

Qualifying Web data to meet expectations

A. Lourenço & O. Belo

Department of Informatics, University of Minho, Portugal

Abstract

The World Wide Web is today's richest medium of information and services. Interchanging data freely "without" any time constraints is much more than appealing. However, the Web has to be restructured in order to address today's needs and expectations, favouring users' searches and navigation and, to do so, Web analysis tools are the order of the day. There is a similarity between the analytical Web data processing and the analytical processing of "conventional" data. The truth is that the enormous volume of data along with the high transactional activity strongly prohibits the use of raw data. In such a scenario, Web server logs emerge as the main sources of information and there are many Web analysis tools today that work over these raw elements. However, the obtained results are far from satisfactory, since they do not help to understand users' navigational patterns. In the near future, data webhouses will be in charge of any analytical work, requiring well-defined extraction, transforming and loading processes, capable of reuniting all relevant data and ensuring its consistency and quality. Only when these processes are implemented is there room for thinking about Web Mining. This paper aims to target the problem of establishing data webhouses as primary Web Mining sources, highlighting the excellence of the Web domain towards mining and defining the guidelines that populating processes must comply with. In this sense, a case study will be presented, identifying its available data sources, establishing a proper data webhouse and specifying all the processing required to populate it.

1 Introduction

The World Wide Web, Web for short, is today's richest medium of information and services. Interchanging data freely without any time constraints or almost any other type of constraint is much more than appealing, it is a major premise to keep evolving. In fact, there is an actual concern with extending the business

circle and embracing the cybernauts as main stream customers. However, this is not an easy task. The tremendous number of Web sites, their dynamism, heterogeneity, high linkage and diversity turned the medium into an entanglement hardly specifiable. It is not just a question of businesses competing for customers, but also a matter of customers trying to find services, products or simply information. The Web has to be restructured in order to address today's needs and expectations, favouring users' searches and navigation. To do so, Web analysis tools are the order of the day.

In this sense, there is a similarity between the analytical Web data processing and the analytical processing of "conventional" data. Despite the specific purpose of each analysis, the truth is that the enormous volume of quite diverse data along with the high transactional activity strongly prohibits the use of raw data. In the Web scenario, Web server logs emerge as main sources of information and there are many commercial and free Web analysis tools that work over these raw elements. However, the obtained results are far from being satisfying. Most of them may help in the eventual site adjustment, but are not of great help when trying to understand users' navigational patterns. Once again, there must be set a bridging point between the transactional and the analytical worlds. Once again, data warehouses seem to be the perfect candidates for the job. Data warehouses or more specifically, data webhouses (performing the convenient instantiation for the Web), will be in charge of any analytical work, i.e., they will keep the whole spectrum of data thought to be analytically worthwhile. Obviously, this implies well-defined *Extraction, Transforming and Loading* (ETL) processes, capable of reuniting all relevant data (not just logs' data) and ensuring that the data webhouse will be filled in with consistent, high-quality elements. Only when these processes are settled and running there is room for thinking about Web Mining, which will provide the actual gain.

This paper aims to target the problem of establishing data webhouses as primary Web Mining sources, highlighting the excellences and pitfalls of the Web domain towards mining and defining the major guidelines that ETL processes must comply to. In this sense, a case study will be presented, identifying its available data sources, establishing a proper data webhouse schema and specifying all the processing required towards data integration into that repository. Session and user identification as well as robots detection are examples of particular Web server logs data processing issues that are attended to.

2 The wealth and complexity of the Web

The characteristics surrounding a domain constitute both its strength and its weakness. There is no such thing as a perfect domain or problem. Before initiating any data process (in particular, any analytical one), it is crucial to know what it is awaiting. Realising what things must be tackled, what can be avoided or worked around and what simply cannot, and identifying ancillary features and valid approaches are mandatory tasks.

In the case of the Web, the domain is very wealthy and relatively young. Its potential is enormous, opening new perspectives in a large variety of fields. During the last few years, commerce activities have been increasing amazingly over the Web. Organisations can greatly benefit from this exceptional opportunity, establishing stronger, tighter relationships with their customers. However, this is not straightforward. The number of competitors per domain is too large and the overall dynamics gives no margin for second thoughts or hesitations. Decisions must be made on the spot and they must be precise, pointing to clear, well-defined objectives. A bad user experience represents much more than the immediate revenue loss, it affects the brand name and, in consequence, its position on the market.

