In: P. Perner (Ed.), Data Mining in E-Commerce, Medicine, and Knowledge Management, Springer Verlag 2002, lnai 2394

Intelligent E-Marketing with Web Mining, Personalization and User-adpated Interfaces

P. Perner and G. Fiss

Institute of Computer Vision and Applied Computer Sciences
Arno-Nitzsche-Str. 45, 04277 Leipzig
e-mail: ibaiperner@aol.com http://www.ibai-research.de

Abstract. For many people the special attraction of E-commerce is linked to the idea of being able to choose and order products and services directly on-line from home. However, this is only one aspect of the new on-line sales model. As in real sales processes competent counselling, in accordance with the customer's necessities, and also after-sales assistance by help of the web play an important part for the customer faith. This requires precise knowledge of the customer's preferences who, however, in general does not like lengthy questioning and the use of other communication routes. Holders of E-shops have thus to gather the consumer's desires and preferences from his interactions and the data resulting from the sales process, which requires a profound data analysis. In this paper we describe what kind of data can be acquired in an e-shop and how these data can be used to improve advertisement, marketing and selling. We describe what kind of data mining methods are necessary and how they can be applied to the data.

Keywords. E-Commerce, Data Mining, User Profiling, Clickstream analysis, Recommendation, User-Adapted Interfaces

1 Introduction

For many people the special attraction of E-commerce is linked to the idea of being able to choose and order products and services directly on-line from home. However, this is only one aspect of the new on-line sales model. As in real sales processes competent counselling, in accordance with the customer's necessities, and also after-sales assistance by help of the web play an important part for the customer faith. This requires precise knowledge of the customer's preferences who, however, in general does not like lengthy questioning and the use of other communication routes. Holders of E-shops have thus to gather the consumer's desires and preferences from his interactions and the data resulting from the sales process, which requires a profound data analysis. This knowledge has then to be converted into an intelligent and, if possible, entertaining presentation of the information wanted by the customer, where multimedia means of expression are used, and without overstraining or understraining

him. Only in this way can he be motivated to continue the dialogue. As the capacities, preferences and interests of the customers vary considerably in this field of application, intelligent user guidance is indispensable.
In Section 2 of this paper we describe the main problem that are concerned with E-Marketing and Selling and why data mining and user modeling is important. The basic data that can be automatically accessed from a website are describe in Section 3. In Section 4 we give a brief overview about the basic data mining methods and indicate how they can be used for marketing and selling. In Section 5 we describe our idea for intelligent e-marketing with data mining and user-adapted interfaces. Finally we give conclusions in Section 6.

2. E-Marketing/Selling

In order to do e-marketing, it is necessary to know about traditional marketing, computing sciences and also about analytical methods.

E-marketing is the concentration of all efforts in the sense of adapting and developing marketing strategies into the web environment. E-marketing involves all stages of work regarding a web site, such as the conception, the project itself, the adaptation of the content, the development, the maintenance, the analytic measuring and the advertisement. One of the most serious misunderstandings is to face the web as a simple extension of marketing campaigns of the company, or "a cheap" institutional propaganda. When launching a business on the Internet, whether it is an institutional site or a site for online shopping/electronic trade, it is necessary to have in mind that this means dealing with media, with very peculiar characteristics.

E-Business can be any site with commercial purposes that is on the internet, regardless of the characteristics of the site. A classification of these activities according to different objectives leads to four basic forms of the usage of the internet for business [1]:

- Online promotion
- Online Shopping
- Online Service and
- Online Collaboration.

The aim of online promotion is to bring an advertisement message which is targeted to specific customer group quickly and cost-effective to this group. Online-shopping is the selling of products or services via the internet. The basic requirements for an on-line shop are at least a product catalogue and a safe and error-tolerant transaction line for ordering and paying the products and services. Online-service means to provide services via the internet. These services can be free or the user has to pay a fee for it. The important advantage is that these services can be accessed from everywhere in the world at any time. With online-collaboration are named all strategies where users are enabled to get into contact with other users. Very popular are user forums (moderated and free onces). The other very common way are chat

rooms. The aim of online collaboration is to transport a special image to the target group, that is not creatable through classic advertisement.

