STATISTICAL MODELS FOR WEB CLICKSTREAM ANALYSIS

Erika Blanc, Paolo Giudici

Dipartimento di Economia Politica e Metodi Quantitativi, Università di Pavia, Italia.

Abstract

The aim of this work is to show how the information, concerning the order in which the pages of a web site are visited, can be profitably used to foresee the visit behaviour at the site itself. The paper is divided in two sections. In the first section, after describing the type of Web data used in the application, we present, under a statistical point of view, the indexes used in the analysis of the sequences and we illustrate the results obtained by the application of such indexes to our data. In the second section, we suggest using, for the study of the association structure between the Web pages, two models used for the analysis of qualitative data.

1. INTRODUCTION

The aim of this paper is to show how the information, coming from a Web clickstream analysis, can be profitably used to predict the visit behaviour at the site.

Our objective is to show how Web clickstream 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. Such analysis can be very useful to understand, for instance, what is the conditional probability of seeing a page of interest.

The scenario described above is an application in the area of Web Usage mining. Web Usage mining, which is one of the three classes of Web mining, applies data mining techniques in order to discover usage patterns from Web data (J. Srivastava et al., 2000). For more details about the taxonomy of Web mining see, for example, R. Cooley et al., 1997.

2. THE ANALYSIS OF THE TRAFFIC ON THE WEB SITE

Every time a 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 a 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 or, it is more commonly said that they identify a visit (J. Srivastava et al., 2000).

3. DESCRIPTION OF THE DATA

The data set (table 1) is the result of the elaboration of a log file concerning a site of e-commerce. It contains the user id (c_value), a variable with the date and the instant the visitor is linked to a page (c_time) and the Web page seen (c_caller).

Tab. 1: Example of the data set.

c_value	*c_time*	*c_caller*
70ee683a6df...	14OCT97:11:09:01	home
70ee683a6df...	14OCT97:11:09:08	catalog
70ee683a6df...	14OCT97:11:09:14	program
70ee683a6df...	14OCT97:11:09:23	product
70ee683a6df...	14OCT97:11:09:24	addcart

In particular, the data set contains 250711 observations that describe the navigation patterns of 22527 visitors inside the 36 pages, which compose a site of e-commerce. This data set has been used as data matrix in the analysis of the sequences (section 4). Besides we point out that it can serve to calculate a series of derived variables as, for example, the total length of the server session, the total number of the clicks made by the visitor in a session, the inter time which exists between a click and another. For the purposes of Sections 5, 6 and 7 we have taken into consideration, as derived variables, only the columns of the data set concerning the pages that constitute the site (table 2). They are binary variables with values 1 or 0, which show if the visitor has at least shown once that page, or not. Only the 19 pages more frequently seen in terms of marginal visit frequency have been considered.

Tab. 2: Example of the derived data set.

c_value	*c_time*	*length*	*clicks*	*week_day*	*addcart*	*program*	*product*
70ee683a6df...	14OCT97:11:09:01	25	5	friday	0	1	1
705c692142e...	20OCT97:15:22:07	47	8	thursday	1	0	1

4. THE INDEXES OF SUPPORT AND CONFIDENCE IN SAS: THEORY AND EXAMPLES

The indexes commonly used in the Web Usage mining are the indexes of support and confidence.

Consider the not direct sequence A → B and indicate as $N_{A \to B}$ the number of visits, which appear in such sequence, at least once. A and B identify two generic pages that compose a Web site. The left side term of the rule (A) is commonly referred as the antecedent, the body of the rule (P. Cabena et al., 1997) or the condition (M. J. A. Berry and G. Linoff, 1997) while the right side term (B) is named consequent, head of the rule or result. Let N be the total number of the server sessions. Notice that the rule A→B will be counted only once even if it had been repeated several times inside the session. The support for the rule A →B is obtained dividing the number of server sessions, which satisfy the rule by the total number of server sessions:

$$support\{A \to B\} = \frac{N_{A \to B}}{N} \tag{1}$$

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 in sequence:

$$support\{A \to B\} = Pr\{A \to B\} \tag{2}$$

The confidence for the rule A→ B instead is obtained dividing the number of server sessions which satisfy the rule by the number of sessions containing the page A:

$$confidence\{A \to B\} = \frac{N_{A \to B}}{N_A} = \frac{\frac{N_{A \to B}}{N}}{\frac{N_A}{N}} = \frac{\text{support}\{A \to B\}}{\text{support}\{A\}} \tag{3}$$

