Designing a Well-Structured E-Shop Using Association Rule Mining

Asem Omari, Stefan Conrad and Sadet Alcic
Heinrich-Heine University Düsseldorf
Institute of Computer Science
Database and Information System
40225 Düsseldorf, Germany
{*omari, conrad, alcic*}*@cs.uni-duesseldorf.de*

Abstract

Many commercial companies collect large quantities of data from daily operations. For example, customer orders or purchase data are collected daily at the counters of grocery stores. Data mining is applied on such kind of data to extract patterns that could be useful to learn about the purchasing behavior of the customers. Such information are used to support a variety of business related tasks. For example, the investment of that kind of information in building a website for a grocery store. Association Rule Mining is one of the techniques used to mine databases. Association Rule Mining is the discovery of Association Rules showing attribute values that occur frequently together. In this paper, we have discovered Association Rules from a grocery store dataset which represents customer transactions in that grocery store. Those rules have been invested to design a well-structured website prototype for that grocery store. Promising results, that could affect the process of website design, have been found. The experiments showed that our method can reduce the cost up to 90% in some transactions.

Keywords: Association Rule Mining, E-Shop Design, Market Basket Analysis.

1. Introduction

Data Mining is the extraction of useful patterns from large databases. One major application area of Data Mining is mining Association Rules among items in a database of sales transactions which is known as Market Basket Analysis. The rules extracted using such a Data Mining technique can be used to support a variety of business related tasks. In our approach, we get benefit from that extracted patterns to support the design of the company's website that the transactions database belongs to. say that we have a physical grocery store which has no website, but it has a dataset that records the transactions of its customers. Association Rule Mining techniques are applied on this dataset. The extracted interesting Association Rules from the transactions dataset of the grocery store are took into account in the process of designing a website for the grocery store. The extracted Association Rules are invested to support the website design from the beginning of the design process which is the main contribution of our work.

This is done in the design phase through improving the structure of the website depending on the extracted patterns in a way that makes it easy for the website's navigator to find his target products in an efficient time, give him the opportunity to have a look at some products that may be of interest for him, and encourage him to buy more from the available products which will consequently increase the company's overall profit. Many improvements and modifications can be done to the website's design, such as adding/modifying links, and/or creating index pages.
This paper is structures as follows: Related work is presented in Section 2. In section 3, we give an overview about Association Rule Mining, and we discuss the Apriori algorithm as we use it in our experimental work. and we will see how we can use the extracted patterns using the Association Rule Mining to improve website's design structure. The experimental work is presented in section 4. An evaluation of the experimental work is presented in section 5. Finally, in section 6, we summarize our paper and present our future work.

2. Related work

Website design is a growing subject for many years due to the important rule of websites in improving the efficiency of organizations, and supporting company's marketing strategies. A lot of research has been done to cover different website design techniques, and strategies. The work in [3] provides a survey of experts' recommendations

978-1-4244-1841-1/08/$25.00 ©2008 IEEE.

of how to create an effective website from an e-commerce point of view. It investigates the determinants of an effective website. The authors in [8] present a method for designing kiosk websites which are websites that provide information and allow users to navigate through that information. The method is based on the principle that the website should be designed and adapted to its users. It starts by identifying different classes of users and describe manually their characteristics, and their information requirements, and how could they navigate the website. The work in [5] gives some recommendations and remarks on how to design retail websites. For example, stores that offer a FAQs section have more visits than those without such a section, and every web page must have consistent navigation links to move around on the site. [4] presents a technique for redesigning a large and complex website and provides usability practices and techniques. It provides some tips and practical issues and solutions for developing a solid information architecture and for implementing web standards. The authors indicate that the decisions about products and services to offer in the website are influenced by: The organization's strategic goals, the expertise of it, employees, past history, process and infrastructure. An important aspect, with respect to the authors, of redesigning a large and complex websites is knowing the target users. Creating user profiles and keep them in mind help the redesign team focus on the core issues. This facilitates the identification of related pages or navigation patterns which can be used in web personalization.

