Mariusz Łapczyński

Sylwester Białowąs

Cracow University of Economics, Poland

DISCOVERING PATTERNS OF USERS' BEHAVIOUR IN AN E-SHOP – COMPARISON OF CONSUMER BUYING BEHAVIOURS IN POLAND AND OTHER EUROPEAN COUNTRIES

Introduction

Modern forms of commerce and sales support systems are designed to offer businesses a great potential for analyzing purchases. In the past sellers could draw conclusions concerning the sale on the basis of their own observations. Today's supermarkets and hypermarkets generate so many transactions that managers' intuition is insufficient. For example, the number of transactions in the U.S. Wal-Mart chain reaches tens of millions per day[1]. The challenge is to use these huge databases in a way that would help to achieve a competitive advantage. One of the methods whose name often appears in the context of the analysis of shopping carts is data mining. From e-commerce point of view data mining models can be useful in[2]: attracting new visitors to the site, ensuring that the visitors will find the product they need and supporting the sales process.

1. Web usage mining

Web mining can be defined[3] as the discovery and analysis of useful information from the Internet. This is done by automatic searching for the information and

[1] T. Brijs et al.: *Using Association Rules for Product Assortment Decisions: A Case Study*. In: Proceedings KDD-99, San Diego, California 1999, pp. 254–260.

[2] G.S. Linoff, M.J.A. Berry: *Mining the Web. Transforming Customer Data into Customer Value*. John Wiley & Sons, New York 2001.

[3] R. Cooley, B. Mobasher, J. Srivastava: *Web Mining: Information and Pattern Discovery on the World Wide Web*. Proceedings of the 9th International Conference on Tools with Artificial Intel-

resources available on-line (web content mining) and discovering users' behaviour patterns from web logs (web usage mining). From the point of view of marketing research and analytical CRM (Customer Relationship Management), it is most interesting to examine buying behaviours, which can be treated as a part of web usage mining.

This area refers to the discovery and analysis of web clickstreams together with other variables collected or generated during contacts of users with different websites. Patterns and models are usually presented in the form of a set of pages, objects or resources, and are related to those with the highest access frequency. As far as possible they should pertain to homogenous groups of visitors in terms of their needs or interests.

In the literature one can find a wide variety of applications of data mining. The analysis was used for targeting and predicting customer behaviour related to new telecommunications services[4]. In the food sector market the basket analysis was utilized to determine the price elasticity[5]. Data mining was also used for shelf management in the retail sector[6] and for building promotional campaigns[7].

1.1. Market basket analysis

In a broader sense, the market basket analysis means monitoring shopping patterns in order to increase consumer satisfaction. In this sense, analyses include among others demographic variables, brand switching, loyalty factors, brand penetration rates and descriptive statistics of the shopping cart.

A narrower meaning of that method means searching for sets of products that are bought together. The purpose is to find frequent and strong rules regarding purchases, and subsequently use them to create an offer that would eventually increase the sales volume.

The use of association rules in traditional trade is usually reduced to the adequate shelf placement of products and their promotion, but rarely results in pricing

ligence, IEEE Computer Society 1997, p.558.

[4] S. Sohn, Y. Kim: *Searching Customer Patterns of Mobile Service Using Clustering and Quantitative Association Rule*. "Expert Systems With Applications" 2008, No. 34(2), pp. 1070–1077.

[5] G. J. Russell, A. Petersen: *Analysis of Cross Category Dependence in Market Basket Selection*. "Journal of Retailing" 2000, No. 76, pp. 367–392.

[6] M.-C. Chen, C.-P. Lin: *A Ddata Mining Approach to Product Assortment and Shelf Space Allocation*. "Expert Systems With Applications" 2007, No. 32(4), pp. 976–986.

[7] D. Van den Poel, J. D. Schamphelaere, G. Wets: *Direct and Indirect Effects of Retail Promotions on Sales and Profits in the Do-it-yourself Market*. "Expert Systems With Applications" 2004, No. 27(1), pp. 53–62.

decisions. The application of the rules in e-commerce provides many more opportunities. The knowledge of buying patterns enables to recommend new products precisely to the visitors on the website. In this study, data mining is treated in the narrower meaning and refers to an e-shop selling women's clothing.

