Knowledge Mining Convenience Store Sales Data

Crockett HOPPER, Charleen XU, Matt STANTON

Submitted in Partial Fulfillment of the Requirements of ESE380L, Spring 2004, Dr. J. Ghosh

I. INTRODUCTION

The intent of this project is to use freely available data mining tools to perform knowledge discovery on market basket transactions in a convenience store database. The chain is headquartered in San Antonio, Texas and consists of 130+ retail locations the Dallas and San Antonio areas. The dataset consists of customer purchases for 40 retail locations in San Antonio that occurred between January and mid-March of 2004. There are two major objectives for this project:

1. Profile customers who purchase *lotto* tickets and *lottery* tickets to determine what other types of items they purchase, if any.

2. Employ the OPPOSSUM[1] algorithm and CLUSION[2] toolkit to identify and visualize clusters of transactions similar in revenue, clusters of transactions similar in margin, and clusters of customers making purchases that are similar in terms of their items' contribution to total basket revenue.

II. INDUSTRY BACKGROUNDER

A. Overview of the C-Store Industry

The convenience store industry is relatively mature, having been in existence for well over 50 years. It is a crowded and highly competitive market space where competitors find that shrinking margins on fuel require that profits be realized on the merchandise inside the store. In fact, what separates a successful c-store chain from those that fail fails is the ability to generate higher-margin inside sales. These inside items typically have a *gross* profit margin of 25% or more (amounting to approximately 13% *net* profit), compared to *gross* gasoline margins that have hovered at less than 8% *gross* profit margin for the last two years (less than 1% *net* profit). The market forces that drive fuel margins so low are beyond the scope of this paper, but it is sufficient to recognize that c-stores realize little profit in the retailing of fuel – relying instead on adding higher margin merchandise to the volume of transactions generated by the fuel stop.

B. The Lottery Customer

Another interesting aspect of the c-store industry is the lottery. There are two types of state-sanctioned legalized games of chance; one is termed *lotto,* referring to the games in which players pick combinations of numbers and feed their selections into a centralized computer system, and the other is termed *lottery,* where the player can win or lose instantly by scratching off areas of game cards.

Virtually all c-stores sell both lotto and lottery tickets as another method of getting traffic into the store. Lottery proponents argue that selling lottery and lotto increases the tendency of customers to buy on impulse and that selling lotto and lottery attracts customers who would otherwise have shopped at a grocery store to make a single stop at the c-store for food and lottery tickets.

While this may be true, there are considerable costs associated with selling lottery and lotto tickets. Significant overhead is required to track the inventory and sale of the game pieces, and the tickets are a common target of theft. The low margins on lottery sales coupled with the level of security required to maintain the inventory require that a notable corollary benefit be present to justify their sale.

The generally accepted industry position is that lottery sales drive high-margin, inside-the-store purchases, but the lack of detailed transactional information makes it difficult to prove such a position. And a counter case can readily be made that often lottery purchasers buy *only* tickets and then consume an inordinate amount of time in line requiring the cashier to check the winnings of those tickets. It can be reasonably argued that sales are actually lost when customers in a long line held up by a lottery customer defect from the queue and go elsewhere.

C. Maximizing Potential

Given that the two highest-volume transactions in a c-store are both low-margin items, it is imperative that c-store chains maximize the potential of inside sales. To this end, c-store chains must do a better job of identifying customer purchasing patterns and strategically applying this information to better select and place products in retail locations. Moreover, this information can then be employed to create innovative, targeted marketing campaigns that entice customers into the stores to purchase high-margin items.

The task of analyzing customer buying habits is made difficult by the fact that most c-stores in the industry use the retail method of accounting, that tracks inventory and sales only by generalized categories. It is fast and inexpensive, but imprecise compared to the item level method of accounting where each item in the store is tracked from receipt at the loading dock to final sale. Even the few chains that use the item level method of accounting do a poor job of tracking sales, instead choosing to aggregate the data for an entire day.

In this experiment, the data from a pilot program presented a rare opportunity to knowledge mine a dataset at a transactional level for 40 stores participating in a proof-of-concept loyalty program. This dataset provides transactional data for all patrons (both loyalty and non-loyalty customers) in the 40 pilot stores.

III. DATA ACCUMULATION AND PREPROCESSING

A. Data Integrity

For the 40 stores in the pilot project, all customer purchases were electronically captured for four months and transmitted near-real-time to a central data warehouse in Virginia.