In our approach, instead of using Association Rule Mining in maintaining the already built websites, we go one step back and invest the extracted interesting Association Rules from a grocery store or a company to design a well-structured website for that grocery store. In that way, we can better plan marketing and advertising strategies which will consequently increase the grocery store overall profit. Furthermore, the main advantage of our method is that it reduces maintenance time and budgetary costs for websites if they are built without taking into account the associations between different products, and customer buying habits that can be found in the transactions database in almost every shop or grocery store. It also permits the sales manager to focus on the core business and gives him a better view about his products and customers which is very helpful in designing retail websites. This method also participates in improving customer satisfaction and encourages him to be a frequent buyer.

3. Association rule mining

In our previous work [2] [1] we investigated the usage of Data Mining to support website's designers to have better designed websites. We explained briefly different Data Mining techniques and we showed how we can use them to support website design. Association Rule Mining is one of the Data Mining techniques that plays an important role in our approach. An Association Rule is an expression of the form $X \Rightarrow$Y, where X and Y are sets of items and have no items in common. This rule means that given a database of transactions D where each transaction $T \in$D is a set of items. $X \Rightarrow$Y denotes that whenever a transaction T contains X then there is a probability that it contains Y, too. The rule $X \Rightarrow$Y holds in the transactions set T with confidence c if $c\%$ of transactions in T that contain X also contain Y. The rule has support s in T if $s\%$ of the transactions in T contains both X and Y. Association Rule Mining is finding all Association Rules that are greater than or equal a user-specified minimum support ($minsup$), and minimum confidence *(minconf)*. In general, the process of extracting interesting Association Rules consists of two major steps. The first step is finding all itemsets that satisfy minimum support (known as *Frequent-Itemset* generation). The second step is generating all Association Rules that satisfy minimum confidence using itemsets generated in the first step.

3.1 The apriori algorithm

The Apriori algorithm [7] generates all frequent itemsets, called also large itemsets, by making multiple passes over the transactions database D. The algorithm makes a single pass over the data to determine the support of each item which results in the set of 1-itemsets. Next, the algorithm will iteratively generate new candidate k-itemsets using the frequent $(k-1)$-itemsets found in the previous iteration. An additional pass over the data set is made to count the support of the candidates. After counting their supports, the algorithm eliminates all candidate itemsets whose support count are less than *minsup*. The algorithm eliminates some of the candidate k-itemsets using the support-based pruning strategy. If any subset of the k-itemset X is not frequent then X is pruned. The algorithm terminates when there are no new frequent itemsets generated. Association rules are generated by generating all non-empty subsets of each frequent itemset and outputs its rule if its confidence is $\geq$ *minconf*.

3.2 Supporting website's design using association rule mining

After generating frequent items, Association Rules that are $\geq minconf$ are generated. Those rules are called interesting Association Rules. They can be invested in many different applications. One of that applications is improving the structure of the company's website that the mined database belongs to. This is done in the website's design phase by creating links between items that seem to be sold

together, or highlight that links if they already exist, and/or create index pages which are pages that have direct links to some products that may be of interest for some group of customers. Consequently, such modifications done to the website's design help customers to find their target products in an efficient time, encourage them to buy more from the available products, and give them the opportunity to have a look at some products that may be of interest for them, which will consequently increase the company's overall profit.

4. Experimental work

For the experiments, we used a dataset that represents customer transactions in a grocery store. This dataset consists of 15 attributes. 10 attributes represent the available products: *Readymade, Frozenfood, Alcohol, Freshvegetables, Milk, Backerygoods, Freshmeat, Toiletries, Snacks,and Tinnegoods*. The remaining 5 attributes: *Gender, Age, Martial, Children, and Working*, represent the gender of the customer, his/her age, his/her marital status, having children or not, and if the customer is a worker or not, respectively. In the mining step, we applied the Apriori algorithm implemented within the Association Rule miner of the WEKA tool [6]. Before running any Association Rule Mining algorithm, we designed the prototype of the website that represents the grocery store and the products available in it. Figure 1 shows that prototype, where Boxes circles represent web pages, and arrows represent links between pages. Of course, this prototype does not represent all products available in the grocery store, it just represents the set of products that are available in the transactions of the tested dataset. A website for the grocery store need to be built in order to give a good view about it, to support and facilitate services introduced to customers, and to encourage customers to buy more from the available products. The website design process starts by identifying the goal of building this website, looking at the transactions database and trying to understand its structure, and checking out what kinds of information is available in the transactions database and whether this information should be presented in the website or not [4]. Beside this information, we considered some standards and recommendations adopted by website designers to have a well-structured website such as every page should be connected directly to the home page, and every parent page should have direct links to its descendants [3].