Therefore, the confidence expresses the probability that in a server session in which has been seen the page A is subsequently required page B. We now examine the most general case and of greater interest in which the aim is to identify some "navigation chains" constituted from a greater number of pages than two.

Consider the sequence $\{A \rightarrow B \rightarrow C \rightarrow D\}$; remember that we are examining indirect sequences therefore after having visualize page A, for example, other pages may have been required before visualizing page B. Besides notice that, in this case, the rule is logically interpreted as "If $A \rightarrow B \rightarrow C$, then D".

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:

$$support\ \{A \rightarrow B \rightarrow C \rightarrow D\} = \frac{N_{A \rightarrow B \rightarrow C \rightarrow D}}{N} \tag{4}$$

The confidence for the rule $\{A \rightarrow B \rightarrow C \rightarrow D\}$ is instead obtained as the ratio between the number of visits, which satisfy the rule and the number of visits that verify only the "body of the rule", namely:

$$confidence\ \{A \rightarrow B \rightarrow C \rightarrow D\} = \frac{N_{A \rightarrow B \rightarrow C \rightarrow D}}{N_{A \rightarrow B \rightarrow C}} \tag{5}$$

The analysis of the sequences produces as results the sequences of pages with the highest indexes of support. Besides, the indexes of confidence allow to evaluate, subordinately to a certain body of departure, what are the most likely transitions from one page to another.

The principal limit of such indexes, 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.

We have elaborated the support and confidence indexes for the data set presented with SAS Enterprise Miner. For interpretative simplicity, we have considered only those pages, which appear in the server sessions a number of times at least equal to the 5% of the frequency of the page most visited. Therefore, the examined pages have been reduced to 19. The results obtained with Enterprise Miner for indirect sequences of two pages are shown in table 3; notice that the sequences have been ordered on basis of their support.

Tab. 3: The most frequent indirect sequences with two pages.

#	*Support (%)*	*Confidence (%)*	*Transaction Count*	*Rule*
1	68.88	89.57	15517	program → product
2	55.38	65.34	12476	product → product
3	48.19	56.85	10856	product → p_info
4	39.56	88.08	8912	catalog → program
5	38.54	50.11	8681	program → p_info

5. THE INDEXES OF SUPPORT AND CONFIDENCE FOR ASSOCIATIONS

The literature on Data Mining suggests using the support and confidence indexes also for the analysis of the associations among the pages. However, in this case, we use the odds ratio. Remember that in a two-way contingency table containing the joint distribution of A and B the odds ratio equals the ratio of the products $\pi_{11}\pi_{22}$ and $\pi_{12}\pi_{21}$ of probabilities from diagonally opposite cells:

$$\theta = \pi_{11}\pi_{22} / \pi_{12}\pi_{21} \qquad (6)$$

We have computed odds ratio for all the couples between the 19 pages considered. In general, a line should be traced between two pages when the confidence interval of the odds ratio is higher than 1. In our analysis this situation happened for 114 combinations of pages so we decided, in order to better understand the association structure among the Web pages, to increase this threshold to 4. In this way we obtained the association structure represented in figure 1, in which the associations having an odds ratio superior than 8 are highlighted with a bold line.

Now, we consider opportune to illustrate the calculation of the index of confidence for associations, considering both modalities of the variables and with reference to a two-way contingency table. To such purpose, consider a two-way contingency table concerning the variables freeze and pay_req:

Tab. 4: Two-dimensional contingency table for freeze and pay_req.

