Web Mining pattern discovery

Paolo Giudici
(Università di Pavia)

Claudia Tarantola
(Università di Pavia)

# 156 (09-03)

Dipartimento di economia politica
e metodi quantitativi
Università degli studi di Pavia
Via San Felice, 5
I-27100 Pavia

Settembre 2003
Web mining pattern discovery

P. Giudici and C. Tarantola

Data Mining Laboratory,
Department of Economics and Quantitative Methods,
University of Pavia, Italy

September 12, 2003

Abstract

The aim of this paper is to show how web click stream data can be used to understand the most likely path of navigations in a Web site. The information deriving from such analysis can be usefully employed to efficiently design the Web site. We consider and compare both statistical methodologies (odds ratios and graphical models) and computational methodologies (association and sequence rules). These methodologies are applied to the analysis of a real set of data regarding an e-commerce web site of a company that sells hardware and software products.

1 Introduction

In the last few years the number of people that have used the Internet has enormously increased. Companies promote and sell their products on the Web, institutions provide information about their service and single individuals exploit personal Web pages to be introduced to the whole internet community.

Every time an user links up at a web site, the server keeps track of all the actions accomplished in the log file. What is captured is the “click flow”, click-stream, of the mouse and the keys used by the user during the navigation inside the site. Usually at every click of the mouse corresponds the visualization of a web page. Therefore, we can define a click-stream as the sequence of the requested pages. The succession of the pages shown by a single user during his navigation inside the Web identifies an user session. Typically, the
analysis only concentrates on the part of each user session concerning the access at a specific site. The set of the pages seen, inside a user session, coming from a determinate site is known with the term server session.

All this information can be profitably used to efficiently design a Web site. A web page is well designed if it is able to attract users and address them easily to other pages within the site. This is the main problem faced by Web Mining Techniques. In fact, a very important area in Web Mining is the application of data mining techniques to discover usage patterns from Web data, in order to optimally design a web site and to better satisfy needs of different visitors.

In this paper we shall show how web click-stream data can be used to understand the most likely paths of navigation in a web site, with the aim of predicting which pages will be seen, having seen a specific path of pages in the past. The methodologies we shall employ to achieve this purpose are either statistical (odds ratios and graphical models) or computational (association and sequence rules). The comparison between the methods introduced in the paper allows to shed some new light on the fields of data mining of association structures. In particular, we compare two kinds of graphical models that we believe well suited to model sequential patterns: bayesian networks and dependency networks.

The structure of the paper is the following. In Section 2 we present the data-set that will be used as a running example throughout the paper and we provide a brief description of the preprocessing stage. In Section 3 we introduce odds ratios and sequence rules, in the context of exploratory statistical analysis, and we present a graphical representation of them. Section 4 is concerned with graphical models and their application to web mining. Finally Section 5 contains some concluding remarks.

2 Description of the data and pre-processing

For the analysis we consider a data-set resulting from the elaboration of a log file concerning a site of e-commerce. The source of the data cannot be specified for privacy reason, the only information that we can provide is that it is the website of a company that sells hardware and software products. The period of observation is of two years, from 30 September 1997 up to 30 June 1999.

The original logfile has been processed to produce a data-set, containing the user id (c.value), a variable with the date and the instant the visitor has linked to a specific page (c.time), the web page seen (c.caller) and the order of visit (c.order). Table 1 regards the click stream of the visitor identified by the (cookie)
This visitor has entered the site on the fourteenth of October, 1997, at 11:09:01, and has visited, in sequence, the pages home, catalog, program, product, program, leaving the website at 11:09:24.

TABLE 1 ABOUT HERE.

The whole data-set contains 250711 observations, each corresponding to a click, and it describes the navigation paths of 22527 visitors among the 36 pages which compose the site. The visitors are taken as unique, that is, no visitors appears within more than one session. On the other hand, we remark that a page can occur more than once in the same session. We now briefly describe the content of the most seen pages, which here correspond to the statistical variables under consideration.

**HOME**: the home page of the web site;

**LOGIN**: where a user has to enter its name and other personal information, during the first registration, in order to access to certain services and products, reserved to the customers;

**LOGPOST**: prompts a message that informs whether the login has been successful or if it has failed;

**LOGOUT**: on this page the user can leave the personal characterization given in the login page;

**REGISTER**: in order to be later recognized, the visitor has to prompt a user-id and password;

**REGPOST**: shows the partial results of the registration, asking for missing information;

**RESULTS**: once the registration is accomplished, this page summarizes the information given;

**REGFORM1**: here the visitor has to insert data that enable him/her to buy a product, such as a personal identification number;

