# Uncovering Useful Patterns in Shopping Cart Data

Ali Haider Hussein Ghazala[1], Naeem M. Asif[2], Baig Mirza Farhaan[3], Jamil Noreen[4]
Auckland University of Technology
New Zealand
ahaiderh@aut.ac.nz
mnaeem@aut.ac.nz
em8803@aut.ac.nz
njamil@aut.ac.nz

***ABSTRACT:*** *Understanding the shopping and purchasing behaviours of customers is an essential task for business and retail organizations. While customers look for useful information from retailers as they shop, businesses seek to collect increasing amounts of data in order to deliver added value to their customers. This requires an intensive analysis of sales data. Extracting shopping patterns across the many levels of information is a non-trivial task as datasets on sales transactions can contain many levels of information such as item category, brand name, colour, and price. This paper examines the use of multi-level association rules to uncover purchasing patterns at multiple levels of detail. It shows how different kinds of purchasing patterns can emerge at different association levels of analysis. This type of analysis is indeed helpful in assisting retailers to make wise decisions for their customers.*

**Keywords:** Shopping cart, Patterns recognition, Retail data

**Received:** 13 June 2016, Revised 17 July 2016, Accepted 21 July 2016



## 1. Introduction

Identifying which items, product types, and brand names are purchased together is valuable information for retailers; it can be used to plan sales promotions and loyalty programs, and design stores and discount plans [4]. To better understand how consumers associate shopping items, a deeper analysis into product identifiers such as category, subcategory, and brand names, is required. Many studies have tried to identify such associations by studying customers attitudes through surveys and interviews. However, little attention has been paid to investigating such associations using transactional shopping data. This paper fills this gap by studying the associations between items purchased via multiple levels of data and a real shopping dataset.

Different data mining and analysis techniques such as *association rules mining* can be used to identify relationships among purchased items based on their characteristics. For instance, a low-level association may find that customers who purchased laundry detergent also purchased fabric care products.

Mining purchasing data at multiple levels of detail usually requires a hierarchical data representation of the items. Thus, a concept taxonomy should be provided to generalize lowlevel details to higher-level concepts. Exploring transactional data across different concept levels may highlight interesting patterns such as the association between *Downy* fabric softener and *Tide* washing powder brands. The latter association is expressed at a higher conceptual level and contains more specific semantics than the former example. In many sales applications, the taxonomy information is either stored implicitly in the database, such as *Wonder* wheat bread is a wheat bread, which is in turn a bread, or computed elsewhere [5].

Although mining multilevel association rules using taxonomy information has been studied by many researchers [9], [10], there has been little work from the viewpoint of practitioners to answer questions such as: What kind of sales patterns can be revealed using multilevel association mining? On a cautionary note, the process of mining multilevel association rules requires careful examination of the discovered rules to avoid misleading results. Fig. 1 shows an example of a multilevel taxonomy in transactional data.

The goal of this paper is to investigate the process of applying multilevel association on real shopping dataset. It underlines the different types of shopping patterns that emerges from many levels of information. It also highlights similarities with earlier qualitative studies that analysed shopping patterns using interviews and questionnaire responses.

## 2. Data

A real life transactional database from Sam's Club, a division of Wal-Mart Stores in the USA, was used for this analysis [7]. Each item is classified within a hierarchical system of categories and sub-categories. Higherlevel categories such as *Clothing* and *Furniture* are further subdivided into smaller sub-categories such as *Swimming Attire* and *Beach Chairs*. The dataset also

