1^{st} Nordic Workshop on Electronic Commerce

Halmstad, Sweden, May 28-29, 2001

Are all e-customers alike?

Jens Hørlück

University of Aarhus, Dk-8000 Aarhus C
Phone: + 8942 1556, Fax: 8613 5132
E-mail: Jhorlyck@econ.au.dk

Jan Larsen, Lars Kai Hansen, Torben Christiansen

Department of mathematical modeling, Technical University of Denmark
Phone: + 4525 3923
E-mail: jl,lkhansen@imm.dtu.dk

Abstract

Contemporary marketing is based on an idea that life-style determines activity patterns. Based on these ideas, customers are segmented and marketing activity is based on these segments. In relation to e-commerce, the question is, whether these activity patterns also are distinctive in the same peoples click-pattern on a particular web site. If so, it would be possible to detect any new entrant and after a few clicks decide that, she with a certain probability belongs to a specific segment. Based on this segment specific portal can be presented. After a few more clicks this portal can be refined even further. This article describes a project, which aims to test this idea, using a specific statistical method on log-files from a commercial web site.

1. Introduction

The use of interactive media in marketing and selling changes buyer-seller relations. There are many reasons for this. The first set of change stems from the lack of social presence and change of communication mode from synchronous to asynchronous. Consequently, buyers have more control of contact and of time and mode of contact.

The second set stems from the general change of push contact to pull contact, which means that buyers have greater control of the content of the information, they acquire.

The third set of changes comes from sellers utilizing technology via database marketing to increase knowledge about the individual buyer for the purpose of creating lock-in effects for frequent customers and by creating virtual communities combined with attractive web sites for selling.

In e-commerce, there is good theoretical evidence [Shapiro & Varian, 1998] that there will be far fewer shops, because marginal cost is very low on the electronic part of the transactions. Therefore, economies of scale will mean fewer and larger shops than in traditional trade and consequently less choice for the customer.

Traditional shops are designed for target groups, based on segmentation of potential customers. Store layout, presentation of goods, adds, uniforms etc. are all carefully designed to suit the target group. Customers choose shops, not only for rational reasons, but also because they prefer the style of some shops to the style of others. The choice is made in the store design / concept design process and success depends on the ability to evaluate the potential customers buying power. It is therefore good salesmanship not to treat customers as if they were alike.

In e-commerce, it is also good salesmanship to evaluate and segment buyers on their attitudes and behavior, if we want to sell efficiently. However, this conflicts to some extent with economies of scale. One could say that in order to be attractive to a large audience, e-commerce retail shops must offer more than one design of shop. An alternative is to utilize the dynamics of the media, evaluate the customer fast after entry to the site, and then adapt the site to her likings.

In traditional trade, lock-in effects are used to tie customers to a specific supplier. They are often a combination of geography, industry, trade agreements and tradition. Apart from trade agreements, none of these is viable in electronic commerce; so one goal is to create new types of lock-in effects. When competitors are a "few clicks" away, lock-in effects can either be created by offering a better service than competitors or by creating switch-cost, which are considered by the buyer to be significant.

In consumer sales this can be done by using the concept of "flow experiences"[Hoffmand, 1996], i.e. to increase the feeling of exhilaration and joy during the buying process. However, these flow experiences are personal and thus require the ability to adapt to the buyer's behavior.

In high frequent sales relations, like most business-to-business relations, it is important to compensate for the lack of synchronous communication and social contact by providing direct, timely and specific information to the buyer. This is called one-to-one marketing [Peters, 1998]. There is ample theoretical and practical evidence that the decline of the cost of marketing has lowered the optimal segment size, perhaps as far down as one [Varki, 1998].

The logical conclusion is to make presentations on the web site more dependent on the customer as a person. This can be done by segmenting visitors into groups and specify web-page design for each group. Amazon.com does this by presenting combining data on buying behavior. For example, when a customer looks at a book, Amazon.com presents other books based on the most frequent combinations of books, sold to customers, who also bought the first book.

This is often combined with person specific data from previous transactions, so the web site is created for each individual person.

2. Traditional segmentation methods, used in e-commerce

2.1 Segmentation, based on demographic data.

Using demographic data is a traditional way to segmentation. In e-commerce, a similar segmentation is possible by creating segment specific portals, i.e. by creating several "parallel shops", using the same basic technology.