**REGFORM2**: here the visitor has to subscribe a contract in which he/she accepts the conditions for on-line commerce;

**HELP**: it answers questions that may arise during the navigation through the web site;

**FDBACK**: a page that allows to go back to the previous one visited;

**FDPOST**: a page that allows to go back to one previously seen page, in determined areas of the site;

**NEWS**: it presents the last novelties available;

**SHELF**: it contains the list of the programs that can be downloaded from the website;

**PROGRAM**: gives detailed information on the characteristics of the software programs that can be bought;

**PROMO**: gives an example (demo) of the peculiarities of a certain program;
**DOWNLOAD:** it allows to download software programs of interest;

**CATALOG:** it contains a complete list of the products on sale in the web site;

**PRODUCT:** shows detailed information on each product that can be purchased;

**P_INFO:** a page on which detailed information on the terms of payment of the products can be found;

**ADDCART:** the place where the virtual basket can be filled with items to be purchased;

**CART:** shows the current status of the basket, that is, which items it contains;

**MDFYCART:** allows to modify the current content of the basket, for instance taking off items;

**CHARGE:** indicates the amount due to buy the items contained in the basket;

**PAY_REQ:** a page which visualizes the amount finally due for the products in the basket;

**PAY_RES:** here the visitor agrees to pay, and data for payment are inserted (for example, the credit card number);

**FREEZE:** where the requested payment can be suspended, for instance to add new products to the basket.

We have performed a preliminary screening of the available data in order to identify possible outliers. For sake of parsimony we do not report here the results of the analysis but we refer to Giudici [4] for the details. The original data-set has been rearranged into a new one organized by sessions; for each session we consider a set of characterizing variables, such that the total time length of the server session (*length*), the total number of clicks made in a session (*clicks*), and the time in which the session starts (*start*, setting at 0 the midnight of the preceding day). Furthermore, such data-set contains binary variables that describe whether each page is visited at least once (modality 1) or not (modality 0). Table 2 shows an extract from this new data-set, that corresponds to the session in Table 1.

**TABLE 2 ABOUT HERE**

After a preliminary screening of the variables we have decided to eliminate all observations above the 99-th percentile of the distribution of the variables *clicks* and *length*. On the other hand, we have decided not to remove observations possibly outlying with respect to the variable *start*. This because of the nature of the variable itself as well as the observed distribution. We have obtained in this way a new data-set containing 22152 observations, in place of the initial 22527.

On the cleaned data-set we can now calculate some descriptive indices regarding the 28 most frequently
visited pages, see Table 3.

Table 3 about here.

Consider the relative frequency of visit for each of the pages (fourth columns of Table 3). From this, it derives that the most visited pages are, in decreasing order, **product** (84.62%), **program** (76.74%), **home** (48.71%), **p_info** (48.29%) and **catalog** (48.71%). Furthermore in the table, for e-commerce purposes, we also indicate the frequency distribution of the each variable, conditionally on the values of the variable purchased. We remark that the overall proportion of purchasers is 7.21% and, therefore, the fourth column can be obtained as a weighted average of the other two, with weights equal to 7.21% and 92.79%.

3 Association indexes

In Web Mining pattern discovery it is useful to investigate the strength of relationship between two web pages. We would like to identify those web page that are often visited jointly. For this purpose we first propose to use descriptive measures of association, more precisely odds ratios and sequence rules.

For this analysis we have binarized the c.caller variable assigning value 1 or 0 if the visitor has at least seen once the page or not. Furthermore, from here on we shall represent the joint distribution of a couple of variables A and B, in our case two different web pages, by a two way contingency table (see Table 4).

Table 4 about here.

The first methodology we consider is based on the calculation of the odds ratios, for more details see Agresti [1]. The odds ratio equals the ratio of the products $p_{11}$ $p_{22}$ and $p_{12}$ $p_{21}$ of probabilities from diagonally-opposite cells: $\theta = p_{11} p_{22}/p_{12} p_{21}$. It can equal any non negative number. We can distinguish three different situations:

1) When all cell probabilities are positive, independence of A and B is equivalent to $\theta = 1$. In our case this means that knowing that the user has visited (or not) page A does not give us any additional information regarding page B.

2) If $\theta > 1$ there is a positive association between A and B, that is the visualization of one page is relatively more frequent when also the other page has been visited. On the other hand, when one page is not visited, it is likely that also the other page will not be visited.

3) If $0 < \theta < 1$ there is a negative association between A and B, that is the visualization of one page is
relatively more frequent when the other page has not been visited. On the other hand, when one page is not
visited, it is likely that the other page will be visited.

For an exemplification consider Table 5 containing the joint distribution of the variables freeze and
pay_req, the odds ratio is equal to $\theta = \frac{3814 \times 16845}{17 \times 1851} = 2041.721$ and indicates a strong positive
association association between the two variables, which means that, in most of the cases if pay_req is visited
also freeze is visited and if pay_req is not visited freeze will not be visited either. In a similar way we have
calculated the odds ratios between all the couples of pages under consideration, see Section 3.1.

