Analyzing the footsteps of your customers

- A case study by ASK|net and SAS Institute GmbH -

Christiane Theusinger[1] Klaus-Peter Huber[2]

Abstract

As on-line presence becomes very important in today's e-commerce age, companies focus on two major aspects: on the one hand they are interested in optimizing their web site design to meet the needs of their customers, on the other hand they aim at user profiling to find out who visits their web site.

Analyzing generated and collected data in this context is crucial as the competition is only one mouse click away.

This paper describes a case study that was carried out by ASK|net together with SAS Institute Germany with the aim to improve web site presence and gain knowledge about users. First, we give an introduction of the data that is essential to any company for these types of analyses to optimize on-line presence. Then, the different phases and the results of the analyses of the case study are described in detail. In the conclusion, an outlook on future developments to efficiently use the results will be given.

1. Introduction

In today's electronic age, most companies have established on-line presence. This presence ranges anywhere from promoting a company's image to introducing its products and services to allowing on-line selling of those products and services.

As visitors navigate through a company's web site, their interactions are captured in web logs. Analyses of these web logs provide valuable insight into what products, services and offerings are of interest to visitors, how many percent of those visitors become on-line purchasers, and how and if those purchasers can be turned into loyal customers. Path analysis in particular deals with navigational behavior of its visitors, see also [1], [2] and [3].

The order in which visitors choose to view pages indicates their steps through the buying process. The similarities and differences in navigational behavior of various classes of visitors, such as new visitors vs. repeat visitors, purchasers vs. non-purchasers, 1st time purchasers vs. repeat purchasers, could hold clues towards improving the web site design, offer personalization opportunities, and help streamline the e-commerce environment. Therefore, analyzing so-called web log data plays the central role to support this strategy as the competition is only one mouse click away. The aim is to extract navigation patterns (as done with path analysis) to personalize web sites through online scoring based on the results of the analysis.

Chapter 2 gives an overview of the data that serve as a basis for path analysis. The web mining project carried out by ASK|net is described in chapter 3 before further work and a conclusion is presented in chapter 4.

2. Data model for web mining analysis

A web site consists of a hyperlinked set of pages. The figure below represents such a web site where the nodes are web pages and lines are hyperlinks.

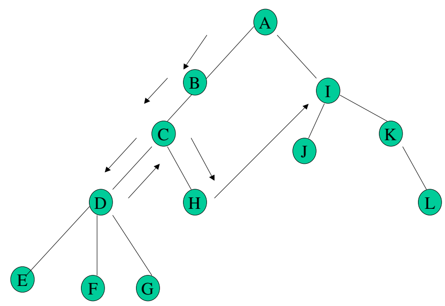

Figure 1: A web site with a sample path

During his visit to a web site, a visitor navigates through the site by either clicking on the hyperlinks, or performing internal searches, or using his/her bookmarks to jump to pages and areas of interest. In the example shown above, a visitor performs different types of moves: forward steps (e.g. from node A to B), backward steps (e.g. from

[1] SAS Institute GmbH, Heidelberg, Germany **Christiane.Theusinger@ger.sas.com**

[2] SAS Institute GmbH, Heidelberg, Germany **Klaus-Peter.Huber@ger.sas.com**

node D to C) as well as forward jump steps indicated by the arrow from H to I. Each of the pages viewed by this visitor is captured in the web log as a separate record.

This sequence is known as a “path” or sometimes called “clickstream“ and is recorded in the form of a log file. Paths can be analyzed to determine the sequences in which users have navigated the web site. This information is also known as “e-intelligence” and is very advantageous to organizations that make the most use of it. The real challenge arises when many visitors navigate through a site that contains thousands and thousands of pages scattered across hundreds of web servers.

2.1 Data model

In general, the data model for clickstream data can be divided into two different types of data: *transaction-based data* and *customer-based data* (aggregated data according to the customer id).

Most of the transaction-based data can directly be retrieved from the log file. In this log file usually only one string for each click is reported with ip address, time stamp, referer address, clicked page and server address. As users behind a firewall can all have the same ip address, an ip address is not suitable as an identification variable so other techniques like cookies or JAVA servlets must be used. Otherwise it is very difficult to distinguish if two clicks belong to the same visitor, see example [4]. A first step may be to distinguish transactions based on the purpose [5], the next step could be to perform association and sequence analysis to get insight in visited web sites and customer paths.

