Web: www.witpress.com Email witpress@witpress.com
Paper from: *Data Mining III*, A Zanasi, CA Brebbia, NFF Ebecken & P Melli (Editors).
ISBN 1-85312-925-9

Dependency Networks and Bayesian Networks for Web mining

C. Tarantola & E. Blanc
Department of Economics and Quantitative Methods,
University of Pavia, Italy.

Abstract

Following the approach described by Heckerman et al. ([5]), we present an application of Dependency Networks and Bayesian Networks to the analysis of a clickstream data set. Our target is to discover which paths are more often followed by the users. The relation between one web page and another one is represent by a direct graph. Whereas Bayesian Networks use direct acyclic graphs, Dependency Networks may contain cyclic structures. The analysis will be performed with the WinMine Toolkit software.

1 Introduction

Web mining is a methodology that applies data mining techniques to discover usage patterns from Web data in order to optimally design a web site and better satisfies needs of different visitors.

The aim of this work is to use different graphical models: Bayesian Networks (BN hereafter) and Dependency Networks (DN hereafter), to analyse click stream data. A click stream is a sequential series of page view requests. The click stream of page views for a single user across the entire Web is a user session. Typically, only the portion of each user session that is accessing a specific site can be used for analysis (for more details see e.g. J. Srivastava et al. [8]).

The methodology presented will be applied to the analysis of a real set of data regarding an e-commerce web site. All computations will be performed with the WinMine Toolkit software.

WinMine is a free software developed by the Machine Learning and Applied Statistics group of Microsoft Research that permits to learn BN and DN from data. From our experience we think that this software is quite user friendly, even

Paper from: *Data Mining III*, A Zanasi, CA Brebbia, NFF Ebecken & P Melli (Editors).
ISBN 1-85312-925-9

if some features should be improved. It can be downloaded from the web site $http://research.microsoft.com/\ dmax/WinMine/ContactInfo.html$.

The plan of the work is the following: in section 2 the structure of the data is described; in section 3 we introduce briefly DN and BN; finally, in section for 4 we present the results of our application to web data.

2 Description of the data

We consider a transactional data set regarding the navigation paths of 22527 visitors to a site of e-commerce. This site consists of 36 pages; for the analysis we take into account only the 26 pages that have been visited more frequently.

Unfortunately, for privacy reason it was not available to us the full description of the meaning of the web pages; this affected negatively the quality of the explanation of our result. For an analysis of the same data set from a different perspective see E. Blanc and P. Giudici [1] and L. Di Scala and L. La Rocca [4].

In Blanc and P. Giudici [1], an attempt is done to explain the meaning of the variables, in the following we will use their description. The main important variables are:

HOME: the home page of the web site;
LOGIN: in order to enter in this particular web site the user must have been previously registered (page register) with a user name and a password that will be used for future sections.
LOGPOST: prompts a message that informs whether the login has been successful or if it has failed;
REGISTER: in order to be later recognized, the visitor has to insert a user-id and password;
REGPOST: shows the partial results of the registration, asking for missing information if there are any;
SHELF: it contains the list of the programs that can be downloaded from the web site.
PROGRAM: gives detailed information on the characteristics of the software programs available;
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: more detailed information about a particular product are provided ;
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;
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);

Paper from: *Data Mining III*, A Zanasi, CA Brebbia, NFF Ebecken & P Melli (Editors).
ISBN 1-85312-925-9

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

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 1)

We have reclassified the numbers of time each web page has been visited according to 4 levels. In particular we have assigned value 1 if a page has been visited one time, 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.