1.2. Searching for sequential rules in the web

An interesting example of an analysis of behavior patterns of customers to an e-shop offering computer equipment can be found in the work of Zhang et al[8]. The purpose of the experiment was to build a model of users' behaviors on the website and implement it in a real-time recommendation system. Particular sequences of clicks were grouped by self-organizing maps in 20 clusters. The results obtained in this way proved to be better than those obtained by using the k-means algorithm. Dividing visitors into separate homogeneous clusters enabled to provide potential customers with tailored information about the offered products.

Munk and Drlík used sequence rules while analyzing sequential behavior patterns of students visiting an educational website[9]. Typically, the purpose of such research is optimization of the website content, website personalization based on past users' behaviors and recommendation of an educational path. The authors focused their efforts on the phase of data preparation for the analytical process. They studied how the different data pre-processing methods influenced the number and quality of sequential rules.

Analyzing the behaviour of Internet users in e-commerce also related to food products, and more specifically olive oil[10]. The main goal of the study was focused on the design of the online retailer website. The authors used directed and undirected data mining models in their study. The clustering was conducted by using the k-means algorithm. The *Apriori* algorithm was used while searching for associations. In the subgroup discovery phase the NMEEF-SD algorithm was applied.

[8] X. Zhang, J. Edwards, J. Harding: *Personalised Online Sales Using Web Usage Data Mining*. "Computers in Industry" 2007, No. 58, pp. 772–782.

[9] M. Munk, M. Drlik: *Impact of Different Pre-Processing Tasks on Effective Identification of Users' Behavioral Patterns in Web-based Educational System*. "Procedia Computer Science" 2011, No. 4, pp. 1640–1649.

[10] C.J. Carmona et al.: *Web Usage Mining to Improve the Design of an e-commerce Website: OrOliveSur.com*. "Expert Systems with Applications" 2012, No. 39, pp. 11243–11249.

2. Research methodology

2.1. Association rules

Searching for associations rules is a part of unsupervised learning (also referred to as undirected data mining). The most popular and the first algorithm (Apriori) was introduced in 1993[11]. The rule takes the form "if condition then result", where the condition can also be named left hand side (LHS), antecedent or body of a rule, and the result is also named right hand side (RHS), consequent or head of a rule. In general both of them are the sets of items that appeared together. The most popular area of application of association rules in the marketing field is the market basket analysis. For example, the rule "if bread then eggs" means that the customer who bought bread also bought eggs. There are several measures that help to choose the appropriate rule, the most popular ones among which are the support, confidence and lift. The rule "if bread then eggs" (support = 10%, confidence = 40%, lift = 2) can be interpreted as follows: 10% of all customers bought both these products, 40% out of all who bought bread also bought eggs. Lift equals 2 indicates that the probability of buying eggs among bread buyers is twice higher than in the entire set of transactions. In this experiment, the authors use qualitative (Boolean) association rules, namely those that include merely information about the categories of purchased products.

2.2. Sequence analysis

Discovering sequences[12], similarly to looking for associations and clustering, is an example of unsupervised learning (also referred to as undirected data mining). Sequential rules have a similar form to association rules, that is "if antecedent then consequent." The difference between them lies in the order of appearance of items in the body or head of the rule. In the analysis of associations that was enough merely to identify the two objects occurring together with a specified frequency (support) and specified strength (confidence). As far as the market basket analysis is concerned, the rule "if A and B then C" is usually interpreted in the same way as a the rule "if B and A then C". The task for researchers is to identify

[11] R. Agrawal, T. Imielinski, A. Swami: *Mining Association Rules Between Sets of Items in Large Databases*. In: Proceedings of the ACM SIGMOD Conference on Management of Data. Washington D.C., May 1993, pp. 207-216.