The transactions were captured by electronically "sniffing" the Point of Sale (POS) terminal receipt printer queues. As a result, item identifiers were limited to a simple 10-character description field and the associated retail price. The vendor of the sniffing hardware then employed a custom algorithm to parse the item data from the other information appearing in the registers' print streams, such as totals, store identifiers, etc. One problem with this method is that the parsing algorithm sometimes failed and misidentified "junk" data as valid. There were approximately 2,800 records in the initial dataset where the item data showed invalid transaction entries such as date, total and tax amount of the transaction. Microsoft T-SQL routines were executed on a Microsoft SQL Server 2000 instance to remove these erroneous basket entries.

B. Data Refining

A profound difficulty with this method of printer sniffing to capture transactional data is that often there are multiple items that have the same description and retail price. An example is a description that reads *FL RUFFLES* with a price of $.99. Linking this description to the master price book server returned 10 unique products (e.g. Frito Lays Ruffles Chips, Frito Lays Ruffles Jalapeno Dip, Frito Lays Ruffles Bar-B-Que Chips, etc.) with this exact description and retail. To identify the correct product in a statistically maximal manner, a rudimentary filtering algorithm was implemented in T-SQL using data from previous months for each store. Thus, given the appearance of the *FL RUFFLES* description and retail price in a particular basket, the generic item description was mapped to a specific item in proportion to the actual sales of the given item. This resulted in a unique product id, weighted average cost per item and retail price per item.

C. Sampling

After the initial dataset was cleaned and the retail prices and weighted average costs were incorporated, the resulting dataset contained roughly 2.8 million records, corresponding to roughly 1.79 million market baskets. The number of unique inventory identifiers in this dataset totaled 2,794. In order to reduce the scope, the Java Math.Random libraries were used to randomly sample records from the database resulting in a working dataset of approximately 68,000 line items in 22,541transactions .

D. Dimension Reduction

It is possible to generate a similarity matrix from a $n \times d$ matrix where n=23,000 and d=2,794, but initial testing revealed that this was beyond the computational capabilities available for this project. It was therefore necessary to roll-up the 2794 individual products into more general categories. This resulted in 52 distinct categories [Table1], yielding an order of magnitude reduction in dimensionality.

E. Data Representation

After cleaning, sampling and dimension reduction, we created two datasets. One tracked extended revenue per basket for each of the 52 categories, while the other tracked weighted gross margin for the categories.

TABLE 1. *Dimensional Roll-Up from 2,794 Products to 52*

AUTO SUPPLIES	GENERIC CIGARETTES	SCHOOL/OFFICE SUPPLIES
BEER	GUM	SNUFF
BREAD_& PASTRY	HBA (HEALTH AND BEAUTY AIDS)	SOAP
CANDY	ICE CREAM	SODA
CANNED FOOD	JERKY	SPORT DRINKS
CAPPUCINO	JUICE	SUPPLIES
CARWASH	LOTTERY	TEA
CHEW	LOTTO	VALUE PRICED CIGARETTES
CHIPS	MAGAZINES/BOOKS	WATER
CIGARS	MINTS	WINE
COFFEE	NEWSPAPERS	WINE COOLERS
COOKIES	NUTS	
CRACKERS	OIL	
DAIRY, EGGS, MEAT	OTHER	
FILM AND BATTERIES	OTHER_GROCERY	
FLAVORED WATER	PET SUPPLEIES	
FOOD	PHONE CARDS	
FOUNTAIN	PLU	
FROZEN FOOD	PREMIUM CIGARETTES	
FUEL	PREPARED SANDWICHES	

Since the convenience store industry typically sells merchandise that varies from around $.10 to at most $5.00 we thought it might be interesting to look at margin as our feature of interest. Our baskets, because of the nature of the industry we were tracking, did not exhibit a wide spectrum of possible prices. Furthermore, we found that roughly half of the baskets in our final dataset contained two items or less. Therefore, most of our baskets were quite small, which is indicative of the type of buying that occurs at most convenience stores. We did not attach much importance to fuel purchases since fuel contributes virtually nothing to profit for the reasons explained earlier.

F. Table Rotation

To effectively mine our data, we needed to transform our ≈68,000 records so that each row represented a single transaction, with the columns representing the revenue or margin for each of the 52 categories. To "pivot" our table we wrote a custom routine in T-SQL.

