FROM LOG FILES TO WEB MINING: THE "ST. VALENTINE'S DAY" EFFECT

Alessia Cavallotti

PiTRE Consulting S.r.l., Via Conca del Naviglio,18 – 20123 Milano

Paolo Mariani

Departimento di Statistica, Università degli Studi Milano Bicocca,Via Bicocca degli Arcimboldi, 8 – 20126 Milano

Abstract

In general terms the Data Mining activity is characterised by an orientation to corporate requirements and by the opportunity of action within an industrial scale decision-making environment, besides featuring a quantity of elaborated information and availability of a relevant number of techniques. The supply of analysis services for marketing purposes fits with this culturally and technologically innovative environment, whereby it contributes to corporate change, from the set-up need of interdisciplinary projects to the professional figure's transformation along with the pertinent operating methodologies. A specific corporate experience description will show how "Database Marketing" and "Data Mining" are something more than simple neologisms indicating, in this specific case, respectively Direct Marketing and Data Analysis activities.

1. INTRODUCTION

Explaining a concept with the relative techniques of handling huge quantities of data and proposing, with a practical application, the use of graphic models for analysing purposes in marketing activities, is what companies need to be effective and efficient in managing their business.

The opportunity of accessing large databases, accrued in many years of activity or coming from outside, regarding the different aspects of company life can offer new solutions to management needs and objectives.

The novelty offered by new technologies and Data Mining (DM) lies with integrating decision-making processes and rules which have been set by synthesizing complex and extended information patrimonies.

This research phase is made easier when a Data Warehouse is present, organized by subject order and containing certified information. Data Warehouses are generally based on the managing systems but they also use information coming from external informative systems: they represent an innovative solution to data storage and management.

The main foreseen steps for Data Mining activities range from the preparation of input data to the enhancement of rules within decision-making processes.

In order for these tools to be effective, it is necessary to understand and to fully exploit the Knowledge Discovery Database techniques (KDD), the complex process of valid model definition, descriptive, potentially useful and readable, thus explaining relationships between figures.

The case described in this work relates to figures[1] provided by an Italian Internet Service Provider (ISP) which released a file created by the Microsoft Internet Information Server (IIS5.0) with log data coming from four different Web Servers, analysed in the same week as well as partial databases with reference to registered users, containing personal details.

2. FROM WEB LOG TO WEB MINING

Web Mining has a premise in Web Housing since that is where it draws data and meta-information from. Web Housing represents a development of the Data Warehouse aimed at e-Intelligence applications.

The creation of a Webhouse as well as the conception of a Data Warehouse is a process aimed at managing, organising and storing all business-related information (e-business especially) in order to facilitate their analysis; the quality of information, both in consistency and informative richness, depends on it.

The Webhouse has been implemented according to the following structure:

1. An Operational Data Definition (ODD) Group has been defined for the different figure sources;
2. An intermediate working tables storage area (STAGING) has been set up so as to make input data homogeneous and so as to elaborate them in further steps;

---

[1] *The management of this huge quantity of data and their subsequent analysis has been achieved thanks to the use of statistical analysis software systems produced by SAS Institute Inc, such as the Warehouse Administrator (version 2.1) for the Webhousing process development and the Enterprise Miner 4 (SAS System V8.1) for the Web Mining process.*

3. DETAIL TABLES have been created for the subsequent E-intelligence applications.

In Figure 1 below, the structure of the whole Webhousing process is shown and the different objects that have been created can be observed.