More sophisticated analyses need more information about transactions so new variables have to be calculated from the log file: number of clicks, average time per site (as well as maximum and minimum), weekday of visit, transaction time (business time, free time, night time), server, type of browser, number of sites visited and which sites were visited in which order, first, second page as well as last, second to last page etc. Based on this data about each customer cluster analyses like k-means or Kohonen are well suited methods to find segments of customers with similar behavior. If in addition one event, like the visit of a special site, can be defined as a business-relevant target, predictive modelling methods (decision trees, neural networks, and regression models, an overview is given in [6] and [7]) can also be used.

To obtain good customer profiles, variables describing the characteristics of the customer should be added. If available, this information is given in a data warehouse where all customer characteristics and historical information about clickbehavior etc. are stored. To combine this information with the transactional data the users must identify themselves when visiting the web site so the cookie id could be matched with their names and the transactional data can be merged with customer-relevant data. Having an e-commerce application, combining all this data will allow to answer questions like: “What kind of customers do I have?“ “How can customers be recognized who are interested in a special product for cross-selling? Predictive modelling techniques will also be used in this context.

3. *Web mining at ASK|net*

ASK|net GmbH is a software vendor that sells its software solely via the world wide web at www.softwarehouse.de. The company began business in 1995 and is based in Karlsruhe, Germany. Like ASK|net, many organizations operate a retail web site or “Internet Shop“ these days. As visitors to these sites have different motivations and display various behavioral patterns “Internet shop keepers” want to know which customers are or will be the most profitable ones. Therefore, information about these web site users is needed.

The present project was centered on the improvement of the Internet shop operated by ASK|net GmbH and on gaining knowledge about user profiles (for more details see [8]).

The aim of the project was to gain e-intelligence from user navigation records. This information was essential for optimizing the web site. Optimization is necessary to promote an increase in customer purchase turnover. Optimization, such as in the present case, should meet the following basic criteria.

- Performance of the web site should be optimal, i.e. potential customers should not have a problem accessing and viewing pages on the web site.
- The web site format should be attractive and appealing.
- Navigation through the web site should be relatively simple.
- Few clicks should be required to successfully make a purchase.
- As users click through the web site, “clickstream” data should be generated and automatically warehoused for data mining later on.

Four questions were significant for determining the methodology which the project was conducted with:

- How can user profiles be established from the web log files for clickstream analysis?
- Which web site factors influence the purchase behavior of customers ?
- How can the Internet shop be optimized for ease of navigation?
- How can the Internet shop be modified to increase turnover?

3.1 Data and Workflow

The data necessary to perform the analysis came from two main sources:

- the web logs recording the clickstreams
- the customer purchase database.

Raw data was written to the log files for each user click on the web site and was stored in an Oracle® data warehouse. Figure 2 shows the data mining process with the different stages. First a clickstream analysis was performed to gather information for the optimization of the Internet shop. In the next step, a second clickstream analysis produced and defined ten user profiles from the log files. The resulting variables were transformed and merged with the information if a recorded web session led to a purchase, which was necessary input for building predictive models. The predictive analysis based on the SEMMA data mining methodology was performed. Clickstreams were compared to determine customer behavior patterns. One point of interest was whether or not a particular path led to a purchase. Several predictive models including regression analysis, neural networks and decision trees were produced using Enterprise Miner™ software (For data mining methodology see [9], for Enterprise Miner™ see [10]).

3.2 Web Mining

Two different path analyses based on sequence analysis techniques were performed on the log data. Predictive modelling was performed after merging the results of the sequence analysis with the customer data.

3.2.1 First Path Analysis

The first clickstream analysis dealt with purchase efficiency or the simplicity of the actions required by the customer in order to make a purchase.