IV. RESULTS

A. Lottery Results

With respect to our first objective—to analyze the purchasing behaviors of lottery customers—we found that historical empirical observations are confirmed by the data. Customers playing games of chance (lottery and lotto) do not tend to purchase other items and, in the event that they do, they tend to purchase low margin items.

To process our lottery data we first extracted from our refined dataset all records where lotto or lottery tickets were purchased. Of the original 22,541 transactions, 3295 contained these purchases. Of those 3295 "game" baskets, 1449 contained only game purchases and 1846 contained purchases of both games and other items. We then discretized our margin data in the following fashion: LOW MARGIN baskets were identified as those having gross weighted margins less than 20%; MID MARGIN baskets where those falling in the range 20% to 39%; and HIGH MARGIN baskets were those whose margins were 40% and above. We preserved the customer sex dimension but discretized the age dimension as follows: AGE < 30 = youth; AGE 30 to 55=middle age; AGE > 55 = mature age.

Using the Weka *a priori* classifier we discovered the rules appearing in [Figure 1].

We found that, generally, middle age males tend to purchase lottery and lotto games. When those customers purchase the games in combination with other items, the other items tend to be low margin. As a result of this finding, it would seem that selling games of chance in convenience stores benefits only the vendors of the games. Of course, social benefits attach to the selling of lotto and lottery, such as public school financing, etc., but from the standpoint of the retailer, selling games of chance does not appear to drive additional purchases. In fact, there exists perhaps a punitive relationship since the all too frequent scenario of long lottery lines results in lost sales. Quantifying this loss would be an interesting subject for further investigation.

B. Cluster Generation and Analysis

Understanding how to execute the Matlab routines to perform the steps in the OPPOSSUM clustering algorithm and CLUSION visualization toolkit was more challenging than anticipated. As time grew short our team investigated another tool by the author of MeTIS to provide a clustering and visualization implementation that would enable us to meet our proposed objectives. Ultimately we did get the Matlab routines to work on a subset of the data (3000 rows x 52 item categories), but the Cluto and gCluto results are worth presenting on their own.

gCluto is a graphical front-end that drives the Cluto clustering toolkit, and it provides a useful visualization for understanding the contents of a cluster very rapidly. The necessity of generating a similarity matrix as an intermediate step in using CLUSION made the task of mapping the cluster rows back to item categories much more time consuming. This was greatly amplified by our lack of experience using Matlab and the "Excel Link Tool" designed to export the generated matrices into Excel for simpler manipulation.

Figure 2 is a selection from the list of clusters generated by gCluto. Strangely, though the program claims to provide all of the functionality of the Cluto toolkit, we could find no way to use an Extended Jaccard coefficient as the similarity measure. Based upon observations using OPPOSSUM against market baskets, we opted for a Cosine distance measure as a second-best choice.

FIGURE 1.

ASSOCIATION RULES REULTS

BASKET MARGIN

Basket Margin	Purchase Item	Support
Low—>	LOTTO/LOTTERY	.824
Medium—>	LOTTO/LOTTERY	.162
High—>	LOTTO/LOTTERY	.014

AGE

Age	Purchase Item	Support
Youth—>	LOTTO/LOTTERY	.072
Mid. Age—>	LOTTO/LOTTERY	.694
Mature—>	LOTTO/LOTTERY	.023

SEX

Sex	Purchase Item	Support
Male—>	LOTTO/LOTTERY	.065
Female—>	LOTTO/LOTTERY	.035

AGGREGATE RESULTS

∴ General Rule: Middle Age Males Tend to Purchase Lotto/Lottery

Basket margin =Low and AGE=Middle Age —>LOTTO=Yes
support = 0.45 conf = 1

Gender=Male and AGE=Middle Age —>LOTTO=Yes
support = 0.44 conf = 1

Basket margin =Low and GENDER=Male—>LOTTO=Yes
support = 0.422 conf = 1

The program accepted a defined number of clusters in the generation phase, but appeared to ignore that setting when generating clusters with the Graph Partitioning option selected. We had attempted to get 10 clusters out of our dataset of 24,000 transactions, but the program generated 120 clusters instead – many of them extremely small. There were still several interesting clusters however, and that is what is displayed in Figures 2 and 3.