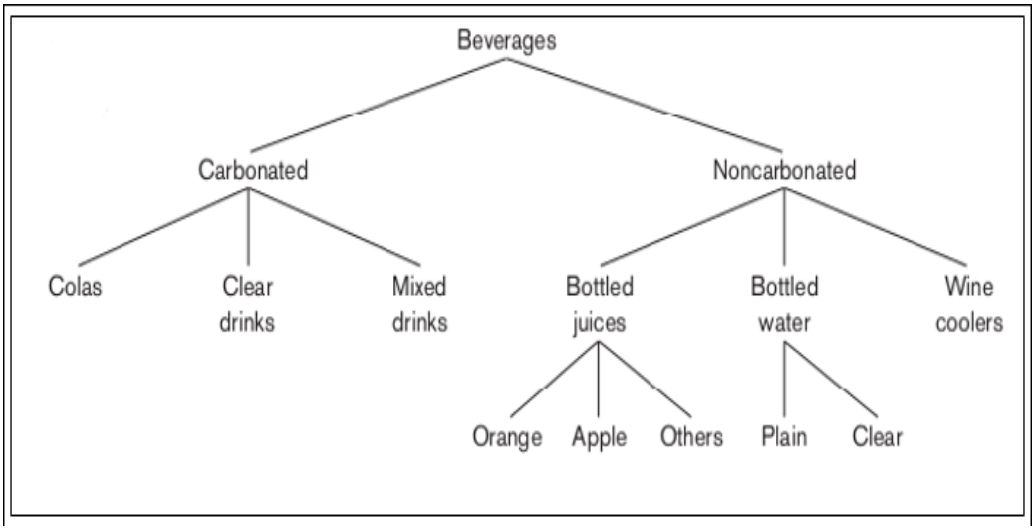


Figure 1. Taxonomy of Beverages

includes other types of metadata, such as Brand Name, for each item. The data cleaning process involved removing duplicate records and extracting missing brand name values from supporting attributes. The final dataset contains 2,698 individual items distributed among 10,405 unique transactions. For each item in the dataset, 3 layers of hierarchical details are used (Sub-Category, Category, and Brand Name). Brand names are missing for some items such as fresh fruit and meat.

### 2.1 Data Analysis

The data mining technique of association rules, also known as market basket analysis, was used to analyse the shopping data

across multiple levels. The method uses the Apriori algorithm [2] for frequent item set mining and association rule learning over transactional databases. The algorithm works by identifying the individual items that are the most frequent in the database and extending them to larger and larger item sets as long as those itemsets appear sufficiently often in the database. The patterns that are discovered are presented in the form of if-then association rules based on attribute-values [1]. An association rule X Y implies there is a high probability of having Y in transactions when X occurs. X and Y are called the antecedent and consequent of the rule respectively. Despite their effectiveness and popularity, Apriori algorithms tend to deliver a large number of rules, and their results should be interpreted with caution [6]. The three most widely used measures for evaluating the usefulness of rules are: support (s), confidence (c), and lift (l).

Support (X Y) = P (X U Y) Confidence (X Y) = P (Y—X)
Lift (X — Y) = P (X U Y)/P (X). P (Y)

The support of a rule is the percentage of transactions in the database that contain both the antecedent and the consequent; i.e. eggs and bread appear in 10 percent of the transactions. The confidence of a rule is the percentage of transactions with X in the database that contains the consequent Y also; example i.e. 40 percent of customers who purchased pretzels also purchased soft drinks. The lift is the ratio of confidence to the percentage of cases containing Y [3], [11]. If purchases of products X and Y are perfectly independent, the lift L (XY) = 1; however, if X and Y appear together more often than we would expect under independence, the lift is greater than 1; otherwise, it is less than one. For example, if a rule for two products has a low support value and a high lift value, it can be interpreted that although these two products are not purchased frequently, they are mostly purchased together.

The Apriori algorithm was run in a loop at support values ranging from 1 to 0.0001 with the condition of a minimum of two items per rule. The amount of support values needed to generate association rules varied at different levels of analysis. Higher-level concepts such as food and clothing can generate rules with support values of more than 10 percent, while lower-level concepts such as a box of eggs, require lower support values. Therefore, the algorithms were tuned at different support and confidence settings at each level of analysis to generate a manageable number of rules. Fig. 2 shows the number of rules discovered at different support value levels. The analysis was performed using the "*arules*" package within the *R* statistical language version 3.2.3 [8].

**3. Results**

An exploratory analysis on the number of items per transaction showed that the data are heavily skewed toward the lower end of the distribution, with most transactions having only a few items (min 1, max 53, median 4). Around 30 percent of the transactions (3,084 records) have a single item only. Fig. 3 shows the distribution for the number of items per transaction (Skewness 1.95). The following sections discuss frequency and association analysis results at the single item, category and sub-category, and brand name levels respectively.

