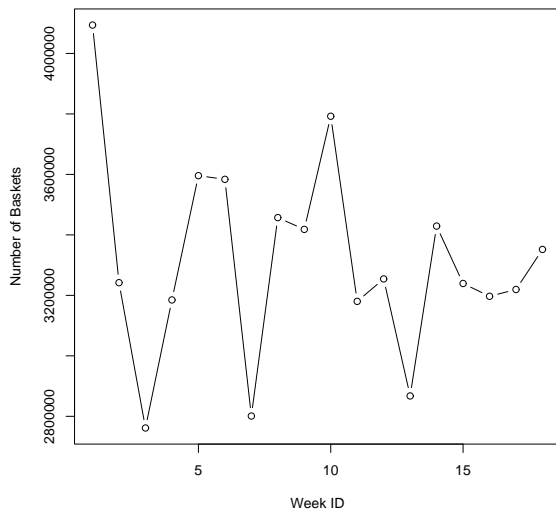
In a typical data mining setting, it is assumed that there is a finite set of literals (usually referred to as items) and each example is some subset of all the literals. The Apriori algorithm performs efficient exhaustive search by using dynamic programming and pruning the search space based on the parameters given by the user for minimum support and confidence of rules. This algorithm has been widely used in data mining for mining association rules over "basket data", where literals are all the items in a supermarket and examples are transactions (specific items bought by customers).

An association rule is an implication of the form $X \rightarrow Y$, where X and Y are subsets of literals and $X \cap Y = \phi$. We say that the rule holds with *confidence* c if $c\%$ of examples that contain X also contain Y . The rule is said to have *support* s in the data if $s\%$ of examples contain $X \cup Y$. In other words, we can say that for the rule $X \rightarrow Y$, its support estimates the joint probability of the rule items $P(X, Y)$ and its confidence estimates the conditional probability of the rule's implication $P(Y|X)$.

We have generated association rules on 18 weeks of data collected in year 2000 from mid April through mid October. There are a few weeks that we are missing due to some technical difficulties in the first phase of data processing needed for obtaining the data from the original format on the tape to our computers (see Section 1). The number of baskets and the number of generated rules for each week are shown in Figure 6.

Some of the rules repeat in different numbers of weeks, Figure 7 shows the number of different triples of items over the number of the triple repetitions. For each week the top 3000 highly ranked rules were selected as described in Section 5.2), meaning that we have 54 000 rules selected in 18 weeks. Since many of the rules actually repeat in different weeks, there are 5856 different rules in the union of all the highly ranked rules of all 18 weeks. In these rules, 1952 different triples of items occur and 1059 of them occur exactly in one week, 377 repeat in exactly two weeks, etc (see Figure 7). Notice that the number of rules is exactly three times the number of triples, since each triple occurs in three distinct

(a)



(b)

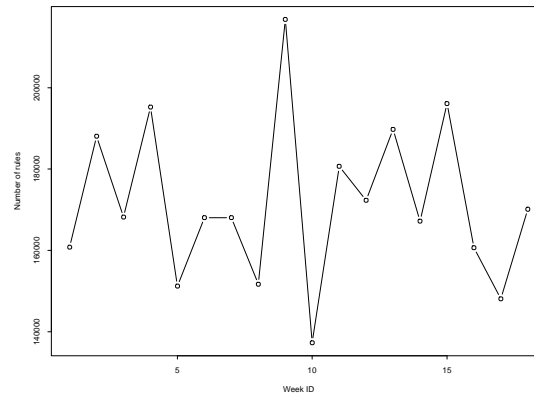


Figure 6: We show for all the weeks ordered in time starting with mid April through mid October, (a) the number of baskets for each week and (b) the number of association rules generated from that baskets when the minimum support was set to 0.1%.

rules each having one of the triple items on the left side of the rule and the remaining two items on the right side of the rule. The reason for that is that our minimum confidence used by the rule generation algorithm was set to a very low value (0.1%). There are 780 rules with 260 different triples of items that repeat in the 3000 highly ranked rules (see Section 5.2) of all the weeks. For illustration, we show some of these rules in Table 1. Some of the rules are formed from a frequent pair, such as SPAGHETTI SAUCE and PASTA or SHAMPOOS and CONDITIONER-RINSES, that is combined with different frequent single items, such as BANANAS or TOMATOES. As can be seen from Figure 9, the same rule has a different support in different weeks. In our illustration, support and confidence are shown for the mid May week in 2000.

