Ertek, G., Demiriz, A., Çakmak, F. (2012) "Linking Behavioral Patterns to Personal Attributes through Data Re-Mining" in Behavior Computing: Modeling, Analysis, Mining and Decision. Eds: Longbing Cao, Philip S. Yu. Springer.

Linking Behavioral Patterns to Personal Attributes through Data Re-Mining

Gürdal Ertek[1], Ayhan Demiriz[2], and Fatih Cakmak[3]

[1]Sabancı University, Faculty of Engineering and Natural Sciences

Orhanli, Tuzla, 34956, Istanbul, Turkey.

[2]Department of Industrial Engineering

Sakarya University, 54187, Sakarya, Turkey.

[3]Sabancı University, Faculty of Arts and Social Sciences

Orhanli, Tuzla, 34956, Istanbul, Turkey.

Linking Behavioral Patterns to Personal Attributes through Data Re-Mining

Gürdal Ertek[1], Ayhan Demiriz[2], and Fatih Cakmak[3]

[1] Sabancı University, Faculty of Engineering and Natural Sciences
Orhanli, Tuzla, 34956, Istanbul, Turkey. ertekg@sabanciuniv.edu
[2] Department of Industrial Engineering
Sakarya University, 54187, Sakarya, Turkey. ademiriz@gmail.com
[3] Sabancı University, Faculty of Arts and Social Sciences
Orhanli, Tuzla, 34956, Istanbul, Turkey.

Abstract. A fundamental challenge in behavioral informatics is the development of methodologies and systems that can achieve its goals and tasks, including behavior pattern analysis. This study presents such a methodology, that can be converted into a decision support system, by the appropriate integration of existing tools for association mining and graph visualization. The methodology enables the linking of behavioral patterns to personal attributes, through the re-mining of colored association graphs that represent item associations. The methodology is described and mathematically formalized, and is demonstrated in a case study related with retail industry.

1 Introduction

This study aims at understanding the behavioral patterns exhibited by people in relation to their personal attributes. The research is conducted in the context of retail industry, where consumers engage in purchase transactions at retail shops and stores. The traditional data mining technique for identifying the patterns in these transactions is association mining, which enables the discovery of interpretable and actionable results related with item associations. However, straightforward application of association mining returns only item purchase patterns. An important question, whose answer has been ignored in literature, is how these patterns are related to consumer attributes, such as demographic attributes and physical state of the consumer during the purchase. In other words, the link between the behavioral pattern (consumer purchase behavior) and the person (consumer) him/herself is missing. Establishing this link requires a methodology, as well as domain knowledge to enable domain-driven data mining [5].

A graph-based visualization methodology, namely AssocGraphRM, is proposed for presenting association mining results, together with summary statistics regarding the associations. The methodology suggests a visual data *re-mining* process, based on the results generated by association mining. In the graph-representation, items and itemsets are represented as vertices, set membership are represented through edges, and attribute statistics are linearly mapped to the colors of vertices. The applicability and usefulness of the methodology is demonstrated through a market basket analysis (MBA) case study where data from a consumer survey is analyzed. The survey contains a multitude

of personal attributes, as well as preferences for items at Starbucks coffee stores. Several actionable insights are derived regarding the relationship between the behavioral patterns (item purchases) and the personal attributes, and their policy implications are discussed.

The remainder of the chapter is organized as follows: In Section 2, an overview of the basic concepts in related studies is presented through a concise literature review. In Section 3, the AssocGraphRM methodology for visual re-mining on colored association graphs is described, and framed as an algorithm using mathematical formalism. The methodology and its applicability is then demonstrated in Section 4, using survey data from food retail industry. The validity of the methodology is discussed in Section 5. Finally, in Section 6, the study is summarized and future directions are discussed.

2 Literature

2.1 Behavior Informatics

The field of *behavior informatics* is introduced by Cao [4], and suggests the analysis of behavioral patterns and impacts following *behavior explicitation*, through the extraction of behavior elements masked in transactional data. The main goals and tasks of behavior informatics are listed in [4] as behavior modeling and representation, construction of behavioral data, behavior impact modeling, behavior pattern analysis, and behavior presentation. The main idea in behavioral informatics is to organize the transactional data into a new form that is constructed in terms of behavior, rather than entity relationships. With the explosion of data that is collected electronically in massive amounts, the main challenge in behavioral informatics is the development of methodologies and systems that can achieve its goals and tasks. This study presents such a methodology, that can be converted into a decision support system, by the appropriate integration of existing tools for association mining and graph visualization.

2.2 Association Mining

Association mining is an increasingly used data mining and business tool among practitioners and business analysts [10], due to the interpretable and actionable results it generates. Association mining results can be classified based on several criteria, as outlined in [18]. In this chapter, we focus on frequent itemset, which are the sets that define single-dimensional, single-level boolean association rules. Efficient algorithms such as Apriori [1] enable the analysis of very large transactional data, frequently from transactional sales data, resulting in a large collection of frequent itemsets.

Association mining is typically presented and discussed in the context of one of its most common applications, namely market basket analysis (MBA), which can be used in product recommender systems [10]. Let $I = \{i_1, i_2, ..., i_m\}$ be a set of items considered in MBA. Each transaction (basket) t will consist of a set of items where $t \subseteq I$. Let $D = \cup t$ be the database of all transactions. The *support* $sup(f)$ of an itemset (and also of the rule that contains the items in that itemset) is defined as the percentage of the transactions in D that contain all the items of the itemset f:

$$sup(f) = \frac{\sum_{t \in D} 1_{\{f \subseteq t\}}}{|D|} \quad (1)$$

A *frequent itemset* is an itemset that has support value greater than or equal to a given minimum support threshold: $sup(f) \geq min\_sup$.