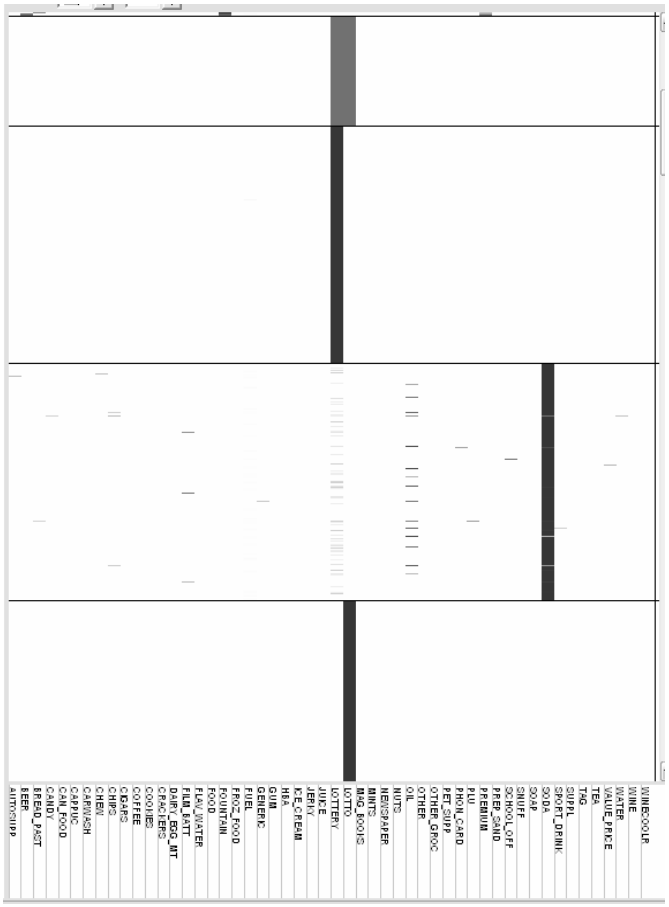
FIGURE 2.

Descriminating	SODA	24.4%	PREMIUM	20.1%	NEWSPAPER	15.2%	BEER	10.3%
Cluster 113	Size: 46	ISim: 1.000	ESim: 0.113					
Descriptive:	BEER	99.9%	FUEL	0.1%	AUTOSUPP	0.0%	MAG_BOOKS	0.0%
Descriminating	BEER	31.6%	SODA	26.3%	PREMIUM	7.7%	CANDY	4.5%
Cluster 114	Size: 244	ISim: 0.517	ESim: 0.109					
Descriptive:	PREMIUM	64.8%	BEER	18.5%	FOUNTAIN	14.5%	FOOD	1.3%
Descriminating	SODA	33.7%	PREMIUM	25.9%	CANDY	5.8%	LOTTO	4.2%
Cluster 115	Size: 374	ISim: 1.000	ESim: 0.071					
Descriptive:	LOTTERY	50.0%	LOTTO	50.0%	MAG_BOOKS	0.0%	MINTS	0.0%
Descriminating	SODA	22.1%	LOTTERY	19.8%	LOTTO	18.6%	BEER	9.4%
Cluster 116	Size: 812	ISim: 1.000	ESim: 0.023					
Descriptive:	LOTTERY	100.0%	FUEL	0.0%	LOTTO	0.0%	MAG_BOOKS	0.0%
Descriminating	LOTTERY	46.3%	SODA	18.4%	BEER	7.8%	PREMIUM	5.4%
Cluster 117	Size: 814	ISim: 0.983	ESim: 0.145					
Descriptive:	SODA	100.0%	OIL	0.0%	LOTTERY	0.0%	FILM_BATT	0.0%
Descriminating	SODA	25.7%	BEER	15.6%	PREMIUM	10.8%	CANDY	6.3%
Cluster 118	Size: 896	ISim: 1.000	ESim: 0.025					
Descriptive:	LOTTO	100.0%	AUTOSUPP	0.0%	MAG_BOOKS	0.0%	MINTS	0.0%
Descriminating	LOTTO	46.0%	SODA	18.6%	BEER	7.9%	PREMIUM	5.5%
Cluster 119	Size: 812	ISim: 0.508	ESim: 0.060					
Descriptive:	FOUNTAIN	95.2%	BEER	1.6%	SUPPL	1.2%	SODA	1.0%
Descriminating	FOUNTAIN	47.5%	SODA	16.0%	PREMIUM	6.4%	BEER	4.1%
Cluster 120	Size: 921	ISim: 0.910	ESim: 0.056					
Descriptive:	PREMIUM	99.8%	AUTOSUPP	0.2%	NUTS	0.0%	LOTTO	0.0%
Descriminating	PREMIUM	40.5%	SODA	21.8%	BEER	9.2%	CANDY	3.7%
Cluster 121	Size: 1507	ISim: 0.739	ESim: 0.063					
Descriptive:	BEER	98.1%	DAIRY_EGG_M	1.2%	PREMIUM	0.7%	CIGARS	0.1%
Descriminating	BEER	41.0%	SODA	24.6%	CANDY	4.2%	FOUNTAIN	3.7%
Cluster 122	Size: 2085	ISim: 0.241	ESim: 0.042					
Descriptive:	GENERIC	47.7%	COFFEE	29.8%	NEWSPAPER	14.2%	CHIPS	3.6%
Descriminating	GENERIC	31.0%	SODA	16.3%	COFFEE	12.5%	BEER	7.2%
Cluster 123	Size: 1939	ISim: 0.179	ESim: 0.046					
Descriptive:	BREAD_PAST	57.7%	CAPPUC	9.9%	ICE_CREAM	9.4%	JUICE	5.1%
Descriminating	BREAD_PAST	30.5%	SODA	19.8%	BEER	9.1%	CAPPUC	7.0%
Cluster 124	Size: 2839	ISim: 0.221	ESim: 0.060					
Descriptive:	WATER	58.9%	SPORT_DRINK	16.0%	SODA	5.4%	SNUFF	4.6%
Descriminating	WATER	47.4%	SPORT_DRINK	11.7%	SODA	10.4%	PREMIUM	4.0%
Cluster 125	Size: 3161	ISim: 0.194	ESim: 0.093					
Descriptive:	SODA	50.9%	BREAD_PAST	10.7%	COFFEE	8.1%	CHIPS	7.3%
Descriminating	BEER	26.8%	LOTTO	8.7%	SODA	7.0%	LOTTERY	6.9%
Cluster 126	Size: 4528	ISim: 0.213	ESim: 0.095					
Descriptive:	SODA	44.5%	CANDY	21.8%	GUM	8.4%	FOUNTAIN	5.6%
Descriminating	CANDY	19.5%	LOTTO	10.3%	GUM	9.8%	PREMIUM	9.5%