The first drawback is the value of demographic data. The Wharton Virtual Test Market from the Wharton School of the University of Pennsylvania conduct academic research regarding online issues. The site was conceived by a group of professors and some of the resent results of the panel testing are presented in [Bellman, 1999] and [Lohse, 1999]. There is statistics of who and how much clients are buying from retailers. They also do forecasting of sales volume in the years to come. Some of the surprising results where that demographics and time spent on-line does not influence whether someone buys online. Whether a client is buying online and how much this person spend is more related to whether they like being online, or whether they have little time to buy things anywhere else. Demographics data alone predict only 1.2% of buying versus not buying, and 0.3% of online transactions. This suggests that the most important information, which can be used to predict shopping habits, is measures of past behavior, and not demographics. I.e. learning is an important factor to consider, when analyzing web-activity.

Even if demographic data were useful, there would be the problem of lack of sufficient data. Customers are rarely willing to give up unneeded information, so at most e-retailers get an address from actual sales transactions. (How many customers are willing to submit martial status and age, when buying a CD?).

Some e-retailers like banks, insurance companies and other financial retailers have demographic data on their current customers to use demographic segmentation.

2.2 Segmentation, based on stored transactions.

The most common method is to link the web-visitor to her previous transactions on the site. Once the visitor has been identified, for example though a cookie, this information is linked to a database, which stores this visitor's history on the site. There are a number of companies, which specializes in collecting data on web-visitors (Humanclick.com, DoubbleClick.com, Engage.com etc). Any web site can contract with for example DoubleClick and when a customer visits the site, the browser reads a DoubleClick cookie and sends information to Double Click, which stores data in their database. Companies like DoubleClick collect data from many sites and use the data to place customer specific banner-ads on web sites. Companies, delivering data to these companies have normally access to database data.

The strategy of using database marketing, based on previous customer history has some major drawbacks. One of the crucial problems relates to identification of the customer. Sometimes the customer is a person, for example when customers identifies themselves directly or uses credit cards. In other instances, identification is done via a cookie, placed on the PC. In these cases, the identified unit is a piece of hardware. Data is therefore less reliable when customers use several computers or when a computer is used by several persons.

Database marketing requires that the web-server knows the customer and it is only possible to personalize web pages for visitors, who have sufficient data stored in the database. This creates problems both for new and unknown visitors as well as for customers with a low trading frequency. If customers only rarely visits the site, previous behavior is not necessarily related to current behavior. Industries with very seasonal trade like toys have similar problems in interpreting data.

If learning, as the above mentioned data from Wharton suggests, is important and activities for each customer is stored in a database over an extended period of time, then a simple use of these data will mix data from different points of a learning curve. Therefore, the customer will not be treated according to her actual state of learning.

Finally: database marketing only registers actual transactions, not the behavior that led to these transactions. The underlying assumption is the idea that all customers behave alike on the web and consequently traditional database marketing uses the same kind of adaptation of sites for all customers. Amazon.com is well known for its use of previous transactions to present exactly the type of material, that best fits with these transactions. However: The portal is identical for all customers and the customer specific material is presented in the same way for all customers. Only the content differs. Layout, navigation, help screen etc. is basically the same for all customers.

2.3 Ethical and legal obstacles to data collection.

On top of these issues, there are legal and ethical problems with data collection. In several countries for example, there are strict laws on consumer protection, which bans or restricts the use of non-specific data collection on visitors. The legal options for data collection depend in some instances on the customer's country of residence in others on the country of server placement.

Another aspect of this is visitor's active prevention of data collection. Some visitors dislike the idea of being "spied upon" and use various measures to prevent this. For example by blocking the use of cookies, by deleting cookies, by using anonymizing services etc. Therefore, it will be impossible to get full coverage of the customer population.

Partly related to this is a more indirect consequence of legal and ethical issues, which is related to the company's image building / preservation. Although it is technically possible to set up a server on a remote island with flexible laws, this option will often conflict with an image of high standard of customer care.