2.3 Visualizing Frequent Itemsets

Commonly, finding the frequent itemsets and association rules from very large data sets is heavily emphasized, since it is considered as the most challenging step in association mining [6, 17, 41, 42]. Results are typically presented in a text (or table) format with certain degree of querying and sorting functionalities. However, the real objective of the association mining analysis is to foster the discovery of insights, and there exists considerably less work that focuses on the interpretation of the association mining results [14, 14].

Information visualization is the branch of computer science that investigates how data and information can be visualized to obtain significant, deep, actionable insights [9, 19, 23, 27]. Within information visualization, graph visualization can be a significant source of insights, as demonstrated by numerous case studies in a multitude of disciplines [8, 26, 30–33, 35, 38, 36]. There is a broad literature on graph visualization, including the literature on graph drawing, but the use of graphs for the visualization of association mining results is not well-formalized in academic literature. Still, data analysis systems such as MS SQL Server [28] and SAS [34] can generate association graphs.

This study is an extension of earlier work by Ertek and Demiriz [14], where a graph-based methodology is proposed to visualize and interpret the results of well-known association mining algorithms as directed graphs. According to the methodology in [14], the items (also referred to as 1-*itemsets*) and the itemsets are represented as vertices on an association graph. The vertices that represent the items are shown with no color, whereas the vertices that represent the itemsets are colored reflecting the cardinality of the itemsets. The sizes (the areas) of the vertices show the support levels. The directed edges symbolize which items constitute a given frequent itemset.

The main idea in the methodology is to exploit already existing graph drawing algorithms [37] and software in the information visualization literature [19] for visualizing association mining results which are generated by already existing algorithms and software in the data mining literature [18].

In the current study, the methodology in [14] is extended to incorporate additional attributes, and mathematical formalism is introduced for describing the methodology. Color is now used to represent the values of a selected additional attribute, instead of the cardinality of the itemset. The cardinality of the itemset is instead reflected by the thickness of the vertices. So, this extended study enables linking association mining results with attributes of the person that carried out the transaction. In the context of behavior informatics, the methodology establishes the critical link between behavioral patterns and personal attributes.

2.4 Re-Mining

The re-mining methodology was first introduced by Demiriz et al. [11]. *Re-mining* process is defined as "combining the results of an original data mining process with a new additional set of data and then mining the newly formed data again". Re-mining is fundamentally different from post-mining [7, 22, 25, 43, 44]: post-mining only summarizes the data mining results, such as visualizing the association mining results [14, 21]. The re-mining methodology extends and generalizes post-mining. Re-mining can be considered as an additional data mining step of Knowledge Discovery in Databases (KDD) process [24] and can be conducted in explanatory/exploratory, descriptive, and predictive manners.

In another study, Demiriz et al. [12] elaborate on the re-mining concept, introduce mathematical formalism and present the algorithm for the methodology. [12] also extends the application of predictive re-mining in addition to exploratory and descriptive re-mining, and presents a complexity analysis.

Quantitative and multi-dimensional association mining (QAM&MAM) are well-known techniques within association mining [18] that can integrate additional attribute data into the association mining process. The associations among the additional attributes, and among them and itemsets are computed. However, both techniques introduce significant additional complexity, since association mining is carried out with the complete set of attributes rather than just the market basket data. The techniques work directly towards the generation of *multi-dimensional* rules. They relate all the possible categorical values of all the attributes to each other, which is *NP-hard*.

Re-mining, on the other hand, conveniently expands *single dimensional* rules with additional attributes. In re-mining, attribute values are investigated and computed only for the associated item pairs, with much less computational complexity that can be solved in polynomial running time. Running time of QAM&MAM increase exponentially with the number of additional attributes and the number of transactions, and re-mining is even more preferable in such situations.

Demiriz et al. [11, 12] propose a practical and effective methodology that efficiently enables the incorporation of attribute data (e.g. price, category, sales timeline) in explaining positive and negative item associations, which respectively indicate the complementarity and substitution effects.

The work closest to re-mining is by Yao *et al.* [40], where a framework of a learning classifier is proposed to explain the mined results. Unlike in [40], the re-mining [11, 12] and visual re-mining approaches (this study) are applied to real world datasets as a proof of their applicability.

3 Methodology: Re-Mining on Association Graphs

The fundamental idea in re-mining is to exploit the domain specific knowledge in a new analysis step. Thus, re-mining is a recipe for domain-driven data mining [5]. The AssocGraphRM methodology proposed and described in this chapter is a special type of re-mining: Re-mining is conducted through visually mining colored association graphs. By introducing mathematical formalism the methodology is presented in the form of an algorithm.