The result of the sequence analysis is a set of rules with statistical measures to interpret the importance of each rule (figure 3). Support means *percentage of users who visited all pages given in the rule* while confidence stands for *percentage of users who followed the specific sequence of the rule.*

For example, consider the first sequence in figure 3. You can see that 15.5% (support) of all recorded clickstreams involved the user viewing both the *login* page and then, at a later time, the *register* page. Considering only the users that viewed the login page, it can be seen that 36.8% (confidence) of these users later viewed the register page.

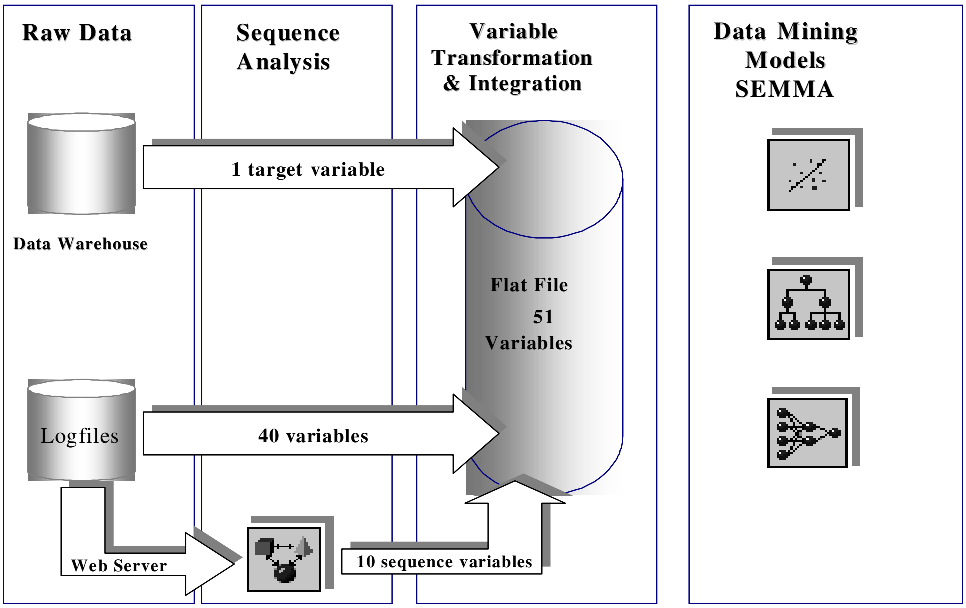

Figure 2: Project sequence

Looking at the second rule, you can see that 32% of all visitors who logged in forgot their password and there was no easy way to reset the password. This point is crucial as this situation leads to new registrations of the same users which would falsify the analysis! The fifth rule shows that paying was difficult to accomplish and required the help screens. So ASK|net had to look for other visitors or had to change the concept, which was done a few weeks after these results were presented.

The findings from this first clickstream analysis suggested that a reorganization of the Internet shop was necessary and were consequently very helpful for the marketing department.

No.	Supp.(%)	Confid. (%)	Rule
1	15.5	36.8	login => register
2	13.4	31.9	login => login
3	12.3	38.5	addcart => login
4	11.2	28.1	addcart => register
5	0.7	4.6	pay_req => help
6	0.3	3.6	news => pay_res

Figure 3: Results from the first sequence analysis

3.2.2 Second Path Analysis

The second clickstream analysis dealt with typical user navigation behavior and one question was central: How can user profiles be defined and derived from the log files ?

The first step was to determine the length of the user profiles. In this case a length of four was chosen based on discussions with ASK|net. We are aware of the fact that this approach depends very much on the specific application. In this context, sequences of a length of four were most suitable for interpretation.

The user profiles as typical navigation patterns were defined based on the results of the sequence analysis. Selection criteria included high support and confidence values for the sequences. This approach was chosen because we wanted to make sure that the defined profiles were typical click sequences .

Consequently, ten user profiles were defined based on paths with high support and confidence values which served as new input variables for the predictive analyses. This means that for each customer the information was available if he or she followed a specific path (user profile) or not. One important sequence was “home => catalog => program => product“ with a support value of 17.1% and a confidence of 82.6%.

