

Association and classification models for web mining

E. Blanc & P. Giudici

*Department of Economics and Quantitative Methods,
University of Pavia, Italy*

Abstract

We present methods for the search of association rules in the analysis of web clickstream data. We first show association rules, and show how to draw data mining conclusions on the basis of them. We then compare association rules, which are local models, with tree-based models, and show how association rules can be deduced from them. Our analysis have been conducted on a real e-commerce dataset.

1 The data

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.

The data set that we consider for the analysis is the result of the elaboration of a log file concerning a site of e-commerce. The source of the data cannot be specified; however it is the website of a company that sells hardware and software products; it will be referred to as "a webshop". The accesses to the website have been registered in a logfile for a period of about two years, since 30 september 1997 to 30 june 1999. The logfile has then been processed to produce

a dataset, named "sequences". Such dataset contains the user id (c_value), a variable with the date and the instant the visitor has linked to a specific page (c_time) and the web page seen (c_caller). Table 1 reports a small extract of the available dataset , corresponding to one visit.

Table 1: The considered dataset.

c_value	c_time	c_caller	c_order
70ee683a6df...	14OCT97:11:09:01	home	1
70ee683a6df...	14OCT97:11:09:08	catalog	2
70ee683a6df...	14OCT97:11:09:14	program	3
70ee683a6df...	14OCT97:11:09:23	product	4
70ee683a6df...	14OCT97:11:09:24	program	5

Table 1 describes that the visitor corresponding to the identifier (cookie) 70ee683a6df... 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.

The whole data set contains 250711 observations, each corresponding to a click, that describe the navigation paths of 22527 visitors among the 36 pages which compose the site of the webshop. The visitors are taken as unique, that is, no visitors appears with more than one session. On the other hand, we remark that a page can occur more than once in the same session.

This data set is a noticeable example of a transactional dataset. It can be used directly, in a form as in Table 1, with as many rows as the number of total clicks, to determine association and sequence rules. Alternatively, a derived dataset can be used, named "visitors". It is organised by sessions, and contains variables that can characterise each of such sessions. These variables include important quantitative ones, such as 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 preceeding day). More importantly for our analysis, such dataset contains binary variables that describe whether each page is visited at least once (modality 1) or not (modality 0). Finally, there is a bianry variable, named purchase, that indicates whether the session has led to (at least) one e-commerce transaction.

Table 2 shows part of a row from the visitors dataset, that corresponds to the session in Table 1.

Table 2: The derived dataset.

C_value	c_time	length	Clicks	time	home
0ee683a6df...	4OCT97:11:09:01	24	5	11:09:01	1

c_value	Catalog	addcart	program	product
70ee683a6df...	1	0	1	1

While the rows in Table 1 correspond to clicks, in table 2 the rows correspond to sessions (or, equivalently, visitors, as they are unique). There are as many rows as the total number of visits to the web site. In particular, looking at the last five columns, it is obtained a binary data matrix that expresses which pages, among the 36 considered, have been visited at least once in each session.

To give an idea of the type of considered pages, we now list some among the most frequent ones.

- 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 characterisation given in the login page;
- REGISTER: in order to be later recognized, the visitor has to prompt a userid 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;

940 *Data Mining III*

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.

2 Pre-processing

Our main aim is to discover the most frequent association patterns of visit between the 36 webpages that form the website.

In order to obtain valid conclusions, the considered data have to be homogeneous. In order to assess whether the dataset is homogeneous we have deemed necessary to run first an exploratory analysis on the “visitors” dataset, which can be desumed from the transactional dataset in Table 1 by building a data matrix with the visitors in the rows and the web pages binary variables in the columns. To such a reduced dataset we have appended additional variables, desumable from the logfile, such as the length of each visit (length), the time of the visit (time), the number of clicks (clicks) per visit and whether the visit has led to a purchase of an (e-commerce) product (purchase). Such additional variables have been used for exploratory (indeed pre-processing) purposes.

