A Visual Analytics Workflow for Investigating Customers' Transactions in Convenience Stores

Ivana Kocanova* Muhammad Adnan† Georgios Aivaliotis‡ Roy A. Ruddle§

University of Leeds, UK.

ABSTRACT

Convenience stores stock a small range of everyday products. When a customer makes a transaction, each product they buy can be treated as an 'event' and the overall contents of the customer's shopping basket is a set of events. Unfortunately, existing event set mining and visualization techniques are poorly suited to the typical complexity of the data (e.g., 140,000 product combinations in 360,000 transactions). To address that we have developed a visual analytic workflow, which starts by using high utility itemset mining to reduce the complexity of the data by a factor of 1000, and then allows users to investigate the composition of transactions and create transaction sketches. Examples are a product-perspective sketch, which reveals sets of products that are often bought together, and a time-perspective sketch that shows how those sets change from breakfast to lunch, dinner and then night.

Index Terms: Human-centered computing—Visualization—Visualization application domains—Visual analytics; Human-centered computing—Visualization—Visualization design and evaluation methods

1 INTRODUCTION

Event data is common in many application domains [7] and may be analyzed by set mining and visualization techniques. However, existing techniques are poorly suited to the scale of real-world applications such as the analysis of customer transactions in convenience stores, which are characterized by many combinations of event but few, if any, that are dominant [2].

This poster describes a visual analytics workflow that combines visualization with two complementary set mining methods (for exclusive and non-exclusive intersections, respectively) for the analysis of transaction data. The work was conducted with a major retailer and makes two main contributions. First, we describe a visual analytic workflow that combines multiple event set mining methods and visualization to analyze large-scale event data. Second, we evaluate the workflow with a convenience store dataset, which is an exemplar of the complexity and scale of real-world data.

2 RELATED WORK

Set mining methods may be divided into two groups. The first contains methods that calculate exclusive set intersections (ESIs), meaning that each intersection matches the whole of a customer's transaction.It is straightforward to exhaustively compute ESIs, but the addition of one product to a transaction generates a new intersection. The second group contains frequent itemset mining (FIM) methods [4], and identifies cross-transaction patterns. However, the number of itemsets increases exponentially with the number of different events, and there is overlap and repetition between the itemsets and transactions. The present research uses high utility itemset mining (HUIM) [5] because it takes into account the quantity of each product in a transaction.

There has been considerable research into techniques for visualizing set data [3]. A number of research tools have been developed, but they each only cater for one type of set mining (e.g., UpSet [6] only uses ESIs). Commercial tools such as Tableau can create visualizations from high-volume data, but do not provide a data mining capability. Instead users have to integrate their own data mining via Python or R interfaces.

3 DATASET

This research used a 1,169,665-record dataset of customer transactions in four convenience stores. Each record contained a Transaction ID, Date, Time, Product and Quantity. The dataset contained 417 distinct products (e.g., bread and milk) and 365,756 transactions.

4 VISUAL ANALYTIC WORKFLOW

Even though 88% of the transactions contained five or fewer different products, ESI showed that there were 140,986 unique combinations of products in the transactions. The challenge was to develop a visual analytic workflow to analyze data of that scale.

Choose utility threshold: The workflow's first step involved choosing a utility threshold for HUIM, which was conducted using the EFIM algorithm [8]. The threshold is a trade-off between the number of itemsets and the percentage of the dataset that is covered by those itemsets (see Figure 1a). We chose a threshold of 1068, because that gave 95% coverage with only 1191 itemsets (coverage is the percentage of items in the transactions that also occur in the itemsets). The high coverage and 100-times reduction in the number of sets (ESI combinations to HUIM itemsets) shows how effective HUIM is for simplifying analysis.

Transaction composition: The next step involved iterative analysis of the composition of transactions in terms of HUIM building blocks (the itemsets). Out of 1191 itemsets, 201 contain one product, 745 contain two products, 234 contain three products and 11 contain four products (see Figure 1b).

The first iteration flagged transactions that exactly matched an identical length itemset. This identified that 96% of single-product transactions are an exact match for a single-product itemset. This coverage progressively drops to 53%, 12% and 2%, for transactions that contain two, three and four products, respectively (see Figure 1b).

Successive iterations analyzed the remaining unflagged transactions of each length to determine whether they contained other shorter-length itemsets. This highlighted that, in addition to 53% length two transactions that exactly match to a same length itemset, 46% transactions contain a length one itemset. Furthermore, 59% and 29% of length three transactions contain a length two or a length one itemset, respectively. Figure 1b indicates a similar level of coverage for length four and length five transactions as well.

Transaction sketches: The final step lets users identify sketches [1], which characterize transactions from one of several possible perspectives to understand the purpose of customers' visits to the