3.2.3 Predictive Modelling

Several data transformations of the clickstream analysis were necessary to add user profile information to the data set used for predictive modeling. The original format of the transformation results of the second clickstream analysis was used. First, the original data set was collapsed across a single user session to produce a single row for each customer. Secondly, the clickstreams were concatenated into one variable and a check procedure was used to determine if the user session included one of the ten specific user profiles or not. This information was then added to the database.

The main goal of the predictive modelling was to find out which factors influence the purchase of software at Softwarehouse. For the following analysis only sessions with more than 5 clicks were used, so the resulting data consisted of 22,527 sessions (= records) with a total number of 1,642 purchases. Thus, the purchase rate was 7.29% in the data.

The input variables for the analysis included:

- number of clicks
- duration of the session
- number of web pages involved
- user profile information
- referrer web address
- language of the web page
- customer purchases, if any (the target variable)
- binary variables for each page (if visited or not)

Enterprise Miner from SAS Institute is a very flexible data mining tool which is based on a project flow where steps can be combined to get predictive models. Decision trees, regression analysis, and neural networks were used to perform predictive modelling. During the project several different approaches were undertaken to improve model performance. Those approaches included different types of data preprocessing and different configurations of the algorithms. Figure 4 shows the final process flow diagram that led to the best results.

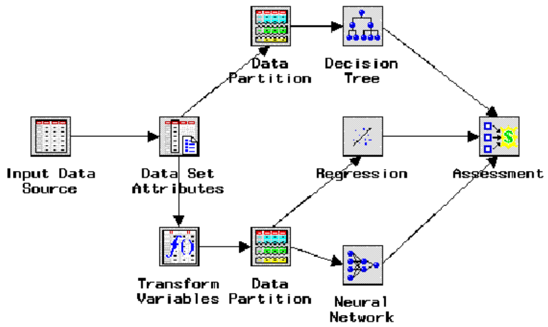

Figure 4: Process flow diagram

In the *input data source* the data file is imported and some statistical measures like average, skewness and the percentage of missing values for each variable are calculated to get a first impression of the data. With *data set attributes,* the role of the response variable is set to target. The data partition divides the whole data set into three parts: training data to develop a model, validation data to optimize the model and test data to investigate the quality of the predictive model on unseen data. For *regression* and *neural network* additionally some numerical variables are normalized to get better suited distributions.

3.3 Results

Figure 5 was created using the assessment node in Enterprise Miner™ software and displays the percentage of the captured response of the predictive models. This chart is one of several ways to compare different models to decide which model is the best.

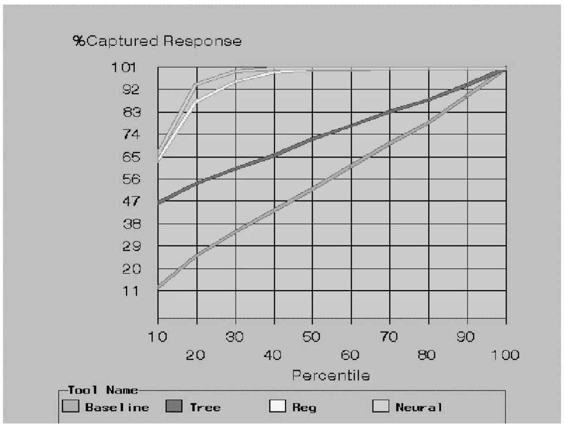

Figure 5: Percent captured response of the predictive models

You can see that in this case in the top 20% with the highest score around 90% of all buyers are captured with the neural network whereas the decision tree only led to 54%. The results of the regression analysis were comparable to those of the neural network.

The findings of the predictive analysis were as follows.

Decision Tree – The optimal tree was produced with the chi-square criteria for splits and a binary tree structure. This was best for analyzing which clickstreams resulted in the purchase of a product. In figure 6 you can see the variables used to build segments of customers. Obviously seq_4 (one of the result paths from the second path analysis) is a good indication for customers, as well as the average length of clicks and other transactional information. This leeds to segments (leaves) with 50% of purchase. So transactional data were really suited to describe buying customers.

While decision trees are especially useful for visualization of the data and for interpreting which combinations of factors are important, figure 5 shows that the regression analysis led to much better results.

