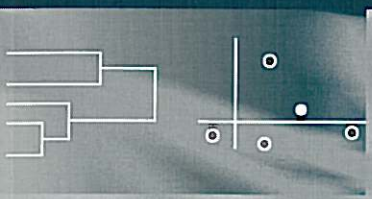
STUDIES IN CLASSIFICATION, DATA ANALYSIS, AND KNOWLEDGE ORGANIZATION

New Developments in Classification and Data Analysis

M. Vichi
P. Monari
S. Mignani
A. Montanari
Editors



Springer

A Different Approach for the Analysis of Web Access Logs

Gabriella Schoier[1] and Giuseppe Melfi[2]

[1] Dipartimento di Scienze Economiche e Statistiche,
Università di Trieste, Italy
gabriella.schoier@econ.units.it
[2] Groupe de Statistique
Université de Neuchâtel
giuseppe.melfi@unine.ch

Abstract. The development of Internet-based business has pointed out the importance of the personalisation and optimisation of Web sites. For this purpose the study of users behaviours are of great importance. In this paper we present a solution to the problem of identification of dense clusters in the analysis of Web Access Logs. We consider a modification of an algorithm recently proposed in social network analysis. This approach is illustrated by analysing a log-file of a web portal.

1 Introduction

The analysis of usage behaviour on the Web has acquired an always greater importance in the development of Web strategies, mainly in the field of e-commerce. Web personalisation can be defined as any action whose aim is to optimise the profitability of a site both from the owner and from the user viewpoint. Personalisation based on Web Usage Mining or Web Log Mining has several advantages comparing with more traditional techniques: the type of input is not a subjective description of the users, since log-files contain detailed information about the usage of a Web site. Web Usage Mining has been developed in order to extract interesting patterns in Web access logs (Srivastava et. al. (2000); Mobasher et al. (2002)).

Many statistical packages nowadays contain useful tools for handling web data. In this paper we illustrate a new approach. The data consist of a set of units (the I.P. addresses) on which one relational variable is measured. This forms a network, i.e., a set of units and relation(s) defined over it (see Wasserman et al. (1994)). The huge increase in the amount of data available on Internet has outstripped our capacity to meaningfully analyse such networks and run into significant computational barriers in large networks.

Some help in the analysis may derive by two classical social network theories. First the small-world literature has shown that there is a high degree of

local clustering in the networks (see e.g. Kochenet al. (1989)); this suggests that an approach for studying the structure of large networks would involve first the identification of local clusters and then the analysis of the relations within and between clusters. Second, literature on peer influence shows that, based on an endogenous influence process, close units tend to converge on similar attitudes (see e.g. Friedkin (1998)) and thus clusters in a small-word network should be similar along multiple dimensions.

In this paper we present a solution to the problem of identification of dense clusters in the analysis of Web access logs, by considering a modification of an algorithm proposed by Moody (2001) in the field of social network analysis. The principal advantage regards the possibility of handling a huge amount of data in a short time; in fact, in order to build up the groups only influence variables, which represent the position of the units, are considered. This results in a reduced and more flexible structure on which different techniques such as blockmodeling (see Schoier (2002)) may be used. The advantage of blockmodeling is that a differentiated structure for the degree of similarity within and between cluster is allowed.

2 Identifying dense clusters in large networks

We illustrate our approach on the basis of the log-files of the Web site www.girotondo.com, a non dynamical portal for children comprising seven different sections containing 362 jhtml pages are. The period of observation is from the 29th November 2000 to the 18th January 2001.

...
212.75.0.22–[08/Jan/2001:10:14:39+0100]GET / HTTP/1.1
212.75.0.22–[08/Jan/2001:10:14:41+0100]GET /picts/index_03.gif HTTP/1.1 304
212.75.0.22–[08/Jan/2001:10:14:41+0100]GET /picts/index_02.gif HTTP/1.1 304
212.75.0.22–[08/Jan/2001:10:14:41+0100]GET /picts/index_01.gif HTTP/1.1 304
212.75.0.22–[08/Jan/2001:10:14:41+0100]GET /picts/index_04.gif HTTP/1.1 304
213.136.136.60–[08/Jan/2001:22:50:01+0100] GET /favolando/11_00/picts/09.swf HTTP/1.0
209.55.1.99–[08/Jan/2001:22:50:01+0100] GET /picts/index_31.gif HTTP/1.0
...

Table 1. Example of a log-file

The original file contained 300'000 records, each of which corresponds to a line in the log-file. Records of log-files containing auxiliary information (e.g. .gif, .jpeg. files) are ignored. This first cleaning produces a file as in Table 2.

We then proceed with a re-codification of the Web pages by associating, according to their alphabetical order, the URLs with numbers 1 to 362 for easier handling.