*e-mail:I.Kocanova@leeds.ac.uk

†e-mail:m.adnan1@hotmail.com

‡e-mail:G.Aivaliotis@leeds.ac.uk

§e-mail:R.A.Ruddle@leeds.ac.uk

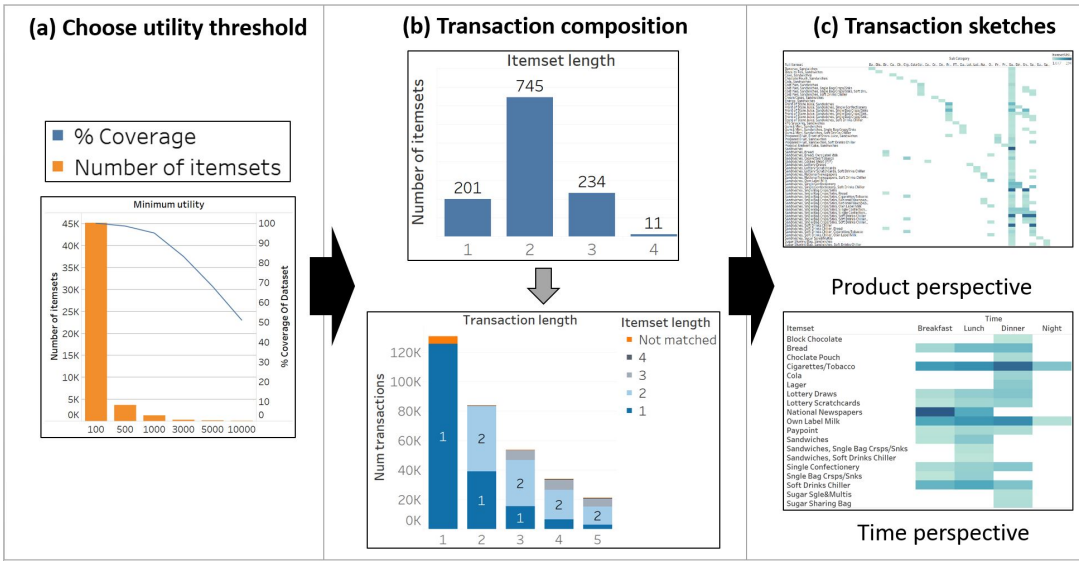

Figure 1: The visual analytics workflow: (a) Choose HUIM utility threshold, (b) Transaction composition, and (c) Transaction sketches.

stores. One perspective places a specific product (e.g., sandwiches) at the center of the sketch. Sandwiches were the most common product in itemsets of length four, and were most often bought with a soft drink, bag of crisps and one of six other products (e.g., confectionery). The same products occurred in shorter itemsets with sandwiches (see Figure 1c).

Alternatively, users may adopt a time perspective. Filtering with a frequency threshold allows users to substantially reduce the number of itemsets and reveal temporal patterns. For example, sandwiches are often bought alone at breakfast (07:00–10:00) and at lunch (11:00–14:00). However, sandwiches only tend to be bought with a soft drink or bag of crisps at lunch, even though both drinks and crisps are often bought by themselves at other times of day (see Figure 1c).

5 Comparison with a Traditional Approach

The traditional approach for analyzing transaction data is to generate an association rules network from the output of frequent itemset mining. Our dataset produces a network with two clusters, one which is around sandwiches, a soft drink and a bag of crisps. Although the network gives a user the impression that a variety of other products are bought with those three, the product-perspective sketch (see Figure 1c) is considerably more informative because it reveals patterns between products and the transaction lengths. The time-perspective sketches are also likely to be more informative than the traditional approach, because the sketches directly show differences that occur between customers' transactions at different times of day. To identify the temporal patterns, a separate network would have to be generated for each time period, requiring users to compare multiple visualizations.

6 Conclusions

This poster describes a visual analytic workflow for investigating customers' transactions in convenience stores. The three-stage visual analytic workflow starts by using high utility itemset mining to identify groups of products, which act as building blocks to reduce the complexity of the data by a factor of 1000. Stage 2 uses the itemset building blocks to iteratively analyze the composition of transactions, which were computed using a complementary set mining method (ESIs). Stage 3 then creates transaction sketches from product, time and other perspectives. The workflow is illustrated using a 1-million record dataset from a major retailer, and comparisons made with a traditional analysis approach that involves generating association rules networks.

7 Acknowledgements

This research was supported by the EPSRC (EP/N013980/1, EP/K503836/1), MRC (ES/L011891/1) and ESRC (ES/L011891/1). We gratefully acknowledge the data and feedback provided by our project partner. The license agreement limits the dataset's use to that funded project.

References

[1] Z. Abedjan, L. Golab, and F. Naumann. Profiling relational data: a survey. *The VLDB JournalThe International Journal on Very Large Data Bases*, 24(4):557–581, 2015.

[2] M. Adnan and R. Ruddle. A set-based visual analytics approach to analyze retail data. In *Proceedings of the EuroVis Workshop on Visual Analytics (EuroVA18)*. The Eurographics Association, 2018.

[3] B. Alsallakh, L. Micallef, W. Aigner, H. Hauser, S. Miksch, and P. Rodgers. The state-of-the-art of set visualization. In *Computer Graphics Forum*, vol. 35, pp. 234–260. Wiley Online Library, 2016.

[4] P. Fournier-Viger, J. C.-W. Lin, B. Vo, T. T. Chi, J. Zhang, and H. B. Le. A survey of itemset mining. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 7(4):e1207, 2017.

[5] W. Gan, J. C.-W. Lin, P. Fournier-Viger, H.-C. Chao, T.-P. Hong, and H. Fujita. A survey of incremental high-utility itemset mining. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(2):e1242, 2018.

[6] A. Lex, N. Gehlenborg, H. Strobelt, R. Vuillemot, and H. Pfister. UpSet: Visualization of intersecting sets. *IEEE Trans. Vis. Comput. Graph.*, 20(12):1983–1992, 2014.

[7] Z. Liu, H. Dev, M. Dontcheva, and M. Hoffman. Mining, pruning and visualizing frequent patterns for temporal event sequence analysis. In *Proceedings of the IEEE VIS 2016 Workshop on Temporal & Sequential Event Analysis*, 2016.

[8] S. Zida, P. Fournier-Viger, J. C.-W. Lin, C.-W. Wu, and V. S. Tseng. Efim: a fast and memory efficient algorithm for high-utility itemset mining. *Knowledge and Information Systems*, 51(2):595–625, 2017.