For a successful web presence it is useful to combine these different models. An online shop will also do some promotion of products via e-mail or provide services to the customer which will help to keep the customer. Successful web presentations are a full integrated part of the hole marketing and communications strategy, requiring on the general principles of e-marketing:

- Interactive and Flexible
- Informative
- Instantaneous
- Measurable
- Affordable and
- Intuitive navigation

It is important to set up the e-business model in such a way that it uses the 6 principles of e-marketing on the one hand and on the other hand meets the customer's or user's personal requirements for services, products and information. This requires to know and to understand the behavior, needs and expectations of the target group.
Apart from traditional marketing based on market research e-marketing can and should use the data given by users while they are navigating through the web site. Focusing on permission marketing where the user allows you to use his data for further communication, you are able to build up a long-lasting customer relationship. By this you can see customer data as the real treasure. To use it requires a good e-business model, data warehousing and especially data mining techniques.
A good user model is the basis for all activities. However, such customer groups are not static, they will chance over time. The internet is a fast medium. The interest of customer groups will change quickly. Customers will move away from one provider to another one. Besides that arises the question: Are your services or products appropriate to get sold via the internet? What should your web pages look like? What technological facilities has the customer who is visiting your web site?
To keep the customer's attention for your web presence requires to build up a strong customer relationship and to offer services which attract the customer to visit the web site frequently and purchase products and services. The need to develop specific marketing strategies for the internet implies that some traditional principles are adapted, or even reinvented. An online-shop offers all the incriedents (for e.g. marketing/selling data, server data, web meta data) to solve and to automate these tasks successfully (see Fig. 1). While a customer is visiting a web site he leaves a trace of data which can be used to understand the customers needs, desires and demands as well as to improve your web presence.

3 The Data

In an e-commerce site are available data across the merchandising data, marketing data, server data, and web meta data. These data can be used for data mining to better understand the marketing and selling processes, the site organization and the server itself (see Fig. 1).
There are different types of data: user entry data, server and cookie logs, web documents and web meta data.

Fig. 1 Internet Retailer Web Site and available Data

3.1 Server and Cookie Data

Web server logs (see Fig. 2) are automatically generated by the server when a user is visiting an URL at a site. In a server log are registered the IP address of the visitor, the time when he is entering the website, the time duration he is visiting the requested URL and the URL he is visiting. From these information can be generated the path the user is going on this website [2]. Web server logs are important information in order to discover the behavior of the user at the website. However, the IP address stored in the server log does not always lead to the particular user. The address might have been changed by the proxy server and the heuristic used for the identification of a user session does not always hold. Therefore cookie logs might be more preferable.

Cookies are short text files that are generated by the server on the client site while his browser is visiting the website. Cookies allow to set a special identification number or code for a particular user. Each time a user is visiting the website he can be identified by this identification code. However to set a cookie requires that the user

has given permission for that which is not always the case. Therefore only the combination of server logs and cookie log will be a good basis for data mining.

In the example given in Figure 2 a typical server log file is shown. Table 1 shows the code for the URL. In Table 2 is shown the path the user is taking on this website. The user has been visiting the website 4 times. A user session is considered to be closed when the user is not taking a new action within 20 minutes. This is a rule of thumb that might not always be true. Since in our example the time duration between the first user access starting at 1: 54 and the second one at 2:24 is longer than 20 minutes we consider the first access and the second access as two sessions. However, it might be that the user was staying on this website for more than 20 minutes since he was not entering the website by the main page.

```
hs2-210.handshake.de - - [01/Sep/1999:00:01:54 +0100] "GET /support/ HTTP/1.0" - -
                        "http://www.s1.de/index.html" "Mozilla/4.6 [en] (Win98; I)"
Isis138.urz.uni-duesseldorf.de - - 01/Sep/1999:00:02:17 +0100] "GET /support/laserjet-support.html
                        HTTP/1.0" - - "http://www.s4.de/support/" "Mozilla/4.0 (compatible; MSIE
                        5.0; Windows 98; QXW0330d)"
hs2-210.handshake.de - - [01/Sep/1999:00:02:20 +0100] "GET /support/esc.html HTTP/1.0" - -
                        "http://www.s1.de/support/" "Mozilla/4.6 [en] (Win98; I)"
pC19F2927.dip.t-dialin.net - - [01/Sep/1999:00:02:21 +0100] "GET /support/ HTTP/1.0" - -
                        "http://www.s1.de/" "MOZILLA/4.5[de]C-CCK-MCD QXW03207 (WinNT;
                        I)"
hs2-210.handshake.de - - [01/Sep/1999:00:02:22 +0100] "GET /service/notfound.html HTTP/1.0" -
                        - "http://www.s1.de/support/esc.html" "Mozilla/4.6 [en] (Win98; I)"
hs2-210.handshake.de - - [01/Sep/1999:00:03:11 +0100] "GET /service/supportpack/ in
                          dex_content.html HTTP/1.0" - - "http://www.s1.de/support/" "Mozilla/4.6
                        [en] (Win98; I)"
hs2-210.handshake.de - - [01/Sep/1999:00:03:43 +0100] "GET /service/supportpack/kontakt.html
                        HTTP/1.0" - - "http://www.s1.de/service/supportpack/index_content.html"
                        "Mozilla/4.6 [en] (Win98; I)"
cache-dm03.proxy.aol.com - - [01/Sep/1999:00:03:57 +0100] "GET /support/ HTTP/1.0" - -
                        "http://www.s1.de/" "Mozilla/4.0 (compatible; MSIE 5.0; AOL 4.0; Windows
                        98; DigExt)"
```