5.2 Ranking Association Rules

The number of association rules derived from the data is usually very high. The basic way of filtering potentially interesting rules is performed by the algorithm itself, where only rules with sufficiently high support and confidence are considered. Association rules can be sorted according to their support and confidence and presented to the user, but their number is still usually rather high. In order to further reduce the number of association rules and enable the user to focus on “interesting” rules only, a number of approaches have

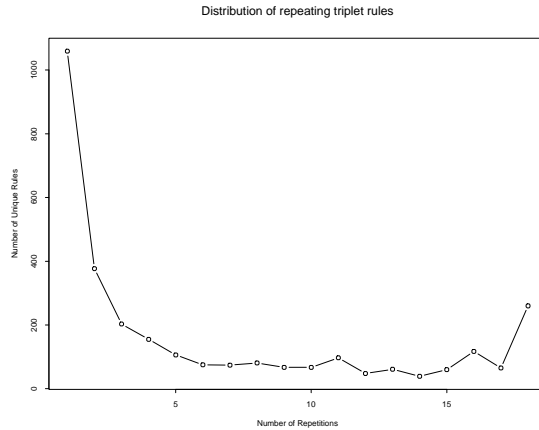


Figure 7: Number of repeating triples of items over the number of the triple repetitions.

been proposed. One of the first approaches to finding interesting rules is based on the user providing *templates* specifying which literals occur in the antecedent and which in the consequent [10]. A special case of that is work on association rules for classification [7], where only one literal can occur in the consequent.

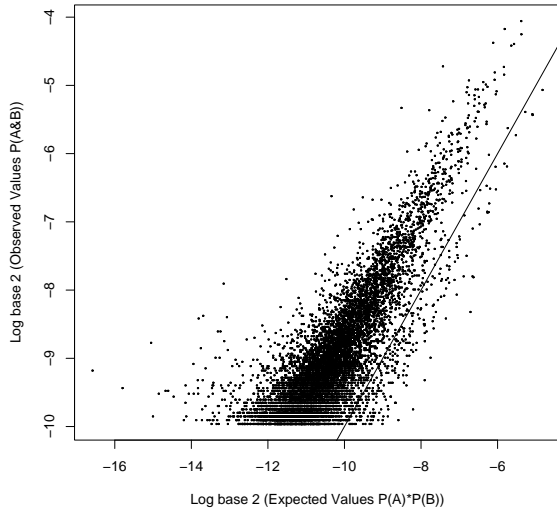
We reduced the number of rules by imposing a ranking on the rules and keeping only the highly ranked rules. For each rule we calculated its *unexpectedness* by comparing the support of the rule with the estimate of support based on the item independence assumption. Specifically, we compared the squared difference between support and estimated support with estimated support. Large values of this statistic make large contributions to the chi-squared test proposed in [2]; see also, [5]. However, we didn't use the chi-squared test for cutting off the top rules since we had no evidence that our data would follow the chi-squared distribution. Instead, we kept the top 1000 rules out of between 40 000 and 400 000 rules, depending on the store and week. We also considered keeping all the rules with the statistic above some threshold, but that resulted in very large differences in the number of rules between the stores.

Figure 8 shows the relationship between observed and expected probability estimates of items for one store, considering pairs and triples of items.

6 Meta Mining

Most data mining approaches concentrate on extracting interesting properties directly from the collected data. There are different proposals on how to post-process the results in order to focus only on interesting properties, as described in Section 5.2. An other way

(a)



(b)

