

Mining massive Data Sets from web

Bruno Scarpa ¹

Dipartimento di Scienze Statistiche, Università di Padova e-mail: scarpa@stat.unipd.it

Abstract: Motivated by the problem of the analysis of the behaviour for visitors of a business web site, classical and new web mining procedures are presented by specifying the characteristics of data from Web. After a survey of very popular “web analytics” tools, where simple descriptive statistics are offered, showing the visiting behavior of a site, more sophisticated models are presented in order to describe, predict and cluster sequences of web pages and time on site, by analyzing raw data.

Keywords: web mining, random effect, non parametric regression

1. Introduction

The growth of the World Wide Web in the last years, and its success and usage in research, business and daily life, makes it the largest publicly accessible data source in the world. This huge source of data is a very challenging ‘world’ for statisticians and data analysts, to whom is required to extract knowledge from all the available data.

The Web is particularly fascinating since it presents many unique characteristics (see for example Liu, 2007): on the Web (a) there is a huge amount of data and information, which is growing very fast. The coverage of information is also very wide; (b) data are available with any format and type (structured tables, unstructured texts, multimedia files, etc.); (c) the information is quite heterogeneous. (d) a large amount of information is linked. (e) information is very *noisy*. Given the large amount of data available, for a given purpose only part of the information is useful, all the rest is considered noise. Also, a large amount of information on the Web is of low quality, erroneous and misleading; (f) many commercial and free services are available (purchases, pay bills, fill informs, etc.); such as purchases, pay bills, fill informs, etc.; (g) data change dynamically. The changing information by itself and the activity of monitoring the change may be very useful; (h) people can be traced and followed in whatever they do.

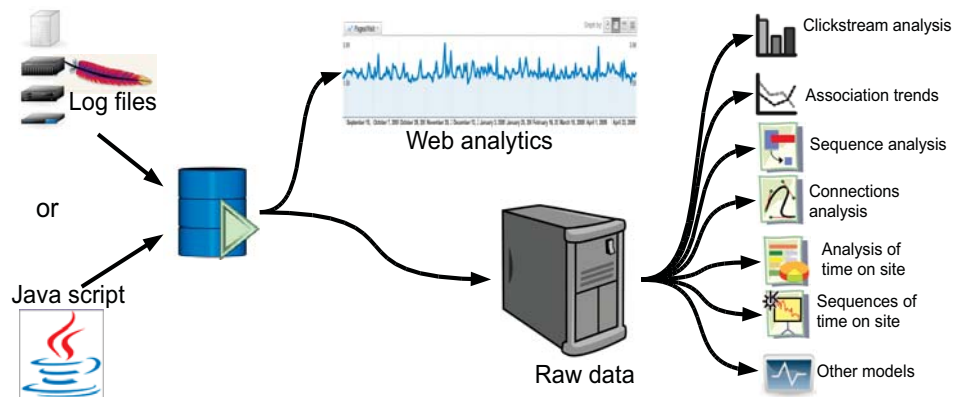
All these characteristics give to the data analyst a very fruitful framework where he can find opportunities and challenges in order to discover useful knowledge from the Web. In this context the applications of data analysis and data mining techniques to discover patterns from the Web are often called *web mining*, and they are generally divided into three main families, according to the primary type of data used: (i) Web usage mining, to discover user access patterns from information about the visits and the “clicks” of the users (often called click-stream data); (ii) Web content mining, to extract useful knowledge from Web page contents (text, image, audio or video data); and (iii) Web structure mining, to discover useful knowledge from hyperlinks and document structure.

In the following, we will consider how statistical modeling can be used in web usage mining in order to achieve some specific typical targets.

¹Address of correspondence: Dipartimento di Scienze Statistiche, Università di Padova. Via Cesare Battisti, 241, I30121 Padova

There are at least a couple of alternatives in order to record data on web usage: a) Web log files, that are organized text files recorded on each Web server collecting all the relevant information on each session and visit of a page, or b) data recorded by a java script embedded in each web page that sends all the relevant information about a site to an external repository recording all data. Advantages and disadvantages of both alternatives have been discussed widely, and both the solutions are useful for some specific goal. As motivating example we consider the Web site of a consultant company and we will use data about the web usage of the different pages of the site in order to better understand the customers and the visitors, having in mind some marketing purposes. An image of the structure of web usage analysis is plotted in Figure 1.

Figure 1: *Web usage mining.*



2. “Web Analytics”

The huge amount of data available for each Web site needs to be transformed into knowledge. Most of the knowledge that companies need from Web usage is obtained by referring to simple aggregate descriptive statistics.

In fact the most diffuse Web mining tools are software that obtain averages or counting on data, and report them in tables and plots (examples are Google Analytics or Net-ratings). These services offer pre-processed data and some descriptive analysis such as sophisticated graphics and tables about *page views* or *page impressions* (the number of requests to load a single page of an Internet site), *unique visitors* (number of units of traffic to a Web site, counting each visitor only once in the time frame of the report), the *average time on site* (average number of seconds a user visits a Web site), and many other aggregate variables of interest.