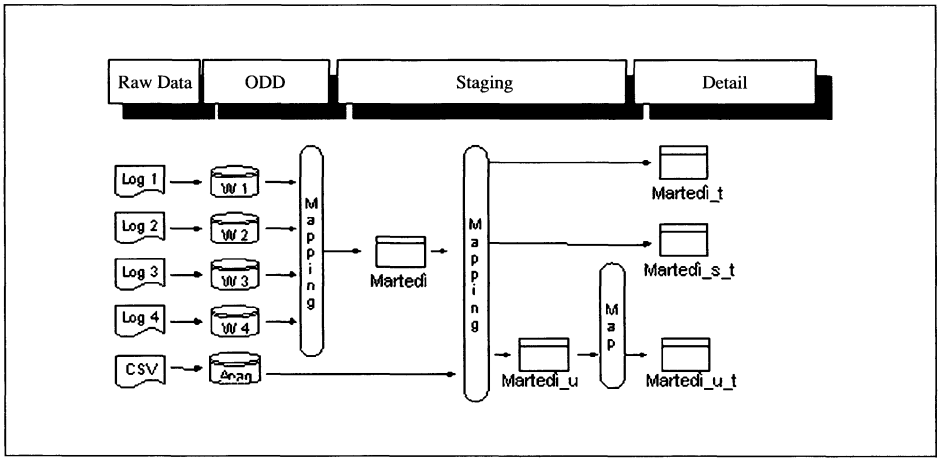

Fig. 1: Webhouse logic structure.

Web Mining applications generally depend on data availability. When the only available data is represented by the log files given by web site traffic, the only possibilities are web traffic mining applications. Data Mining systems give their best results when they look for clear and measurable objectives. It is necessary to exploit as best as possible the information contained in the data. Data Mining algorithms allow to highlight visitor behavioural models and to generate reports or implement actions based on such models. In order to use Data Mining techniques on a web site it is necessary to recognize and memorize visitor characteristics and interactions.

3. THE MARKET BASKET ANALYSIS

One of the most common unidirect Data Mining techniques concerns the analysis of frequent patterns in a set of figures (Market Basket Analysis). The problem can be illustrated as follows: given a binary database (any record contains a variable number of items) such as

$$r1: A, B, D, F$$
$$r2: B, C, D$$
$$r3: A, F, D$$

(where every r record represents a session and the visited pages are the A, B, C, D, F items) the desired result is a group of rules which allows to establish some associations on the basis of the page frequency in the database records. In other words, the analysis of the records in the previous example shows that page D is present whenever page A is, while page F is present with page A only in a certain record fraction.

In general, the aim of this analysis is to define rules such as:

$$if\ (A\ and\ D)\ then\ F \tag{1}$$

It is possible to define quantities which measure the validity of each such rule.

In particular, support can be used as the possibility to observe any item that is present in the rule in a database record; higher *support* indexes will therefore be associated to more welcomed items. In the specific case of the above mentioned rule, it would be:

$$support = p(condition+result) = p(ADF) \tag{2}$$

where the probability p(ADF) to observe items A, D and F in a record at the same time is given by the record fraction where these items appear simultaneously.

A measure of a rule's reliability is given by confidence, defined as follows

$$confidence = support\ /\ p(condition) = p(ADF)\ /\ p(AD) \tag{3}$$

For a rule to be useful, its confidence level must be higher than the possibility to observe the single result.

Results are shown as in Table 1 (below) and a part of the results which were obtained with the Market Basket Analysis application are shown, too. It is evident how the sequences are ordered following the *support* index and the most frequent one is */sms/public/default.asp=>/sms/members/formslogin.asp;* the confidence index tells us that the visitors who visualized the page */sms/public/default.asp* will also visualize, with a 57.2% probability, the page */sms/members/formslogin.asp*.

Using the graph theory it has been possible to represent relationships among visited pages. Figure 2 features the results of such representation, where it is possible to notice how complex interpretation becomes when following an increasing number of arches (links). The graph contains all possible combinations: each knot represents a visited page and becomes bigger when the number of visits increases; the presence of an arch identifies a dependency among knots, its dimension represents the level of associations whereas its direction, in the sense of the arrow, defines the sequence of the visited pages.

Tab. 1: Page sequence.