Figure 3 is the matrix plot of the rows and columns that makes the type of items present in the cluster readily apparent. The more common item categories are displayed as solid red bars, and it is clear here that there are groups of transactions consisting only of lottery and lotto tickets.

These clusters represent approximately 10 percent of the transaction volume for the c-stores, but very little margin. The gCluto tool provided us a second set of evidence (in addition to the association rules mentioned before) that the argument that lottery drives high-margin purchases is not absolute, and perhaps not even typical. The lottery clusters identified using gCluto account for nearly two-thirds of all lottery purchases.

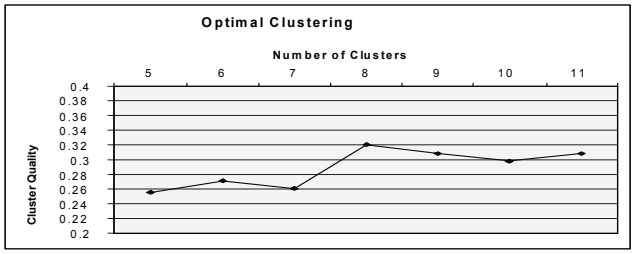
FIGURE 3.



Even though gCluto makes it easy to identify the items in a given cluster, it provides little aid in the way of determining an optimum number of clusters. The evaluation routines provided with the CLUSION toolkit make it simple to determine the optimal number of clusters for a given dataset because the large penalty of generating the similarity matrix is done once and many different clusterings can then be rapidly calculated using MeTIS.

Figure 4 Illustrates the cluster quality measure calculated on a series of MeTIS runs using the evaluation routines in the CLUSION toolkit. Value-based clusterings with a number of clusters from 5 to 11 were tried, with the optimal value (highest quality at lowest number of clusters) at 8. Similar runs were performed for the sample-balanced clusters and the optimum was found at 10 clusters.

FIGURE 4.



Once the optimum number of clusters k was determined, the similarity matrix and the cluster array (map of rows to clusters)

were passed to CLUSION to generate the matrix plot of cluster similarity. Figure 5 shows the value-balanced clusters and figure 6 shows the sample balanced clusters.