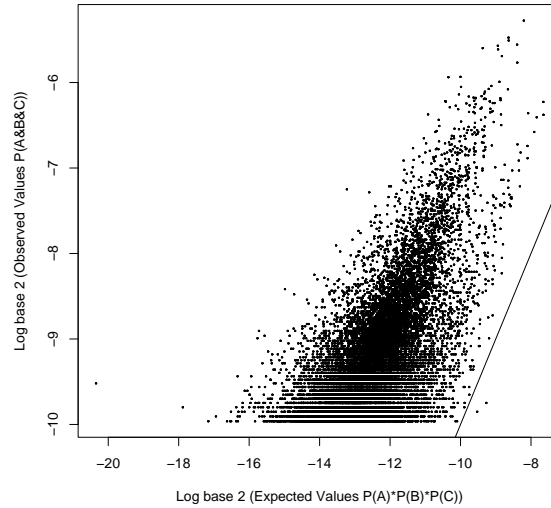


Figure 8: Relationship between observed and expected probability estimates of items for one store, considering (a) pairs of items and (b) triples of items occurring in at least 0.1% of baskets.

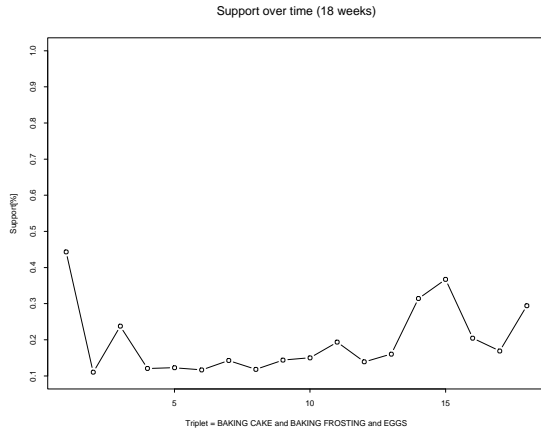
of post-processing is mining the results of other data mining algorithms. For instance, finding spatio-temporal information by mining the induced rules [8]. Inferring higher order rules from association rules as proposed in [11] involves applying time series analysis on support and confidence of an association rules, that is repeating over time. This assumes that we have enough data collected over time and that the same association rules repeat over time. In our data, we found 260 triples repeating over all the weeks. Support for two of them over all the weeks is shown in Figure 9, (a) shows the changing support of triple BAKING CAKE-BROWNIE-COOKIE and BAKING READY-2-SPREAD FROSTING and EGGS while (b) shows the support of triple SPAGHETTI SAUCE and PASTA and FROZEN POULTRY.

6.1 Intersection of rules sets over time and stores

In our efforts to determine how similar our data was both across time and between stores we computed a varying strictness intersection between stores and weeks, which we called a squishy intersection. For this intersection we used a subset of our ranked rules. See Section 5.2 We took the top 1000 rules from each store for each of nine consecutive weeks. Then the squishy intersection was done in two ways.

Method A involved varying levels of intersection between all of the stores within a week followed by varying levels of intersection across 9 consecutive weeks. Such that we have all of the possible intersections that included the rules which occurred in at least

(a)



(b)

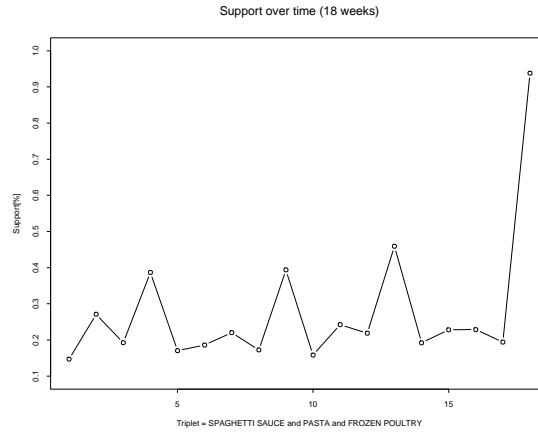


Figure 9: Support of a rule changing over time (all 18 weeks from mid April through mid October). (a) For the first triple from Table 1 and (b) for the fourth triple from Table 1.