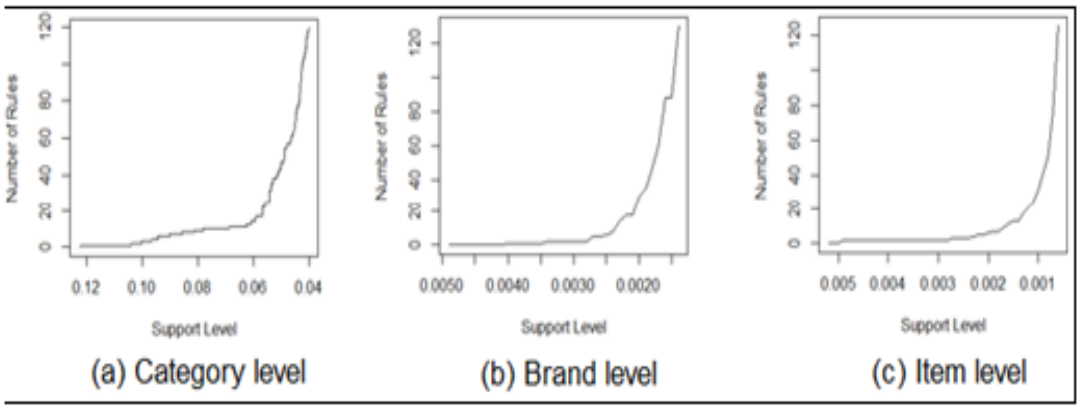


Figure 2. Number of rules at different support value levels

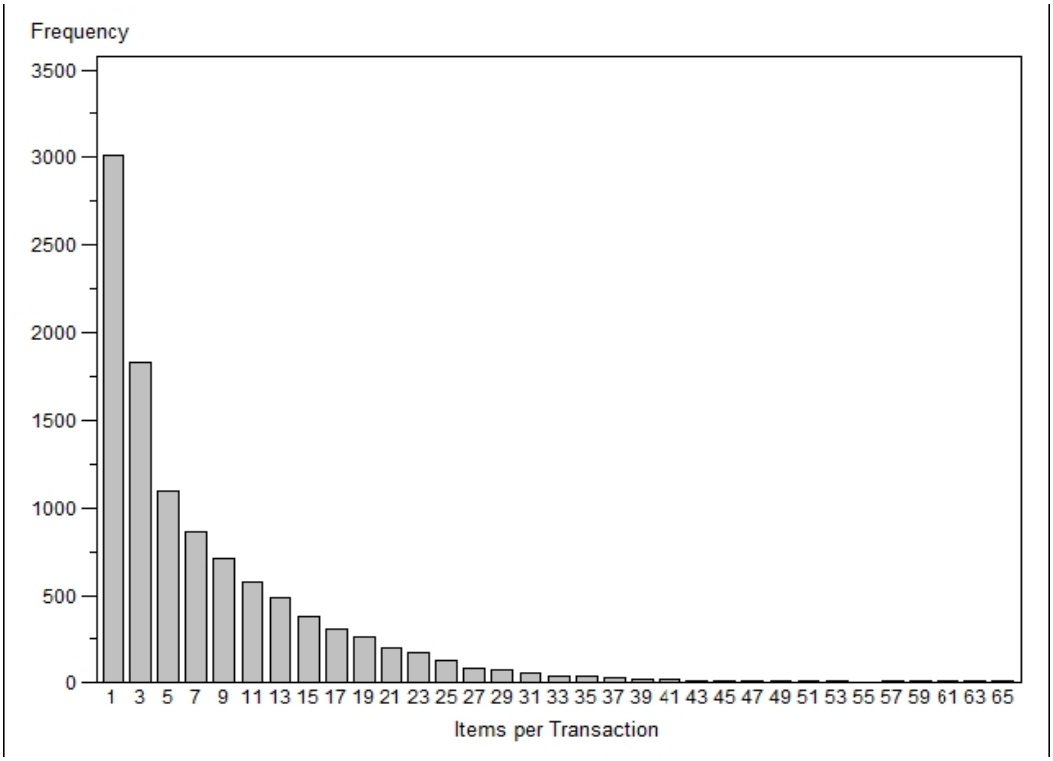


Figure 3. Number of Items per Transaction

| Item | Category | Brand | Frequency |
|---|---|---|---|
| LARGE EGGS | Eggs | CAL-MAINE | 0.0807 |
| NEW MEMBER CARD | Sam's Membership | SAM'S MEMBERSHIP | 0.0548 |
| 15$ COUPON | Sam's Coupon | SAM'S Club | 0.0384 |
| PAPER TOWELS 15 ROLL | PAPER TOWELS And Napkins | BOUNTY | 0.0329 |
| TISSUE | PAPER TOWELS And Napkins | KIMBERLY-CLARK | 0.0329 |
| CHICAGO TRIBUNE | Publications | CHICAGO TRIBUNE | 0.0262 |
| GLAD TRASH BAG | Disposer Bags | GLAD | 0.0257 |
| MAC & CHEESE | Box Dinners | KRAFT | 0.0229 |
| FABRIC SOFTNER | Fabric Care | BOUNCE | 0.0227 |
| TORTILLA CHIP | Chips And Cookies | ON THE BORDER | 0.0219 |
| GO-GURT YOGURT | Dairy - YOGURT | YOPLAIT | 0.0202 |
| BUTTER | Dairy - BUTTER | MID AMERICA FARMS | 0.0198 |
| CHEX MIX TRADITIONAL | Snacks | GENERAL MILLS | 0.0186 |
| TOSTITOS TORTILLA CHIPS | Chips And Cookies | TOSTITOS | 0.0184 |
| DISHWASHING POWDER | Cleaning - Dish Detergent | CASCADE | 0.0181 |
| WHITE BREAD | Baking - BREAD | MOTHER'S COOKIES | 0.0175 |
| ALL NATURAL JUICE | Beverages Juice | CAPRI SUN | 0.0169 |
| POPCORN CHICKEN | Branded Food - Nuggets | TYSON FOODS | 0.0161 |
| PAPER TOWELS 6 ROLL | PAPER TOWELS And Napkins | PROCTER & GAMBLE | 0.0156 |
| COOKIES | Chips And Cookies | PEPPERIDGE FARMS | 0.0151 |

Figure 4. The top 20 most frequent items in the dataset

### 3.1 Item-Level Analysis

The analysis at the item level looks for patterns among items that are purchased together. The dataset at this level contains 2,698 unique items such as fresh food, clothing, books, and so on. The frequency analysis shows that a small number of items appear repeatedly in the dataset, while the majority of items are grouped at the lower end of the frequency distribution. The item that was purchased most often was "*Eggs Box Large Size*" with a support value of 0.0807, while the least frequently purchased item was the "750*ML Taylor Fladgate*" bottle of red wine, with a support value of just 0.000099. Fig. 4 shows the support values, and the category and brand names for the top 20 most popular items.