FREEZE	PAY_REQ		
	1	0	Tot.
1	3814	1851	5665
0	17	16845	16862
Tot.	3831	18696	22527

The index of confidence for the association between freeze and pay_req is given by:

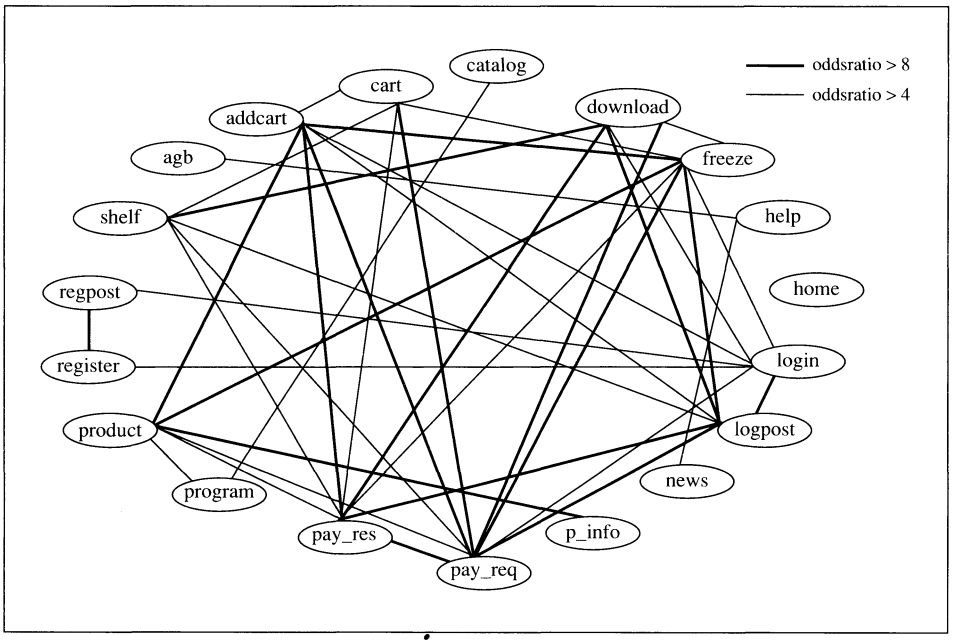

Fig. 1: Couples of pages having an odds ratio higher than 4.

$$confidence(freeze - pay\_req) = \frac{3814}{5665} = 0,6733 \tag{7}$$

Reversing the terms of the rule the support remains unchanged, while the confidence changes and is given by:

$$confidence(pay\_req - freeze) = \frac{3814}{3831} = 0,9956 \tag{8}$$

Since the 99,56% of the server session that contain page pay_req contain also page freeze it is logical to think that page freeze is usually requested before accessing the page pay_req. Therefore, in figure 1 an arrow has been placed from freeze to pay_req. A similar reasoning has been then extended to other couples of pages, which have an odds ratio superior than 8 (figure 1). So we have obtained the graph in figure 2 that should be interpreted in the following way: an arrow which, starting from a page points toward other pages highlights what are the most frequent descendent for that page; similarly, the arrows which point to a page, coming from other pages, show what the most frequent antecedents are for that page.