Some years ago, [1] questioned whether the Web should be viewed as a quagmire or a gold mine. Several issues were discussed, namely, overload, quality and discipline problems, concluding that a well-designed analysis platform would actually be capable of discovering gold in the Web. To do so, certain desiderata should be satisfied. Some of the imposed requirements are hard to be met in most real-world problems, but the fact is that the Web seems to work with them just fine [2]. In fact, E-commerce, sustained by proper design, is able to satisfy many of these desiderata [3].

Clickstream data is wealthy. Information about page views during each user's session is vast (easily reaching the figure of one or two millions per month for a small site), and can be quickly gathered. Besides that, the presence of numerous attributes (along with the large size of the dataset) justifies the need for data mining. Simple techniques, such as bar graphs, scatter plots and spreadsheet tables would do just fine for small problems.

On the other hand, there is data quality. Sustaining that it is possible to get clean data (in a 100% view) is unrealistic. However, the quality of the data is an indubitable asset that must be enforced at all cost. Electronic collection at the source is a step forward as the likelihood of noisy and corrupt data is reduced, ensuring superior quality and highly reliable data. Pages demanding human-entered information, such as registration forms, can still contain errors, but on-line validation may prevent against many unintentional errors that would be hardly caught on traditional formats.

In many traditional businesses, the outputs of *Knowledge Discovery in Databases* (KDD) are most of the times depreciated on behalf of legacy systems and inflexible domains that make it very hard and expensive to actually apply the new knowledge and improve processes. In E-commerce, the situation turns around and most of the mining results can actually be tried out, testing the users' acceptance and general behaviour. The layout and design of Web sites can be modified, adjusting and redirecting personalisation procedures and executing and automating e-mail campaigns. These campaigns can be tracked all the way through the line, analysing the user's behaviour and generating valuable reports.

All the above items are excellences that make the Web a very appealing domain for mining, but not all of it is so pretty. Many interesting information lays in it and there is certainly the need and the will for acquiring and exploiting it. Nevertheless, the format, location and availability of these data are still

complex matters. For instance, it is a common practice to presume that web server logs are the main data source for web mining and web analysis in general [4] [5]. Notwithstanding, this is not quite true. It is a fact that web server logs should be considered and faced as the basis for any analytical structure. However, they are not the only available source and, depending on the level of outsourcing of the organisation, they may not even be the primary source of data.

Web server logs detach from other sources, because, by definition, they are the diary of all journeys, i.e., they record every move made by every user that accessed the site [6]. So, it is likely that they keep primary elements like: the referrer, the query (if there is one), the date and time of the request, the requested page, etc. The problem is that they were not design for analytical purposes, but rather to debug web servers and, on account of this, their elements are too limited to act on their own in the analytical arena.

Another problem is the stateless nature of the HyperText Transfer Protocol that “refuses” the notion of a session. It is undisputable that being able to identify the users and circumscribe the requests that conform each user’s session are essential for further mining. The use of multiple web servers for Web load balancing, the presence of proxy servers, the IP reassignment and the rejection of cookies by browsers, among others, make it very hard to track and reunite the data concerning each session and to identify each visitor [7] [5]. Besides these two problems, there are still other kinds of flaws in logging. For instance, hitting the browser’s Back button will not have a correspondence in terms of the server log, i.e., the server will not be aware of the revisiting of an earlier page. Unless it is intentionally prevented, the page will be loaded from the browser’s cache and it will not be any log record about it. More important, events like “add to cart”, “delete item” or “change quantity” are not available in web logs. So, inferring these important events, while trying to understand the usage patterns, becomes very complicated. Information related to form fulfilment is also desired and it is also absent. For example, in order to improve a Web site, it would be relevant to have access to search forms’ information. If the keywords that were used are known, organisations can update their synonyms and their product mix towards future compliance. Finally, and despite the inclusion of other kinds of data, there are also open research issues to attend to, such as the detection of robots [8] [9] and the scalability of data mining algorithms [3], sustaining and enforcing ETL processes and being able to apply the new knowledge.

3 The ETL process

One of the biggest challenges facing any data warehouse project is extracting the data from the data sources and transforming it into a form suitable to use [10]. In fact, the usefulness of the data warehouse is dependent on the availability and quality of data. A large content *per se* will not mean a thing. Its value is relative and comes with the worth of its contents.