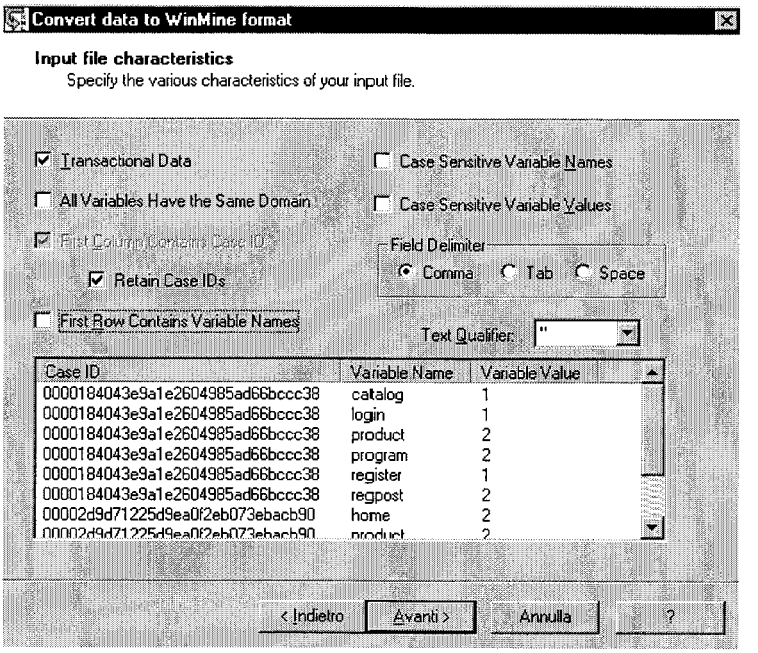

Figure 1: Format of the data required by WinMine

3 Dependency Networks and Bayesian Networks

In this paper we present two different typologies of graphical models that can be used to represent the relationship between a set of discrete variables: Bayesian Networks and Dependency Networks. We will not review here the general theory of graphical models, for an introduction see Lauritzen [6].

BN are well known and have been widely used in different fields, see e.g. Jensen [7]. On the other hand DN are not so popular and have been proposed only recently

Paper from: *Data Mining III*, A Zanasi, CA Brebbia, NFF Ebecken & P Melli (Editors).
ISBN 1-85312-925-9

by Heckerman et al [5] as an alternative to BN. In both cases we use a graph to represent the probabilistic relationship between a set of data.

More precisely both models are characterised by:

i) a directed graph,
ii) a set of local probabilities; we consider a set of conditional distributions, one for each node x_i in the graph given its parents, $p(x_i|pa_i)$, where pa_i are the parents of x_i.

The main differences between these two representations are:

i) in a BN we work with acyclic graphs, whereas in the DP cyclic structures are permitted and multi-directional arrows are present in the graph.
ii) in the BN 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 the case with DN.

Furthermore, as pointed out in Heckerman et al [5], DN are often *inconsistent*, that is the local distributions cannot be obtained via the rule of probability from a common joint distribution $p(x)$. The advantages deriving from the use of inconsistent dependency networks are mainly computational. In section 4 the two methodologies will be compared with reference to the analysis of web data.

Following Heckerman et al [5] we now present with an example which are the advantages that can derive from the use of DN.

DN are easier to be understand by individuals not familiar with the graphical model semantic.

Consider the graphs in Figure 2, they 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 BN, whereas graph (b) is a DN.

Figure 2: A Bayesian network (a) and the corresponding Dependency network (b).

As we have already said, the first difference is that in the DN there is a cyclic structure and there are arrows with multiple directions.

Both BN and DN encode a set of conditional independence, but for untrained individuals the second ones result easier to understand.

Consider for the example the graph represented in Figure 2.a. When shown a graph like this one and told to the individual that it represents causal relationships, an untrained person often gains an accurate impression of the relationships.

The problem arises when the person can be only told that the relationships are "predictive" or "correlated". In this case, in fact an untrained user will correctly

Paper from: *Data Mining III*, A Zanasi, CA Brebbia, NFF Ebecken & P Melli (Editors).
ISBN 1-85312-925-9

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 them to understand that A and B are dependent given C.

The solution proposed by Heckerman et al [5] is to substitute a BN structure with one where the parents of each variable render that variable independent of all the other ones (see Figure 2.b).

Note that consistent DN shares the Markov properties of a Markov network with the same set of adjacencies. This is no more true when the network is inconsistent.

4 Analysis of web data

In this section we present the results obtained by applying BN and DN to the analysis of web data using WinMine toolkit.

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 setting this is done automatically by the program, for more details see Heckerman et al [5] and Chickering et al. [3].

Figure 3 and 4 show a BN and a DN structures learned from the data.