In AssocGraphRM, following the execution of conventional association mining and the generation of the frequent itemsets, a new database R is formed from the frequent itemsets F and additional attributes A, and then exploratory visual analysis is performed. Visual re-mining consists of the following main steps:

Step 1:	Carry out association mining
Step 2:	Compute the statistics for the frequent itemsets
Step 3a:	Represent the frequent itemsets as a directed graph
Step 3b:	Map the computed statistics linearly to colors
Step 3c:	Color the graph with respect to this coloring scheme
Step 4:	Visually explore the colored graphs and discover actionable insights

The graph is constructed by following the design specifications below:

a. Each item(set) is represented as a vertex.
b. Area of each vertex is proportional to the support of the corresponding item(set).
c. For itemsets with more than single item, edges are drawn from the (vertices of the) items of that itemset to the (vertex of the) itemset.
d. Line thickness of each vertex is proportional to the number of items in the corresponding itemset.
e. Color of each vertex reflects the value of a selected attribute for the corresponding itemset.

The inputs for the algorithm, parameters to be decided before the algorithm run, and additional definitions involving sets and functions are presented below:

Inputs

I:	set of items; $i \in I$
D:	set of transactions, containing only the itemset information
A_n:	additional numerical attribute n introduced for re-mining; $n = 1 \ldots N$. Any non-numerical (categorical/ordinal) attribute can be converted into a numerical attribute by computing the percentage of transactions for a specific value of the categorical/ordinal attribute. For example, if the original attribute is the gender of the person involved in the transaction, then let A_n be the percentage of transactions containing the given itemset and involving a female.
A:	set of all attributes introduced for re-mining; $A = \cup A_n$

Parameters

min_sup:	minimum support required for frequent itemsets
min_items:	minimum number of items in the itemsets
max_items:	maximum number of items in the itemsets

min_diameter: diameter for the item with *min_sup*. Note that the diameter of an itemset may be smaller than this value, but that of an item can not.

Definitions

F_1: set of frequent items (1-itemsets); $f \in F_1$

$F_{>1}$: set of frequent k-itemsets ($k > 1$) that have positive association; $f \in F_k$

F: set of all frequent itemsets (k-itemsets, with $k = 1 \ldots max\_items$); $f \in F$

min_value_n: minimum value for attribute n, over all frequent itemsets F

max_value_n: maximum value for attribute n, over all frequent itemsets F

R: set of records for re-mining, that contain *positive* associations; $r \in R$

$G(V,E)$: association graph that represents the frequent itemsets with vertices V and edges E

$\mathscr{G}$: graph collection that will be used for visual exploration in the re-mining phase

Functions

apriori$(D, min\_items, max\_items, min\_sup)$:

apriori algorithm that operates on D and generates frequent itemsets with minimum of *min_items* items, maximum of *max_items*, and a minimum support value of *min_sup*

sup(f):

support of itemset f

$\psi(A_n, f)$:

function that computes the value of attribute A_n for a given frequent itemset f, $f \in F_k$. Once the function runs for a particular itemset, it stores the information in its corresponding record **record_of**(f), and returns from the record next time it is run.

create_new_vertex(v) :

function that creates a new vertex v

vertex_of(f) :

function that returns the vertex that corresponds to itemset f

create_new_edge(e) :

function that creates a new edge e

record_of(f) :

function that returns the record that corresponds to itemset f

clone_graph(G) :

function that creates a clone of graph G

compute_color(δ)

function that computes color based on darkness $\delta \in [0,1]$. Besides the argument δ, the RGB value for the computed color depends on the base RGB values for $\delta = 0$ and $\delta = 1$.

The complete methodology is formalized as an algorithm below:

Algorithm: AssocGraphRM

1. *Perform association mining.*
   $F = \textbf{apriori}(D, 1, max\_items, min\_sup)$
2. *Define frequent items and itemsets.*
   $F_1 = \{f \in F : |f| = 1\}$
   $F_{>1} = F - F_1$
3. *Label the item associations accordingly and append them as new records.*
   $R = \{r : r = (f), \forall f \in F\}$
4. *Expand the records with the cardinality value for the itemsets, and additional attributes for re-mining.*
   for all $r = (f) \in R$
   $r = (f, |f|)$
   for $n = 1 \ldots N$
   $r = (r, \psi(A_n, f))$
5. *Create the vertices of the graph, with area being linearly proportional to the support, and thickness being proportional to the cardinality of the itemset.*
   $V = \{\}$
   for all $f \in F$
   create_new_vertex(v)
   $v.itemset = f$
   $v.record =$ **record_of**(f)
   $v.diameter = min\_diameter\sqrt{\frac{sup(f)}{min\_sup}}$
   $v.thickness = |f|$
   vertex_of$(f) = v$
   $V \bigsqcup v$
6. *Create the edges of the graph, emanating from the items in the itemsets and terminating at the itemsets.*
   $E = \{\}$
   for all $f \in F_{>1}$
   for all $i \in f$
   create_new_edge(e)
   $e.from =$ **vertex_of**(i)
   $e.to =$ **vertex_of**(f)
   $E \bigsqcup e$

7. *Apply organic layout on G.*
8. *Compute the minimum and maximum values for each of the attributes.*
   $min\_value_n = min_{n=1...N, f \in F} \psi(A_n, f)$
   $max\_value_n = max_{n=1...N, f \in F} \psi(A_n, f)$
9. *Color the vertices in G with respect to each additional attribute, and add to the graph collection $\mathscr{G}$. The closer the value for attribute n gets to the maximum value of that attribute max_value_n, the darker the vertex will be colored.*
   for $n = 1 \ldots N$
      $G' =$ **clone_graph**(G)
      for all $v \in V$
         $f = v.itemset$
         $v.darkness = \frac{\psi(A_n, f) - min\_value_n}{max\_value_n - min\_value_n} \in [0,1]$
         $v.color =$ **compute_color**$(v.darkness)$
      $\mathscr{G} \bigsqcup G'$
10. *Perform visual re-mining through human-involved exploratory examination of the graphs in $\mathscr{G}$.*

---

4 Case Study

The proposed methodology is demonstrated through a case study using a survey dataset collected from coffee retail industry. Several insights are discovered regarding the relationships among the frequent itemsets and personal attributes, and suggestions are made on how these actions might be used used as operational policies.