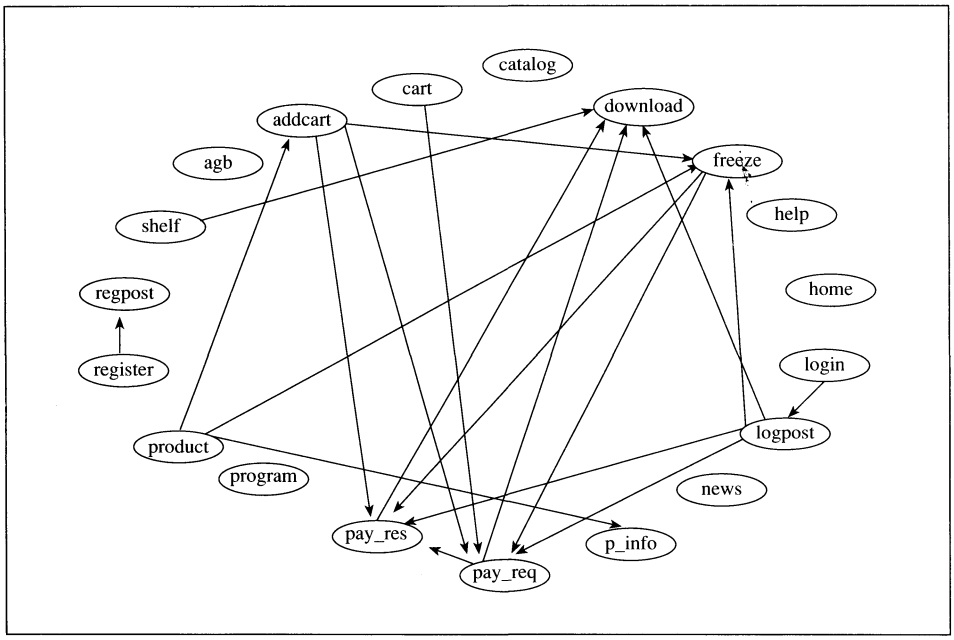

Fig. 2: Graph oriented on the base of the confidence indexes for associations.

6. BUILDING AN UNDIRECTED GRAPHICAL MODEL

The previous conclusions are prevalently of explorative kind. The odds ratio, for which we have provided a confidence interval, have been marginally obtained for every couple of pages, without considering the many interdependences among the 19 considered pages. Therefore, we have decided to build a more general statistical model for the associations.

At first we consider a log-linear model that is an interdependence model whose aim is to study the association structure, namely the interaction among the 19 Web pages more frequented of the site. In order to obtain a simple model, only the interaction terms containing at most two factors have been taken into consideration; therefore, those of superior order have been set to zero. Following this hypothesis, the log-linear model turning out can be made equivalent to a graphical log-linear model. Therefore, the nullity of an interaction term between two factors, which correspond to conditional independence among the corresponding couples of variables, can be translated in the absence of an arc between the corresponding nodes of a graph of the conditional independences.

Besides, the model has been built keeping present that SAS software used for the analysis attributes a positive value to the interaction parameter in case of positive association between the analysed variables; the interaction parameter is instead assigned a negative value in case of negative association between the variables; while the null value of the interaction parameter indicates a statistical independence situation between the corresponding variables. The model represented in figure 3 contains only arcs related to positive association terms.

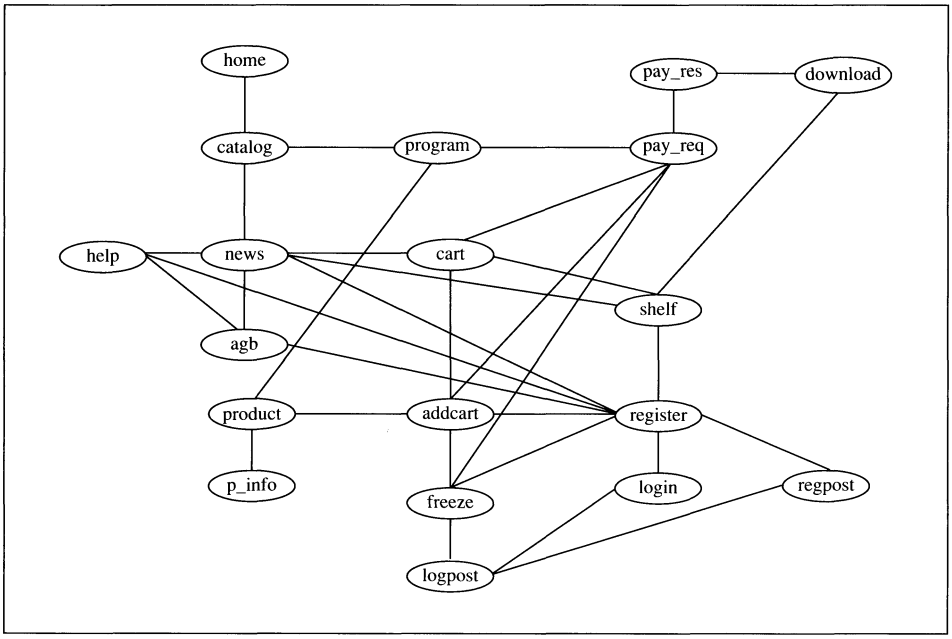

Fig. 3: The graphical log-linear model of the only positive associations.