Then we consider only the pages which have been visited by at least 5 I.P. addresses. This reduces our analysis over 117 pages. Furthermore we consider two fields in the original log-file: the I.P. address and the code corresponding

```
...
130.93.25.19 20/DEC/2000:10:19:44+0100 "GET/mappa/01.jhtml HTTP/1.0"          2472
235.58.54.78 20/DEC/2000:10:19:41+0100 "GET/news/archivio.jhtml HTTP/1.0"     115
267.12.83.56 20/DEC/2000:10:19:40+0100 "GET/news/01/01/01.jhtml HTTP/1.0"     793
241.27.83.61 20/DEC/2000:10:19:37+0100 "GET/favolando/01.jhtml HTTP/1.0"      88
...
```

Table 2. The first cleaning of a log-file

```
...
/links/12_00/02.jhtml  257
/links/12_00/022.jhtml 258
/mappa/01.jhtml        259
/news/01.jhtml         260
/news/02.jhtml         261
/news/archivio.jhtml   262
/news/form.jhtml       263
...
```

Table 3. Part of the index file

to the visited page. Each I.P. address corresponds to at least one viewed page. After a further preprocessing phase, a file of 1000 I.P. addresses with the relative set of viewed pages has been considered.

This allows to build up a matrix X of dimension 1000×117 where the lines represent the I.P. addresses and each column represent one of the 117 pages selected for the aim. The coefficient is 1 if the page has been viewed by the I.P. address, 0 otherwise. According to the social network theory (see Wasserman, (1994)) this corresponds to a 2-mode matrix. The next step is to produce a 1-mode matrix on the basis of a suitable relation between I.P. addresses.

We consider the following relation: the I.P. addresses v_i and v_j are in relation if they have visited at least 35 pages in common. The number defining the relation (in our case 35) must be a number that discriminates nodes that have a relation from nodes that have not a relation. In particular it must be not too small and not too large: 35 appear as a suitable value. The set of nodes together with the above relation can be interpreted as a network that can be represented as a finite graph $G(V, E)$ where V represents the set of nodes (in our case the I.P. addresses) and E the set of pairs of adjacent nodes: v_i is in relation to v_j if and only if $(v_i, v_j) \in E$. In other words, elements of E are edges with adjacent nodes at their extremities.

The set of all nodes adjacent to node v_i is called its neighbourhood. A path in the network is defined as an alternating sequence of distinct nodes and edges which begin and end with nodes and in which each edge is incident with its preceding and following nodes. The node v_i can reach the node v_j if there is a path in the graph starting with v_i and ending with v_j. The length of a path from v_i to v_j is given by the number of edges in the path. The distance between v_i and v_j is the minimal lenght of a path from v_i to v_j. A network is connected if there is a path between all pairs of nodes. When the

ties are concentrated within subgraphs (a subgraph is a graph whose nodes and edges form a subset of the nodes and edges of a given graph G) the network is clustered.

The level of clustering depends on the uniformity of the ties distributed throughout the network. In order to efficiently analyse a large network, as is our case, it is suitable first to individuate local clusters and then to analyse the internal structures of the clusters and relations between them.

Given the adjacency 2-mode matrix X and the relation defined above, we may produce a 1-mode matrix $\tilde{X}$ of dimension 1000×1000 where lines and columns represent I.P. addresses and the coefficient $\tilde{x}_{ij}$ for $i, j = 1, \ldots, 1000$ is the number of pages which have been viewed by both I.P. addresses v_i and v_j.

This matrix can be transformed into a binary matrix X^* representing an adjacency matrix, by setting

$$x^*_{ij} = \begin{cases} 1 \text{ if } \tilde{x}_{ij} \geq 35 \\ \\ 0 \text{ otherwise} \end{cases} \qquad i, j = 1, \ldots, 1000.$$

The 1-mode matrix X^* (I.P. addresses $\times$ I.P. addresses) can be obtained via the program UCINET (Borgatti et al. (1999)). Of course the diagonal coefficients are 1, but this information is not of interest and will be ignored.

	138.222.202.11	151.15.169.130	151.2.15.154	151.20.111.0	.....
138.222.202.11	1	0	1	1	...
151.15.169.130	0	1	0	0	...
151.2.15.154	1	0	1	1	...
151.20.111.0	1	0	1	1	...
.....	...	...	...	...	

Table 4. Adjacency matrix

At this point we introduce a matrix Y, called influence matrix, of dimension $N \times m$ where N is the number of I.P. addresses (in our case $N = 1000$), and m represents the number of components describing the reciprocal influences. A reasonable assumption is to set $m = 3$. This corresponds to assume that each user associated to an I.P. address may be influenced by the behaviour of three other users, identified by their I.P. addresses.

In order to build the matrix Y we use a modified version of the *Recursive Neighbourhood Mean* (RNM) algorithm , proposed by Moody (2001). Our algorithm has been implemented in SAS. The *Modified Recursive Neighbourhood Mean* (MRNM) algorithm consists in the computation of a suitable weightted mean by iteration, and generalises the RNM algorithm. This can be described as follows:

1. We assign to each I.P. address of the network, corresponding to a line of the influence matrix Y, a random number issued from a uniform distribu-

tion in $(0,1)$ for each of the m coefficents of the line. We obtain a matrix $Y^{(0)}$ $(N \times m)$ made of random numbers.