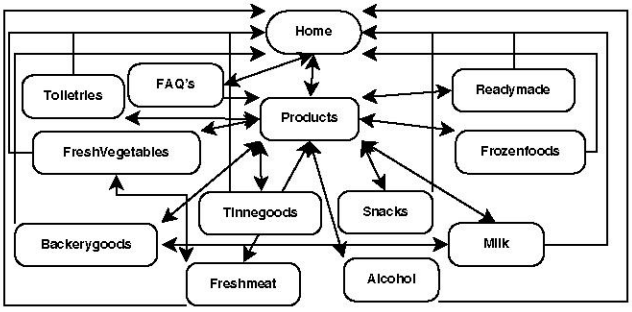

Figure 1. Initial website prototype of the grocery store

4.1 The investment of the extracted interesting association rules

As we mentioned before, we used the Apriori algorithm to mine for interesting Association Rules. In the mining process, we used different $minsup$ and $minconf$ values ranging from 8% to 20% for $minsup$, and from 70% to 90% for $minconf$. The best extracted Association Rules have been studied and analyzed in order to decide how to invest them in the process of designing the website of the grocery store. The website prototype is represented in figure 2. The following rules are an example of the extracted interesting Association Rules:

```
1. alcohol=1 milk=1
  ==> working=Yes conf:(0.96)
2. alcohol=1 bakerygoods=1 tinnedgoods=1
    ==> readymade=1  conf:(0.81)
```

Both rules are extracted by setting the minimum support value $minsup$ to be 20% for the first rule and 8% for the second rule. The $minconf$ was set to 90% and 80% respectively. The first rule says that 96% of customers who buy *alcohol* and/or *milk* are workers, and 20% of all customers buy *alcohol* and/or *milk* and they are *workers*. From this rule, and other similar extracted interesting Association Rules, an index page is decided to be created. That index page have direct links to products that may be of interest for the customers who are *workers*. As we see in figure 2, we called this index page *offers for workers*. It has direct links to *Readymade, Frozenfood, Snacks, Alcohol, and Milk*. The second rule means that 81% of transactions which contain *Alcohol* and/or *bakerygoods* and/or *tinnedgoods* contain also *readymade*, and 8% of all transactions contain all (i.e *Alcohol*, *bakerygoods*, *tinnedgoods*, *readymade*). From this rule and other similar interesting rules, direct links from *Alcohol*, *bakerygoods*, and *tinnedgoods* to *readymade* are created. Some arrows in the prototypes are bidirectional, that means that the pair of products connected by a bidirectional arrow are frequently bought together. In other words, customers who buy the first product buy the second product with, and vice versa. For example, customers who buy *backerygoods* buy also *readymade* with, and vice versa. On the other hand, one