4.1 Retail Industry

Recent research has positioned association mining as one of the most popular tools in retail analytics [3]. Market basket analysis is considered as a motivation, and is used as a test bed for these algorithms. Additional data are readily available either within the market basket data or as additional data, thanks to loyalty cards in retailing, which enable linking transactions to personal data, such as age, gender, county of residence, etc.

As of November 2010, the retail industry in US alone runs on a monthly sales of $377.5 billion, with food services and food retail industry constituting %10 share in it [39]. Due to its gigantic size, retail industry has been selected as the domain of the case study. Starbucks is one of the best-known brands in food retail, and the best-known brand in coffee retail / specialty eateries industry, with 137,000 employees and a global monthly revenue of nearly $1 billion [39]. Due to the company's visibility, the products of Starbucks have been considered for constructing the survey data.

4.2 The Data

A survey has been conducted with 644 respondents, that contain nearly equal distribution of working people vs. students (all students were assumed non-working), women

vs. men, and a multitude of universities and working environments. Each respondent was questioned for 22 attributes that reflect their demographic characteristics and life style preferences.

The fields in the dataset include demographic attributes (YearOfBirth, Gender, EmploymentStatus, IncomeType, etc.), attributes related with life style (FavoriteColor, SoccerTeam, FavoriteMusicGenre, etc.), educational background (University, EnglishLevel, FrenchLevel, etc.), perceptional and intentional information (ReasonForGoing, etc.), and physical status at the time the survey was conducted (HungerLevel, ThirstLevel). The number of additional attributes to be used in Step 8 of the algorithm totalled to 21.

As the transaction data, each respondent was also asked which items they would prefer from the menu of Starbucks Turkey stores if they had 15 TL Turkish Liras (approximately $10). The menu considered was the menu as of October 2009, and the respondents were limited to select at most four items without exceeding the budget limit. While the original survey distinguished between the different sizes (tall, grande, venti), in the data cleaning and preparation phase, the sizes for each type of item (such as Cafe Latte or Capuchino) were aggregated and considered as a single item (group). The original survey also contained preferences under a budget of 20 TL, but those preferences were not analyzed in the case study.

Even though market basket analysis is carried out in retail industry with transactions data that is logged through sales, preference data was collected in the survey and used instead of the transactions data. This is due to the well-known difficulty of obtaining real world transactions data from companies, which they consider highly confidential, even when masked.

4.3 The Process

Data was assembled in MS Excel and cleaned following the guidelines in the taxonomy of dirty data by Kim et al. [20]. The transactions were given as input into Borgelt's apriori software [2], and the apriori algorithm was run with a support value of 2%. The sizes of the vertices (based on the support values), the statistics for the frequent itemsets, and the corresponding vertex colors were computed in MS Excel spreadsheet software through distributed manual processing by 30 Sabancı University students, and assembled through the EditGrid [13] online spreadsheet service. The association graph was manually drawn in yEd Graph Editor software and an organic layout was applied. yEd implements several types of graph drawing algorithms, including those that create hierarchical, organic, orthogonal, and circular layouts, and allows customization of the layouts through structure and parameter selections. Past experience with the software in applied research projects has shown that in Classic Organic Layout in yEd is especially suitable for displaying associations. Organic layout is generated based on force-directed placement algorithms [16] in graph drawing. This layout selection ends up placing items that belong to similar frequent itemsets and in close proximity of each other. After constructing the base graph in yEd, the yEd graphml file was cloned, and each copy was colored manually according to a different additional attribute. Then the resulting collection of graphs were analyzed through brain-storming sessions and actionable insights, together with their policy implications, were determined.

4.4 Analysis and Results

Visual re-mining was performed on the colored association graphs in the collection $\mathscr{G}$ through human-involved exploratory examination of the graphs. Figures 1-4 are selected graphs from $\mathscr{G}$ that are constructed for this case study. In the figures, regions of the graphs are highlighted for illustrating the insights and policies.

Figure 1 shows only the frequent item preferences of the participants in the case study. In this figure, the large region shows that Mosaic Cake and White Chocolate Mocha are selected by a large percentage of the people, and they are purchased together frequently, as represented by the large size of the corresponding vertex F11. The small region suggests that Chai Tea Latte and Lemon Cake are purchased frequently with each other, but not with other items. In the context of retailing, these are referred to as *complementary items*, and the classic operational policy is to use each of these items for increasing the sale of the other(s):

Policy 1: "Bundle Mosaic Cake, White Chocolate Mocha and Water together, and/or target cross-selling by suggesting the complementary item when the other is ordered."

Policy 2: "Bundle Chai Tea Latte and Lemon Cake together, and/or target cross-selling by suggesting the complementary item when the other is ordered."