For instance, we have decided to proceed with outliers removal, and eliminate all observations above the 99-th percentile of the distribution of the variables clicks and length, which indeed behave similarly. On the other hand, we have decided not to remove observations possibly outlying with respect to the

time variable. This because of the nature of the variable itself as well as of the observed distribution.

The resulting visitors dataset contains 22152 observations, in place of the initial 22527.

In consideration of the heterogeneous nature of the navigators of the site, confirmed by the exploratory stage, we have then decided to perform a cluster analysis, in order to find homogeneous clusters of behaviours. We remark that our primary goal here is not cluster analysis per se and, thus, cluster analysis can be seen as somehow exploratory or, better, preliminary, to the association models that we are seeking.

As clustering variables we have considered the three quantitative variables start, length, clicks, as well as the binary variable purchase, which can all be seen as instrumental to our objective, of understanding navigation patterns. For the cluster analysis, we have decided to first run a hierarchical method, to find the optimal number of groups, and then a non-hierarchical method, to allocate observations in the determined number of groups. As distance function we have considered the Euclidean distance; as hierarchical method the method of Ward (after some comparative experiments). Finally, to allocate observations we have chosen the K-means non-hierarchical method.

The results obtained from cluster analysis confirm heterogeneity of behaviours, and allocate the visitors in four distinct clusters. To the purpose of finding navigation patterns, we have decided to concentrate the analysis on only one cluster. This choice is obviously subjective but, nevertheless, has two important peculiarities. First, the visitors in the chosen cluster stay connected for long time, and visit many pages. Both these occurrences allow to better explore the navigation sequences between the web pages. Second, the chosen cluster has a high probability of purchase; it seems important to consider the typical navigation pattern of a group of high purchasers.

Therefore, in the following sections, we shall consider a reduced dataset, corresponding to examined cluster of visitors, with 1240 visits and 21889 clicks.

Before starting with the modelling results we remark that we have appended, in the cluster dataset, a start_session and an end_session page, respectively, before the first page and the last page visited in each session. Obviously, such pages are fictitious, and are not random, but certain. However, they will serve to find the most common entrance and exit pages.

Finally, we point out to the reader that an analysis of the entire dataset described here, and not of one cluster alone, has been performed by Blanc and Giudici (2001).

3 Association rules

An association rule is a statement between two itemsets A and B, that can be written in the form $A \rightarrow B$. If the rule is ordered in time we have a sequence rule: in this case A preceeds B. For an introduction see, for example, Hastie, Tibshirani and Friedman (2001).

An association rule model is, essentially, an algorithm that searches for the most interesting rules in a database.

The indexes commonly used in Web Mining to evaluate the importance of an association rule are the indexes of support and confidence.

Consider the rule $A \rightarrow B$ 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:

$$\text{Support}\{A \rightarrow B\} = \frac{N_{A \rightarrow B}}{N}$$

Therefore, it 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:

$$\text{Support}\{A \rightarrow B\} = \Pr\{A \rightarrow B\}$$

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:

$$\text{Confidence}\{A \rightarrow B\} = \frac{N_{A \rightarrow B}}{N_A} = \frac{\frac{N_{A \rightarrow B}}{N}}{\frac{N_A}{N}} = \frac{\text{support}\{A \rightarrow B\}}{\text{support}\{A\}}$$

Therefore, the confidence approximates the conditional probability that in a server session in which has been seen the page A is also required page B.

What just said has been referred to itemsets A and B containing one page each; however, each itemset can contain more than one page, and the previous definition still hold.

The main limit of the indexes of support and confidence, 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 some reliable behaviour forecasts for new users.

In particular, we have implemented our analysis in the software IBM Intelligent Miner for data, which is based on the calculation of the support as main interestingness measure of a sequence rule. Furthermore, the software orders rules on the basis of their confidence as well as on the sign of the association (type).

