Data mining of baskets collected at different locations over one year

Dunja Mladenić*+, William F. Eddy+, Scott Ziolko+
*J.Stefan Institute, Ljubljana, Slovenia
+Carnegie Mellon University, Pittsburgh, PA, USA
e-mail: Dunja.Mladenic@ijs.si, http://www-ai.ijs.si/DunjaMladenic/

Keywords: Data Mining, Meta Mining, market basket analysis, association rules, decision trees

Received: June 30, 2001

This paper describes the first steps in analysis of millions of baskets 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. 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. 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 form 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 obtain are in EBCDIC (rather than ASCII) and contain fields which are packed decimal and other non-standard formats. 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 the names 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

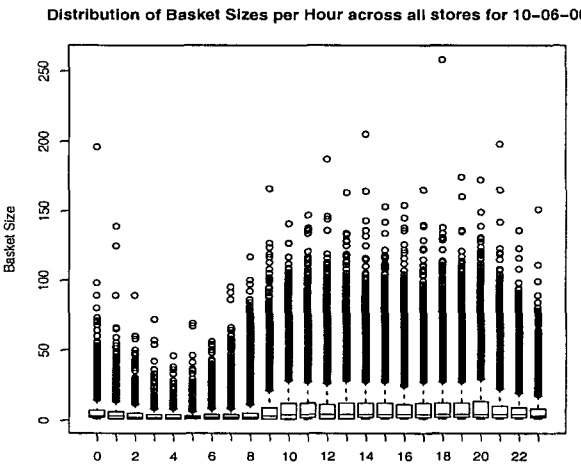

Figure 1: Side-by-side boxplots showing the distributions of basket sizes for each hour of one day across all stores. As expected the most items are purchased during the daytime (Hour 10 to 20 meaning 10 a.m. to 8 p.m.). Notice also that a considerable number of baskets have around 100 items.

"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.

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.

3 Decision Trees

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 decision tree algorithm (an extension of C4.5 proposed by Quinlan[8]) deriving a rule set from a decision tree. 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 some of the rules generated by the program. Figure 2 shows some of the rules derived (with pruning set to severity 75 and at least 10 examples per node) from 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.

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 2: Some of the decision tree rules for the hour from 4 A.M. to 5 A.M. The two numbers in brackets show the number of baskets covered by the rule and the fraction of baskets for which the rule holds.

The rules for the hour with the highest number of the baskets for the same day (5 P.M. to 6 P.M.) are given in Figure 3. 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 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.

Rules for Medium (a subset of 93 rules):

```
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)
```

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)
```

Figure 3: Some of the decision tree rules for the hour from 5 P.M. to 6 P.M.. Notice that UNK is used for for Unknown. 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.

4 Association Rules

In order to find associations between the items in the data, we used association rules. 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 [2] using the publicly available implementation [4], a version of which is incorporated in the commercially available data mining package "Clementine-SPSS".

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 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 [3]; 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.

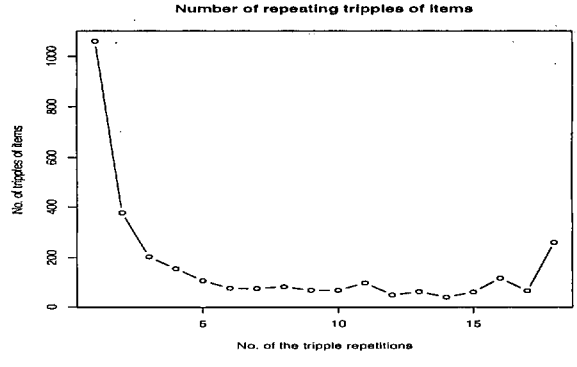

Figure 4: Number of repeating triples of items over the number of repetitions of that triplet.

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. Some of the rules repeat in different number of weeks, Figure 4 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, meaning that we have 54 000 rules selected in 18 weeks. Since many of the rules actually repeat in different weeks, there is 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 4). Notice that the number of rules is exactly three times the number of triples, since each triple occurs in three distinct rules each having one of the triple item 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 agorithm 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 of all the weeks. For illustration, we show some of that 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. The same rule has a different support in different weeks.

5 Discussion

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. An other way of post-processing is mining the results of other data mining algorithms, applying *Meta-Mining*. For instance, finding spatio-temporal information by mining the induced rules [1]. Inferring higher order rules from association rules as proposed in [7] 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.

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

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

By applying a simple intersection on the sets of the most "interesting" rules found for different weeks and different stores, we additionally reduced the number of rules a person may want to check when searching for "stable rules" repeating either over time or/and over different locations (stores). More sophisticated methods are needed to obtain additional information from the generated rules, such as clusters of similar stores or time series analysis of the selected rule support or confidence. These are part of our on-going work investigating items in our every-week growing database of transactions.

Acknowledgement

The work reported in this paper was supported by the Slovenian Ministry of Education, Science and Sport.

References

[1] 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.

[2] 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.

[3] Brin, S., Motwani, R., and Silverstein, C. (1997), Beyond Market Baskets: Generalizing Association Rules to Correlations, In *Procedeengs of the ACM Conference on Management of Data (SIGMOD-97)*, pp. 265-276, Tucson, Arizona, USA.

[4] C. Borgelt. Apriori. http://fuzzy.cs.Uni-Magdeburg.de/~borgelt/.

[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 impelmentation, *Technical Report IJS-DP-7824*, Department of Intelligent Systems, J.Stefan Institute, Slovenia, January, 1998.

[7] Spiliopoulou, M. and Roddick, J. F. (2000), Higher Order Mining: Modelling and Mining the Results of Knowledge Discovery, *Proc. 2nd Int. Conference on Data Mining Methods and Databases*, WIT Press. (Ebecken, N. and Brebbia, C. A., Eds.), pp. 309-320.

- 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.

[8] J. Ross Quinlan (1993), C4.5: Programs for Machine Learning, Morgan Kaufmann Publishers, Inc.