3. Behavior dependent segmentation

3.1 Web behavior as segmentation method.

We assume that web visitors behave differently, when visiting a web site. The reasons might be personality dependent behavior, experience with IT in general or experience with the specific site. This is in line with contemporary marketing, whereby consumers are segmented by their "lifestyle". Each person's lifestyle is a based on a coherent set of values, partly guided by self-referral, partly by the practical implications of our daily life and partly by our background and upbringing.

There are several models of lifestyle segmentation, which basically all are based on asking a large number of customers of their attitudes and/or their daily practices.

The basic theory behind these types of segmentation is that lifestyle determines our daily practice and therefore our buying behavior. In other words: our practice is guided by our lifestyle and contemporary marketing is based upon this fact. As mentioned above: Shops are designed with these types of segments in mind; sales training put emphasis on segments.

We assume, as an analogy to traditional life-style segmentation, that web-visitors have distinctive web-behavior, not only buying behavior in terms of selection of goods, but also in the way users behave on the web, measured by their click-stream. Data collection from click stream can be combined with stored data to get a richer picture, but we assume that this is not necessary.

Our hypothesis is as follows.

The individual e-consumer have a consistent pattern of web-behavior, which is governed by a "web-lifestyle" and which can be used for segmentation to guide non-identified e-consumers into a segment-specific portal.

3.2 Consequences for web-behavior segmentation

If the hypothesis is valid, it should be possible to use historical data on customer behavior to create segments. Each of these segments will be defined by a specific behavior, described by their click stream.

If our hypothesis is valid, then it will be possible to determine the segment, to which the customer belongs, after a limited number of clicks. When that is done, we can change the design of the web pages according to the segment she belongs and then we can guide her more directly and more appropriately.

This means that it is possible to make a solid guess of a customer after a few clicks and segmentation can be done without prior knowledge of the customer. This can also be done dynamically and the customer can be "moved" to a different segment during the visit, if her click stream pattern tells us to do so.

Ethical and legal problems diminish considerably, since segments are created by looking on statistics, i.e. historical behavior data, and the actual classification of the individual customer is based on anonymous click streams.

This segmentation method takes learning into consideration. If learning is important, web-behavior will change as the customer learns to use the specific site. Analysis of customers past behavior will therefore create different segments for states of learning. When entering, it will, after a limited number of clicks, be possible to determine the learning state of the web-visitor and she will therefore be treated accordingly. As an individual, the visitor will move from segment to segment during a sequence of visits and the customer specific portal will change accordingly for the same person.

If web-behavior data is combined with stored data, segmentation will be based on a richer description of the group of customers and it will therefore be possible to create segments that are more precise. It is also possible to combine segmentation based on web-behavior alone and segmentation based on a combination. Known customers get a different treatment than unknown, but the web design is based on segments in all cases.

Stored data can also be combined with behavior dependent segmentation by filling in customer specific content in segmented web pages.

4. The project

4.1 Purpose

As a first part of an empirical test of this hypothesis, we set up a project to test the basic ideas behind the hypothesis and to test the necessary tools. The idea behind this project is to use historical data of web-behavior to construct meaningful segments, which can be used for guiding the "non" identified users in a way, which more closely corresponds to their web-behavior.

The purpose of this project was to analyze web-behavior to test:

1) Whether individuals have different patterns of behavior, i.e. whether their click streams differ in a consistent way. This is done by deploying unsupervised learning on historical data

2) Whether this behavior can be grouped in a meaningful way. I.e. classification of new users by supervised learning using expert/vendor annotations of identified user segments.

The ultimate goal of the project is to be able to classify a new entrant quickly after entrance, based on their click-stream on the site so far. If we, after the first few clicks, could make a rough segmentation and then gradually refine the segments, we could provide users with a segment specific portal and present the e-customer with a more appropriate web site. The next step would therefore be to test

3) Whether this grouping can be used to predict commercial behavior. This requires commercial interpretation of various segmentation parameters to test the feasibility of the model.

4) Finally, construct interactive web pages using adaptive behavior modeling (model-in-the-loop). The structure and content on the web site is continuously adapted to the user’s behavior.

We concentrated on level 1 and 2 partly because we had to rely on off-line data, partly because we wanted to advance one step at a time.