50% of the stores and occurred in any number of weeks.

$$\tilde{\Pi}_w \tilde{\Pi}_s \mathcal{R}_{ws}$$

Method B reversed the order of intersection that was used in method a producing the intersection of each store over the 9 consecutive weeks at varying levels of strictness then intersecting all of the store sets at the levels of 50%-100% inclusion:

$$\tilde{\Pi}_s \tilde{\Pi}_w \mathcal{R}_{ws}$$

give the graphs of intersection size depending on the threshold on the minimum number of weeks and the minimum number of stores that the rule should occur in to get in the intersection.

a) that are repeating in different stores

The following four rules repeated in 86% of the stores in 9 consecutive weeks (15-May through 10-July):

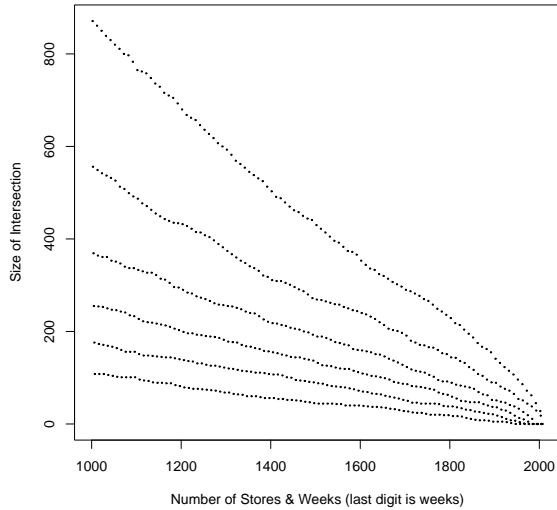
- (RTE KIDS or RTE WHOLESOME FAMILY) and SPAGHETTI SAUCE and PASTA
- SPAGHETTI SAUCE and PASTA and TOILET TISSUE
- SPAGHETTI SAUCE and PASTA and TOMATOES
- MEXICAN FOODS and SHREDDED CHEESE and TOMATOES

If we change our restrictions slightly, such that a rule needs to appear in at least 86% of the stores in 8 of the 9 weeks (15-May through 10-July) we get 48 rules such as:

- MEATS DELI and LETTUCES and TOMATOES
- TOILET TISSUE and PAPER TOWELS and FACIAL TISSUES
- TOILET TISSUE and PAPER TOWELS and SOAPS - HAND & BATH
- VEGETABLES and CARROTS and PEPPERS

Similarly, if we reduce our restrictions further, for rules that appear in at least 80% of the stores in each of 8 weeks we have 84 rules containing items similar to:

(a)



(b)

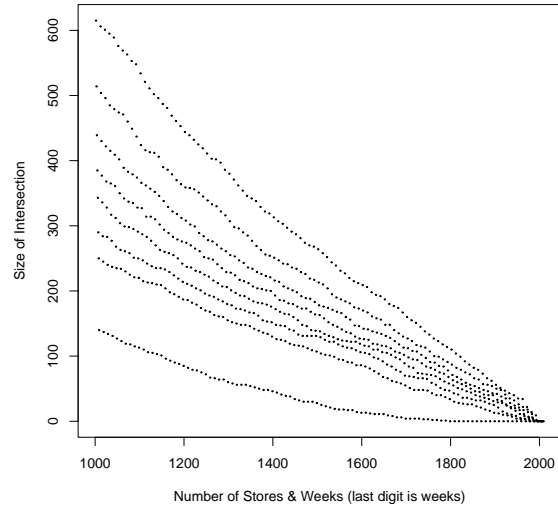


Figure 10: (a) Squishy Intersection Across Weeks then Between Stores (b) Squishy Intersection Within Weeks then Across Weeks Each line represents an increase in the strictness of w from the equation in 6.1

- CHICKEN and SPAGHETTI SAUCE and PASTA
- WIENERS and POTATO CHIPS and BUNS
- NO BAKE CAKE & PIE MIX and MEATS DELI and BREADS-ROLLS

b) Switching the order of the intersection we then have more rules that occur with greater regularity.

The most complete coverage are the two rules which are in 96.5% of the stores in all 9 weeks. These rules are:

- .GROUND BEEF and SPAGHETTI SAUCE and PASTA
- VEGETABLES and LETTUCES and TOMATOES

When reducing the strictness of the intersection to 8 weeks instead of 9, the greatest coverage we get is 98% of the stores all contain the rule:

- .GROUND BEEF and SPAGHETTI SAUCE and PASTA

If we look at the same levels as in the first type of intersection we have: 86% of stores, 9 weeks = 24 rules 86% of stores, 8-9 weeks = 46 rules 80% of stores, 8-9 weeks = 70 rules

6.2 Clustering Stores

Possible approach to mining association rules is by clustering sets of rules. In our case, we have the data collected over about a year showing changes in time, as well as over different

stores showing changes over locations. We applied hierarchical clustering on the stores for each week separately, using the matrix of distances between the stores calculated as described in Section 6.2.