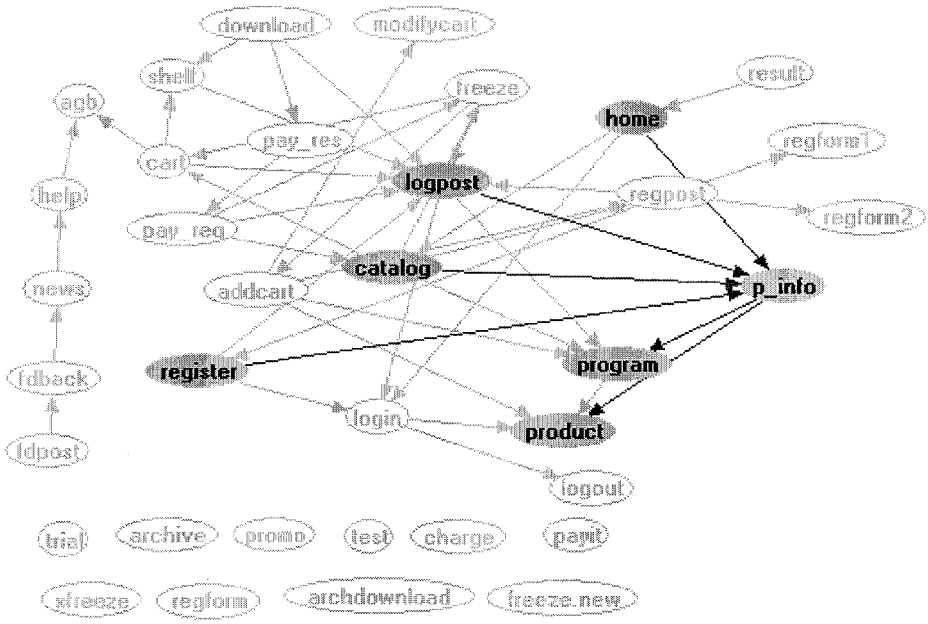

Figure 3: A Bayesian Network for the web data

Paper from: *Data Mining III*, A Zanasi, CA Brebbia, NFF Ebecken & P Melli (Editors).
ISBN 1-85312-925-9

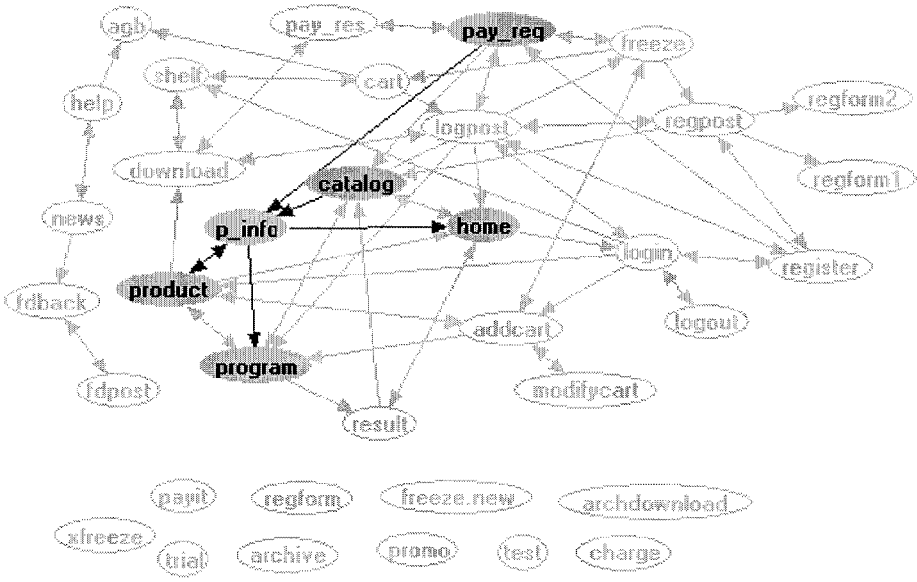

Figure 4: A Dependency Network for the web data

After only a short inspection, we note that both BN and DN determine the same path from *fdpost* to *cart*. It is interesting to note that BN and DN may attribute different roles to the same variable; for example in the BN *home* is predictive of *pinfo* whereas in the DN the role of these variables is reversed. More investigation is needed.

When learning a BN and a DN from data the local distributions are estimated by means of probabilistic decision trees (see Buntine [2]).

Winmine allows us to display the tree associated with each variable. To view a decision tree for a variable, it is only necessary to double clicks on the corresponding node in the dependency network. Figure 5 represents the decision tree for the variable *pinfo* in the dependency network.

Note that there is a split on variable X in the decision 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 *pinfo* of assuming values $1, 2, 3, 4$ or *not visited*.

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 [5].

Paper from: *Data Mining III*, A Zanasi, CA Brebbia, NFF Ebecken & P Melli (Editors).
ISBN 1-85312-925-9

Note that DN are sligthtly less accurate then BN; in our case the score for a DN is -0.374718 whereas that of the BN is equal to -0.373111.