The web server log transactions and these log-files contain the click stream/access patterns of the web-site visitors. Learning about advertising from click streams including a time variable is suggested in [Winer, 1997] and links to companies are listed in [Bauer, 1998]. Considerations of designing adaptive hyperspaces are presented in [Brusilovski, 1995]. Inference about web pages from the number of "hits" is made in [Brusilovski, 1998].

4.2 Log files as basis for project data

Information about behavior is divided into two categories: direct and in-direct measurements. The in-direct measures are recorded without any interaction with the client, by simply recording the physical behavior browsing the web pages and from log files. As mentioned above, some companies go beyond this, collect data from many sites, and thus acquire massive data on each individual's behavior. Direct measurements are typically feedback from the client on previous actions and decisions of the agent, and feedback on quality of and satisfaction with the purchased goods. The project only planned to use in-direct behavioral measures for segmentation.

In-direct behavioral measures (on-line click stream)	Direct behavioral measure
IP address	Purchase history
Host information	Product evaluation (profiling)
Which browser the visitor is using	Feedback on agent actions (satisfaction)
Which page the visitor is on	
How much time the visitor spent on each page	
Date	
Where the visitor came from if they clicked on a link	

From historical data, we extracted data for construction of features from the log file to construct description of behavior, formed as a vector. We decided to use only in-direct measures to see whether there are consistent behavioral patterns. A log-file includes one registration for each URL. A Web page consists typically of several URL's: One for each frame requests, one per pictures etc. I.e. a click normally generates several log lines. Some of these URLs are of no importance for this purpose and can be removed automatically (Pictures, frame requests etc.).

For each request a typical log file registers:

- IP number of requestor
- Time of request.
- Specific URL requested
- A few codes and information about the browser
- URL of the home pages from which the request originated.
- Site specific information such as parameters for database calls (if defined at design time)

By using this information, we can put together a pattern of behavior for each time a visitor looks at a site, located at one particular server, called a session click-stream. More precisely: a session is defined as entrance, search/purchase and leaving a site.

4.3 Events and sessions

We used the term "event" as the basic building block for modeling sessions. An event is a willful act from the user and it differs from browser clicks:

- A specific event might always take several specific clicks.
- One specific click can participate in several events.

First, among some of the activities in the sessions there might be a 100% dependency, i.e. if log line <A> occurs the web server produces automatically log line <B> hence the transaction of <A,B> is merged into one single event <E> [Zaiane, 1998]. Secondly, the very product specific activities may be possible in a limited period within the scope of the analysis and therefore merged together with activities at a higher level. As an example, imaging a one-week sale of some groceries that is not included in the normal assortment an apple for example. The user activity: "select apple" might rather be logged as "select special grocery" than "select apple".

Other Examples:

- Entrance to a site may take several clicks, but it is only one event.
- Selection of a particular good and adding it to a shopping cart is one event, which may take several clicks.
- Visiting a specific page can be used in a "search" event and a "buy" event.

Going through a specific site and defining "events" is a necessary prerequisite for modeling user session. This requires analysis of the web site and defining sequences of clicks that combines to events. If there is no automation involved in the tools for web design, this can be a very tedious affair, since all routes on the web site have to be tried manually to find the combinations of clicks for each event.

The crucial point of this modeling is the extraction of features Finding descriptions, which are manageable and meaningful was a matter of experimentation: which data to include and how to represent the data for modeling. For example:

- Should a vector consist of a set of all pages visited (time spent / URL) in one session, whether it is short or long?
- Alternatively, should there be a vector for each visited page?

We decided to model each session as a vector. A session consist of all events, a visitor creates from entering the site to leaving the site again.

Web sites are inherently dynamic in nature. A redefinition of a web page typically creates new URLs, which changes loglines without changing the basic content. Since a vector is defined from these data, which often changes, it will make it difficult to compare behavior over time. The best solution to this problem is to make logging based on events and not URL requests. Events are more stable than URL requests, because events are independent of the specific URL.

Very short sessions are likely to be mistakes e.g. following a search result from a search engine and quickly leave the site again. The minimum session length strongly depends on the construction of the entrance to the site and no general recommendation can be given. The minimum session length in [Tak, 1996] is by inspection set to five pages, leaving only 20% of all sessions for analysis in that particular case.