The whole process is a major source of problems. The fact that many heterogeneous data sources are involved is a synonym of difficulties. Furthermore, there are numerous data problems that are unknown or unexpected

at first and that will raise up complexity even more. Bad or missing data, different representations of the same data in different applications and a multiplicity of applications implemented in different system technologies come with the territory. The overall architecture of ETL processes embraces all the programs related to the process of the data (extraction and transfer, loading, clickstream processing, transformation, cleansing and validation), the database and data models that store the queryable data, the analytical applications that make use of the information, and the underlying data and process infrastructure that supports all these components. Moreover, it should be highlighted that the data staging area is the working basis of the process, supporting all the ETL processing steps prior to the final data loading into the data webhouse.

3.1 Evolving the ETL process to meet the clickstream data

The clickstream has a unique nature, requiring additional processing. The diversity of existing data sources increases the level of complexity. There are internally controlled infrastructure components like firewalls, proxy servers, Web servers and even, media servers, but there are also outside data like content caching servers, search engines and online advertising engines. Moreover, the nature of the data involved is quite different from the traditional data, requiring a closer look over the data sources during the schema design.

It is not usual that a “normal” data warehouse relies on data coming from external sources. This situation changed with the appearance of Web data sources, as many of the Web sites outsource functions. Besides that, Web serves and Web application servers are being replicated, and data is now spread throughout multiple sources. Data for a single user’s session has to be extracted from multiple sources. Often, there is the necessity of making the site more clickstream-friendly, changing or adapting the underlying Web applications in order collect the needed data. Due to all these issues, the clickstream processing is handled by a separate set of tasks (Figure 1), within the ETL process, that support the supplementary processing it requires.

Generally speaking, there will be a mechanism that moves the clickstream data to a location where it will be processed. This mechanism does not need to be too elaborated and might be as simple as a file transfer or copy. Next, there will be basic data parsing and loading into the data staging area, place to where the data from the other sources will also converge. Once the data is all in the data staging area, the clickstream data has to be submitted to some programs that map hits to user sessions, identify users and perform the processing associated to pages (for example, linking hits for page elements to pages). Depending on what has been specified there may be a step to filter undesired data such as hits for images or other page elements.

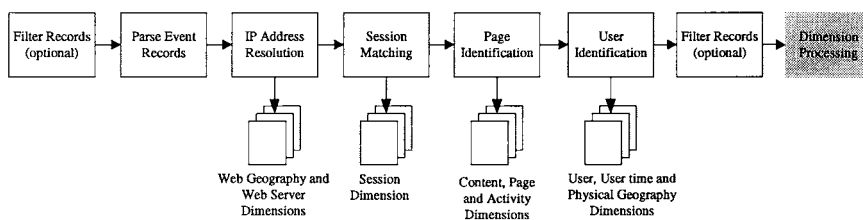


Figure 1: Stages in clickstream processing.

4 Taking advantage of data webhouses' contents

The primary concern in any data warehouse project is the establishment of an adequate set of dimensions. In [4], there can be found a portfolio of available conformed dimensions for data webhousing that give a hint about the exploitable universe. Normally, dimensions authorities are more than willing to help, inviting the attachment of their data. It is up to the analysts to decide what will be the chosen dimensions and facts and the correspondent grain, according to users' requirements and expectations.

In this work, the aim was set on presenting the large spectrum of possibilities presented by the Web and, in particular, by the E-commerce, in terms of analysis. After stating the availability of large amounts of quite diverse data and clarifying the ETL steps involved in Web data processing, it is now time of presenting a case study. For this purpose it was chosen the case study proposed in the KDD-Cup 2000 concerning a retail site, Gazelle.com, that sells legware and legcare products and, for which, there was available clickstream and purchase data [11]. The idea was to take a real case, preferably, already mined and assessed what would be the benefits of imposing a data webhousing step in the process. From the contest's approaches and results we know what can be achieved in terms of mining the raw data and what data processing is required to achieve it.

The participants were judged on the following five questions: given a set of page views, will the visitor view another page on the site or leave; given a set of page views, which product brand will the visitor view in the remainder of the session; given a set of purchases over a period of time, characterize visitors who spend more than \$12 (order amount) on an average order at the site; given a set of page views, characterize killer pages, i.e., pages after which visitors leave; and, given a set of page views, characterize which product brand a visitor will view in the remainder of the session.