FIGURE 5.

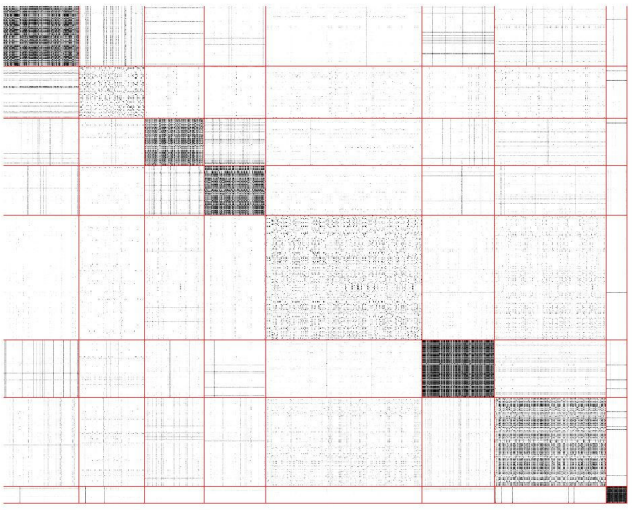
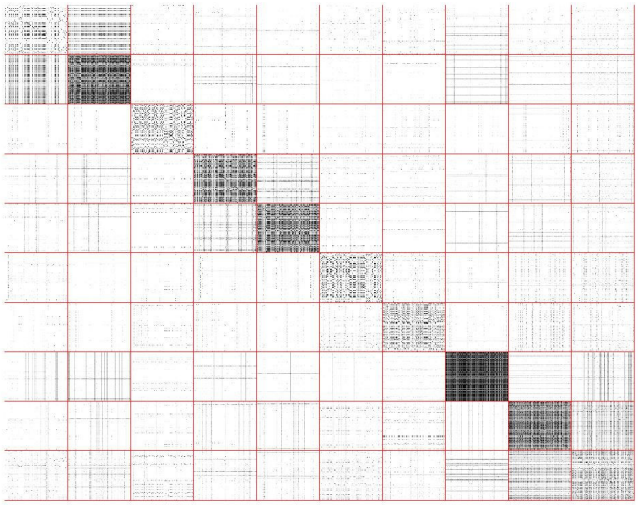


FIGURE 6.



With a pre-calculated similarity matrix, reclustering is very fast and simple. Dealing with the Matlab routines was not without difficulty however. Matlab requires that all operations be performed in main memory; the limits of the 32-bit address space of an Intel Pentium4 computer restricted the size of our dataset to 3000 rows by 52 item category columns. Therefore all results in the CLUSION runs are a further random subset of the initial sample run through the association rules and the gClutotools. The value-balanced clusters consist of a number of transactions in the range [100-752] and revenue in the range [$2395.39-$2419.11]; the sample-balanced clusters each consist of 300 transactions with revenue in the range [$818.48-3623.46]. In general, the value-balanced clustering algorithm produced higher-quality clusters (greater internal and lesser external similarity measures) than the sample-balanced clusters. It is interesting that several types of baskets were illustrated equally well with each. The lotto customer already discussed makes another appearance here in clusters with very few other items. This reiterates the point that the argument that selling lottery tickets is a path to greater profit is invalid.

Another interesting result that arose from using extended Jaccard as a similarity measure was that there are different kinds of lottery customers. They show up distinctly as clusters 3 and 4 in the value-balanced set because the amount spent on each type of game is swapped between the two groups. Had a simple Jaccard coefficient measure been used, the two groups could not have been differentiated. Tracking the total revenue illustrates that there are some lottery players who spend more on the instant win tickets and some who spend more on the longer term game with a larger payout.

One can observe in the CLUSION plot that clusters 3 and 4 are not the darkest among all of them. Had the two not been differentiated, we would expect to see a single, darker cluster for some subset of these transactions; however, the revenue balancing would have placed more of the lotto purchases in other clusters.

It should also be noted that the clusters fell out differently in the sample-balanced case. Here, cluster 4 is primarily lotto tickets, while cluster 5 is a combination of the two. In the value-balanced clusters, it seems that the requirement to keep the revenue similar across clusters forced some of the lotto purchases into other clusters.