The Apriori algorithm was used to identify subsets of items that are usually purchased together. The first association rule between two items (Secondary Membership Renewal => Primary Membership Renewal) was identified at a support level of 0.0049. The number of discovered rules grows exponentially as the support value gets smaller, with 1,129,655 rules discovered at 0.0001 – the minimum level of support. Generating a manageable amount of association patterns requires tuning the algorithms support and confidence values. The algorithm was configured so the minimum support and confidence values were 0.0005 and 0.5 respectively. These settings generated 214 association rules with support values between 0.00059 and 0.0049 and lift values between 6.1 and 715.06. Evaluating the quality and usefulness of the discovered rules requires investigating the confidence and lift values, and applying human common sense.

The high lift values for certain rules indicate that despite their overall low frequency, there is a high confidence these items are mostly purchased together. Examples of discovered rules include:

- the Harry Potter novel series (Chamber of Secrets, Prisoner of Azkaban => Philosopher's Stone) s=0.00069 c=.078 l=715.06
- clothing and style purchases: (Cotton Ash T-Shirt => Cotton Navy T-Shirt) s = 0.0007 c= 0.58 l = 368.70, and (Cotton T-shirt => Cotton Fleece Short) s = 0.0008 c = 0.8889 l = 473.12

As the confidence and lift values decrease, the discovered association rules tend to describe more generic shopping behaviour. For example, everyday shopping items such as a box of eggs and bottles of milk are usually associated with a wide range of other food and household items. "*Grade A Large Eggs*" appears in 160 different rules with a maximum support lift value of (45.38), but none of those rules are among the highest 20 lift rules. Another example of generic shopping behaviour with a relatively high support value and low lift value is (Low Fat Milk, Split Top Cookies => Grade A Large Eggs) s=0.0024, c=0.57, l=7.07. Fig. 5 shows the association rules with the highest 20 lift values.