Fig. 2 Excerpt from a Server Logfile

URL Address	**Code**
www.s1.de/index.html	A
www.s1.de/support/	B
www.s1.de/support/esc.html	C
www.s1.de/support/service-not found.html	D
www.s1.de/service/supportpack/index_content.html	E
www.s1.de/service/supportpack/kontakt.html	F

Table 1. URL Address and Code for the Address

User Name	Time	Path
USER_1	1:54	A
USER_1	2:20 -2:22	B → C
USER_1	3:11	B
USER_1	3:43 - 3:44	E → F

Table 2. User, Time and Path the User has taken on the Web-Site

3.2 User Entry Data / Profiles

On-line forms on a website are a very popular media for the acquisition of data from visitors. Usually the website visitor is requested to fill in into these forms information like name, address, telephone number etc. but also life style information and user interests are stored. This information can be directly stored into a data base which can be taken later on for data mining. However, for a user it is often boring to answer all these questions. Therefore, on-line forms or questionnaires should be set up in such a way that they do not take too much of the user's time and that he is motivated to give all the requested answers.

A newer trend is the Open Profiling Standard (OPS) [3] which allows to automatically access user profiles from the browser of the client site. The OPS standard defines the data format and the transaction rules for electronic profiles [11]. The user can set up his profile on a voluntary basis and by doing so keep track of what information he likes to provide. The other advantage of an electronic profile for the user is that he only needs to define his basic profile once and not whenever he is entering a web site.

3.3 Web Documents and Web Meta Data

The web documents (HTML documuent, see Fig. 3) contain information such as text, images, video or audio. They have a structure that allows to recognize for e.g. the title of the page, the author, keywords and the main body. The formatting instruction must be removed in order to access the information that we want to mine on these sides. An example of an HTML document is given in Figure 3. The relevant information on this page is marked with grey color. Everything else is HTML code which is enclosed into brackets <>. The title of a page can be identified by searching the page for the code <title> to find the beginning of the title and for the code </title> to find the end of the title. Images can be identified by searching the webpage for the file extension .gif, .jpg.

Web meta data give us the topology of a website. This information is normally stored as a side-specific index table implemented as a directed graph. Usually, these web meta data are specified manually by the website administrator. This can become hard for large websites. Therefore, recently methods have been developed to annotate this documents automatically.

```
<html>
<head>
        <title>welcome to the homepage of Petra Perner</title>
</head>

<body bgcolor="#ccffcc" text="black"
background="../images/hint.gif" link="#666699">

<td width="20" valign="top"><img height="5" width="20"
src="../images/nix.gif"></td>

<td width="423" valign="top">
<font face="Arial,Helvetica,Geneva" size="4" color="#666699">

Welcome to the homepage of Petra Perner</b><br></font></br></br>

<font face="Arial,Helvetica,Geneva" size="3"
color="#666699">Industrial Conference Data Mining 24.7.-25.7.2001
</font></br></br> </br>

<font face="Arial,Helvetica,Geneva" size="3" color="black">

In connection with MLDM2001 there will be held an industrial
conference on Data Mining.</br></br>
Please visit our website http://www.data-mining-forum.de for more
information.</br></br>
List of Accepted Papers for MLDM is now available. Information on
MLDM2001 you can find on this site under the link MLDM2001</br>
</br>

        </font></td></tr></table></div>
        </body>
</html>
```

Fig. 3 Example of an Html-Document

4 Data Mining

4.1 Basic Problem Types

Data Mining [4] methods can be distinguished into two main categories of data mining problems:

1. prediction and
2. knowledge discovery.