In addition to a sharp look of these data, often organized as time or spatial series, well known statistical models could be fitted in order to face specific problems, such as the effect of an action on the time pattern of the visits to a Web site, or the relationship over time between visitors of different pages. Analysis of such kind of problems could be faced by classical tools such as simple linear or nonlinear models, models for multivariate time series, state space models (e.g. Durbin and Koopman, 2001), intervention analysis,

transfer function models (see for example Box, Jenkins and Reinsel, 1994) or switching regime models (Kim and Nelson, 1999).

3. Analysis of Raw data

Even if most of the needs of companies and analysts are covered by ‘simply’ analyzing aggregated data available with standardized tools, sometimes, a deeper analysis is needed. Analysis of raw data (information about each single visit and visitor) may be very useful and rich of knowledge for decision making. Web usage data sets easily became massive and analysis of these type of information require not only statistical efficiency and good inferential properties, but also computational practicability.

In the following we face some typical problem of interest for decision makers that needs to analyze Web usage raw data, by starting from real data. The dataset contains data about the web site of a consultant company visited by 26 157 anonymous visitors. For each visitor the pages of the site that have been visited in a fixed time interval are available. Visitors are identified with an identification number and no personal information is given. The number of pages of the Web site is 231 and the total number of page views on the entire site is 47 387, so that every page has been visited in average 205 times and each visitor visited in average 1.81 pages. Some of the pages are similar and they have been aggregated in 9 categories (home, contacts, events, fun, map, about us, publications, services, sectors) with respect to the content of every single page. Day and hour of the visit and its duration for every single page are also recorded in the data set.

3.1 Sequence rules and clustering behavior

A typical goal in analyzing Web usage consists in studying the association of visits of different pages. A large literature has been produced in order to face such a problem starting from Agrawal et al. (1993). The most known algorithm is the so called apriori algorithm (Agrawal and Srikant, 1994), based on a couple of association measures, the *support* (the percentage of connections between two pages), and the *confidence* (the percentage of connections between two pages with respect to all the visits to the first page).

A more specific analysis of the associations between pages is often performed by studying sequences of visited pages. The pattern of visits or the navigational pattern can be studied as a longitudinal analysis of categorical variables and analyzed for example by using sequence analysis, a methodology developed in bioinformatics that helps to describe and visualize sequences of discrete sequential data (in web mining sequences of visited pages). More probabilistic approaches have been proposed, in particular Markov models, have been widely used in predicting the next user action based on a user’s previous surfing behavior (e.g. Di Scala and La Rocca, 2002) or to discover high probability user navigational trials in a Web site. Recursive graph models and Bayesian network have also been largely used (e.g. Heckerman, 1997).

3.2 Analysis of the time on site

The analysis of the time on site is one way of measuring visit quality. If visitors spend a long time visiting a site or a page, they may be interacting extensively with it. As an example we fit an anova with a skew-normal distribution on the logarithm of the time on

site of the pages for the data of the consultant company, with respect to the category to which the page belongs and the international region from where the page was recalled (Europe, North America, Central and South America, other countries). A hierarchical analysis has also been performed by considering the single user as a random effect.

Sequences of time on site may be analyzed to profile customers by grouping users of the site. In particular visitors spending short time in their first page visited in a session and longer times in the last pages are potential customers (they are looking for something that, eventually, they find in the site). These are target customers for the company, but it could be interesting to discover other specific patterns in the sequences. In order to face such a problem, a flexible approach for incorporating prior information in semiparametric Bayesian analyses of hierarchical functional data has been applied. The proposed approach is based on specifying the distribution of functions as a mixture of a parametric hierarchical model and a nonparametric contamination. The parametric component is chosen based on prior knowledge (the target customer), while the contamination is characterized as a functional Dirichlet process.

4. Discussion

The usage of statistical models in analyzing massive Web data is almost as wide as statistics itself. Some particular characteristics of Web data requires to develop specific models and algorithms. Some example has been shortly shown, and during the talk will be more extensively presented. Many other topics have been left to the reader curiosity such as text mining (analysis of the contents of web pages) or social networking (analysis of the relationship that arises with peer to peer tools or with network programs – e.g. facebook).

References

- Agrawal R. and Srikant R. (1994) Fast algorithms for mining association rules, in: *Proceedings of the 20th International Conference on Very Large Data Bases (VLDB'94)*, 487–499.
- Box G., Jenkins G. and Reinsel G. (1994) *Time series analysis: forecasting and control*, Englewood Cliffs, N. J. : Prentice Hall.
- Di Scala L. and La Rocca L. (2002) Probabilistic modelling for clickstream analysis, in: *Data mining III*, Zanasi A., Trebbia C., Ebecken N. and Melli P., eds., WIT Press, Southampton.
- Durbin J. and Koopman S. (2001) *Time series analysis by state space methods*, Oxford University press.
- Hamilton J. (1994) *Time Series Analysis*, Princeton, NJ: Princeton University.
- Heckerman D. (1997) Bayesian network for data mining, *Journal of Data Mining and Knowledge Discovery*, 1, 79–119.
- Kim C. and Nelson C. (1999) *State-Space Models with Regime Switching*, Cambridge, Massachusetts: MIT Press.
- Liu B. (2007) *Web data mining: exploring hyperlinks, contents, and usage data*, Springer, Berlin.
- R. A., Imielinski T. and A. S. (1993) Mining association rules between sets of items in large databases, in: *Proceedings of the ACM SIGMOD Conference*, 207–216.