From this set of questions it is clear that the whole point is to analyse what happens in each page view, trying to interpret the sequence of clicks made by the visitor and predicting what he will be doing next. Mainly, the questions were "solved" recurring to classification and association rule mining. Obviously, as data was available in a two text files form, some efforts had to be made in order to process and arrange it conveniently. Sessions identification and page views

grouping, visitors' identification and robots detection were some of issues that the contestants had to deal with. Moreover, any sort of aggregation required additional processing.

By sustaining a proper data webhouse schema, most of the data processing would be already done and data would be structured in such a way that would allow these sorts of questions and many others. Next, it is presented a possible schema considering the page view level (Figure 2).

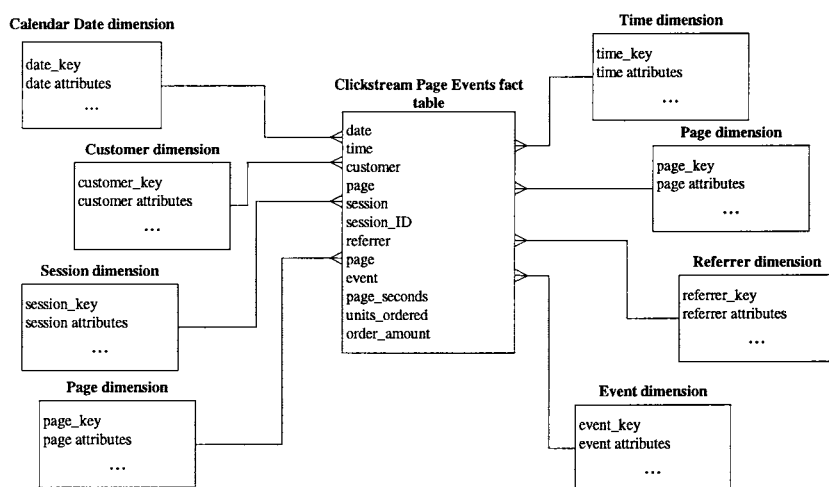


Figure 2: A data webhouse concerned with page events.

Targeting events implies knowing when each one of them happened and therefore, a calendar date and a time of day are in order. The first has one record for each day of the calendar and keeps track of many special attributes including seasons, holidays, local workdays and other elements that are specific to the business. The other records the moment of day in which the event occurs. Maybe administrators want to know what happens on a certain hour or interval in time or, for a given situation, an analysis in seconds is relevant. Once again, the data webhouse must provide flexibility towards users enquiring necessities.

Entities are other crucial assets. If an event occurred, someone had to deploy it and it is related to or it affects a certain product or service. Therefore, tracking entities will allow both product /service prospecting and customer/visitor behaviour analysis. The Products dimension emerges from the store's catalog, i.e., a database of all available products is maintained for transactional purposes. So, it is a matter of extracting all valuable products' characterization elements and conform them into the correspondent dimension. Information about customers is always welcome. Although, Web Server logs are not exactly rich on this resource and additional data, from transactional systems or other sources, is required.

The Page dimension describes the context for a Web page event and each record is related to an individual page type. Static pages have their own records and when the definition of one of them eventually changes, the correspondent record is either overwritten or duplicated (techniques normally applied to slowly changing dimensions). Independently from the type of page considered, it is crucial that during the website design developers assign descriptive codes and attributes for each page. Well-defined functions are needed in order to describe the page, fulfilling the different fields. Moreover, as Web design continues to evolve, it is important to sustain a flexible definition of a page, embracing both static and dynamic types.

The Event dimension reports the incident, i.e., it describes what happened on a certain page at a particular point in time. Situations like Open Page, Refresh Page, Click Link or Enter Data are possible. The context within which events took place is also relevant, especially if more than one level of diagnosis is included. For instance, the local context of a given user's session may be Requesting Product Information, but the overall session context might be Ordering a Product. Therefore, it is valuable to capture both contexts along with the related success status in a Session dimension. This dimension is very important, because it brings light to group session's studies. With its help, it is possible to answer to questions concerning subjects like: the number of customers that consult product information pages before ordering or dropping the order or began an ordering process, but never completed it.