While prediction is the strongest goal, knowledge discovery is the weaker approach and usually prior to prediction.

The classification of a customer into a customer who is highly likely to buy a product belongs to predictive data mining. In this example, we have to mine a data base for a set of rules that describes the profile of a customer who has a preference for a certain product. The shop assistant can use this classification knowledge to identify a customer as a potential buyer.

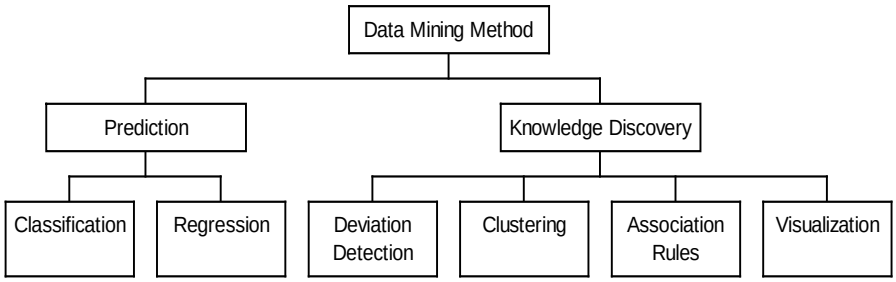

Fig. 4. Types of Data Mining Methods

For that kind of data mining, we need to know the classes or goals our system should predict. In most cases we might know these goals a-priori. However, there are other tasks were the goals are not known a-priori. In that case, we have to find out the classes based on methods such as clustering before we can go into predictive mining. Furthermore, the prediction methods can be distinguished into classification and regression while knowledge discovery can be distinguished into: deviation detection, clustering, mining association rules, and visualization. To categorize the actual problem into one of these problem types is the first necessary step when dealing with Data Mining.

Note that Figure 4 only describes the basic types of data mining methods. We consider for e.g. text mining, web mining or image mining only as variants of the basic types of data mining which need a special data preparation.

4.2 Prediction

4.2.1 Classification

Assume there is a set of observations from a particular domain. Among this set of data there is a subset of data labeled by class 1 and another subset of data labeled by class 2. Each data entry is described by some descriptive domain variables and the class label. To give the reader an idea, let us say we have collected information about customers, such as marital status, sex, and number of children. The class label is the information whether the customer has purchased a certain product or not. Now we want to know how the group of buyers and non-buyers is characterized.

The task is now to find a mapping function that allows to separate samples belonging to class 1 (e.g. the group of internet users) from those belonging to class 2 (e.g. the group of people that do not use the internet). Furthermore, this function should allow to predict the class membership of new formerly unseen samples.

Such kind of problems belong to the problem type "classification". There can be more than two classes but for simplicity we are only considering the two-class problem. The mapping function can be learnt by decision tree or rule induction, neural networks, discriminate analysis or case-based reasoning. We will concentrate in this paper on symbolic learning methods such as decision tree induction. The decision tree learnt based on the data of our little example described above is shown in Figure 5. The profile of the buyers is: marital_status = single, number_of_ children=0. The profile of the non-buyers is: marital_status = married or marital_status = single and number_of_children > 0. This information can be used to promote potential customers.

age	sex	marital_status	number of_children	purchaised
33	female	single	0	0
34	female	single	1	1
35	female	married	2	1
36	female	married	0	1
29	female	single	0	0
30	male	single	0	0
31	male	single	1	1
32	male	single	2	1
33	male	married	0	1

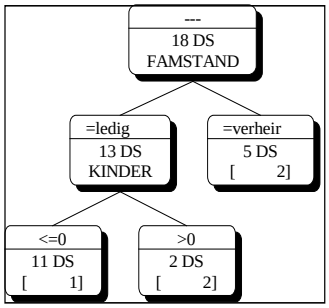

Fig. 5. Example Data Base and Resulting Decision Tree for Campaign Management

4.2.2 Regression

Whereas classification determines the set membership of the samples, the answer of regression is numerical. Suppose we have a CCD sensor. We give light of a certain luminous intensity to this sensor. Then this light is transformed into a gray value by the sensor according to a transformation function. If we change the luminous intensity we also change the gray value. That means the variability of the output variable, will be explained based on the variability of one or more input variables.

4.3 Knowledge Discovery

4.3.1 Deviation Detection

Real-world observation are random events. The determination of characteristic values such as the quality of an industrial part, the influence of a medical treatment to a patient group or the detection of visual attentive regions in images can be done based on statistical parameter tests.