SET SIZE	SUPPORT	CONF	COUNT	RULE
2	13.035	57.192	9527	/sms/public/default.asp ==> ==> /sms/members/formslogin.asp
...	...	...	...	...
4	0.034	23.148	25	/error/error.asp ==>/default.asp ==> ==> /ciaoamici/default.asp ==> ==> /servizi_hp/ /default.asp
...	...	...	...	...

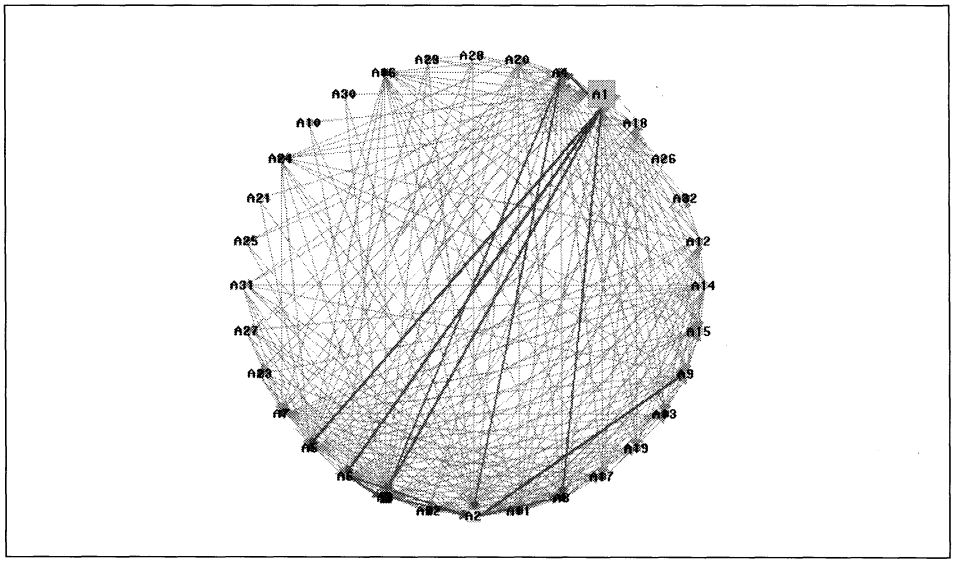

Fig. 2: Complete graph.

4. THE "ST. VALENTINE'S" EFFECT

At first, only the pages with a strong associative level have been taken into account (higher *support* values – Fig. 3) so as to highlight the explanatory power of the above graph, and then only the pages with a weak associative level (with low *support* values – Fig. 4). The result shows that the most visited pages are those related only to some of the services presented by the site such as the horoscope, the search engine and the SMS services in particular. This is not casual, as St. Valentine's Day falls during the analysed week therefore all the related pages have obtained higher visiting levels.

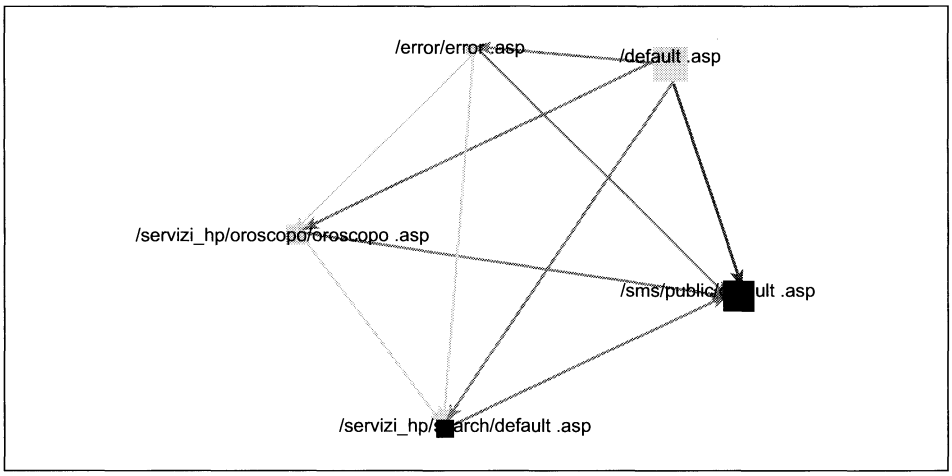

Fig. 3: Graph of the most cross-referenced pages.