Regression Analysis – Before building a regression model it turned out to be useful to normalize the variables *length, number of clicks, and average length of clicks* to achieve better model results. The method used was a stepwise regression model with validation error as the optimization criterion. The regression model was useful for analyzing relationships and for verifying the results of the decision tree model. Important factors in the regression analysis also tended to be important in the decision tree. As shown in figure 5, the regression analysis was almost as effective as the neural network as a predictive model. The influence of the different factors are shown in figure 6. You can see that variables of great significance in the decision tree were also important in the regression analysis. As a measure of importance the T-Scores are added to the decision tree. A high T-Score shows that the variable is very important for the regression analysis.

Neural Network – The neural net was the best overall predictive model. The best results were achieved with a net structure with two hidden layers with three hidden neurons in the first layer and two hidden neurons in the second layer. The only drawback of this model was that the results are not as easy to interpret as the results from the decision tree or the regression analysis.

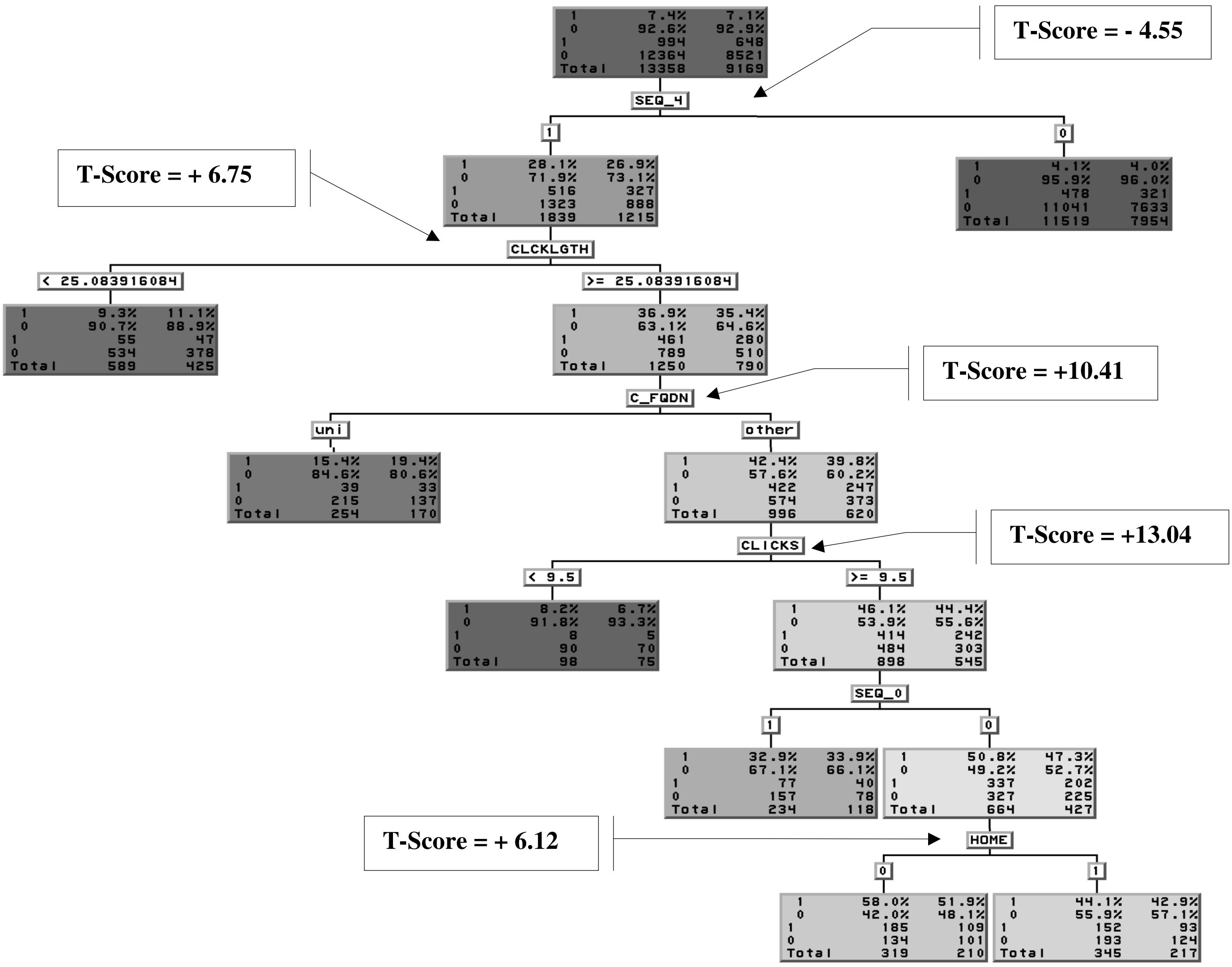

Figure 6: The decision tree

The primary focus was to use information derived during the project to make changes to the web site. Several lessons were learned:

- The purchase procedure was too long and complicated. Seven clicks minimum were needed to make a purchase.
- The registration and login procedure was not optimal. Often, the same customer would complete multiple registrations. This complicated basket analysis.
- The purchase procedure did not include the submission of demographic data. Thus it was not possible to profile customer purchase characteristics.
- User visits to the "News" page were not associated with an increased probability of purchase. So the page did not add extra value.

Points covered as a deployment of the results included:

- Special offers could be added to the News page to increase the influence on customer purchases.
- Develop a system for users to replace forgotten passwords easily.
- Develop a simple login procedure. This is important to prevent the "shopping cart" from becoming lost.

As a result of the different analyses it was possible to optimize the web site structure and detect drawbacks that had not been obvious in the past. So this project showed how web mining can improve a company's website.

4. Conclusion and future work