4.3.2 Cluster Analysis

A number of objects that are represented by a n-dimensional attribute vector should be grouped into meaningful groups. Objects that get grouped into one group should be as similar as possible. Objects from different groups should be as dissimilar as

possible. The basis for this operation is a concept of similarity that allows us to measure the closeness of two data entries and to express the degree of their closeness.

Once groups have been found we can assign class labels to these groups and label each data entry in our data base according to its group membership with the corresponding class label. Then we have a data base which can serve as basis for classification.

4.3.3 Visualization

The famous remark "A picture is worth more than thousand words." especially holds for the exploration of large data sets. Numbers are not easy to overlook by humans. The summarization of these data into a proper graphical representation may give humans a better insight into the data. For example, clusters are usually numerically represented. The dendrogram illustrates these groupings, and gives a human an understanding of the relations between the various groups and subgroups.

A large set of rules is easier to understand when structured in a hierarchical fashion and graphical viewed such as in form of a decision tree.

4.3.4 Association Rules

To find out associations between different types of information which seem to have no semantic dependence, can give useful insights in e.g. customer behavior. Marketing manager have found that customers who buy oil in a supermarkt will also by vegetables. Such information can help to arrange a supermarket so that customers feel more attracted to shop there.

To discover which HTML documents are retrieved in connection with other HTML documents can give insight in the user profile of the website visitors.

4.3.5 Segmentation

Suppose we have mined a marketing data base for user profiles. In the next step we want to set up a mailing action in order to advertise a certain product for which it is highly likely that it attracts this user group. Therefore, we have to select all addresses in our data base that meet the desired user profile. By using the learnt rules as query to the data base we can segment our data base into customers that do not meet the user profile and into those that meet the user profile. The separation of the data into those data that meet a given profile from those that do not met a given profile is called segmentation. It can be used for a mailing action where only the address of the customers who meet a given profile are selected and mailed out an advertising letter.

5 Intelligent E-Marketing with Data Mining and User-Adapted Interfaces

5.1 Objectives

In the following we describe the recent work we are developing for an on-line sales and advertisement model methods and processes for integrated Data Mining and the ensuing user-specific adaptation of the web contents. The result shall be tools comprising the following essential steps:

- Identification and recording of web data that are in the following steps the base for building up user profiles.
- Analysis of the data by data mining processes in order to discover and build up user profiles, e.g. correlation of activities like purchase of associated products or combining the purchase of some products with certain delivery options.
- Integration and visualization, respectively, of the analyzed data for the Web Content Management and the author process, respectively.
- Conversion of the user profiles and the set of rules into user-adapted multimedia presentations and interaction forms.

The solutions achieved in this project are to be new Data Mining processes, oriented on the Web-Mining, and allowing at the same time the feedback according to the knowledge drawn from the Data Mining process to the web contents and the Content Management, respectively. In the framework of this project we want to investigate among other things the connection of temporary user modeling with long-term models, as well as the reciprocal influence of both models in different application contexts and we want to realize them with the help of component technologies. As a result there will be prototypical software components at disposal.

5.2 The Architecture

Figure 6 describes the architecture of the E-shop system with integrated data mining components. These components include components to access data, clean data, data mining components, components to visualize the results of the data mining process and components for the direct usage of the mined knowledge in the e-shop.
In detail these functional components are:

- The user interface client components,
- The user modeling components
- The data mining components,
- The knowledge repository,
- The visualization component for the web usage mining and user profiles and

- The data base with the different media elements and templates.

The data are collected by on-line forms, by server logs and by cookie logs or java agents in a history list. This information is given as an input to the data mining component where it can be used for different purposes. The data can be used to learn the user model and the user preferences as well as the usage of the website. In the data mining component are realized data mining methods such as decision tree induction and conceptual clustering for attribute-value based and graph-structured data. Decision tree induction requires that the data have a class label. Conceptual clustering can be used to learn groups of similar data. When the groups have been discovered the data can be labeled by a group name and as such it can be used for decision tree induction to learn classification knowledge.

Based on the user model the presentation style and the content of the web site (adaptive multimedia product presentation) are controlled. Besides that the user model is used to set up specific marketing actions such as e.g. mailing actions or cross-selling actions. The results of the webusage mining are used to improve the website organization as well as for monitoring the impact rate of the advertisement of particular events. Product models and preferences are used to control the content of the website. The preferences can be learned based on the user´s navigation data. Besides that an intelligent dialogue component allows to control the dialogue with the user [9].