4.4 Empirical data

The historical data, used for testing the hypothesis, came from two companies:

Company A is today in the energy sector and has created a new department for E-commerce.

One initiative is to use their many petrol stations, located on main traffic routes. The company has created a site, where customers can order groceries during the day and pick them up on their way home from work. Their traditional marketing is based on long opening hours and on their physical location and they do not compete on prices. Since each petrol station had its own selection of goods, customers had to select a specific shop and then they could see the goods available in this particular shop. Customers had to go through a registration process, if they wanted to buy, but they could also search the site as guests. The site had a service for registered users, called "My market". It consisted of goods specifically selected by the individual customer. These goods were presented on a special page.

The company was chosen for the following reasons:

1. Their web marketing is based on convenience and to utilize this effectively, regular customers ought to have a very specific opening page, modeled after their previous behavior. This would make them feel more comfortable and would increase reselling.

2. Company A have many different and unknown visitors and the challenge would be to present the right information for unknown customers in a way, meaningful to them.

3. Customers, who actually ordered, could be precisely identified.

Company B is a wholesaler in steel and related metals. They have a site, where everybody can see their product catalogue and very specific information on steel. Regular customers have access to their extra-net and can see customer specific prices; the customers own product numbers etc and regular customers can order via the extra-net. They have a quite successful extranet with a relatively high number of users and a steady growth in sales via the Internet.

This site was chosen for the following reasons:

1. Buyers belong to a known number of high frequent users, who all log on as persons and therefore can be identified precisely.

2. Other visitors mainly come from educational institutions that use the site for collecting data on steel for educational purposes.

Contrary to company A, the assumption was that users belong to two distinctive groups, each quite "homogenous".

Data from Company B was selected for the ability to invalidate the hypothesis. If, for example, the number of segments exceeds the number of high frequent users, the hypothesis is clearly invalid. Alternatively, if segments, based on web-behavior, were tested against precise identified persons afterwards and customers were spread over several segments, then the hypothesis was invalidated.

In other words: proper use of data from company B could be used to find out whether clusters in web-activity were arbitrary or dependent on persons actual behavior.

As the project went on, we decided to abandon data from company B. Their site was based on an older version of Lotus-Notes and it was very difficult to interpret the log-files in order to define events at a sufficiently detailed level. In addition, within the limits of the project, we could not combine the log-data with transaction data in order to have a precise personal identification. This was crucial to the project, since Company A's buyers (as companies) often have 2-4 persons, who used the system and they often entered the Internet via firewall, or via dynamically allocated IP addresses.

4.5 Data analysis

For this project, we only used data, available from log-files, which included:

1. Registration of sequence of visited web pages (URLs)

2. Time spent on each web page.

3. Summary of content of visited pages in parameters of log lines.

4. User information obtained through client’s web browser, e.g., IP address, computer name, etc.

A more thorough analysis could also include

- User identification using client cookies.
- User data obtained via prompting and by use of interrogative strategies.

As mentioned above, this additional information can be of value in determining segments from historical data, especially for secure identification of individuals. It cannot be used for the final level four from above.

The first step was definition of events. Which sequence of URLs in the log lines belong to an event? Some URLs can be removed automatically (Pictures, frame requests etc.). The rest have to be combined manually to rules: If a specific sequence of URLs is used for a well-defined purpose, this sequence is combined into a rule. The rules are then used to search the log lines for events. Definition of rules was done by a combination of extraction of web-site features and manual inspection.

The second step was identification of sessions. We had to rely on the IP address and the recorded time for each event. Based on this, sessions for each visit were defined.

This process presented us with a couple of problems, which were solved by defining limits in data analysis. First: How do you interpret a long delay between two URL-requests? Do they belong to one session or to two separate sessions? We choose the industry standard of 30 minutes as the longest possible delay [Cooley, 1999]. Another decision taken was to delete very short sessions, because they are likely to be mistakes, e.g. a short visit after using a search engine. We chose four events as the minimum, because this was the shortest path to actual trading.

Then we had a set of problems that requires log-decisions to be defined at web-design time. Neither company A nor Company B had considered these problems in their web-design, so we had to rely on traditional logged data.

Traditional server logging only logs URLs, which means that calls to a database is logged as a specific URL without the content of the database call being presented fully in the log-file, either because keyed data do not appear on the log-file or insufficient logging of choices made in a Java applet.