The type of a rule indicates the validity of a rule based on the analysis of statistical dependency between the rule body and the rule head. The rule type is determined using the Chi-squared test for statistical independence and can be negative, neutral or positive. If the rule type is negative, the item set X has a negative influence on the occurrences of the item set Y; in this case the lift of the rule, that expresses the distance from the situation of statistical independence, takes values between 0 and 1. If the rule type is neutral, the lift of the rule takes is equal to 1 and the item sets X and Y are statistically independent. If the type is positive, the item set X has a positive influence on the occurrences of the item set Y; in this case the lift of the rule is greater than 1.

The obtained results, that we do not report here for lack of space, indicate that, among the rules having a support higher than 40% and a confidence higher than 25% there are only sixteen rules having a positive type. For e-commerce purposes it is of particular interest to examine the associations that contain either the Program or the Product page. Furthermore, the rules with highest support contain one or both of these two pages. In the sequel we thus concentrate our analysis on these two pages. Table 3 describes the association rules with the highest confidence for the Program page.

Table 3: Association rules with the highest confidence, for the Program page.

Type	Support (%)	Confidence (%)	Itemsets
+	53,06	94,54	Catalog → Program
+	74,51	83,92	Product → Program
.	49,52	79,74	Home → Program

From Table 3, Program results to be positively associated with only two pages: product (conf. 83,92%) and catalog (conf. 94,54%). On the other hand, Home and Program, which have a high confidence index, have a neutral type (namely, they are independent).

Consider now the strongest associations with the Product page, represented in Table 4.

Table 4: Association rules for product with the highest confidence.

Type	Support (%)	Confidence (%)	Itemsets
+	44,51	99,46	P_info → Product
+	48,95	98,38	Addcart → Product

+	43,63	98,36	Freeze → Product
+	74,51	95,95	Program → Product
+	53,38	95,11	Catalog → Product
.	54	87	Home → Product
-	46,04	86,52	Login → Product
-	41,2	85, 17	Logpost → Product
-	25	82,14	Shelf → Product

From Table 4, Product results to be strongly associated to program (conf. 95.95%), catalog (conf. 95,11%), addcart (conf. 98,38%), p_info (conf. 99,46%) and freeze (conf. 98,36%). We have also a high value of the confidence index with the variable home (conf. 87,01%), but in this case since the lift is very close to 1, on the basis of the Chi-square test of statistical independence, the rule body and the rule head seem to be independent. Notice also the existence of a negative association with the variables shelf (conf. 82,14%), login (conf. 86,52%) and logpost (conf. 85,17).

4. Classification trees

We now consider the application of classification trees (Breiman et al, 1984) to the dataset examined in the previous section. Classification trees require the specification of a target variable, to be predicted on the basis of all the others, that are considered as potentially explanatory. This because classification trees are, differently from association rules, predictive models. It is thus necessary to specify a response variable to be predicted.

Classification trees work on the derived dataset, as in Table 2. A classification tree can produce rules of the associative type: one rule, of a variable length, for each terminal node (leaf) of the tree. We remark that a tree usually produces rules on the basis of the majority rule in the terminal leaf, namely, the head of each rule corresponds to the class in the leaf with greatest probability.

Here we consider two types of predictions: the prediction of the most visited pages which, in our case, are the pages program and product. These predictions can be accomplished by setting, respectively, program and product as target variables.

We now present the optimal classification trees obtained with the software IBM Intelligent Miner applying an entropy-based CART method, for both target variables. In both cases we shall compare the rules obtained from the classification trees with the association rules discussed above.

We have obtained two classification trees, respectively taking, as target variables, the pages program and product. The binary decision tree consists of the root node on top, followed by non terminal nodes and terminal nodes (leaves). From the two decision tree we can derive the rules in tables 5 and 6, respectively for program and product; each rule has, as measure of

interestingness, its purity, namely, the percentage of records correctly classified in the corresponding leaf.

Table 5: Association rules for program, induced by the classification tree.