7. BUILDING A DIRECTED GRAPHICAL MODEL

At last we present a directed graphical model that differently from the previous graph allow us to represent the knowledge of causality between the variables. In fact, in a directed graphical model the link between the variables is not symmetric, as in precedence, but asymmetrical. The graphical model built by us takes its starting point from these models. However, we signal that assuming a causal not reversible sequence among the binary variables of visit to the single pages is not fully a realistic hypothesis. In fact, since the "feedbacks" are rather

frequent and the input data set does not contain information about the visit order, it is not evidently possible to assume that the causality relations are determined by the time visit order.

The directed graphical model represented in figure 4 has been obtained on the basis of 19 models of logistic regression, one model of logistic regression for each of the 19 pages most frequented, which were acting from time to time as response variables. In each of them, the 18 remaining binary variables have been all used as explanatory variables.

In particular, the directed graphical model has been built supposing that, if a significantly positive odds ratio had occurred, this would have corresponded to the existence of a causal link from the explanatory variable at issue to the target variable. Graphically, we have represented the single causal links, so identified, by an one-way arrow which leaves from the explanatory variable and arrives to the target variable.

This model determines the pages, which condition more the visit to a determinate Web page; naturally, since considering the visit order is not possible, these pages can precede or follow the visit to the target page.

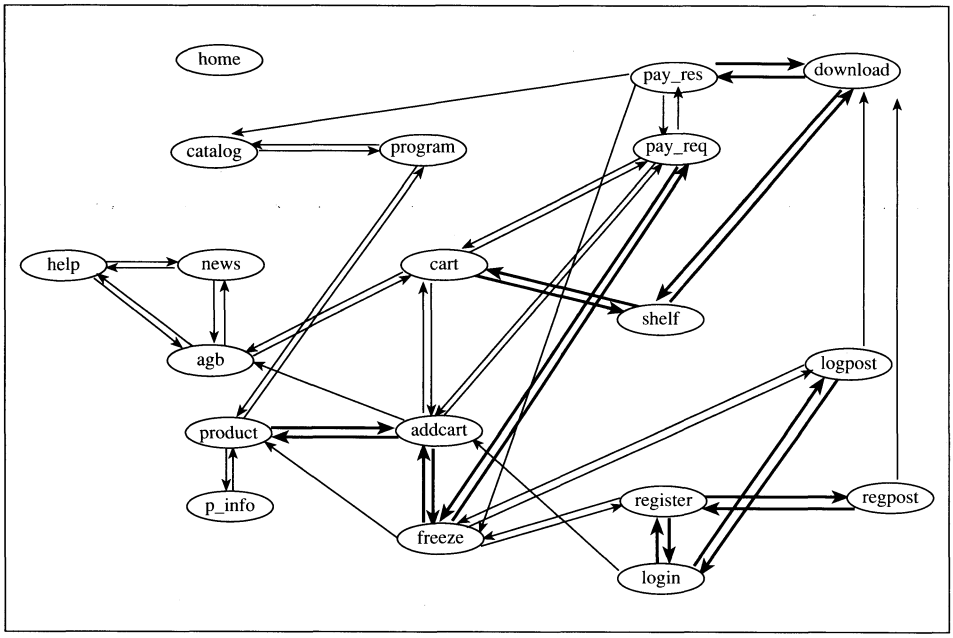

Fig. 4: Oriented graphical model of the only positive associations.

8. CONCLUDING REMARKS

We have considered a collection of statistical methods to model Web mining data. It is quite difficult to choose between them. Here the situation is complicated by the fact that we have to compare local models (such as sequence rules) with global models (such as graphical models).