When the user uses the browsers cache by clicking the [back] button, it creates discontinuities in the logged click stream, because users navigate the site without being logged. This surely belongs to individual behavior, but it is not traceable in the click-stream. This problem cannot be solved with traditional log-files, since it requires that logging is done on the client and not the server.

The last problem, we could not solve relates to precise identification, as mentioned above.

Different users, entering the site sequentially and passing the same gateway may appear with the same IP address.

Both companies had a stable Web site throughout the period of analysis, so the problem of redefinition of URLs was not a problem in this project.

4.6 Hierarchical Probalistic Clustering

The project used a method called Generalizable Gaussian Mixture Model. The mathematical techniques are described in detail in other papers from the some of the authors [Hansen, 1999] [Larsen, 2000/2001].

The objective is to develop a user behavior clustering model using unsupervised learning. From a training set of historical user behaviors, the clustering model is able to cluster new users into a number of clusters. Behavior clustering is considered as an explorative tool, which identifies behavior structure while interpretation of the clusters to define segments requires annotations, i.e. each cluster have to be interpreted in the context of the web site and its purpose. Only clusters that have such a commercial interpretation are here called segments. Clusters that do not have such an interpretation are labeled clusters. Once annotations are provided an automatic classification of users into segments is possible. When that is done, there still is a rest group of web-users whose cluster has no commercial interpretation.

The model will be able to estimate the probability that a new visitor belongs to a certain segment/cluster. The user will be identified as belonging to the segment/cluster with the highest probability. Furthermore, the model will be able to display prototype visitors for each segment and to identify outliers, i.e. visitors that do not comply with the current clustering.

Issues in the testing of the model were:

- Clarification of the number of records necessary to learn the clustering (learning curve).
- How many pages does the user need to visit before a reliable clustering is obtained?
- What is the structure and reliability/robustness of the hierarchical clustering (generalization error)?
- Are the industrial collaborators able to annotate the presented clustering?

A problem of completely different nature was interpretation of results. For example: how would one interpret results, which are very sensitive to variations in the modeling of feature extraction? Alternatively, how to interpret clustering where the differences in probabilities are small, but insensitive to differences in feature extraction.

5. Results

From company A, we e obtained log lines from half a year's activity and this resulted in 31.700 well-defined sessions. Of these 4.339 sessions had more than the minimum of four events.

We defined 60 unique events from the structure of the web site. Thus, the event space was the numbers 1-60. Each session can be described as a sequence of numbers. For example, a user could use two different log-on techniques to the site: either as a member login with personal password or as a guest. This was defined as the first two events and if a user tried to log-on two times and failed the first, this would count as two sessions. We then mapped the sessions onto the event space.

Part of the data was used for user behavior model construction where the remaining part was used for testing and evaluation of the resulting model, hence simulating the visit of new users.

Repeated training of the unsupervised GGM model resulted in the most generalizable model contained 17 segments/clusters (arbitrarily numbered from 1 to 17). For each cluster, we obtained the number of sessions in the cluster and a "typical user", i.e. a description of the users closest to the center.

From these data we gave a commercial / user centered interpretation of each cluster (with number of sessions in parenthesis).

The segments/clusters, we found was grouped according to the commercial interpretation of the typical user.

The first two groups of segments relates to whether users are registered or visiting as guests.

The third group of segments relates to the process of entering the site and to registration. Registration is for cluster 6 and 17 are done in a separate session with practically no ordering.

The fourth groups with one cluster relates to ma faulty web design

The last group includes the remaining clusters.

Segments of registered users (i.e. users which had been through a registration process)

12	(531)	Uses search function, some ordering.
5	(395)	Uses "My market" (goods of special interest for this person). Some ordering.
9	(234)	Registration as a new customer, some browsing
13	(213)	Uses "My Market", but this cluster have a scattered selection of events and doubtful interpretation.
2	(20)	Gets help to remember password, no ordering.

Segments of guests (non registered users)

3	(708)	Efficient guests, who quickly finds their local shop and browses among available in this shop. Approximately 200 leaves after seeing the local shop's top-level page.
14	(109)	Browses around on several shops.
17	(84)	As 3, but differs, because they look at the registration page without finishing registration.
6	(32)	Uses a form to contact the local shop, very little browsing.