This difference is not surprising since the number of parameters in the BN are fewer than the number of parameters in the corresponding DN.

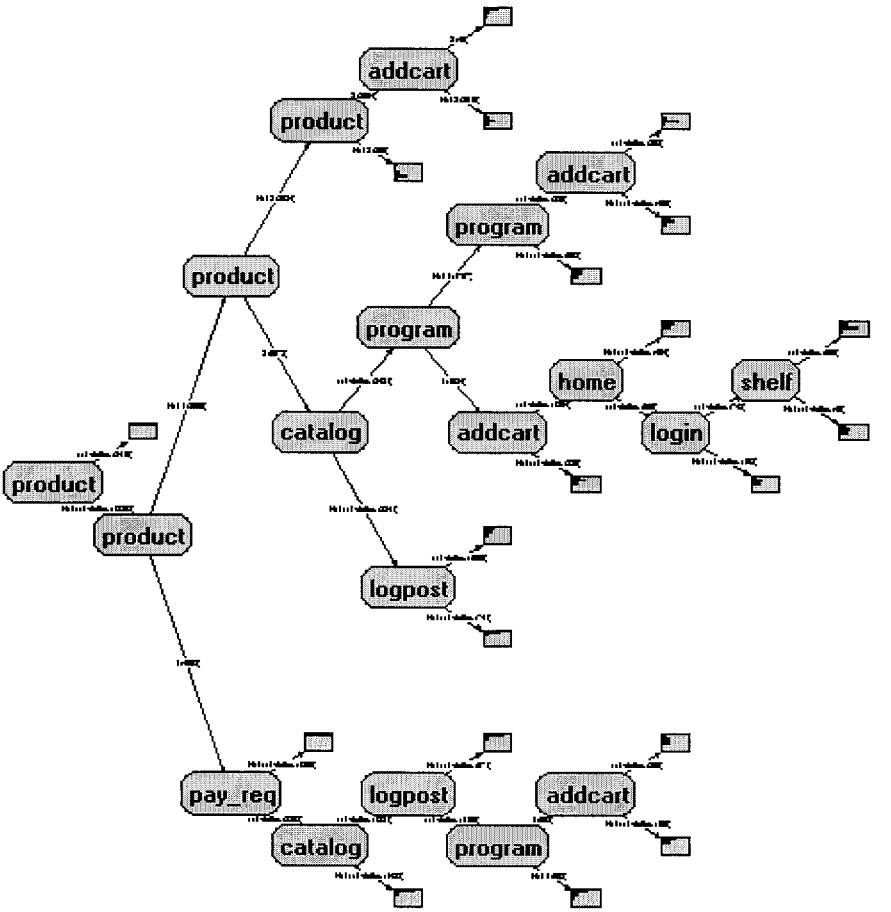

Figure 5: A Decision tree for pinfo

Paper from: *Data Mining III*, A Zanasi, CA Brebbia, NFF Ebecken & P Melli (Editors).
ISBN 1-85312-925-9

Acknowledgement

The first author acknowledge support from "Progetto Giovani Ricercatori" year 2001, University of Pavia. The second author acknowledge support from the project "Statistical methods for data mining" university of Pavia.

References

[1] Blanc, E & Giudici, P, Statistical models for Web Clickstream analysis, submitted, 2002.
[2] Buntine, W.,Theory refinement on Bayesian netwroks. *Prooceedings of Fourtheen Conference on Uncertainty in Artificial Intelligence*,Morgan Kaufman, pp. 52–60, 1991.
[3] Chickering, D.M, Heckerman, D.,Meek, C., A Bayesian approach to learning Bayesian networks with local structure. *Prooceedings of Thirtheen Conference on Uncertainty in Artificial Intelligence*, Morgan Kaufman ,pp 80–89, 1997.
[4] Di Scala, L., La Rocca, L., A Markov Model for Web Data, submitted, 2002.
[5] Heckerman, D., Chickering, D.M, Meek, C, Rounthwaite, R & Kadie, C., Dependency Networks for Inference, Collaborative Filtering and Data Visualization. *Journal of Machine Learning Research*, **1**, pp. 49–75, 2000.
[6] Lauritzen, *Graphical models*. Claredon Press, Oxfrod 1996.
[7] Jensen, F. V, *An introduction to Bayesian Networks*. UCL Press, London 1996.
[8] Srivastava, J.,Cooley, R.,Deshpande, M & Tan, P. Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data. *SIGKDD Explorations*, **V. I., 2**, pp. 12-23.