Figure 3.a shows different page associative levels of the most visited service, i.e. SMS. Present alternative paths to access the service have brought or intrigued the user to its use.

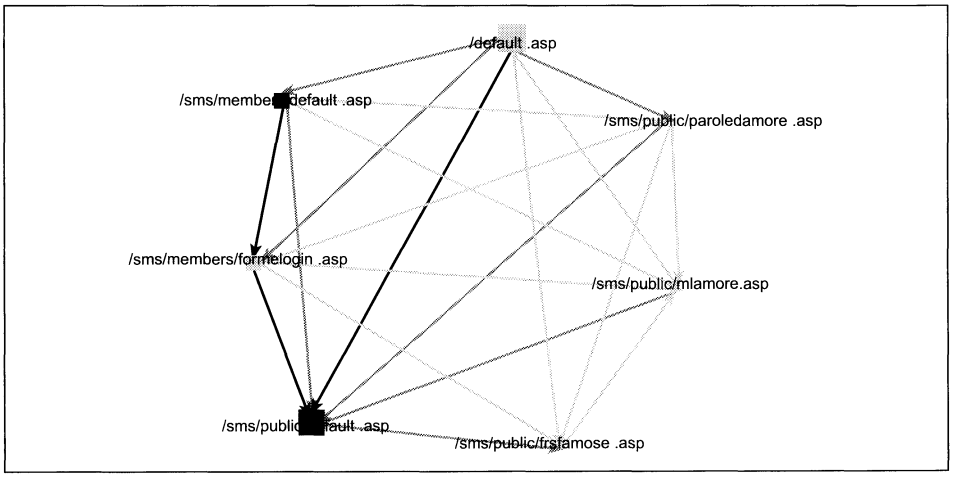

Fig. 3.a: Graph of SMS service related pages.

On the other hand, Figure 4 shows how, in this particular situation, services that are generally highly appreciated (such as information and entertainment) acquire a secondary role.

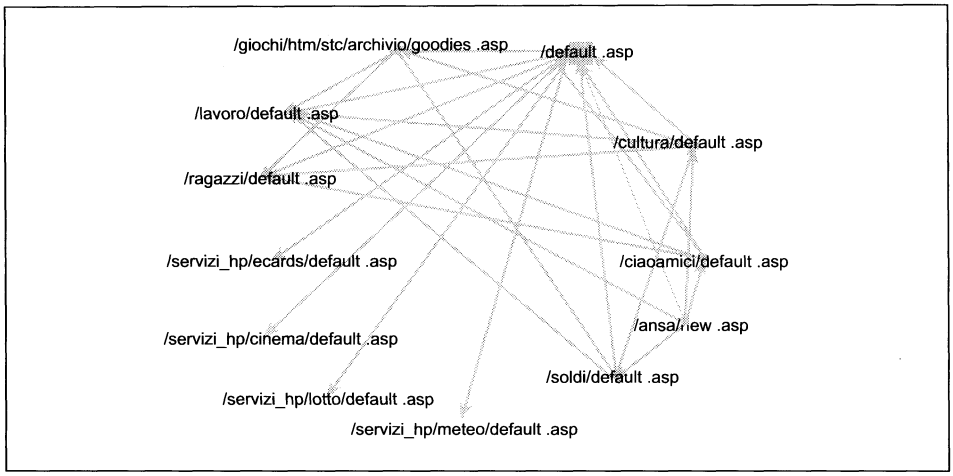

Fig. 4: Graph of the weakest cross-referenced pages.

5. CONCLUSIONS

Web Mining applications, one of the many methodologies within e-intelligence applications, mainly depend on data availability and give their best when they have clear and measurable objectives. The described case is a clear example of how web sites "respond" to events.