Purity (%)	Itemsets
94,5	Catalog=1 → Program=1
65,4	Catalog=0 & Product=1 → Program=1

Table 6: Association rules for product, induced by the classification tree.

Purity (%)	Itemsets
96	Program=1 → Product=1
92,1	Program=0 & Addcart=1 → Product=1
96,6	Program=0 & Addcart=0 & P_info=1 → Product=1
91,2	Program=0 & Addcart=0 & P_info=0 & Shelf=1 → Product=0
81,3	Program=0 & Addcart=0 & P_info=0 & Shelf=0 & Home=1 → Product=1

Comparing Tables 5 and 6 with tables 3 and 4 note that the rules extracted from the association rules and those extracted from the decision tree are rather similar. In fact, when program is the target variable, the pages most positively associated result to be catalog and product, as in Table 3. On the other hand, when product is the target variable, in both Tables 4 and 6 we discover a positive association with the pages program, addcart and p_info and a negative association with the page shelf. However, for the home page, the decision tree discovers a positive association while the association rules finds a situation of independence between the body rule and the head rule.

5 Conclusions

In this paper we have presented two methodologies for the analysis of web mining data. The first one is based on association rules by means of association rules one is able to understand local associations between visited pages. Association rules are relatively easy to extract and interpret. However, it is difficult to have a global picture of what is going on.

The second methodology, on the other hand, is based on classification trees, which are global in nature, and they thus allow an overall interpretation of the association rules that can be derived from them. On the other hand, the specification and interpretation of such models may be quite difficult.

We thus conclude that both methodologies should be considered in practical applications, and the choice accomplished on the basis of the objectives of the analysis.

We remark that other statistical methods can be applied to web clickstream data, in order to detect association rules. For instance, Blanc and Giudici (2002) consider, besides sequence rules, odds ratios and graphical loglinear models, while Blanc and Tarantola (2002) consider bayesian networks and dependency networks. Furthermore, in a recent paper, Di Scala and La Rocca (2002) also consider the application of Markov chain models to web data, with main emphasis on assessing homeogeneity of the considered Markov chain. Finally, Rognoni, Giudici e Polpettini (2002) consider using Markov chains to estimate directly the transition probabilities from one page to another.

References

- [1] Blanc, E. & Giudici, P., Statistical Models for web clickstream analysis. Technical Report, 2002.
- [2] Breiman, L., Friedman, J.H., Olshen, R. & Stone, C. J., *Classification and regression trees*. Wadsworth: Belmont, 1984.
- [3] Blanc, E. & Tarantola, C., Dependency Networks and Bayesian Networks for Web Mining. Technical Report, 2002.
- [4] Cabena, P., Hadjinian, P., Stadler, R., Verhees, J. & Zanasi, A., *Discovering Data Mining from Concept to Implementation*. Prentice-Hall: New York, 1997.
- [5] Di Scala, L. & La Rocca, L., A Markov Model for Web Data. Technical Report, 2002.
- [6] Giudici, P., *Metodi statistici per le applicazioni di Data Mining*, McGraw-Hill: Milano, 2001.
- [7] Han, J. & Kamber, M., *Data Mining: Concepts and Techniques*, Morgan Kaufmann, 2000.
- [8] Hand, D. J., Heikki, M. & Smyth, P., *Principles of Data Mining*, MIT Press, 2001.
- [9] Hastie, T., Tibshirani, R., & Friedman, J., *The elements of statistical learning: data mining, inference and prediction*, Springer-Verlag: Berlin, 2001.
- [10] Lauritzen, S., *Graphical Models*, Clarendon Press: Oxford, 1996.
- [11] Rognoni, M., Giudici, P., & Polpettini, P., Statistical models for the forecast of the visit sequences on web site, Technical Reports, 2002.
- [12] Whittaker, J., *Graphical Models in Applied Multivariate Statistics*, Wiley: Chichester, 1990.