TABLE 5 ABOUT HERE

In Web mining pattern discovery we are also interested in the order in which the pages have been visited.
The nature of the odds ratio does not permit to take into in account this information; in fact it allows us to
identify if two pages, say A and B, are likely to be seen together, but not in which order. For this reason
we suggest to consider sequence rules, a particular type of association measure with rules ordered in time.
In this case $A \rightarrow B$ indicates not only that the two pages are both visited but also that A precedes B. For
an introduction see, for example, Giudici [4].

We can distinguish between indirect and direct sequence rules (see, for example, [5]). A sequence rule is
indirect if between the visit of page A and the visit of page B other pages can be seen. On the other hand,
in a direct sequence rule A and B are seen consecutively. In web click stream analysis, a sequence rule is
typically indirect, for this reason here we shall consider only this case.

Let us consider the indirect sequence $A \rightarrow B$ and indicate as $N_{A \rightarrow B}$ the number of visits which appear in
such sequence, at least once. Let $N$ be the total number of the server sessions. Notice that the rule $A \rightarrow B$
will be counted only once even if it had been repeated several times inside the session. The support for the
rule $A \rightarrow B$ is obtained dividing the number of server sessions which satisfy the rule by the total number of
server sessions:

$$support\{A \rightarrow B\} = \frac{N_{A \rightarrow B}}{N}. \quad (1)$$

The support is a relative frequency that indicates the percentage of the users that have visited in succession
the two pages. In presence of a high number of visits, as it usually happens, it is possible to state
that the support for the rule expresses the probability an user session contains the two pages in sequence:

$$support\{A \rightarrow B\} = Pr\{A \rightarrow B\}.$$
The confidence for the rule $A \rightarrow B$ instead is obtained dividing the number of server sessions which satisfy the rule by the number of sessions containing the page $A$:

$$\text{confidence}\{A \rightarrow B\} = \frac{N_{A \rightarrow B}}{N_A} = \frac{s\text{upport}\{A \rightarrow B\}}{s\text{upport}\{A\}}$$ (2)

It approximates the conditional probability that in a server session in which page $A$ has been seen is subsequently required page $B$.

The results obtained by means of SAS Enterprise Miner are reported in Table 6, where the sequences have been ordered on basis of their support. Note that the sequence $\text{program} \rightarrow \text{product}$ is the one with the highest frequency ($\approx 69\%$) it also corresponds to the sequence with the highest value of the index of confidence.

TABLE 6 ABOUT HERE.

The rules discussed so far are sequence rules of order two. In general we are interested in identifying “navigation chains” constituted from a greater number of pages than two. Consider, for example, the undirected sequence $\{A \rightarrow B \rightarrow C \rightarrow D\}$. Remember that we are examining indirect sequences therefore after having visualized page $A$, for example, other pages may have been required before visualizing page $B$. In this case, the rule is logically interpreted as “If $\{A \rightarrow B \rightarrow C\}$, then $D$”.

Analogously to the case of order two, the support for the rule $\{A \rightarrow B \rightarrow C \rightarrow D\}$, that expresses the relative frequency with which the sequence $\{A \rightarrow B \rightarrow C \rightarrow D\}$ occurs on the set of the considered server sessions, is given by:

$$\text{support}\{A \rightarrow B \rightarrow C \rightarrow D\} = \frac{N_{A \rightarrow B \rightarrow C \rightarrow D}}{N}$$ (3)

While, the confidence rule for $\{A \rightarrow B \rightarrow C \rightarrow D\}$ obtained as the ratio between the number of visits which satisfy the rule and the number of visits which verify only the “body of the rule”, namely:

$$\text{confidence}\{A \rightarrow B \rightarrow C \rightarrow D\} = \frac{N_{A \rightarrow B \rightarrow C \rightarrow D}}{N_{A \rightarrow B \rightarrow C}}$$ (4)

Before ending this section we present in Table 7 the ten sequences of any order (up to a maximum of 10) having the highest value for the index of confidence.

TABLE 7 ABOUT HERE
3.1 Graphical representation of association rules

In this section we show how the association rules introduced in the previous section can be useful represented in terms of a graph. Let first consider the odds ratio approach. For each couple of pages we have constructed a $2 \times 2$ contingency table and calculated the corresponding odds ratios. We have then focalized our attention only on the positive associations since we are interested in couple of pages that have been effectively seen by the visitor. Furthermore, we have considered only the strongest associations, namely such that $\theta > 4$. The marginal association between the different web pages can then be represented by means of the graph in Figure 1.

**FIGURE 1 ABOUT HERE**

The structure of this graph can be usefully enriched using the sequence rules presented in the previous section. More precisely, we can consider the confidence index, and for those variables having a particularly strong association, described by an odds ratio greater than 8, we have included a directed edge from $A$ to $B$ if $\text{confidence}(A \rightarrow B) > \text{confidence}(B \rightarrow A)$, see the graph in Figure 2.