Besides identifying the most significant and interdependent items, one can also observe items that are independent from all items but one. The central item is an *attractor*,

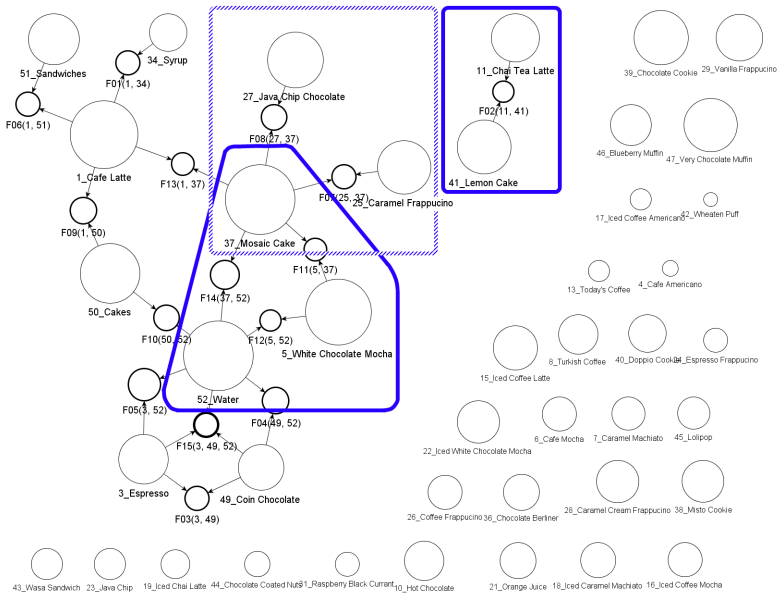

Fig. 1. Association graph that displays the frequent itemsets.

drawing attention to less popular items related to them. In retail industry, attractor items are placed at visible locations (ex: the aisle ends) to attract customers to the items in nearby but less visible locations (ex: inside the aisles).

Another type of insight that can be derived from Figure 1 is the identification of items which form frequent itemsets with the same item(s) but do not form any frequent itemsets with each other. One such pattern is indicated with the dashed borderline. Caramel Frappucino and Java Chip Chocolate each independently form frequent itemsets with Mosaic Cake, but do not form frequent itemsets with one another. These items may be *substitute items* and their relationship deserves further investigation.

While Figure 1 illustrates behavioral patterns with regards to item purchases, it does not tell us how these patterns relate to attributes of people. The mapping of attributes to colors on the graph in the proceeding figures solves this problem, and enables the discovery of deeper additional insights in more dimensions.

Figure 2 displays the items and itemsets together with gender attribute (percentage of females selecting that itemset, computed from the data field Gender) mapped to color. Darker vertices (itemsets) indicate that among the people that selected that itemset, the percentage of females is higher compared to males. Even though the vertex sizes and locations are exactly the same as the earlier figure, new insights are derived due to the coloring. The selected region shows that Mosaic Cake is purchased with either Java Chip Chocolate, as represented by the itemset F08, or with Caramel Frappucino, as represented by the itemset F07. However, F08 and F07 are colored clearly differently, with F08 being darker. This means that the percentage of women among those that prefer the itemset F08 (Mosaic Cake and Java Chip Chocolate) is higher.

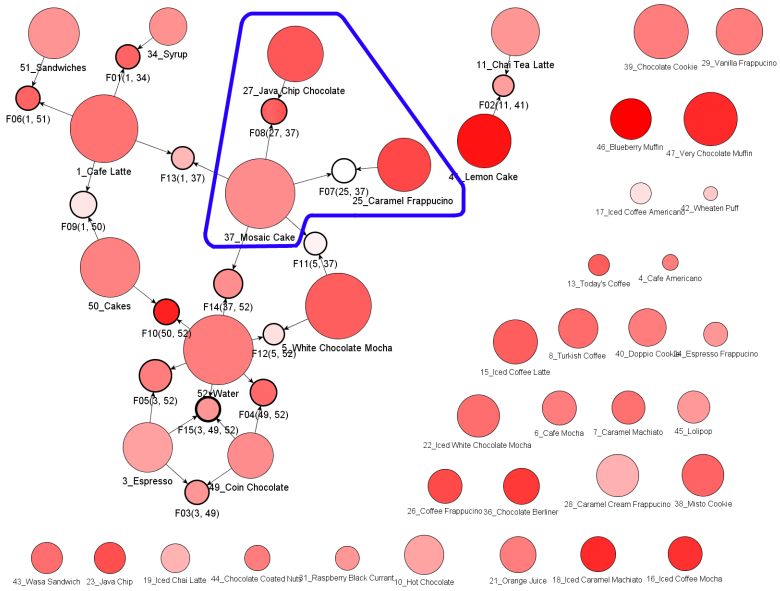

Fig. 2. Gender attribute (percentage of females selecting each itemset) mapped to color.

So, if only the content of the selected region is considered, the following policy can be applied:

Policy 3: "If a male customer orders Mosaic Cake, try to cross-sell to him Caramel Frappucino by suggesting that item; otherwise, if a female customers orders it, try to cross-sell to her Java Chip Chocolate."

Figure 3 displays the knowledge of French language (FrenchLevel) mapped to color. Darker vertices indicate that among the people that selected that itemset, the level of knowledge for the French language is higher on the average. The k-itemset ($k > 1$) with the darkest color is F07, which is the preference for the items Mosaic Cake and Caramel Frappucino together. Items that have the darkest colors, in order of decreasing support, are Very Chocolate Muffin, Doppio Cookie, Orange Juice, Caramel Machiato, and Java Chip. This discovery can be used in conjunction with geographical location of the stores, as the next policy suggests:

Policy 4: "If the store is located near a university or high-school where the language of instruction is French, then emphasize Very Chocolate Muffin, Doppio Cookie, Orange Juice, Caramel Machiato and Java Chip as stand-alone products, and the item pair Mosaic Cake and Caramel Frappucino as a bundle."