Calculating distance between rule sets

In order to perform clustering, we have to define a distance between rule-sets. We adapted Hausdorf distance between two sets as follows.

$$dist(A, B) = \max[\sum_i dist(a_i, B), \sum_j dist(b_j, A)]$$

$$dist(a_i, B) = \min_j dist(a_i, b_j)$$

where $dist(a_i, B)$ is the distance between a rule from set A and the whole set B and $dist(a_i, b_j)$ is the distance between two rules. For distance between two rules we used commonly used symmetric distance between association rules known as syntax-based or item-based distance [9] defined as

$$X \ominus Y = (X - Y) \cup (Y - X)$$

We calculate the distances between the stores using the reduced set of rules based on the ranking, as described in 5.2 and selecting the top 1000 rules. We also performed some initial experiments selecting the rules with quality higher than $Threshold = 2000$ which resulted in distances spread between 52 and 20 500 as shown in Table 2. For illustration, we show the selection of top rules according to the quality threshold or rules rank influences distances for one of the stores. In case of using the quality threshold, the closest store to store 0002 is store 0617 with distance 533 and the least similar store is 0010 with distance 10850. The large distance was mainly due to the fact that store 0010 has much more rules than the other stores. Originally store 0010 has 236 760 rules that are reduced to 22 009 using the cut-off, while store 0002 has 135 303 rules that are reduced to 3099. Taking a fixed number of 1000 best rules for each store resulted with distances between 383 and 1382 (see Table 2). Store 0002 is the closest to store 0060 with distance 646, while the distance to store 0617 is now 769. The least similar store is now 0040 with distance 1199, while store 0010 is now at distance 546. Notice that in that way most of the selected rules contain three items, since the two items rules are less *unexpected* and thus ranked lower.

- BAKING CAKE-BROWNIE-COOKIE <- BAKING READY-2-SPREAD FROSTING and EGGS (0.11%, 71%)
 - EGGS <- BAKING READY-2-SPREAD FROSTING and BAKING CAKE-BROWNIE-COOKIE (0.11%, 32%)
 - BAKING READY-2-SPREAD FROSTING <- BAKING CAKE-BROWNIE-COOKIE and EGGS (0.11%, 24%)
- SPAGHETTI SAUCE <- PASTA and FROZEN POULTRY (0.27%, 55%)
 - PASTA <- SPAGHETTI SAUCE and FROZEN POULTRY (0.27%, 52%)
 - FROZEN POULTRY <- SPAGHETTI and SAUCE PASTA (0.27%, 11%)
- PASTA <- SPAGHETTI SAUCE and TOMATOES (0.67%, 47%)
 - SPAGHETTI SAUCE <- PASTA and TOMATOES (0.67%, 44%)
 - TOMATOES <- SPAGHETTI SAUCE and PASTA (0.67%, 28%)
- PAPER TOWELS <- TOILET TISSUE and PAPER NAPKINS (0.22%, 50%)
 - TOILET TISSUE <- PAPER TOWELS and PAPER NAPKINS (0.22%, 45%)
 - PAPER NAPKINS <- TOILET TISSUE and PAPER TOWELS (0.22%, 14%)
- SHAMPOOS <- CONDITIONER-RINSES and BANANAS (0.14%, 61%)
 - BANANAS <- SHAMPOOS and CONDITIONER-RINSES (0.14%, 28%)
 - CONDITIONER-RINSES <- SHAMPOOS and BANANAS (0.14%, 26%)
- VEGETABLES <- LETTUCES and CARROTS (0.81%, 68%)
 - LETTUCES <- VEGETABLES and CARROTS (0.81%, 38%)
 - CARROTS <- LETTUCES and VEGETABLES (0.81%, 26%)

Table 1: Example rules that repeat in all 18 weeks, if we generate association rules ignoring the store information. We show all three rules for the selected triples of items. For each rule we give its support and confidence from the mid May week. Support is showing the fraction of baskets containing all the rule items, thus it is the same for all three rules containing the same tripped of items. Confidence is showing the proportion of baskets containing the pair of items in the right side of the rule that also contain the item in the left side of the rule.