**FIGURE 2 ABOUT HERE**

Figure 2 should be interpreted in the following way: arrows starting from a page pointing towards other pages highlight what are the most frequent descendent for that page; similarly arrows which point to a page, coming from other pages, show which are the most frequent antecedents for that page.

For an exemplification let us consider the relation between *freeze* and *pay.req*; see Table 5. The index of confidence for the association between *freeze* and *pay.req* is given by:

$$\text{confidence}(\text{freeze} \rightarrow \text{pay.req}) = \frac{3814}{5665} = 0.6733.$$  

Whereas, reversing the order of the terms we obtain:

$$\text{confidence}(\text{pay.req} \rightarrow \text{freeze}) = \frac{3814}{3831} = 0.9956.$$  

Since 99.56% of the server session that contain *pay.req* contain also *freeze* it is logical to think that page *freeze* is usually requested before accessing the page *pay.req*. Therefore, in Figure 2 an arrow has been placed from *freeze* to *pay.req*. A similar reasoning has been applied to the other pairs of pages examined, depicted in Figure 2.
4 Graphical models

The main limit of the indexes presented in Section 3, for other aspects extremely flexible and informative, is that, as descriptive indexes, they allow only to draw valid conclusions for the observed data-set. In other terms, they do not allow to obtain reliable behavior forecasts for new users. A more general model is needed. Different solutions have been presented, for a review see Giudici [4]; the solution we present here is based on the use of graphical models.

A graphical model is a family of probability distributions incorporating the conditional independence structures represented by a graph. The vertex of the graph correspond to the variables under examination (in our case web pages), whereas the edges in the graph indicate a relation between the examined variables. In general, a direct influence of a variable on another one is indicated with a directed edge (graphically described by an arrow) while an undirected edge (described by a line) represents a symmetric association. Furthermore, using a set of rules that we shall not present here, we can use a graph to identify conditional independence between the examined variables. For more details on graphical models see e.g. Lauritzen [8].

We can distinguish four different types of graphs and correspondingly different types of graphical models:

a) **Undirected graphs**, containing only undirected edges used to model symmetric relations among the variables. They give rise to Markov Networks.

b) **Acyclic Directed graphs**, containing only directed edges. They are used to model asymmetric relations among the variables. They give rise to Acyclic Directed Graphical models also called Bayesian Networks.

c) **Graphs with multi-directional edges**, they are a particular type of directed graphs that may contain multi-directional arrows. They give rise to Dependency Networks.

d) **Chain graphs**, they contain both undirected and directed edges, and, therefore, can model both symmetric and asymmetric relationships. They give rise to Chain models.

In Section 4.1 we consider the application of Bayesian Networks and Dependency Networks to discover web mining patterns. More details on these two types of models will be given there.

4.1 Bayesian Networks and Dependency Networks

*Bayesian Networks* and *Dependency Networks* are two alternative representations of a multivariate problem in terms of a directed graph. Bayesian Networks are well known and have been widely used in different fields,
see e.g. Jensen [7]. On the other hand Dependency Networks are not so popular and have been proposed only recently by Heckerman et al [6] as an alternative to Bayesian Networks.

Both models are represented by a directed graph describing the ordered relationship between the variables, and such that the distribution of each variables $x_i$ is fully specified by its parents, $pa(x_i)$. More precisely, both models are defined in terms of the set of conditional distributions $p(x_i|pa_i)$ one for each node $x_i$ in the graph given its parents.

On the other hand, while Bayesian Networks are represented by acyclic graphs, Dependency Networks may contain multi-directional edges and potentially cyclic structures. The second and more important difference is that whereas in Bayesian Networks the joint distribution is obtained directly from the local ones as $p(x) = \prod_{i=1}^{V} p(x_i|pa_i)$, this is not always the case with Dependency Networks.

The main advantage of Dependency Networks is in terms of data visualization. They are in fact easier to be understood by people not familiar with graphical modelling semantic.

Consider the following example by Heckerman et al [6]. The graphs in Figure 3 represent the relation between three demographic variables (age, gender, income) that will be indicated for simplicity with $A$, $B$ and $C$ respectively. Graph (a) is a Bayesian Networks, whereas graph (b) is a Dependency Networks.

FIGURE 3 ABOUT HERE

As we have already said, the first difference is that in the Dependency Networks there is a cyclic structure and there are arrows with multiple directions. Consider for the example the graph represented in Figure 3.a. When a graph like this is shown and explained as representing causal relationships to an untrained person, such a person often gains an accurate impression of the relationships involved. The problem arises when the person can be only told that the relationships are “predictive” or “correlative”. In this case, in fact an untrained user will correctly conclude that $A$ and $B$ are predictive of $C$, but will wonder why there are no arcs from $C$ to $A$ and to $B$. Furthermore it will be quite difficult for him to understand that $A$ and $B$ are dependent given $C$.