The following processes can be handled with these components:

- Web-Site Administration,
- Advertisement,
- Marketing and Selling,
- Adaptive Multimedia Product Presentation,
- Event Recognition, and
- Learning Ontology Knowledge.

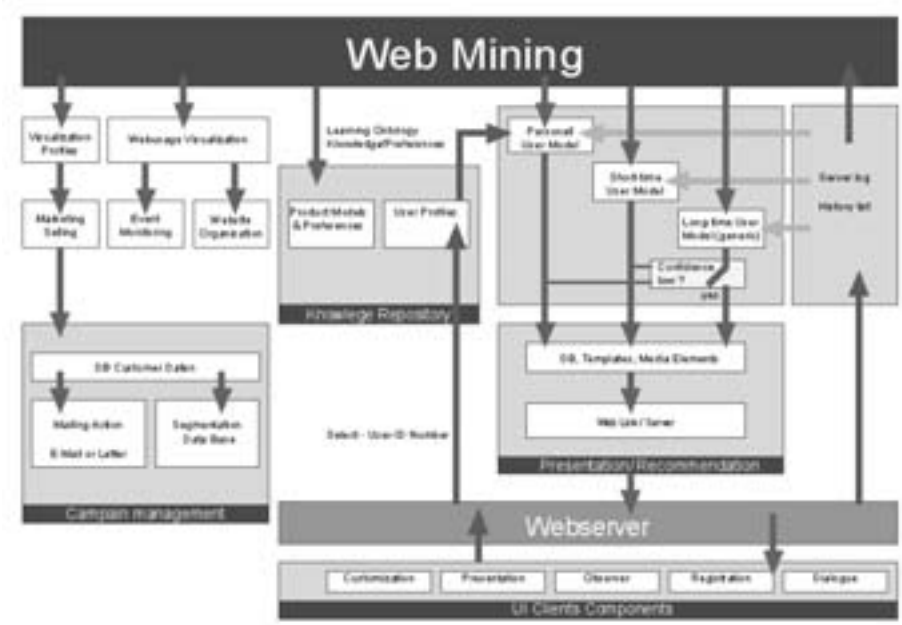

Fig. 6. Architecture of an intelligent E-shop

5.2.1 The Data

The data are the server log, the history list and the user data obtained by the registration component. The server log contains data described in Sect. 3.1. Data for the history list are collected by an observation component which is installed in the browser on the client site when the user is visiting the website. These observation components observe actions of the user on the internet pages [17].

5.2.2 User Models

Different user models are contained in our architecture: the individual user model, the short-term user model, and the long-term user model.
The individual user model can be obtained by setting up electronic user profiles. The users needs and preferences can be collected by a questionnaire [12][16] when the user is entering the website or by accessing electronic profiles which can be accessed from the browser of the client site [11]. Theses profiles can be stored in a knowledge

repository as individual user profile and each time the user visits the web site and identifies himself it can be used to present the user the content depending from his preferences and the used hardware and software facilities.

To fill out forms or questionnaires requires considerable user effort and the cooperation of the user. Since not every user likes to give out information about himself therefore, it should be possible to categorize the users into several user groups. Each user group represents a significant and large enough group of users sharing several properties together. Based on these user groups should be controlled the functions of the e-shop. The identification of the user groups can be done based on the users navigation behavior. The user's action while browsing a web site should be observed and should be used to learn the user profiles.

The users interest may change over time. Therefore the user model should adapt to this concept drift. A recent trend separates the user model into a short-term and a long term user model [13][14]. The short-term user model is based on highly specific information, whereas the long-term user model is based on more general information. The short-term model is learned from the most recent observations only. It represents user models, which can adjust more rapidly to the user´s changing interests. If the short-term model can not classify the actual user at all, it is passed on to the long-term model which represents stereotypical user groups [19]. The purpose of the long-term model is to model the user's general preferences for certain products that could not be classified by the short-term model.

5.2.3 Mining for the User Model

Webb et al. [18] summarize four major issue in learning user models:

- The need for large data sets;
- The need for labeled data;
- Concept drift, and
- Computational complexity.

The problem of the limited data set and the problem of concept drift has lead to hybrid user models separated into a short-term and a long-term user model.