For global models, statistical evaluation can proceed in terms of classical scoring methods, such as likelihood ratio scoring, AIC or BIC. Or, alternatively, by means of computationally intensive predictive evaluation, based on cross-validation and/or bootstrapping. But the real problem is how to compare them with sequence rules.

A simple and natural scoring function of a sequence rule is its support that gives the proportion of the population to which the rule applies. Another measure of interestingness of a rule is the lift of the rule itself. The lift is the ratio between the confidence of the rule A→ B and the support of the head of the rule (P. Cabena et al., 1997). Recalling the definition of the confidence index, the lift compares the observed absolute frequency of the rule with that corresponding to independence between A and B.

We finally remark that other statistical methods can be applied to Web clickstream data, in order to detect association rules. For instance, Blanc and Tarantola (2002) consider Bayesian networks and dependency networks. Finally, 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 homeogenity of the considered Markov chain.

We finally acknowledge support from SAS Institute, Milan, for having suported the data as well as the software SAS Enterprise Miner on trial.

REFERENCES

AGRESTI A. (1990), *Categorical Data Analysis,* Wiley, New York.

BERRY M. J. A., G. S. LINOFF (1997), *Data Mining Techniques: For Marketing, Sales and Customer Support,* Wiley, New York.

BLANC E., GIUDICI P., (2002), *Sequence Rules for Web Clickstream Analysis,* In: Advances in Data Mining: applications in E-commerce, medicine and knowledge management,Petra Perner (ed.),.Springer-Verlag, Berlin Heidelberg, pages 1-14.

BLANC E., TARANTOLA C. (2002), *Dependency Networks and Bayesian Networks for Web Mining*, Technical Report, Submitted.

CABENA P., HADJINIAN P., STADLER R., VERHEES J. and ZANASI A. (1997), *Discovering Data Mining from Concept to Implementation,* Prentice-Hall, New York.

COOLEY R., MOBASHER B., SRIVASTAVA J. (1997), "Web mining: Information and pattern discovery on the World Wide Web", International Conference on Tools with Artificial Intelligence, *Newport Beach,* pages 558-567.

DI SCALA L., LA ROCCA L., (2002), A Markov Model for Web Data, *Technical Report,* Submitted.

EDWARDS D., *Introduction to Graphical Modelling,* Springer-Verlag, New York, 1995.

GIUDICI P. (2001), *Metodi statistici per le applicazioni di Data Mining,* McGraw-Hill Libri Italia, Milano.

HAN J., KAMBER M., (2001), *Data Mining: Concepts and Techniques,* Morgan Kaufmann, San Francisco.

KOSALA R., BLOCKELL H., (2000), Web Mining Research: A Survey, *SIGKDD Explorations,* vol. II, Issue 1, pages 1-11.

MCCULLAGH P., NELDER J. A. (1989), *Generalized Linear Models,* 2nd ed., Chapman & Hall, London.

SRIVASTAVA J., COOLEY R., DESHPANDE M., TAN P., (2000), "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data", *SIGKDD Explorations,* V. I, Issue 2, pages 12-23.

WHITTAKER J., (1990), *Graphical Models in Applied Multivariate Statistics*, Wiley, Chichester.

MODELLI STATISTICI PER LA WEB CLICKSTREAM ANALYSIS

Riassunto

L'obiettivo di questo lavoro è mostrare come le informazioni, relative all'ordine con cui le pagine di un sito Web sono visitate, possano essere utilizzate proficuamente per prevedere il comportamento di visita al sito stesso. Il lavoro è diviso in due parti. Nella prima parte, dopo aver decritto il tipo di dati Web utilizzati nell'applicazione, presentiamo, da un punto di vista statistico, gli indici utilizzati nell'analisi delle sequenze e illustriamo i risultati ottenuti dall'applicazione di tali indici ai nostri dati. Nella seconda parte, proponiamo di utilizzare, per lo studio della struttura di associazione tra pagine Web, due modelli usati per l'analisi di dati qualitativi.