Segments of users with Problems. Probably lost as customers.

11	(152)	Do not remember / have never got a password. Unsuccessful use of support function
8	(59)	Do not remember / have never got a password. Unsuccessful use of support function. Uses form to contact company.
1	(40)	Tries to register as customer, but do not succeed.

7	(34)	Technically successful registration, but customer leaves without confirming. Site design error
15	(448)	Uses search for goods before selection of local shop. Since shops have different goods, the search function needs to know the selected shop and does not work without. However, the web design does not prevent search before selection of shop. Most leave the site after a few trials, but a substantial number have more than 10 attempts and a few more than 25.

Clusters with no interpretation

4	(76)
13	(219)
16	(49)

5.1 Closer examination of some segments

Segment 12 and 5 represent two different segments of users. Although they are very much alike in large parts of the used event space, there are distinctive differences. These two segments represent users each with their distinct behavior. The first cluster finds goods on the local web site by using a personalized tool. The other cluster uses the search function for the same purpose. If a web site should serve these two types of customers in the best possible way, one would alter the design to accommodate for the type of customer. The distinction between these two segments could be detected after a few clicks.

A completely different type of segment is represented by 8 and 11. This is users, who need help of some sort and who cannot use the help function on the site. One could interpret this as an example of bad web-design, but since the support function is used successfully by some users in clusters 2-5-12 and 13, one ought to accept that no matter how good a design, some users will have difficulties in navigating the site. Users in these two segments can be detected after a few clicks and they could be helped by letting a help page appear automatically to guide them through the remaining process. If they succeeded in getting access to the site, it would probably be possible to redefine their cluster / segment; but as it is, they never enter and we cannot know.

Segment 7 represents a puzzle. Apparently users without problems, but they leave without registration and do not return to browsing as guests within the next 30 minutes. They are comparable to users in segments 17, who browse as guests and looks at registration.

As a side effect, we found a design error. Cluster 15 represents users, who try to use the site as stated, but it does not work. They are trapped by a faulty web-design. Although this is a cluster, which should not exist, it is nonetheless an important finding, which never would appear in the traditional tools for web analysis. The correct handling of such clusters would be to redesign the site immediately. Fourteen percent of the users are probably lost as customers this way.

Our problems were not technical in nature, but we soon found that actual buying from this site was to limited to segment customers based on buying behavior, and among them very few customers who repeatedly bought at the stores. However the results clearly shows that segmentation of users is technical feasible and even with this limited number of sessions, commercially interpretable.

6. Conclusion

The methods showed some clear segments of user behavior and the modeling techniques were good at detecting these segments. However, we could neither falsify, nor support our hypothesis, due to limitations in the data material, especially data about the actual buying behavior. This calls for repeating the project at a web site with a substantial number of customers and with many regular and reoccurring customers, who can be identified in the log file.

The analysis of the design of the specific web site and the definition of events was far more tedious than expected and the analysis of the two company's log lines required a substantial manual effort. If this type of analysis should be used commercially, it has to be taken into consideration at the time of design of the web site.

6.1 The "ideal" design for this kind of data analysis

A good log file should include:

- A "naming" of visitors upon entry, for example by using cookie technology, thereby linking all the visitors clicks in one session.
- It should be possible to link this "name" to previous visits, i.e. some kind of permanent "naming" of the customer. (or the PC if cookie technology is used)
- Each log line should be an event instead of URL. This requires careful analysis of the web design to make a predefined list of events. With every change of web site, the definition of clicks belonging to an event must be reassessed.
- Events should be logged with sufficient parameters, i.e. a database call (defined as an event) should be logged with the necessary parameters from this call.
- Due to the storage of web pages in the local cache in the client's browser, not all of the user actions are recorded in the server log files. The use a Java script acting as remote sensing agent is one way of retrieving the missing logs [Shahabi, 1997, Hamilton, 1996]. Each time a new web page is loaded from either the local cache or the web server, the java script sends a message to the server about the system time of the page view. This method will provide the log server, with all actions of the user and at the same time log the exact time of the page view to be used in analysis of the dynamic. The main drawback is that the user has to accept the download of the Java applet.