The solution proposed by Heckerman et al [6] is to substitute a Bayesian Networks structure with one where the parents of each variable render that variable independent of all the other ones (see Figure 3.b). For example the Bayesian Network in Figure 3.a becomes that of Figure 3.b. Note that to quantify the dependencies we still consider the conditional probability of a variable given its parents. In order to specify
these probabilities one should transform the Dependency Network into a Markov Network with the same set of adjacencies (same set of vertices connect by an edge) and then calculate the required probability from this new network. Unfortunately, when the number of variables is elevated, as in data mining problems, this procedure turns out to be computationally inefficient. For this reason Heckerman et al [6] propose to learn independently the local distribution from the data and then to construct the corresponding Dependency Network. Unfortunately in some case the distributions obtained are inconsistent, that is the local distributions cannot be obtained via the rules of probability from a common joint distribution $p(x)$. In Heckerman et al [6] an algorithm is described that permits to partly overcome this problem.

4.2 Application

Before proceeding with describing the results from the analysis of the data, we deem important, according to the data mining process paradigm (see Giudici [4]) to first describe the software employed. All computations have been performed with the WinMine Toolkit, a software developed by the Machine Learning and Applied Statistics group of Microsoft Research. It can be freely downloaded from the web site http://research.microsoft.com/ dmax/WinMine/ContactInfo.html. From our experience we think that this software is quite user friendly, even if some features could be improved. WinMine requires a specific format of the data. For each visitor we must provide the user id, the name of the web page seen and the number of times this page has been visited (see Figure 4)

DIFFERENTLY FROM SECTION 3 WE HAVE RECLASSIFIED THE NUMBER OF TIMES EACH WEB PAGE HAS BEEN VISITED ACCORDING TO 4 LEVELS. IN PARTICULAR WE HAVE ASSIGNED VALUE 1 IF A PAGE HAS BEEN VISITED ONCE, VALUE 2 IF IT HAS BEEN VISITED TWO OR THREE TIMES, VALUE 3 IF IT HAS BEEN VISITED FROM FOUR TO NINE TIMES AND VALUE 4 IF IT HAS BEEN VISITED TEN OR MORE THAN TEN TIMES. WE REMARK THAT THIS CLEARLY EXTENDS THE RANGE OF POSSIBLE CONCLUSIONS TO BE OBTAINED FROM PATTERN DISCOVERY, WITH RESPECT TO THE COMMON BINARY ASSUMPTION, USED IN SECTION 3.

WinMine allows us to choose only the distribution of the data, in our case we use a multinomial one since we are working with categorical variables. Regarding the prior distribution setting, this is done automatically by the program. For more details see Heckerman et al [6] and Chickering et al. [3].
Figure 5 and 6 show a Bayesian Networks and a Dependency Networks structures learned from the data.

After only a short inspection, we note that both Bayesian Networks and Dependency Networks determine the same path, from \textit{fdpost} to \textit{cart}. It is interesting to note, however, that Bayesian Networks and Dependency Networks may attribute different roles to the same variable; for example in the Bayesian Networks \textit{home} is predictive of \textit{pinfo} whereas in the Dependency Networks the role of these variables is reversed.

When learning Bayesian Networks and Dependency Networks from the data, we remark that the local distributions are estimated by means of probabilistic classification trees (see Buntine [2]). The software Winmine indeed allows us to display the tree associated with each variable. To view a classification tree for a variable, it is only necessary to double click on the corresponding node in the Dependency Network. Figure 7 represents the classification tree for the variable \textit{pinfo} in the Dependency Network.

Note that there is a split on variable X, in the classification tree for Y, if and only if there is an arc from X to Y in the Dependency Network. Furthermore, the histograms at the leaves correspond to probabilities of \textit{pinfo} assuming values 1, 2, 3, 4 or \textit{not visited}. We remark the advantage of this representation, especially for predictive purposes.

A very important aspect, in the application of a data mining model, is the evaluation of its performance (see, for instance, Giudici [4]). In fact, the software WinMine provides a tool to evaluate how well the model constructed predicts out-of-sample data. First of all we have to randomly split the data into two subsets: a training set (70% of the data) and a test set (30% of the data). For each case in the test data, the tool evaluates the log posterior probability of the value for each output variable, given the values of all other variables. The average of these log posteriors across all variables in all cases is reported, see Heckerman et al [6].

Our application shows that Dependency Networks are slightly less accurate than Bayesian Networks; in our case the score for a Dependency Networks is $-0.374718$ whereas that for the corresponding Bayesian Networks (with the same sampled subsets) is equal to $-0.373111$. This difference is not surprising since the number of parameters in the Bayesian Networks are fewer than the number of parameters in the corresponding
5 Conclusions

In this paper we have compared alternative models for pattern discovery in web mining. Our main objective of interest is to infer, on the basis of the observed data, sequences of web pages that are very often seen by the users of a web site. This information can then be used either to device specific direct marketing campaigns or, simply, to redesign the web site, aiming at the more general public of users.