Distances statistics	Cut-off on quality	Best 1000 rules
Min.	52	383
1st Qu.	615	659
Median	1093	742
Mean	1698	760
3rd Qu.	2047	848
Max.	20500	1382

Table 2: Basic statistics of distances between the store pairs when a) taking from each store all the rules with the quality $\frac{(Support-EstimatedSupport)}{EstimatedSupport} > 2000$ or b) taking from each store the best 1000 rules regardless of their quality.

Store	Cut-off		Best 1000	
	Min. distance (store)	Max. distance (store)	Min. distance (store)	Max. distance (store)
0002	533 (0617)	10 850 (0010)	646 (0060)	1199 (0040)
0010	1343 (0624)	20 499 (0223)	383 (0008)	1269 (0038)

Table 3: Example stores showing influence of using the cut-off or best 1000 rules.

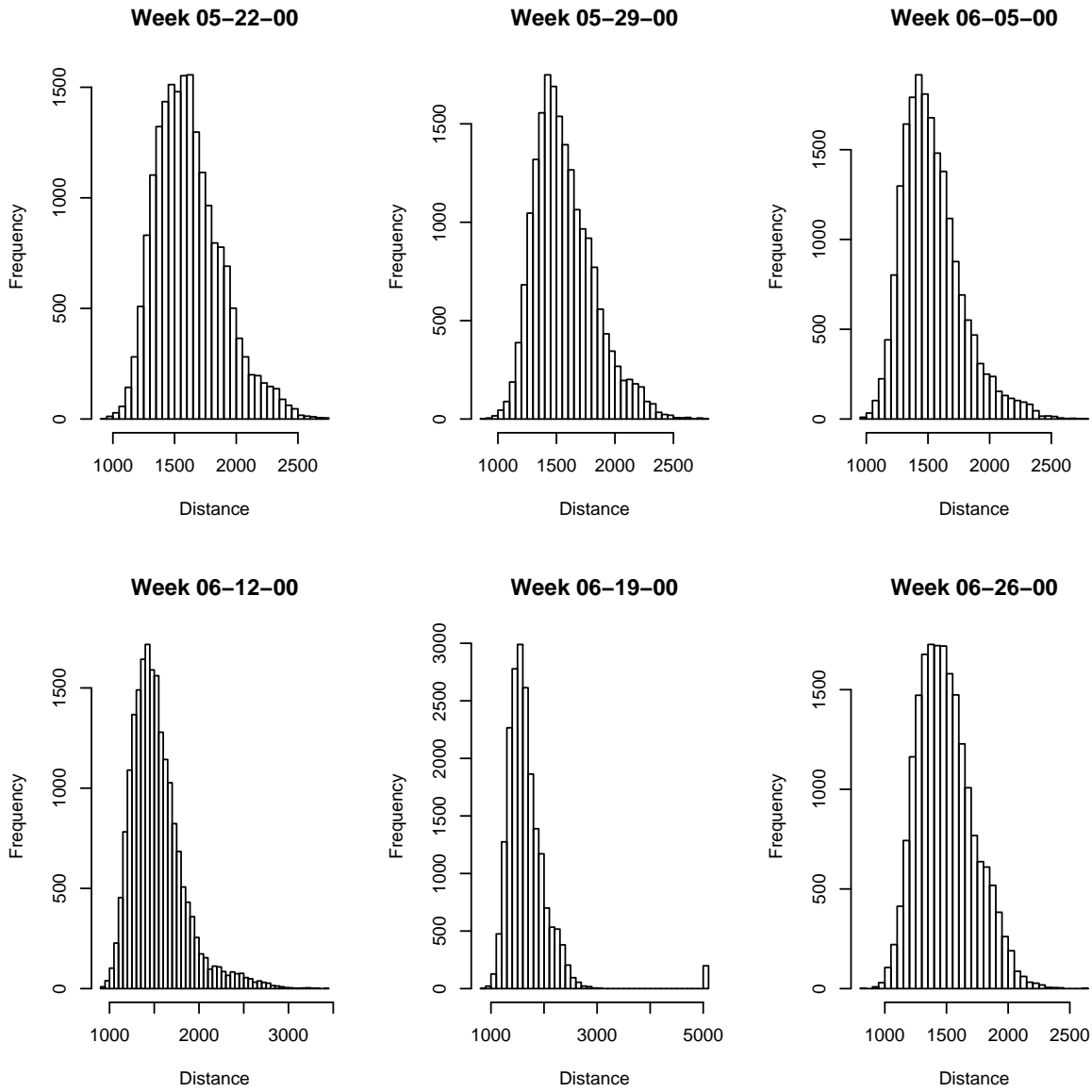


Figure 11: Histogram of distances between the stores in different weeks. We show here the frequency of different distances in weeks of May and June 2000.

Clustering results

Hierarchical clustering of the stores for each week grouped the stores based on the corresponding association rules using the distance between the rule sets as described in 6.2. We used Grouper clustering C++ library, which is a part of Data and Text Mining library [6]. The system was given a matrix with distances between the stores for each week. We found that there is a group of six stores that cluster together in each week. Two of these six stores are always clustered at the first clustering hierarchy level, and these two are located in the same city. The remaining four stores form a hierarchy in different combinations depending on the week and they are all located in the same city but different than the first two stores. Looking to the Minimum Spanning Tree generated for each week, we found that the same six stores are close to each other over most of the weeks.

References

- [1] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, and A. I. Verkamo. Fast discovery of association rules. In U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy (eds.) *Advances in Knowledge Discovery and Data Mining*, AAAI Press/The MIT Press, pp. 307–328, 1996.