[12] B. Liu: *Web Data Mining. Exploring Hyperlinks, Contents, and Usage Data*. Springer, Heidelberg 2007.

transactional patterns, and the products A, B and C are purchased during a single visit to the store, a website, etc. In the case of sequential rules the difference between antecedents (A and B as well as B and A) has its unique interpretation since the element that appears in the rule as the first, has been purchased first.

3. Results

A set of log files from an e-shop offering clothing for women was used in the research. The dataset contained 210,814 records and referred to the apparel from the spring and summer collections in the following categories: skirts and dresses, trousers, blouses, special offers. The website had several language versions including English, German and Czech, so potential customers could originate from outside Poland. The company accepted the possibility of shipping abroad, which also facilitated the execution of orders from distant locations. The limited volume of paper forced the authors to present selected results which were reduced here to the set of blouses and tunics.

IP addresses of visitors were obtained from web logs, and then on the basis of geolocation each address in the database was assigned to a particular country. The identification of the user's country of residence was not fully possible due to several difficulties that emerged. In some cases the domain names ended with the letters .net, .com, .biz, or .org and were also recorded by using text instead of numbers. Sometimes the geolocation system did not indicate a specific country but merely a continent, e.g. Europe. These cases were classified as unidentified (2.5% of entire dataset). Some addresses came from very distant and exotic locations, e.g. Cayman Islands, Christmas Island, Faroe Islands, but this may mean that the user connects to the Internet via a server located in these countries while remaining in Poland. One can also assume that if the service provider is located close to the border with a neighbouring country, it acquires its clients from abroad. One cannot therefore exclude the situation that a user connects from Poland via a Czech server and is classified as a customer from the Czech Republic.

3.1. Descriptive statistics

Considering the models of blouses (Figure 1) the differences in consumer preferences may be noticed. The most popular models among users from Poland are C5, C17 and C14, and the most popular colours are white, grey and colourful. The models C5, C17 and C56 were most frequently chosen by visitors from other EU countries, while the models C56, C6 and C2 were bought by customers from the remaining European countries. The most frequently displayed colours of blouses and tunics were the same for all visitors from Europe. Slightly different preferences can be observed among visitors from outside of Europe. Their favourite models are C35, C38 and C2, and their favourite colours are white, black and colourful.

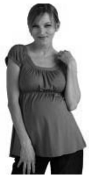
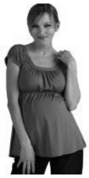
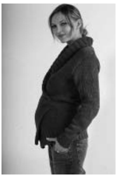
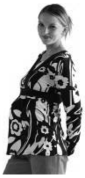
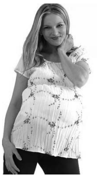
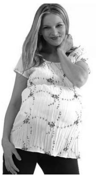
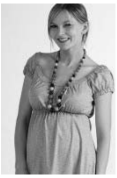
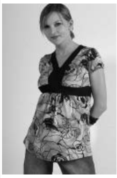
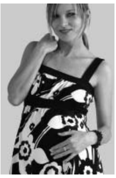
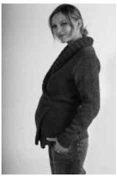
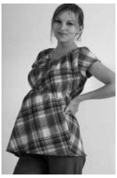
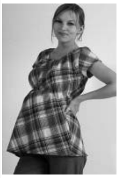
Poland		UE		Non-EU countries		Countries outside Europe	
Type of blouse	Percent	Type of blouse	Percent	Type of blouse	Percent	Type of blouse	Percent
C5 	4.8	**C5** 	4.3	**C56** 	4.9	**C35** 	9.3
C17 	4.0	**C17** 	3.7	**C6** 	4.9	**C38** 	7.4
C14 	3.2	**C56** 	3.6	**C2** 	3.9	**C2** 	7.4

Figure 1. The percentage of users choosing different models of blouses

58% of Polish customers chose blouses the price of which is higher than the average price for this product category (blouses, tunics, sweaters, etc.). Similar choices are made by 55% of users from other EU countries, 54% of visitors from other European (non-EU) countries and 61% of customers from outside of Europe. Visitors from Poland more frequently than others limit their selection of products to the first page. This concerns 45% of the users from Poland, and only 39% of other customers.