The obtained result points out how different services differentiate due to an event; St. Valentine's Day favours services such as the horoscope and SMS due to their strong link with the mentioned event, while the usual top favourites, Entertainment (games) and information/miscellaneous (ANSA), only gain a secondary position.

This and many other types of information can be collected from the interaction of a visitor on a website. It is actually possible to re-organize the website page layout deriving it from the associations' analysis, thus increasing profits through effective promotions.

Promoting a service could mean promoting the associated one, too.

Data Mining techniques are fully exploited on a website when all characteristics of potential visitors (besides their interaction.) are recognized and memorized.

In this way it is possible to study real behavioural models through Data Mining algorithms, generating reports or implementing actions on the basis of the registered models.

BIBLIOGRAPHY

BERRY M., e G. LINOFF, (1997), "Data Mining Techniques for Marketing, Sales, and Customer Support" , Wiley and Sons, New York.

BERRY M.J.A., LINOFF G.S., (1999), "Mastering Data Mining: The Art and Science of Customer Relationship Management", Wiley, New York.

CABENA P., (1997), "Discovery Data Mining: From Concept to Implementation", Prentice-Hally, Englewood Cliffs, New York.

CAVALLOTTI A., MARIANI P., (2001), "Dai log file al Web Mining: l'effetto San Valentino", Modelli statistici per le applicazioni di Data Mining, Atti del convegno SMDM, Pavia.

DEL CIELLO N., DULLI S., SACCARDI A., (2000), "Metodi Di Data Mining per Il Customer Relationship Management", Franco Angeli, Milano.

CUZZOCREA G. e SACCARDI A., (1998), "Metodi per il supporto alle decisioni di marketing", Note del corso SAS.

DI BARTOLO A., VERRECCHIA F., (2000), "Data Warehouse dc Webmining: la Statistica al servizio della New Economy", atti del convegno SAS Campus, Università degli studi di Milano-Bicocca, Milano.

INMON W.H., (1996), "Building the Data Warehouse", seconda edizione, Wiley, New York.

DAI LOG FILE AL WEB MINING: L'EFFETTO "SAN VALENTINO"

Riassunto

L'articolo riporta alcuni aspetti, tratti da specifiche esperienze aziendali, che rendono i termini Database Marketing e Data Mining qualcosa in più di semplici neologismi per indicare, in questo caso specifico, l'attività di Direct Marketing e l'Analisi dei Dati. La possibilità di accedere ad ampie basi di dati, accumulate nel corso di anni di attività o provenienti da fonti esterne, riguardanti diversi aspetti dell'attività aziendale possono fornire una nuova risposta alle esigenze ed agli obiettivi del management.

La novità offerta dalla nuova tecnologia e dal Data Mining (DM) sta nell'integrare i processi decisionali con regole costruite sintetizzando complessi ed estesi patrimoni informativi. Il DM ha come premessa il Data Warehouse (DW) da cui trae i dati e le meta-informazioni. I DW poggiano le basi sui sistemi gestionali ma utilizzano anche dati di sistemi informativi esterni, rappresentando una soluzione innovativa ai problemi d'immagazzinamento e gestione dei dati. Le fasi previste per l'attività di Data Mining vanno dalla predisposizione dei dati di input all'implementazione delle regole nei processi decisionali. Tra le principali tecniche di DM si riporta l'esempio di un'applicazione di Market Basket Analysis. Questa particolare tecnica mira ad identificare delle regole ricorrenti all'interno di un set di dati. In questo contesto si fa riferimento ai dati forniti da un Internet Service Provider (ISP) Italiano relativi alle interazioni dei visitatori con il sito web. Sfruttando la teoria dei grafi è stato possibile rappresentare le associazioni tra le pagine viste, mettendo in evidenza come si differenzino i diversi servizi offerti in relazione ad un evento. Per concludere, il risultato dell'analisi fornisce valide indicazioni per un'eventuale riorganizzazione del layout del sito, permettendo così di aumentare i profitti mediante efficaci promozioni.