If a web site was designed with these elements, it would be possible to automate the clustering technique to a large extent and thereby reassess the segmentation and to experiment to a larger extent with different mathematical modeling tools and with different number of clusters.

References

Bauer, C. and A. Scharl (1999). Acquisition and symbolic visualization of aggregated customer information for analyzing Web information systems. HICSS-3 - 32nd Annual Hawaii International Conference on Systems Sciences, Hawaii.

Bellman, S.; Bradlow, E.; Huber, J.; Johnson, E.; Kahn, B.; Little, J. and D. Schkade (1999). “Agents to the Rescue ?” Marketing Letters 10(3): 285-300.

Bellman, Steven; Gerald L. Lohse; Eric J. Johnsen (1999). “Predictors of Online Buying Behavior.” Communications of the ACM 42(12): 32-38.

Brusilovsky, Peter (1995). “Methods and Techniques of Adaptive Hypermedia” User Modeling and User-Adapted Interaction 6(2-3): 87-129.

Brusilovsky, Peter and J. Eklund (1998). “A Study of User Model Based Link Annotation in Educational Hypermedia.” Journal of Universal Computer Science (Springer Science) 4(4): 429-448.

Cooley, R.; J. Srivastava and B. Mobasher (1999). “Data Preparation for Mining World Wide Web Browsing Patterns.” Journal of Knowledge and Information Systems 1(1).

Hamilton, M. (1996). "Java and the Shift to Net-Centric Computing." Computer 29(8): 31-39.

Hansen, Lars Kai ; Sigurdur Sigurdsson; Thomas Kolenda; Finn Årup Nielsen; Ulrik Kjems and Jan Larsen (1999). “Modelling Text with Generalizable Gaussian Mixtures.” Working Paper, , Department of Mathematical Modelling, Technical University of Denmark.

Hoffmand, Donna L. and Thomas P. Novak (1996). “Marketing in Hypermedia Computer-mediated Environments. Conceptual Foundations.” Journal of Marketing 60(july): 50-68.

Johnson, Stephen Kobrin and Eric J. (1999). “We know all about you: Personal privacy in the information age.” Working Paper, Wharton School, Pennsylvania: 1-13.

Larsen, Jan; Lars.K. Hansen; A. Szymkowiak; Torben Christiansen and Thomas Kolenda (2000). Webmining: Learning from the World Wide Web. Nonlinear Methods and Data Mining 2000, Rome, Italy.

Larsen, Jan; Lars.K. Hansen; A. Szymkowiak; Torben Christiansen and Thomas Kolenda (2001). “Webmining: Learning from the World Wide Web.” Computational Statistics and Data Analysis.

Lohse, Gerald L. , Steven Bellman, Eric J. Johnson (1999). “Consumer buying behavior on the Internet: Findings from panel data.” Journal of Interactive Marketing.

Lohse, Peter Spiller and Gerald L- (1997). “A Classification of Internet Retail Stores.” .

Peters, Linda (1998). “The new interactive media: one-to-one, but to whom.” Marketing Inttelligence & Planning 16(1): 22-30.

Shahabi, C. Z., A.M.; Adibi, J.; Shah, V.: (1997). Knowledge discovery from users Web-page navigation. Seventh International Workshop on Research Issues in Data Engineering. High Performance Database Management for Large-Scale Applications.

Tak, W. Y. J., Matthew; Garcia-Molina, Hector; Dayal, Umeshwar, (1996). "From user access patterns to dynamic hypertext linking." Computer Networks and ISDN Systems 28(11).

Varki, Sajeev (1998). “Technology and Optimal Segment Size.” Marketing Letters 9(2): 147-167.

Winer, R.; Deighton, J.; Gupta, S.; Johnson, E.; Mellers, B.; Morwitz, V.; O'guinn, T.; Rangaswamy, A.; Sawyer, A (1997). “Choice in Computer-Mediated Environments.” Marketing Letters 8(3): 287-296.

Zaiane, O. R. M. X. J. H. (1998). Discovering Web access patterns and trends by applying OLAP and data mining technology on Web logs",. Proceedings from IEEE International Forum on Research an Technology. Advances in Digital Libraries, IEEE.