E-commerce as a selling channel will be one of the most important topics for the next three years. One advantage is that not only the products bought by the customer but also the behavior of the customer can be recorded and analyzed. For this purpose, web mining is needed because it allows to find customer segments with the same behavior, typical paths and correlations between customer characteristics and products etc.

In this paper we have shown which techniques can be used and what results were achieved in a project with real world data: sequence rules were very helpful in detecting structural problems of the web design, while predictive modelling gave good results although only transaction-based data was used.

Furthermore several future developments were planned as a result of the project. The next step will be to develop predictive models with multi-class targets like product category to find suitable products for advertising when an existing customer logs in. Personalizing web sites will be one major step to go in the future. To realize this it is necessary to include more demographic aspects in the register procedure. This additional information can then be used to segment customers and, for example, to customize the web site design using cookie technology according to the segment a customer belongs to.

Acknowledgements

Special thanks to Stefan Weingärtner for the implementation of the data mining models. Thanks also to Helmut Filipp at ASK|net who supported this project from the marketing side.

References

[1] Mena, Jesus: “Data Mining Your Website“. Digital Press, Boston, 1999.

[2] Cooley, R. et al. : “Web Mining: Information and Pattern Discovery on the World Wide Web.“ Proceedings of ICTAI 1997, Newport Beach, California. November 1997.

[3] Büchner, A., Mulvenna, M. D.: “Discovering Internet Marketing Intelligence through Online Analytical Web Usage Mining,“ ACM SIGMOD Proceedings, April, 1998.

[4] Pitkow, J.: “In search of reliable usage data on the WWW.“ Sixth International World Wide Web Conference, pages 451-463, Santa Clara, CA, 1997.

[5] Cooley, R. et al.: “Grouping Web Page References into Transactions for Mining World Wide Web Browsing Patterns.“ Proceedings of KDEX 1997.Newport Beach, California. November 1997.

[6] Berry, M.; and Linoff,G: “Data Mining Techniques“. John Wiley & Sons, New York, 1997.

[7] Berry, M.; and Linoff,G: “Mastering Data Mining“. John Wiley & Sons, New York, 2000.

[8] Weingaertner, S.: “Web-Mining – Ein Erfahrungsbericht“, Handbuch Data Mining, Vieweg Verlag, 2000.

[9] SAS Institute Inc.: “Getting started with Enterprise Miner Software, Version 3.0“, Cary, NC: SAS Institute Inc., 1999.

[10] SAS Institute Inc.: “Data Mining Projects Methodology“, Cary, NC: SAS Institute Inc., 1999.