- [2] Brin, S., Motwani, R., and Silverstein, C. (1997), Beyond Market Baskets: Generalizing Association Rules to Correlations, In *Proceedings of the ACM Conference on Management of Data (SIGMOD-97)*, pp. 265-276, Tucson, Arizona, USA.
- [3] C. Borgelt. Apriori. <http://fuzzy.cs.Uni-Magdeburg.de/~borgelt/>.
- [4] SPSS, Clementine. <http://www.spss.com/clementine/>.
- [5] DuMouchel, W. and Pregibon, D. (2001), Empirical Bayes Screening for Multi-Item Associations, In *Proc. KDD, 2001*, ACM.
- [6] Grobelnik, M., Mladenic, D. (1998), Learning Machine: design and implementation, *Technical Report IJS-DP-7824*, Department of Intelligent Systems, J.Stefan Institute, Slovenia, January, 1998.
- [7] B. Liu and W. Hsu and Y. Ma (1998), Integrating Classification and Association Rule Mining, In *Knowledge Discovery and Data Mining*, pp. 80-86.
- [8] Abraham, T. and Roddick, J. F. (1998), Opportunities for knowledge discovery in spatio-temporal information systems, *Australian Journal of Information Systems* 5(2), pp. 3-12.

- [9] G. Dong and J. Li, (1998), Interestingness of discovered association rules in terms of neighbourhood-based unexpectedness, in *Second Pacific-Asia Conference on Knowledge Discovery and Data Mining: Research and Development* (X. Wu, R. Kotagiri, and K. Korb, eds.), Melbourne, Australia, pp. 72-86, SpringerVerlag.
- [10] Klemetinen, M., Mannila, H., Ronkainen, P., Toivonen, H., and Verkamo, A.I. (1994), Finding interesting rules from large sets of discovered association rules, *Proceeding of the Third International Conference on Information and Knowledge Management (CIKM-1994)*, Gaithersburg, Maryland, pp. 401-407.
- [11] Spiliopoulou, M. and Roddick, J. F. (2000), Higher Order Mining: Modelling and Mining the Results of Knowledge Discovery, *Data Mining II - Proc. Second International Conference on Data Mining Methods and Databases*, Cambridge, UK, WIT Press. (Ebecken, N. and Brebbia, C. A., Eds.), pp. 309-320.
- [12] J. Ross Quinlan (1993), *C4.5: Programs for Machine Learning*, Morgan Kaufmann Publishers, Inc.