Figure 4 displays the hunger level of the customer (HungerLevel) mapped to color, where darker colors denote higher hunger level on the average. It is observed that none

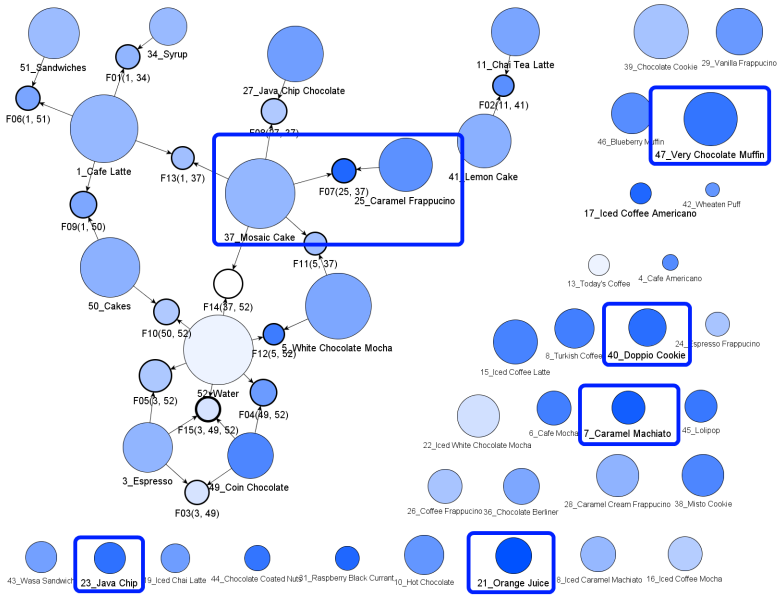

Fig. 3. Knowledge of the French language (average French knowledge of those selecting each itemset) mapped to color.

of the k-itemsets with $k > 1$ have white color. This is consistent with what would be expected, since a person who is not hungry is unlikely to order many items. The items that can be offered to a person, even if he is not hungry at all, are suggested in the next policy:

Policy 5: "If, at the point of sale (POS), a customer does not seem to be hungry, suggest Iced White Chocolate Mocha, Java Chip or Orange Juice."

The coloring of the vertices revealed insights on the behavioral patterns, explaining them through personal attributes. The fact that the association graphs can be interpreted even by the least technical analysts is a big advantage and a great motivation for using the methodology in the real world.

5 Validity

A fundamental question, regarding the validity of the proposed AssocGraphRM methodology, can be posed in the line of the classical dilemma of statistical data analysis [29]: "Is the discovered relation a result of causality, or is it just correlation?" For example, considering Policy 4 in Section 4, does a person who is fluent in French order Orange Juice due to his knowledge of French, or due to some other reason, which would explain both his language proficiency and preference for Orange Juice? For practical purposes, this is not a problem. Even if the underlying attribute for the behavior

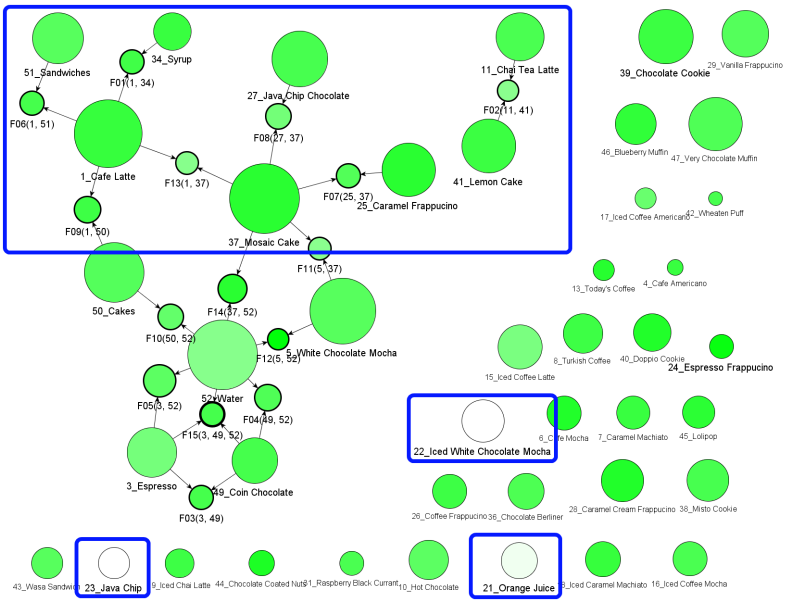

Fig. 4. Hunger level attribute (average hunger level for those selecting each itemset) mapped to color.

is hidden, the visible attribute FrenchLevel signals the presence of the detected specific behavior. The person should still be offered Orange Juice in the store (Policy 4), especially if he does not seem to be hungry (Policy 5).

A notable shortcoming of the methodology, related with the above issue, is that it enables re-mining only on a single additional attribute. However, there may be conflicting outcomes, or deeper interactions at deeper levels of analysis that would make the policies suggested at single-level depth invalid. For example, Policy 4 in the case study suggests Orange Juice to a person who is fluent in French. However, Policy 5 suggest the same item to customers who are not hungry at all. So what should be done when a *hungry* French-speaking person arrives? Due to FrenchLevel attribute, he should be offered Orange Juice, but due to HungerLevel attribute, he should definitely not be offered that item. One way to resolve this conflict would be to offer only "safe" items, which do not create a conflict. For the described example, Doppio Cookie and Caramel Machiato are two items that cater to both French-speaking people and to hungry people. So these items can be suggested to the mentioned customer.