The value-balanced clusters 1, 6, and 7 are the kinds of transactions that keep the c-stores in business. Each of these clusters is high margin and the majority of baskets contain one of a small set of items. Cluster 1, for example consists primarily of premium brand cigarettes and beer. Cluster 6 is also composed of premium brand cigarettes, but these transactions tended to replace beer with soda. Cluster 7 includes the low-margin fuel category, but makes up for that with a notable percentage of soda, chips, and candy.

Value-Balanced Cluster contents (most common items)

1. Premium, Beer
2. Carwash, Fuel, Generic Cigarettes, Snuff, HBA
3. Lottery, Lotto
4. Lotto, Lottery
5. bread_past, candy, coffee, cookies, dairy_egg_mt, fountain, gum, newspaper, premium, sport_dr, water, chips, cookies, ice_cream, juice
6. Premium, Soda
7. Soda, Chips, Fuel, Candy
8. Fuel

Sample-Balanced Cluster contents (most common items)

1. Beer, Premium
2. Beer
3. Fountain, Generic, Snuff, Carwash
4. Lottery
5. Lottery, Lotto
6. Candy, Fuel, Sport_Drink, Water

7. Bread_Pastry, Dairy_Egg_Mt, Coffee
8. Premium
9. Soda
10. Candy, Chips, Premium, Soda

V. KNOWLEDGE GAINED & INDUSTRY APPLICATION

Our conclusions had to meet two criteria in order to be useful. First, we sought to derive novel information from the dataset. Second, we needed to ensure that our conclusions were reasonable given our understanding of the industry domain. Our results passed both of these tests.

With respect to the gaming customers, we proved through association rules, and both sample and revenue balanced clustering that these customer types tend to purchase games in exclusion to other items. In our cluster analyses, we determined that nearly ¼ of the total revenue in the sample was directly attributable to customers purchasing games and nothing else.

This information is very useful to retailers selling games since it confirms the long held belief that lotto and lottery merchandising neither directly nor indirectly benefits the merchandiser. With this knowledge, retailers might be able to negotiate higher margins with the state lottery commission. Alternatively, it might be interesting to pilot a test project in which selected stores do not sell games. The impact on sales, if there is any, could then be quantified and compared to locations selling games.

In terms of our non-gaming cluster analyses, we showed that revenue balanced clustering is a useful tool for eliciting unique groupings of items whose revenue contributions are represented relative to total basket revenue. With this data, a marketer can actively profile for target marketing the types of customers visiting his retail locations, while simultaneously accounting for the margins and quantities of the items she purchases. In one highly defined cluster, we found a strong correlation between the purchase of premium brand cigarettes and drinks (both soft and alcoholic). This information might translate into a marketing promotion offering discounts on one item when multiples of another item are purchased. This would have the net effect of increasing both basket value and margin since premium cigarettes and soda/alcohol are high margin items. Moreover, the marketer might send via direct mail targeted coupons to these customers in an effort to increase the frequency of store visits.

VI. CRITIQUE

Several aspects of our data impacted the quality of our results. First, our reliance on a matching algorithm to identify items based on generic POS descriptions inevitably led to some misidentifications. To more accurately knowledge mine this industry's data henceforth, a procedure should be adopted that prints the product UPC next to the item. That would ensure much more accurate item classifications and greatly reduce by orders of magnitude the time required to pre-process the dataset.

An additional constraint on the accuracy of our data relates to the complex nature of post-purchase rebating that occurs in the c-store industry. Our margin analyses were based on the weighted average cost as indicated in the master price file. However, this file does not encapsulate the numerous rebating schemes constantly occurring in this industry. For example, soda vendors typically conduct three or more rebating strategies simultaneously. These internal promotions may change on a monthly or semi-monthly basis. Further exploration of data from this industry should have accurate weighted average costs with the rebates pre-factored.

We faced computational constraints with respect to computing the similarity matrices using the extended Jaccard similarity metric. This limitation required taking a sub-sample from our initial dataset.

Finally, our experience with MATLAB, METIS and gCluto is minimal. Having a better understanding of these toolkits would greatly aid our attempts at further data analysis.

REFERENCES

[1] Alexander Strehl and Joydeep Ghosh. Relationship-based visualization of high-dimensional data clusters. In Proc. Workshop on Visual Data Mining (KDD 2001), San Francisco, pages 90-99. ACM, August 2001.

[2] Alexander Strehl and Joydeep Ghosh. Value-based customer grouping from large retail data-sets. In Proc. SPIE Conference on Data Mining and Knowledge Discovery, Orlando, volume 4057, pages 33-42. SPIE, April 2000.