Association rules in each segment include some different types of blouses in comparison to those that were selected most often. In the following figures (2-5) transactional patterns with the highest value of support are displayed. Customers from Poland often chose pairs of blouses with short sleeves. 2.97% of all transactions contained items C12 and C17. 37.72% of buyers who chose the top C12 also chose the blouse C17.

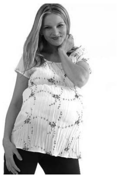
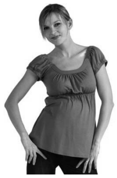
Rule 1			**Rule 2**		
LHS	**==>**	**RHS**	**LHS**	**==>**	**RHS**
C12 	==>	**C17** 	**C17** 	==>	**C5** 
Support = 2.97%, confidence = 37.72%, lift = 2.9			Support = 2.84%, confidence = 21.93%, lift = 1.5		

Figure 2. Association rules with the highest value of support for blouses chosen by users from Poland

In the group of customers from other EU countries, the most popular sets include sweaters (if C57 then C 56) and tops with short sleeves.

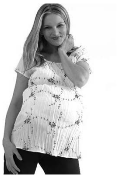
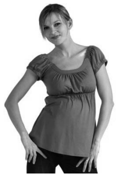
Rule 1			**Rule 2**		
LHS	**==>**	**RHS**	**LHS**	**==>**	**RHS**
C57 	==>	**C56** 	**C17** 	==>	**C5** 
Support = 4.66%, confidence = 42.27%, lift = 3.3			Support = 4.10%, confidence = 28.13%, lift = 1.7		

Figure 3. Association rules with the highest value of support for blouses chosen by users from other EU countries

Customers from other European countries formed sets of tops with short sleeves or blouses with a wide belt at the bottom (if C49 then C50). High values of lift measure indicate that the choice of a product from LHS increases over 6 times the probability of choosing a product form RHS of the rule.

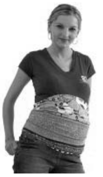
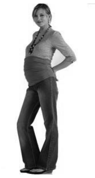
Rule 1			Rule 2		
LHS	==>	RHS	LHS	==>	RHS
C6	==>	C17	C49 	==>	C50 
Support = 8.57%, confidence = 75.00%, lift = 6.6			Support = 8.57%, confidence = 75.00%, lift = 6.6		

Figure 4. Association rules with the highest value of support for blouses chosen by users from other European (non-EU) countries

Visitors from outside of Europe formed baskets containing tunics with colourful flowers and combined them with clothes in black.

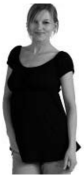
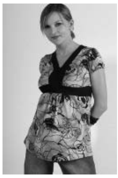
Rule 1			Rule 2		
LHS	==>	RHS	LHS	==>	RHS
C35	==>	C38	C8 	==>	C38 
Support = 12.50%, confidence = 66.67%, lift = 2.7			Support = 12.50%, confidence = 50.00%, lift = 2.0		

Figure 5. Association rules with the highest value of support for blouses chosen by users from outside of Europe

3.3. Sequential rules

The longest sequence of clicks on pages with blouses can be observed among customers from EU. Their paths are the longest and contain many different products. They differ distinctly from the paths of other visitors (from Poland and other countries). This may mean that this group of customers needs more time to make a purchase decision.

The most popular colour of clothing in each segment was white. If a user from Poland chooses a white blouse then he chooses a white blouse once again with the probability of 0.39. If he displays two white blouses he clicks at a white top again with the probability of 38% etc. Grey is an alternative blouse colour to white. Customers from EU are the most determined to find their favourite colour. The longest clickstreams consist of nine white blouses.

Visitors from other European countries who are interested in a white blouse search for a suitable product relatively short. If they do not decide to choose that colour after three steps, they change their preferences to grey.