2. The matrix $Y^{(t+1)}$ is defined by

$$Y_{ik}^{(t+1)} = \frac{\sum_{j \in L_i} Y_{jk}^{(t)} \tilde{x}_{ij}}{\sum_{j \in L_i} \tilde{x}_{ij}} \qquad k = 1, \ldots, m, \qquad i = 1, \ldots, N,$$

where L_i is the subset of $1, \ldots, N$ corresponding to the I.P. adresses which are in relation with the address v_i, and $\tilde{x}_{ij}$ is the number of pages viewed by both v_i and v_j.

3. Repeat n times Step 2.

Remark 1. For $\tilde{x}_{ij} \equiv 1$, the algorithm corresponds to the classical RNM algorithm.

This procedure requires as input the list of adjacences, that is, the pairs of nodes v_i, v_j such that $x_{ij}^* = 1$.

In an ideal situation $Y = \lim_{n \to \infty} Y^{(n)}$. However $n = 7$ suffices to get a stable matrix.

I.P.	$Y_{\cdot 1}$	$Y_{\cdot 2}$	$Y_{\cdot 3}$	cluster
138.222.202.11	0.4881	0.4255	0.5359	1
151.15.169.130	0.4882	0.4259	0.5358	3
151.2.15.154	0.4881	0.4255	0.5359	1
...	.....	.....	.....	.....

Table 5. Summary of MRNM procedure

In Table 5 are reported the three columns of the matrix Y and the results of the Ward's minimum variances cluster analysis, carried out on the basis of the three components of Y. In such a way we obtain a clear clustering that reveals a structure of three groups as one can see from Figure 1.

The first cluster, which contains most of elements, is made up by the I.P. addresses which have a high frequency of relations, the second one, up on the left, is identified by I.P. addresses which have not many relations while the third one by I.P. addresses which have few relations. The two isolated I.P. addresses are referred to those that have no relation with other I.P. addresses.

3 Conclusions

In this paper we have presented a solution to the problem of identification of dense clusters in the analysis of Web access logs, by considering a modification of an algorithm known from social network analysis. Following the cluster analysis eventually block-modelling techniques can be applied. In doing so

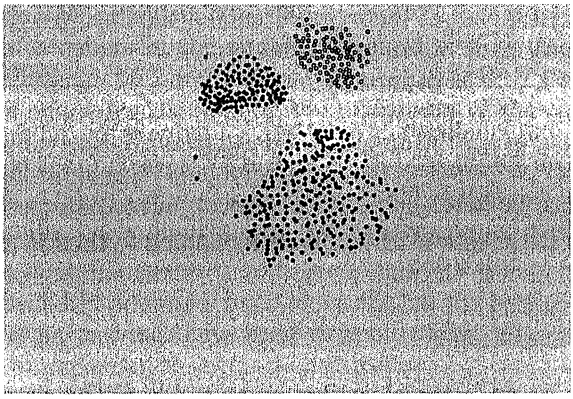


Fig. 1. Clustering according to MRNM algorithm

we have obtained an useful tool to study and profile customers in terms of their browsing behaviour and personal information. This allows us to build up useful business intelligence for the improvement of Web sites and the development of systems when data sets are large or even huge.

References

BATAGELJ, V. and MRAVR, A. (2002): *PAJEK: Program for large Network Analysis. http://vlado.fmf.uni-lj.si/pub/networks/pajek/*

BORGATTI, S.P., EVERETT, M.G and FREEMAN L. C. (1999): *Ucinet for Windows Software for Social Network Analysis, Harvard: Analytic Technologies.http://www.analytictech.com/.*

FRIEDKIN, N. and JOHNSEN, E. C. (1998): Social position in influence networks, *Social Networks, 19, 122–143.*

KOCHEN, M. (1989) : The small World. *Ablex Publishing Corporation*, Norwood, New York.

MOBASHER, B., DAI, H., LUO, T., SUNG Y. and ZHU, J. (2002): Integrating Web Usage and Content Mining for more Effective Personalization. *http://www.maya.cs.depaul.edu/ mobasher/personalization*

MOODY, J. (2001) : Peer influence groups: identifying dense clusters in large networks , *Social Networks, 23, 261–283.*

SCHOIER, G. (2002): Blockmodeling Techniques for Web Mining. In W. Haerdle and B. Roenz, *Proceedings of Compstat 2002*, Springer & Verlag, Heidelberg, 10–14 .

SRIVASTAVA, J., COLLEY, R., DESHPAND, M. and TON P. (2000): Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data, *http://www.maya.cs.depaul.edu/ mobasher/personalization* .

WASSERMAN, S. and FAUST, K. (1994): *Social Network Analysis: Methods and Applications*, Cambridge University Press, New York.