Transactional databases contain a large number of items with some of them being used daily. This leads to the discovery of association rules with low support values. However, further examination of the confidence and lift values highlights association rules that reflect specific needs and preferences. The next section examines association rules when a variety of shopping items are grouped together into common brand names.

### 3.2 Brand-level Analysis

Association analysis at the brand level uncovered brands that are frequently purchased together. The dataset contains 785 unique brand names for a wide range of products including *Kellogg's*, *Kraft*, and *Glad*. Frequency analysis showed that the most purchased brand was *Member's Mark*, a brand name owned by Wal-Mart, with a support value of 0.13, while the least frequent brand is *Weider Nutrition*, a sports nutrition product with a support value of just 0.000123. Fig. 6 shows the support values and products information for the 20 most frequently occurring brand names. The first association rule between two brand names (*Cal-Maine => Member's Mark*) was identified at the support level of 0.004. Generating a manageable number of association rules requires tuning the algorithm support and confidence values. The minimum support and confidence values were thus set to 0.0005 and 0.2 respectively. These settings generated 214 association rules with support values between 0.00059 and 0.023 and lift values between 1.56 and 8.57. Evaluating the quality and usefulness of the discovered rules requires investigating their confidence and lift values as well as applying human common sense.

The high lift values for certain rules highlights how customers associate different brand names when making purchasing decisions. These rules can identify customers? favourite combinations of brand names; i.e., the dataset contains different brand names for more than 10 detergent and fabric care products as well as 15 sweet and candy items. The algorithm identified that the association rule (*Downy Fabric Softener => TIDE Washing Powder*) s = 0.005 c = 0.259 l = 8.579 generated the highest lift value for customer preferences for two laundry products. Another association rule with a high lift value was (*M & M Mars => Hershey's*) s = 0.006 c = 0.209 l = 6.56, which is a combination of 2 brand names for chocolate-loving customers. Identifying

associations among certain brand names may be infeasible at other levels of detail as some brands may be labelled as multiple items based on their size or flavour features.

| Rules | Support | Confidence | Lift |
|---|---|---|---|
| Downy => Tide | 0.0051 | 0.2595 | 8.5792 |
| M & M Mars => Hershey's | 0.0063 | 0.2090 | 6.5621 |
| Bounce => Cascade | 0.0058 | 0.2043 | 5.4991 |
| General Mills, Nabisco => Kellogg's | 0.0054 | 0.3235 | 4.2960 |
| Ziploc => Kimberly-Clark | 0.0075 | 0.2293 | 4.1278 |
| Kimberly-Clark => Procter & Gamble | 0.0163 | 0.2933 | 4.1179 |
| Fresh => USDA Choice | 0.0075 | 0.3547 | 4.1038 |
| Tide => Procter & Gamble | 0.0088 | 0.2898 | 4.0682 |
| Glad => Procter & Gamble | 0.0115 | 0.2835 | 3.9803 |
| Campbell's => Kellogg's | 0.0051 | 0.2971 | 3.9451 |
| Kellogg's, Member's Mark => General Mills | 0.0056 | 0.3435 | 3.9189 |
| General Mills, Kellogg's => Nabisco | 0.0054 | 0.2716 | 3.9076 |
| Cascade => Kimberly-Clark | 0.0080 | 0.2159 | 3.8870 |
| Kraft, Tyson Foods => Cal-Maine | 0.0058 | 0.3821 | 3.7517 |
| Bounce => Procter & Gamble | 0.0075 | 0.2652 | 3.7232 |
| Tide => Kimberly-Clark | 0.0062 | 0.2041 | 3.6735 |
| Cal-Maine, Tyson Foods => Kraft | 0.0058 | 0.2611 | 3.5908 |
| Capri Sun => Kellogg's | 0.0057 | 0.2690 | 3.5720 |
| Maruchan => Cal-Maine | 0.0056 | 0.3516 | 3.4517 |
| Cal-Maine, Member's Mark => Kraft | 0.0058 | 0.2487 | 3.4198 |
| Tyson Foods, USDA Choice => Cal-Maine | 0.0056 | 0.3462 | 3.3986 |
| Quaker => Kellogg's | 0.0091 | 0.2543 | 3.3767 |
| Stella => Cal-Maine | 0.0051 | 0.3417 | 3.3545 |
| Kraft => General Mills | 0.0084 | 0.2931 | 3.3439 |
| Heinz => Kraft | 0.0067 | 0.2422 | 3.3301 |
| Kraft, Member's Mark => Cal-Maine | 0.0058 | 0.3357 | 3.2961 |

Figure 7. Top 20 highest lift brand names association rules

As confidence and lift values decrease, the discovered rules tend to describe brand name associations among the grocery and food items that are purchased most frequently. Popular food and beverage brands such as *Member's Mark* and *Cal-Maine* appear in association with brands such as *MOTT'S* and *Tyson Foods*. Fig. 7 lists the top 20 brand name association rules with the highest lift values.