Most applications use the nearest neighbor method to model the short-term user model. This method searches for similar cases in a data base and applies the action associated to the nearest case to the actual problem. A specific problem of this method is the selection of the right attributes that describe the user profile and/or the set up of the feature weights [13] as well as the complexity. Bayesian classifiers are used for the long-term model [13].

We intend to use incrementally decision tree induction to learn both user models; the short-term and the long-term user model. It allows us to use the same development strategy for learning the models in both cases. This can be an important system feature. To overcome the limited data set problem we use boosting for building the short-term model. Decision tree induction can be used to learn the classification model as well as to cluster data. In contrast to nearest neighbor methods, decision

tree induction generalizes over the data. This will give us a good understanding of the user modeling process [5].

Decision tree induction allows one to learn a set of rules and basic features necessary for the user modeling. The induction process does not only act as a knowledge discovery process, it also works as a feature selector, discovering a subset of features from the whole set of features in the sample set that is the most relevant to the problem solution. A decision tree partitions the decision space recursively into sub-regions based on the sample set. In this way the decision tree recursively breaks down the complexity of the decision space. The outcome has a format, which naturally presents the cognitive strategy of the human decision-making process. This satisfies our need for visualization and reporting the results to the marketing people.

A decision tree consists of nodes and branches. Each node represents a single test or decision. In the case of a binary tree, the decision is either true or false. Geometrically, the test describes a partition orthogonal to one of the coordinates of the decision space. The starting node is usually referred to as the root node. Depending on whether the result of a test is true or false, the tree will branch right or left to another node. Finally, a terminal node is reached (sometimes referred to as a leaf), and a decision is made on the class assignment. Also non-binary decision trees are used. In these trees more than two branches may leave a node, but again only one branch may enter a node. For any tree all paths lead to a terminal node corresponding to a decision rule of the "IF-THEN" form that is a conjunction (AND) of various tests.

The main tasks during decision tree learning can be summarized as follows: attribute selection, attribute discretization, splitting, and pruning. We will develop special methods for attribute discretization [6] that allow to discretize numerical attributes into more than two intervals during decision tree learning and to agglomerated categorical attribute values into supergroups. This leads to more compact trees with better accuracy. Besides that we will develop special pruning methods. Both techniques are necessary for the special kind of data and will be set up for the special needs of learning the user model.

To understand the concept drift, we will develop a method to compare the outcome of the decision tree induction process and to derive conclusions from it. This will give us a special technique to control the user model.

5.2.4 Web Usage Mining

Analyzing the server logs and the history list can help to understand the user behavior and the web structure, thereby improving the design of the website. Applying data mining techniques on access logs unveils interesting access patterns that can be used to restructure sites in more efficient groupings, pinpoint effective advertising locations, and target specific users for specific selling ads.

We intent to develop conceptual clustering technique to understand the user accessing pattern. Classical clustering methods only create clusters but do not explain why a cluster has been established. Conceptual clustering methods build clusters and explain why a set of objects confirms a cluster. Thus, conceptual clustering is a type of learning by observations and it is a way of summarizing data in an understandable

manner [7]. In contrast to hierarchical clustering methods, conceptual clustering methods build the classification hierarchy not only based on merging two groups. The algorithmic properties are flexible enough in order to dynamically fit the hierarchy to the data. This allows incremental incorporation of new instances into the existing hierarchy and updating this hierarchy according to the new instance.

We propose an algorithm that incrementally learns the organizational structure [8]. This organization scheme is based on a hierarchy and can be up-dated incrementally as soon as new cases are available. The tentative underlying conceptual structure of the access pattern is visually presented to the user. We have developed two approaches for clustering access patterns. Both are based on approximate graph subsumption. The first approach is based on a divide-and-conquer strategy whereas the second is based on a split-and-merge strategy which better allows to fit the hierarchy to the actual structure of the application, but requires more complex operations. The first approach uses a fixed threshold for the similarity values. The second approach uses an evaluation function for the grouping of the cases.

5.2.5 Reporting Tools

Although the outcome of the data mining component is a set of rules or a description of the clusters which can be directly used to control the functional components of the website or directly incorporated into the user modeling component. We also prefer to report the results of the data mining process in a form a system administrator or marketing person can use for further review of the results. Therefore we will integrate visualization components into our system that allow to visualize the resulting decision tree, the hierarchical representation of the conceptual clusters, and the statistics for the event marketing.

5.2.6 Knowledge Repository

In the knowledge repository are stored individual user profiles and product models and preferences. The individual user profile is created by the user with the help of the registration component of the user interface. It can be updated by the user itself or electronically by the data mining component after having analyzed the user data when visiting the website.