The models we have compared belong to two main families. On one hand, we have employed, and represented graphically, descriptive (non inferential) statistical measures, aimed at capturing efficiently the most relevant marginal associations between groups of variables (itemsets). The measures we have employed belong either the class of methods developed in the machine learning and computational literature (such as association and sequence rules, direct and indirect) or to those available from the statistical literature on categorical data analysis. Other measures could have been chosen (such as chi-squared based measures) but we believe that the chosen ones do well represent the two main “strategies of action” for web mining pattern discovery. The main advantage of descriptive measures is that they are relatively simple to compute and interpret; on the other hand, they have the disadvantage of being based on the sample at hand. A more serious drawback (which indeed is a computational advantage) is that they are local, namely based on subsets of variables. Correlation between subsets is not fully taken into account. Furthermore, it is rather difficult to develop measures of “overall quality” of an output set of association rules, therefore model assessment becomes rather subjective (see for instance Giudici [4], for a discussion on this matter).

A large part of the paper is concerned with building two inferential models for web mining pattern discovery: bayesian networks and dependency networks. Again, other models could have been considered, such as Markov models and probabilistic decision trees (see the above reference for a comparison). However, we have chosen two graphical modelling methods as they advantageously combine global modelling with local computations (see for instance Lauritzen [8]). The advantages of using such an approach is the obtainment of coherent predictions, whose uncertainty can be easily evaluated, for instance, by means of a confidence interval; furthermore, model diagnosis and comparison can proceed in a rigorous and consistent manner, as we have demonstrated in the previous section. The possible disadvantage attached to them is the necessity to
come up with “a structure” of some kind, that describe relationships between variables in a rather constrained way. This constraint is stronger for Bayesian Networks than for Dependency Networks; this also implies a better predictive performance of Bayesian Networks.

The conclusion is that the choice between different pattern discovery methods depends on the context of analysis. If there is sufficient subject-matter knowledge to entertain a graphical model (in order of requirements: a Dependency Network or a Bayesian Network) then one may go ahead with it, and obtain better interpretation and predictions. Otherwise, if such a knowledge is weak, one may get valuable (although perhaps preliminary) results with the application of sequence rules.

References


Table 1: The considered data-set

<table>
<thead>
<tr>
<th>c_value</th>
<th>c_time</th>
<th>c_caller</th>
</tr>
</thead>
<tbody>
<tr>
<td>70ee683a6df</td>
<td>14OCT97: 11:09:01</td>
<td>home</td>
</tr>
<tr>
<td>70ee683a6df</td>
<td>14OCT97: 11:09:08</td>
<td>catalog</td>
</tr>
<tr>
<td>70ee683a6df</td>
<td>14OCT97: 11:09:14</td>
<td>program</td>
</tr>
<tr>
<td>70ee683a6df</td>
<td>14OCT97: 11:09:23</td>
<td>product</td>
</tr>
<tr>
<td>70ee683a6df</td>
<td>14OCT97: 11:09:24</td>
<td>program</td>
</tr>
</tbody>
</table>

Table 2: The derived data-set

<table>
<thead>
<tr>
<th>c_value</th>
<th>c_time</th>
<th>length</th>
<th>clicks</th>
<th>time</th>
<th>home</th>
<th>catalog</th>
<th>addcart</th>
<th>program</th>
<th>product</th>
</tr>
</thead>
<tbody>
<tr>
<td>70ee683a6df</td>
<td>14OCT97</td>
<td>24</td>
<td>5</td>
<td>11:09:01</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Variable</td>
<td>Frequency of visit</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-----------</td>
<td>--------------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Purchase=0</td>
<td>Purchase=1</td>
<td>Frequency of visit</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ADDCART</td>
<td>26.33%</td>
<td>99.56%</td>
<td>31.61%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AGB</td>
<td>5.12%</td>
<td>8.33%</td>
<td>5.35%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CART</td>
<td>15.67%</td>
<td>55.73%</td>
<td>18.56%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CATALOG</td>
<td>42.68%</td>
<td>68.13%</td>
<td>44.52%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CHARGE</td>
<td>0.06%</td>
<td>1.63%</td>
<td>0.17%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DOWNLOAD</td>
<td>3.96%</td>
<td>65.44%</td>
<td>8.40%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FBACK</td>
<td>2.47%</td>
<td>1.88%</td>
<td>2.42%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FDPOST</td>
<td>0.78%</td>
<td>0.81%</td>
<td>0.79%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FREEZE</td>
<td>19.00%</td>
<td>99.94%</td>
<td>24.84%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HELP</td>
<td>8.49%</td>
<td>8.58%</td>
<td>8.50%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HOME</td>
<td>49.32%</td>
<td>40.89%</td>
<td>48.71%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LOGIN</td>
<td>40.18%</td>
<td>65.87%</td>
<td>42.03%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LOGOUT</td>
<td>1.90%</td>
<td>4.01%</td>
<td>2.05%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LOGPOST</td>
<td>30.42%</td>
<td>3.64%</td>
<td>33.53%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MDFYCART</td>
<td>3.46%</td>
<td>10.33%</td>
<td>3.96%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NEWS</td>
<td>9.60%</td>
<td>4.76%</td>
<td>9.25%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P_INFO</td>
<td>49.52%</td>
<td>32.50%</td>
<td>48.29%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PAY_REQ</td>
<td>10.45%</td>
<td>97.43%</td>
<td>16.73%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PAY_RES</td>
<td>5.05%</td>
<td>90.61%</td>
<td>11.22%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PRODUCT</td>
<td>83.69%</td>
<td>96.49%</td>
<td>84.62%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PROGRAM</td>
<td>76.81%</td>
<td>75.89%</td>
<td>76.74%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PROMO</td>
<td>0.27%</td>
<td>0.88%</td>
<td>0.32%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>REGFORM1</td>
<td>1.24%</td>
<td>2.07%</td>
<td>1.30%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>REGFORM2</td>
<td>1.28%</td>
<td>1.31%</td>
<td>1.29%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>REGISTER</td>
<td>26.17%</td>
<td>35.94%</td>
<td>26.88%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>REGPOST</td>
<td>16.79%</td>
<td>32.25%</td>
<td>17.91%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RESULT</td>
<td>4.15%</td>
<td>45%</td>
<td>4.18%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SHELF</td>
<td>15.90%</td>
<td>74.64%</td>
<td>20.13%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3: Frequency distribution of the visits and of the purchase variable
<table>
<thead>
<tr>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>0</td>
</tr>
</tbody>
</table>