### 3.3 Category-Level Analysis

Shopping items are usually grouped into categories and sub-categories for better management and classification. Analysing association rules at these higher levels can reveal shopping patterns in item categories that are frequently purchased together. The dataset at this level contains 62 main categories that are further divided into 717 sub-categories. Frequency analysis showed that the most frequently purchased product category was Fresh Meat that includes all types of lamb, beef, and pork products with a support value of 0.27, while the least frequently purchased product category was "Toys" with a support value of just 0.0000989. The sub-category "Milk" – a member of the "Beverages" category – was the most frequent sub-category with a support value of 0.14, while the least frequent sub-category was "Eyeglass Cloths and Cleaners" with a support value of just 0.000099.

| Rules | Support | Confidence | Lift |
|---|---|---|---|
| Milk, Vegetables => Small Pack Eggs | 0.0043 | 0.5513 | 6.8239 |
| Milk, Shredded Cheese => Small Pack Eggs | 0.0033 | 0.5500 | 6.8080 |
| Milk, Beef-Roasts => Small Pack Eggs | 0.0034 | 0.5313 | 6.5759 |
| Milk, Pork-Bone In => Small Pack Eggs | 0.0031 | 0.5000 | 6.1891 |
| Local Juice, Breads => Milk | 0.0032 | 0.8421 | 6.0514 |
| Small Pack Eggs, Kids Multi-Serve Juice => Milk | 0.0032 | 0.6957 | 5.0072 |
| Small Pack Eggs, Wings => Milk | 0.0032 | 0.6809 | 4.9007 |
| Small Pack Eggs, Yogurt => Milk | 0.0040 | 0.6780 | 4.8799 |
| Small Pack Eggs, Crackers => Milk | 0.0049 | 0.6757 | 4.8634 |
| Small Pack Eggs, Dry Pasta => Milk | 0.0036 | 0.6316 | 4.5460 |

Figure 8. Highest 10 Lift Rules for Category and Sub-Category Levels

The Apriori association rules algorithm was used to identify shopping patterns at the category and sub-category levels for different support values. The first association rule between two categories (Branded Food Products => Fresh Meat) was identified at the support level of 0.12, while the first rule at the sub-category level (Small Pack Eggs => Milk) was identified at the support level of 0.04. The hierarchical nature between lower-level items and their higher-level categories and sub-categories requires tuning the algorithm to the relevant support value for that level. The minimum support values for the algorithm were set to 0.04 and 0.003 at the category and sub-category levels respectively. A total of 130 association rules were identified for both category and sub-category levels. These rules generated lift values that ranged from 1.8 to 6.8. Evaluating the quality and usefulness of these rules requires investigating confidence and lift values as well as applying human common sense.

The wide variety of food and beverage products at the items level has led to their domination of the association rules at the category and sub-category levels. Higher-level categories such as milk and fresh meat appeared in the majority of the discovered rules and had high lift values. These results show that association rules analysis at the category level represent everyday shopping activities that include groceries and other household items. Rules with high lift values at the item level such as those including clothing and books did not score high support at the category level analysis. Fig. 8 shows the 10 lift rules with the highest values in the category and sub-category levels.

## 4. Discussion

This paper studied the use of *multilevel association rule mining* to discover interesting patterns among shopping items at multiple levels of abstraction. The analysis discovered different types of patterns that highlight relationships based on features such as categories and brand names. This knowledge can help retailers understand customer preferences and design better marketing plans to reflect the discovered associations.