A threat to validity of the analysis in the case study is the validity of the collected data. Preferences for food items are heavily influenced by the time of the day, as well as temperature and other conditions under which the data was collected. A French-speaking person would most probably prefer to drink Orange Juice in a warmer day, and a hot Caramel Machiato on a colder day, but Policy 4 does not currently differentiate between the two situations. The survey did not record all such conditions and thus the threat to the validity of the listed policies is indeed pertinent. However, the main contribution of this study is the AssocGraphRM methodology, rather than the specific policies, and this threat to validity does not affect the main contribution.

It is crucial that the policies obtained through AssocGraphRM be handled through a scientific analysis-based approach: They should be put into computational models for justifying their feasibility quantitatively. For example, Policy 3 in the case study of Section 4 suggests that Caramel Frappucino should be offered to a male customer, rather than Java Chip Chocolate, since he has a higher chance of accepting the former offer. But what if the profit margin of the latter was much higher? Then it might be feasible to offer to the customer the same item that is offered to the female customer. So the superiority of this policy can not be guaranteed by our methodology alone, without a formal fact-based numerical analysis. Thus, the policies should not be applied in isolation, but taking into consideration other critical information, and their interactions.

6 Conclusion and Future Work

A novel methodology was introduced for knowledge discovery from association mining results. The applicability of the methodology was illustrated through a market basket analysis case study, where frequent itemsets derived from transactional preference data were analyzed with respect to the personal attributes of survey participants. The theoretical contribution of the study is the methodology, which is formally described as an algorithm. The practical contribution of the study is the proof-of-concept demonstration of the methodology through a case study.

In every industry, especially food retail industry, new products emerge and consumer preferences change at a fast pace. Thus one would be interested in laying the foundation of an analysis framework that can fit to the dynamic nature of retailing data. The presented methodology can be adapted for analysis of frequent itemsets and association rules over time by incorporating latest research on evolving graphs [15] and statistical tests for measuring the significance of changes over time.

The main motivation of the chapter is the discovery of behavioral patterns in relation to the human that exhibited the behavior. The methodology can be applied to similar data from different fields that study the behavior of agents individually and in relation to each other, including psychology, sociology, behavioral economics, behavior-based robotics, and ethology.

Acknowledgement

The authors thank İlhan Karabulut for her work that inspired this research, Ahmet Şahinöz for creating colored graphs with the earlier datasets, that inspired the final form of the graphs. The authors thank Samet Bilgen, Dilara Naibi, Ahmet Memişoğlu, and Namık Kerenciler for collecting the data used in the study, and to Didem Cansu Kurada for her insightful suggestions regarding the study.

References

1. R. Agrawal and R. Srikant. Fast algorithms for mining association rules in large databases. In *Proceedings of the 20th International Conference on Very Large Data Bases*, pages 487–499. Morgan Kaufmann Publishers Inc., 1994.
2. C. Borgelt. `http://fuzzy.cs.uni-magdeburg.de/~borgelt/apriori.html`, 2011.
3. T. Brijs, G. Swinnen, K. Vanhoof, and G. Wets. Building an association rules framework to improve product assortment decisions. *Data Mining and Knowledge Discovery*, 8(1):7–23, 2004.
4. L. Cao. Behavior informatics and analytics: Let behavior talk. In *Data Mining Workshops, 2008. ICDMW '08. IEEE International Conference on*, pages 87 –96, 2008.
5. L. Cao and C. Zhang. The evolution of KDD: towards domain-driven data mining. *International Journal of Pattern Recognition Artificial Intelligence*, 21(4):677 – 692, 2007.
6. A. Ceglar and J.F. Roddick. Association mining. *ACM Computing Surveys (CSUR)*, 38(2):5, 2006.
7. S. W. Changchien and T.-C. Lu. Mining association rules procedure to support on-line recommendation by customers and products fragmentation. *Expert Systems with Applications*, 20(4):325 – 335, 2001.
8. M. Chatti, M. Jarke, T. Indriasari, and M. Specht. NetLearn: Social Network Analysis and Visualizations for Learning. *Learning in the Synergy of Multiple Disciplines*, pages 310–324, 2009.
9. C. Chen. Information visualization. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(4):387–403, 2010.
10. A. Demiriz. Enhancing product recommender systems on sparse binary data. *Data Mining and Knowledge Discovery*, 9(2):147–170, 2004.