6 Conclusions

We have introduced a new architecture extending an e-shop into an intelligent e-marketing and selling platform which can adapt to user needs and preferences. The data which can be accessed during a user session as well as the method for analysing these data play an important role for achieving this goal. Therefore, we have reviewed the basic data that can be created during a customer session. Based on the kind of data and the wanted output the data mining methods are selected. We have reviewed the basic data mining methods and given an overview on what kind of method is eligible

for the considered result. We have identified two types of data mining methods useful for our first set up of the intelligent e-shop. These are classifications based on decision tree induction and conceptual clustering. With these methods we can solve such problems as learning the user model, web usage mining for web site organization, campaign management, and event monitoring. The data might be labeled or might not have a label. In the latter case clustering is to use to learn similar groups and label them. Recently, we have continued to develop and implement the methods for decision tree induction and conceptual clustering. Each method will be implemented as a component with standard input and output interfaces that allows to assemble the components as far as will be needed for the particular e-shop.

Acknowledgment

Part of this work has been funded by the German Ministry of Economy and Technology. The funding is greatly acknowledged.

References

1. M. Stolpmann, On-line Marketing Mix, Kunden finden, Kunden binden im E-Business, Galileo Press, Bonn 1999.
2. Cooley, R., Mobasher, B., and Srivastava, J., Data Preparation for Mining World Wide Web Browsing Patterns, Knowledge and Information Systems, 1(1), 1999.
3. M. Merz, E-Commerce und E-Business, dpunkt.verlag Heidelberg, 2002.
4. J. Han and M. Kamber, Data Mining, Concepts and Techniques, Academic Press, San Diego, 2001
5. P. Perner, Data Mining on Multimedia Data, Springer Verlag in preparation
6. P. Perner and S. Trautzsch, Multinterval Discretization for Decision Tree Learning, In: Advances in Pattern Recognition, A. Amin, D. Dori, P. Pudil, and H. Freeman (Eds.), LNCS 1451, Springer Verlag 1998, S. 475-482.
7. D.H. Fisher, "Knowledge Acquisition via Incremental Clustering," Machine Learning, 2: 139-172, 1987.
8. P. Perner and G. Fiss, Conceptual Clustering of Graph-structured Data, PKDD2002, submitted
9. P. Cunningham, R. Bergmann, S. Schmitt, R. Traphöhner, S. Breen, and , B. Smyth, WEBSELL: Intelligent Sales Assistent for the World Wide Web, Zeitschrift Künstliche Intelligenz, 1, 2001, p. 28-32.
10. T. Ardissono and A. Goy, Tailoring the Interaction with Users in Web Stores, Journal on User Modeling and User-Adapted Interaction, 10(4) , 2000, p.251-303.
11. S. Shearin and P. Maes, Representation and Ownership of Electronic Profiles, Workshop Proc. Interactive Systems for 1-to-1 E-Commerce, CHI 2000, The Hague (The Netherlands), April 1-6, 2000.
12. S. Shearin and H. Liebermann, Intelligent Profiling by Example, Proc. of Intern. Conf. of Intelligent User Interfaces (IUI2001), p. 145-152, Santa Fe, NM, Jan. 14-17, 2001.

13.D. Billsus and M. Pazzani, A Hybrid User Model for News Story Classification, In: J. Kay, User Modeling: Proceedings of the Seventh International Conference, UM99, Springer Wien New York, 1999, p. 99-108.
14.J. Kay and R. Thomas, Long term learning in the workplace, Communications of the ACM, 38(7), July 1995, 61-69.
15. I. Koychev and I. Schwab, Adaption to Drifting User`s Interests,
16.Henry Lieberman. *Autonomous Interface Agents*. In Proceedings of ACM Conference on Human Factors in Computing Systems (CHI), 1997
17.M. Claypool, Phong Le, M. Waseda, D. Brown, Implicit Interest Indicators, IEEE Internet Computing, Nov./Dez. 2001 http://www.computer.org/internet
18. G.I. Webb, M.J. Pazzani, and D. Billsus, Machine Learning for User Modeling, User Modeling and User-Adapted Interaction, 11: 19-29, 2001.
19.J Kay, Lies, damned lies and stereotypes: pragmatic approximations of users, in Procs of UM94 - 1994 User Modeling Conference, UM Inc, Boston, USA, 1994, pp 175-184.