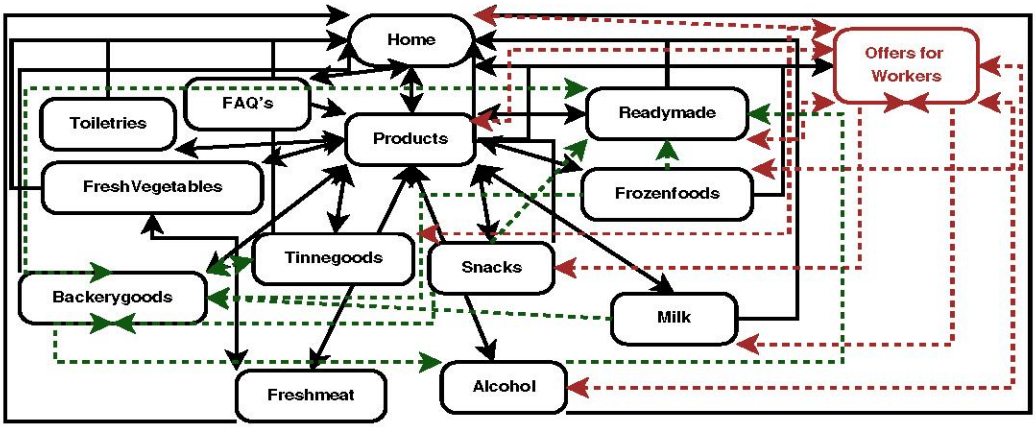

Figure 2. Website prototype with the help of extracted AR's

directional arrows employ that customers who buy the first product buy the second product with, but not vice versa. For example, in the prototype of figure 1, a bidirectional arrow between *bakerygoods* and *milk* was created because we believed that those two products are strongly related to each other, so they needed to be connected. But from the extracted rules, we found that a percentage of customers who buy *milk* buy also *bakerygoods*. In contrast, we found no rule which indicates that the customers who buy *bakerygoods* buy also *milk* with. From that, we modified the arrow between those two products to be one directional arrow from *milk* to *bakerygoods*.

As a result, such interesting Association Rules can be used to design a well-structured website, plan marketing and advertising strategies which will consequently increase the grocery store overall profit. Furthermore, the main advantage of our method is that it reduces maintenance time and budgetary costs for websites if they are built without taking into account the associations between different products, and customer buying habits that can be found in the transactions database in almost every shop or grocery store. It also permits the sales manager to focus on the core business and gives him a better view about his products and customers which is very helpful in designing retail websites. This method also participates in improving customer satisfaction and encourages him to be a frequent buyer.

5. Method evaluation

In order to evaluate our method we implemented a simulation tool in Java. This tool simulates the behavior of customers in both website prototypes. We divide our dataset into two parts. The first part was used to extract Association Rules that have been later used in improving the elementary website prototype as discussed in section 4. The second part was used in the testing process. In every transaction in the dataset, the customer has a list of products he wants to buy. Those products are considered as to be the set of target products. Every product is represented by a page in the website prototype. In the elementary prototype in figure 1, the customer starts at the home page. Then, he starts searching for his first target product. The first target product is chosen randomly from the set of target products presented in every transaction. So, he has to go to *products* and from there he searches for his first target product. After that, if there is a direct link from that target product page to one of the next target product pages, the new target product page is visited, otherwise he has to backtrack to *products* in order to search for the next target products. This process is repeated until all pages of target products are found.
In the prototype in figure 2, the customer starts at the home page. Then, before starting to search for his target products, it is checked if the customer is a *Worker* or not. If he is a *Worker*, then he goes directly to *offers for workers* page. From there, he will start searching for his target products. If there exist a direct link to a product page of a certain product in the set of target products, then it would be followed. If there is a direct link from that target product page to one of the next target product pages then the new target product page is visited, otherwise the customer have to backtrack to *offers for workers* in order to search from there for the next target products. The green columns (titled *Improved*) represent the average costs of the improved prototype in figure 2. Our method reduced the cost up to 90% for some transactions in comparison with the costs of the elementary prototype.

If there are no products of the set of target products presented in the *offers for workers*, then the customer backtracks to *products* in order to search for the rest of his target products until all pages of target products are found. In every step in the process of searching for target products,

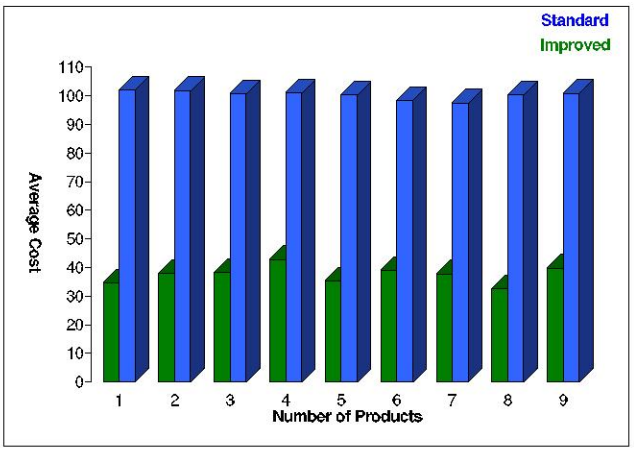

Figure 3. The average costs of both prototypes

all existing links Li from any visited page $i = 1, 2, ..., m$ to any other pages are calculated. We defined the cost C_T of every prototype to finish visiting product pages in some transaction T to be the ratio between the sum of all existing visited links in every transaction $\sum_{i=1}^{m} Li$, and the number of target pages (i.e. products) n, where $n \neq 0$:

$C_T = (\sum_{i=1}^{m} Li)/n$

The lower the value of the C_T is, the higher the efficiency of the prototype is to find target products in transaction T. In other words, the customer will need less time and effort in order to reach his target products as the value of the C_T becomes lower. We ran the simulation tool at both prototypes simultaneously. We assumed that we have a total of 100 products in the grocery store. Figure 3 shows the efficiency of both prototypes to finish 500 transactions. The columns represent the average costs with respect to different number of products in both prototypes. The blue columns (titled *Standard*) represent the average costs of the elementary prototype in Figure 1.

6. Summary and future work

In this paper, with the help of Association Rule Mining, we introduced a method to design an improved and well-structured website design for an E-shop in the design phase. In other words, we have a physical grocery store which has no website, but it has a dataset that records the transactions of its customers. Association Rule Mining techniques are applied on this dataset. The extracted interesting Association Rules from the transactions dataset of the grocery store are took into account in the process of designing a website for the grocery store. The extracted Association Rules are invested to support the website design from the beginning (i.e. in the design phase). Many improvements and modifications are done to the website's design, such as adding/modifying links, and/or creating index pages. We introduced a technique to evaluate our method by comparing the navigation efficiency among different website designs. The experiments showed very promising results of our method that could be very effective in the process of designing websites. Our method will not only reduce the navigation costs but it also reduces the maintenance costs needed in the future.

As a future work, we plan to apply other techniques to evaluate our method, for example by making questionnaires, or allowing a group of users navigate through our website design to test their navigation behavior. We also plan to find other suitable datasets to make more tests and compare the efficiency of our method among different datasets. we plan to use patterns extracted using other Data Mining techniques such as clustering and classification in the process of designing a website for some grocery store or company. The automation of the process of building the improved prototype belongs also to the future work.

References

[1] Asem Omari and Stefan Conrad. Association Rule Mining and Website's Design Improvement. In *(18th GI-Workshop on the Foundations of Databases), 6- 9 June 2006, Wittenberg, Sachsen-Anhalt, Germany*, pages 115–119, 2006.

[2] Asem Omari and Stefan Conrad. On the Usage of Data Mining to Support Website Designers to Have Better Designed Websites. In *Advanced International Conference on Telecommunications and International Conference on Internet and Web Applications and Services (AICT-ICIW'06), Guadeloupe, French Carribian*, volume 0, page 171, Los Alamitos, CA, USA, 2006. IEEE Computer Society.

[3] Dave Gehrke. Determinants of Successful Website Design: Relative Importance and Recommendations for Effectiveness. In *HICSS 1999: Proceedings of the Thirty-second Annual Hawaii International Conference on System Sciences*, volume 50, page 5042, Washington, DC, USA, 1999. IEEE Computer Society.

[4] Elaine Chou. Redesigning a Large and Complex Website: How to Begin, and a Method for Success. In *SIGUCCS 2002: Proceedings of the 30th annual ACM SIGUCCS conference on User services*, pages 22–28, New York, NY, USA, 2002. ACM Press.

[5] G. L. Lohse and P. Spiller. Electronic Shopping. *Commun. ACM*, 41(7):81–88, 1998.

[6] Ian H. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Publishers, San Francisco, 2005.

[7] J. Han and M. Kamber. *Data Mining Concepts and Techniques*. Morgan Kaufmann Publishers, San Francisco, 2001.

[8] O. D. Troyer and C. J. Leune. WSDM: A User Centered Design Method for Web Sites. *Computer Networks*, 30(1-7):85–94, 1998.