At the item level, the associations observed reveal that shopping activities are made in a variety of contexts. Frequently-purchased items such as eggs and milk are usually associated with a wide range of other grocery and household items. This shopping behaviour generated association rules with relatively high support values with low confidence and lift levels. On the other hand, less frequently-purchased items such as clothes, books, and batteries generated association rules with high confidence and lift values. The association between the *Harry Potter* novel series generated the highest lift value despite being part of the books category, and caused the publication category to be ranked $17^{th}$ among other categories. Another high confidence association was highlighted in the rules between: (a) Ash Colour and Navy Colour shirts, and (b) ladies? cotton T-Shirt and Fleece Shorts. This knowledge can help retailers to design special promotions for items that, despite their relatively low frequency, are most often purchased together. The patterns discovered at the item level also support earlier studies that described shopping as a *contextualised* act, where shopping trips for grocery and household items are different in motive from shopping trips for clothes.

The association analysis was performed at the brand level to highlight associations between brand names that are frequently purchased together. The discovered associations highlight customers? favourite combinations of brand names among the many available brand options. The analysis revealed that despite the availability of several brands for similar items, certain brand names were associated with high confidence and lift values. For example, the laundry brands *Downy* and *Tide* generated the highest lift value for brand name associations. Another association between *M & M Mars* and *Hershey's* identified the favourite combination of chocolate brand names for customers with a sweet tooth. In both examples, many other options of chocolates, detergents, and fabric softeners were also available. Identifying these associations at the items level could be unfeasible since the same brand name could be labelled as multiple different products based on attributes like package size and flavour.

Analysing associations at a higher-aggregated level can reveal shopping patterns in item categories and sub-categories that are frequently purchased together. Higher-level concepts such as baking needs and fresh meat are more likely to generate greater support values than low-level items due to the hierarchical relationship between the two levels. Also, the wide variety of food and beverage products at the item level has led to the domination of their association rules at the levels of categories and sub-categories. For example, the association between eggs and milk generated the highest support value for the level of analysis, reflecting a very common combination of groceries. The association analysis at the category and sub-category levels describes the mainstream shopping activities of household items and
groceries.

## 5. Conclusions

This paper presents an empirical study using multi-level association rules mining to extract shopping patterns from shopping datasets. The results show the validity of this approach in discovering association patterns related to items? metadata such as colours, brand names, and so on. The results confirm earlier customers? behavioural studies that describe shopping as a highly contextualized act. However, the study also highlighted challenges and limitations related to evaluating the usefulness of the large numbers of discovered rules. Future work will involve the use of graph data representation for shopping transactions to provide better filtering and evaluation methods of the discovered association rules.

## References

[1] Agrawal, R., Imieli´nski, T., Swami, A. (1993). Mining association rules between sets of items in large databases. ACM Sigmod Record, 22, p. 207–216. ACM.

[2] Agrawal, R. Srikant, R. (1994). Fast algorithms for mining association rules. *In:* Proc. 20th int. conf. very large data bases, VLDB, volume 1215, p. 487–499.

[3] Garc´1a, E. Romero, C. Ventura, S. Calders, T. (2007) . Drawbacks and solutions of applying association rule mining in learning management systems. In Proceedings of the International Workshop on Applying Data Mining in e-Learning (ADML2007), Crete, Greece, p. 13–22.

[4] Gutierrez, N. (2006). Demystifying market basket analysis. DM Review Special Report

[5] Han, J., Cai, Y., Cercone, N. (1993). Data-driven discovery of quantitative rules in relational databases. IEEE transactions on Knowledge and Data Engineering, 5 (1) 29–40.

[6] Huynh, X.-H. (2006). Interestingness measures for association rules in a KDD process: postprocessing of rules with ARQAT tool. PhD thesis, Universit´e de Nantes

[7] Roy Dholakia, R. (1999). Going shopping: key determinants of shopping behaviors and motivations. *International Journal of Retail & Distribution Management*, 27 (4) 154–165.

[8] Team, R. C. (2013). R: A language and environment for statistical computing.

[9] Wan, Y.-b., Liang, Y., Ding, L.-Y. (2008). Mining multilevel association rules with dynamic concept hierarchy. *In:* 2008 International Conference on Machine Learning and Cybernetics, V 1. 287–292. IEEE

[10] Xu, Y., Zeng, M., Liu, Q., Wang, X. (2014). A genetic algorithm based multilevel association rules mining for big datasets. *Mathematical Problems in Engineering*, 2014

[11] Zhao, Y. (2012). R and data mining: Examples and case studies. Academic Press.