The referrer, i.e., the URL that the user clicked assessing the page in question should be the starting point of any investigation. Referring information highlights the site's reaching points. It is relevant to know how the customer got in touch with the site, specially, in terms of marketing purposes. Analysing the disclosure of the site in the Web, identifying who thought that it was worthwhile to link to it and how easy it is to access to it, are some of the issues that the Referral dimension must attend to.

For each page event, the following measures are considered: the `page_seconds`, i.e., the number of seconds elapsed before the next page event; the `units_ordered` and the `order_amount` have an explicit semantic. The last two fields will only have non-empty value when the event in cause places the order. This situation will cause many zero or null field values for a great deal of records. However, they are still considered relevant as they tie all important Web revenue to behaviour.

5 Conclusions

In [3] there is a statement that perfectly resumes the overall picture of the Web: "the Web is an amazing experimental laboratory". In fact, it can be used for conducting experiments, studying users' behaviour and learning trends quickly. Compared with traditional businesses, it is much easier to change a site's contents or organisation according to users' response. Furthermore, the effect of such modifications is much more immediate and visible, and will provide feedback to the site's adaptation process. However, there is a catch. All this

experimentation is not sustainable as a standalone process, but rather it requires strong analytical support, namely Web Mining.

This paper targeted the problem of establishing data webhouses as primary Web Mining sources, highlighting the excellences and pitfalls of the Web domain towards mining and defining the major guidelines that ETL processes must comply to. A case study was introduced, identifying its available data sources, establishing a proper data webhouse star schema and specifying all the processing required towards data integration into that repository. As it was used in the KDD Cup, besides the actual data, there were also available data mining outputs and remarks about the processing that had to be taken. So, it was possible to assess major differences between mining the raw data and the data webhouse's contents, not just in terms of data processing, but also in the flexibility and easiness of constructing datasets to perform tasks of diverse nature. The volume of Web data is so enormous and diverse that trying to do the work on a strike is too risky and might get too sloppy. With this paper it was demonstrated that data webhouses grant structure, quality and availability of data, which are, as is well-known, essential characteristics for any KDD process.

Acknowledgements

We wish to thank to Blue Martini Software for contributing the KDD Cup 2000 data [12].

References

- [1] Etzioni, O., The World-Wide Web: Quagmire or Gold Mine? in Communications of the ACM 39(11), pp. 65-68, 1996.
- [2] Kohavi, R. & Provost, F., Applications of Data Mining to Electronic Commerce, Data Mining and Knowledge Discovery journal 5(1/2), 2001.
- [3] Kohavi, R., Mining E-commerce Data: the Good, the Bad, and the Ugly, an invited talk at PAKDD 2001. Hong Kong, China, 2001.
- [4] Kimball, R. & Merz, R., The Data Webhouse Toolkit: building the Web-enabled data warehouse, Wiley Computer Publishing, John Wiley & Sons, Inc., 2000.
- [5] Cooley, R., Mobasher, B. & Srivastava, J., Data preparation for mining world wide web browsing patterns, in Knowledge and Information Systems, 1(1), 1999.
- [6] Srivastava, J., Cooley, R., Deshpande, M. & Tan, P.N., Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data, in SIGKDD Explorations, (1) 2, pp. 12-23, 2000.
- [7] Berendt, B., Mobasher, B., Spiliopoulou, M. & Wiltshire, J., Measuring the accuracy of sessionizers for web usage analysis, in Workshop on Web Mining at the First SIAM International Conference on Data Mining, pp. s 7-14, 2001.



410 Data Mining IV

- [8] Chakrabarti, S., Berg, M. & Dom, B., Focused crawling: a new approach to topic-specific Web resource discovery, in Proceedings of 8th International World Wide Web Conference, vol. 31, n. 11-16, pp. 1623-1640, 1999.
- [9] Tan, P.N. & Kumar, V., Discovery of web robot sessions based on their navigational patterns, *Data Mining and Knowledge Discovery*, 6: 9-35, 2002.
- [10] Sweiger, M., Madsen, M., Langston, J. & Lombard, H., *Clickstream Data Warehousing*, Wiley Computer Publishing, John Wiley & Sons, Inc., 2002.
- [11] Kohavi, R., Brodley, C., Frasca, B., Mason, L. & Zheng, Z., KDD-Cup 2000 organizers' report: Peeling the onion, *SIGKDD Explorations*, 2(2):86-98, 2000.
- [12] <http://www.ecn.purdue.edu/kddcup>.