11. A. Demiriz, G. Ertek, T. Atan, and U. Kula. Re-mining positive and negative association mining results. *Lecture Notes in Computer Science*, 6171:101–114, 2010.
12. A. Demiriz, G. Ertek, T. Atan, and U. Kula. Re-mining item associations: Methodology and a case study in apparel retailing. *Decision Support Systems*, doi:10.1016/j.dss.2011.08.004. 2011.
13. EditGrid. `http://www.editgrid.com`. 2011.
14. G. Ertek and A. Demiriz. A framework for visualizing association mining results. *Lecture Notes in Computer Science*, 4263:593–602, 2006.
15. C. Erten, P.J. Harding, S.G. Kobourov, K. Wampler, and G. Yee. GraphAEL: Graph animations with evolving layouts. In *Lecture Notes in Computer Science*, volume 2912, pages 98–110. Springer, 2004.
16. T.M.J. Fruchterman and E.M. Reingold. Graph drawing by force-directed placement. *Software: Practice and Experience*, 21(11):1129–1164, 1991.
17. G. Grahne and J. Zhu. Fast algorithms for frequent itemset mining using fp-trees. *IEEE Transactions on Knowledge and Data Engineering*, pages 1347–1362, 2005.
18. J. Han and M. Kamber. *Data mining: concepts and techniques*. Morgan Kaufmann, 2006.
19. I. Herman, G. Melançon, and M.S. Marshall. Graph visualization and navigation in information visualization: A survey. *IEEE Transactions on Visualization and Computer Graphics*, 6(1):24–43, 2000.
20. W. Kim, B.J. Choi, E.K. Hong, S.K. Kim, and D. Lee. A taxonomy of dirty data. *Data Mining and Knowledge Discovery*, 7(1):81–99, 2003.
21. S. Kimani, S. Lodi, T. Catarci, G. Santucci, and C. Sartori. VidaMine: a visual data mining environment. *Journal of Visual Languages & Computing*, 15(1):37 – 67, 2004.
22. B. Liu, W. Hsu, and Y. Ma. Pruning and summarizing the discovered associations. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 125–134. ACM, 1999.
23. H. Ltifi, B. Ayed, A.M. Alimi, and S. Lepreux. Survey of information visualization techniques for exploitation in KDD. In *Computer Systems and Applications, 2009. AICCSA 2009. IEEE/ACS International Conference on*, pages 218–225. IEEE, 2009.
24. O.Z. Maimon and L. Rokach. *Data mining and knowledge discovery handbook*. Springer-Verlag New York Inc, 2005.
25. G. Mansingh, K.M. Osei-Bryson, and H. Reichgelt. Using ontologies to facilitate post-processing of association rules by domain experts. *Information Sciences*, 2010.
26. F. Mansmann, F. Fischer, D.A. Keim, and S.C. North. Visual support for analyzing network traffic and intrusion detection events using treeMap and graph representations. In *Proceedings of the Symposium on Computer Human Interaction for the Management of Information Technology*, pages 19–28. ACM, 2009.
27. R. Mazza. *Introduction to information visualization*. Springer-Verlag New York Inc, 2009.
28. Microsoft. MS SQL Server, Analysis Services. `http://tinyurl.com/6gudq23`. 2011.
29. S. Nowak. Some problems of causal interpretation of statistical relationships. *Philosophy of science*, 27(1):23–38, 1960.
30. S. OHare, S. Noel, and K. Prole. A graph-theoretic visualization approach to network risk analysis. *Visualization for Computer Security*, pages 60–67, 2008.
31. G.A. Pavlopoulos, A.L. Wegener, and R. Schneider. A survey of visualization tools for biological network analysis. *Biodata mining*, 1:12, 2008.
32. A. Perer and B. Shneiderman. Integrating statistics and visualization: case studies of gaining clarity during exploratory data analysis. In *Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*, pages 265–274. ACM, 2008.
33. R. Santamaría and R. Therón. Overlapping clustered graphs: co-authorship networks visualization. In *Smart Graphics*, pages 190–199. Springer, 2008.

34. SAS. http://www.sas.com. 2011.
35. Z. Shen and K.L. Ma. Mobivis: A visualization system for exploring mobile data. In *IEEE Pacific Visualization Symposium, 2008, PacificVIS'08*, pages 175–182. IEEE, 2008.
36. R. Tamassia, B. Palazzi, and C. Papamanthou. Graph drawing for security visualization. In *Graph Drawing*, pages 2–13. Springer Berlin/Heidelberg, 2009.
37. I. Tollis, P. Eades, G. Di Battista, and L. Tollis. *Graph drawing: algorithms for the visualization of graphs*, Prentice Hall. 1998.
38. M. Wattenberg. Visual exploration of multivariate graphs. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*, pages 811–819. ACM, 2006.
39. WolframAlpha. http://www.wolframalpha.com, 2011.
40. Y. Yao, Y. Zhao, and R. Maguire. Explanation-oriented association mining using a combination of unsupervised and supervised learning algorithms. *Lecture Notes in Computer Science*, 2671:527–532, 2003.
41. M.J. Zaki. Scalable algorithms for association mining. *Knowledge and Data Engineering, IEEE Transactions on*, 12(3):372–390, 2002.
42. M.J. Zaki and C.J. Hsiao. Efficient algorithms for mining closed itemsets and their lattice structure. *Knowledge and Data Engineering, IEEE Transactions on*, 17(4):462–478, 2005.
43. Y. Zhao, H. Zhang, L. Cao, C. Zhang, and H. Bohlscheid. Combined pattern mining: from learned rules to actionable knowledge. *Proc. of the 21st Australasian Joint Conference on Artificial Intelligence (AI 08)*, pages 393–403, 2008.
44. P.D. McNicholas and Y. Zhao *Association Rules: An Overview*, in *Post-Mining of Association Rules: Techniques for Effective Knowledge Extraction*, Yanchang Zhao, Chengqi Zhang and Longbing Cao (Eds.), ISBN 978-1-60566-404-0, May 2009. Information Science Reference. pp. 1-10. IGI Publishing Hershey, PA. 2009.