Customers from outside of Europe who are looking for white tops do that in up to four clicks. At the beginning of the path they take into consideration other colours, while at the end of the path they choose a black product.

Results indicate differences between customers from Europe and outside of Europe. The first group is looking for white products relatively long and alternatively takes into account the grey colour. Path of clicks of visitors from outside of Europe is much shorter, and the alternative colour is black.

Conclusion

Web usage mining is becoming a very important research approach to e-commerce. It enables one to gain an insight into the behaviour patterns of users on the website and transactional patterns of e-shop customers by using the statistical and data mining models. The limited volume of the paper allowed for presenting only a minor part of the results that were used to change the appearance and functionality of the website and personalize marketing communication. The segmentation of customers with regard to their residence was based on IP addresses. Although, these data were not always accurate, one can observe the differences in the behaviour and preferences of customers in separate subsets. In the future, it would be worthwhile to extend the experiment by comparing groups of customers using different language versions of the website. It is also worth examining whether other analytical tools, such as Markov chains, would provide more readable results.

Bibliography

Agrawal R., Imielinski T., Swami A.: *Mining Association Rules Between Sets of Items in Large Databases*. In: Proceedings of the ACM SIGMOD Conference on Management of Data. Washington D.C., May 1993.

Brijs T. et al.: *Using Association Rules for Product Assortment Decisions: A Case Study*. In: Proceedings KDD-99, San Diego, California 1999.

Carmona C.J., Ramirez-Gallego S., Torres F., Bernal E., del Jesus M.J., Garcia S.: *Web Usage Mining to Improve the Design of an e-commerce Website: OrOliveSur.com*. “Expert Systems with Applications” 2012, No. 39.

Chen M.-C., Lin C.-P.: *A Data Mining Approach to Product Assortment and Shelf Space Allocation*. "Expert Systems With Applications" 2007, No. 32(4).

Cooley R., Mobasher B., Srivastava J.: *Web Mining: Information and Pattern Discovery on the World Wide Web*. Proceedings of the 9th International Conference on Tools with Artificial Intelligence. IEEE Computer Society, 1997.

Linoff G.S., Berry M.J.A.: *Mining the Web. Transforming Customer Data into Customer Value*. John Wiley & Sons, New York 2001.

Liu B.: *Web Data Mining. Exploring Hyperlinks, Contents, and Usage Data*. Springer, Heidelberg 2007.

Munk M., Drlik M.: *Impact of Different Pre-Processing Tasks on Effective Identification of Users' Behavioral Patterns in Web-based Educational System*. "Procedia Computer Science" 2011, No. 4.

Russell G.J., Petersen A.: *Analysis of Cross Category Dependence in Market Basket Selection*. "Journal of Retailing" 2000, No. 76.

Sohn S., Kim Y.: *Searching Customer Patterns of Mobile Service Using Clustering and Quantitative Association Rule*. "Expert Systems With Applications" 2008, No. 34(2).

Van den Poel D., Schamphelaere J.D., Wets G.: *Direct and Indirect Effects of Retail Promotions on Sales and Profits in the Do-it-yourself Market*. "Expert Systems With Applications" 2004, No. 27(1).

Zhang X., Edwards J., Harding J.: *Personalised Online Sales Using Web Usage Data Mining*. "Computers in Industry" 2007, No. 58.

DISCOVERING PATTERNS OF USERS' BEHAVIOUR IN AN E-SHOP – COMPARISON OF CONSUMER BUYING BEHAVIOURS IN POLAND AND OTHER EUROPEAN COUNTRIES

Summary

The goal of this article is to analyze behaviour patterns of customers to an e-shop offering clothing for women. The authors discovered the sequences of selected products from the store (web clickstream analysis) and conducted market basket analysis by using popular association rules. IP addresses of visitors that were obtained from web logs enabled identification of the user's country of residence. Geolocation was the basis of comparison of consumer buying behaviours in Poland and other European countries.

Keywords: consumer buying behaviour, web mining, e-shop