Table 4: Example of $2 \times 2$ table

\[
\begin{array}{c|c|c}
\text{Pay\_req} & 1 & 0 \\
\hline
\text{Freeze} & 3814 & 1851 \\
\hline
1 & 5665 \\
\hline
0 & 17 & 18696 \\
\hline
tot & 3831 & 18696 \\
\end{array}
\]

Table 5: Two dimensional contingency table for Freeze and Pay\_req

<table>
<thead>
<tr>
<th>#</th>
<th>Support (%)</th>
<th>Confidence (%)</th>
<th>Transaction Count</th>
<th>Rule</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>68.88</td>
<td>89.57</td>
<td>15517</td>
<td>\text{program} \rightarrow \text{product}</td>
</tr>
<tr>
<td>2</td>
<td>55.38</td>
<td>65.34</td>
<td>12476</td>
<td>\text{product} \rightarrow \text{product}</td>
</tr>
<tr>
<td>3</td>
<td>48.19</td>
<td>56.85</td>
<td>10856</td>
<td>\text{product} \rightarrow p_{info}</td>
</tr>
<tr>
<td>4</td>
<td>39.56</td>
<td>88.08</td>
<td>8912</td>
<td>\text{catalog} \rightarrow \text{program}</td>
</tr>
<tr>
<td>5</td>
<td>38.54</td>
<td>50.11</td>
<td>8681</td>
<td>\text{program} \rightarrow p_{info}</td>
</tr>
</tbody>
</table>

Table 6: The most frequent indirect sequences with two pages
<table>
<thead>
<tr>
<th>#</th>
<th>Chain Length</th>
<th>Support (%)</th>
<th>Confidence (%)</th>
<th>Trans. Count</th>
<th>Rule</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5</td>
<td>4.45</td>
<td>96.62</td>
<td>1002</td>
<td>( \text{logpost} \rightarrow \text{cart} \rightarrow \text{catalog} \rightarrow \text{program} \rightarrow \text{product} )</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>2.80</td>
<td>96.48</td>
<td>631</td>
<td>( \text{product} \rightarrow \text{logpost} \rightarrow \text{catalog} \rightarrow \text{program} \rightarrow \text{product} )</td>
</tr>
<tr>
<td>3</td>
<td>6</td>
<td>2.80</td>
<td>96.48</td>
<td>630</td>
<td>( \text{product} \rightarrow \text{login} \rightarrow \text{logpost} \rightarrow \text{catalog} \rightarrow \text{program} \rightarrow \text{product} )</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>5.69</td>
<td>96.32</td>
<td>1281</td>
<td>( \text{home} \rightarrow \text{logpost} \rightarrow \text{catalog} \rightarrow \text{program} \rightarrow \text{product} )</td>
</tr>
<tr>
<td>5</td>
<td>6</td>
<td>5.68</td>
<td>96.31</td>
<td>1280</td>
<td>( \text{home} \rightarrow \text{login} \rightarrow \text{logpost} \rightarrow \text{catalog} \rightarrow \text{program} \rightarrow \text{product} )</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>2.31</td>
<td>96.76</td>
<td>520</td>
<td>( \text{login} \rightarrow \text{logpost} \rightarrow \text{cart} \rightarrow \text{catalog} \rightarrow \text{program} \rightarrow \text{product} )</td>
</tr>
<tr>
<td>7</td>
<td>4</td>
<td>13.56</td>
<td>95.68</td>
<td>3054</td>
<td>( \text{logpost} \rightarrow \text{catalog} \rightarrow \text{program} \rightarrow \text{product} )</td>
</tr>
<tr>
<td>8</td>
<td>7</td>
<td>4.60</td>
<td>95.66</td>
<td>1037</td>
<td>( \text{home} \rightarrow \text{logpost} \rightarrow \text{catalog} \rightarrow \text{program} \rightarrow \text{product} \rightarrow \text{addcart} \rightarrow \text{freeze} )</td>
</tr>
<tr>
<td>9</td>
<td>6</td>
<td>4.60</td>
<td>95.66</td>
<td>1037</td>
<td>( \text{home} \rightarrow \text{logpost} \rightarrow \text{catalog} \rightarrow \text{program} \rightarrow \text{addcart} \rightarrow \text{freeze} )</td>
</tr>
<tr>
<td>10</td>
<td>7</td>
<td>4.60</td>
<td>95.66</td>
<td>1036</td>
<td>( \text{home} \rightarrow \text{login} \rightarrow \text{logpost} \rightarrow \text{catalog} \rightarrow \text{program} \rightarrow \text{addcart} \rightarrow \text{freeze} )</td>
</tr>
</tbody>
</table>

Table 7: The 10 sequences with the highest values of confidence
Figure 1: Graph of the marginal associations
Figure 2: Oriented graph constructed on the base of the confidence indexes of association
Figure 3: A Bayesian network (a) and the corresponding Dependency network (b).
Figure 4: Format of the data required by WinMine
Figure 5: A Bayesian Network for the web data
Figure 6: A Dependency Network for the web data
Figure 7: A Decision tree for pinfo
List of the lately published Technical Reports
(available at the web site: "http://economia.unipv.it/Eco-Pol/quaderni.htm").

### Quaderni di Dipartimento

<table>
<thead>
<tr>
<th>#</th>
<th>Date</th>
<th>Author(s)</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>142</td>
<td>04-02</td>
<td>M.Balconi, S.Borghini, A.Moisello</td>
<td>Ivory Tower vs Spanning University: il caso dell’Università di Pavia</td>
</tr>
<tr>
<td>143</td>
<td>05-02</td>
<td>L.Bottolo, G.Consonni, P.Dellaportas, A.Lijoi</td>
<td>Bayesian analysis of extreme values by mixture modelling</td>
</tr>
<tr>
<td>144</td>
<td>06-02</td>
<td>I.Epifani, A.Lijoi, I.Pruenster</td>
<td>Exponential functionals and means of neutral to the right priors</td>
</tr>
<tr>
<td>145</td>
<td>06-02</td>
<td>P.Bertoletti</td>
<td>A note on third-degree price discrimination and output</td>
</tr>
<tr>
<td>146</td>
<td>12-02</td>
<td>P.Berti, L.Fratelli, P.Rigo</td>
<td>Limit theorems for predictive sequences of random variables</td>
</tr>
<tr>
<td>147</td>
<td>01-03</td>
<td>A. Lijoi, I. Pruender</td>
<td>Practicable alternatives to the Dirichlet process</td>
</tr>
<tr>
<td>148</td>
<td>02-03</td>
<td>A. Lijoi, Pruender, S.G.Walker</td>
<td>Extending Doob's consistency theorem to nonparametric densities</td>
</tr>
<tr>
<td>149</td>
<td>02-03</td>
<td>V.Leucari, G.Consonni</td>
<td>Compatible Priors for Causal Bayesian Networks</td>
</tr>
<tr>
<td>150</td>
<td>02-03</td>
<td>L. Di Scala, L. La Rocca, G.Consonni</td>
<td>A Bayesian Hierarchical Model for the Evaluation of a Web Site</td>
</tr>
<tr>
<td>151</td>
<td>02-03</td>
<td>G.Ascari</td>
<td>Staggered Prices and Trend Inflation: Some Nuisances</td>
</tr>
<tr>
<td>152</td>
<td>02-03</td>
<td>G.Ascari, P.Gagnepain</td>
<td>How inefficient are football clubs?</td>
</tr>
<tr>
<td>153</td>
<td>03-03</td>
<td>P.Dellaportas, C.Tarantola</td>
<td>An evaluation of the Spanish arms race combining factor levels</td>
</tr>
<tr>
<td>154</td>
<td>05-03</td>
<td>A. Lijoi, Pruender</td>
<td>A note on the problem of heaps</td>
</tr>
<tr>
<td>155</td>
<td>09-03</td>
<td>P.Berti, P.Rigo</td>
<td>Finitely additive uniform limit theorems</td>
</tr>
</tbody>
</table>