Clickstreams, the basis to establish user navigation patterns on web sites

R. Alves, O. Belo, F. Cavalcanti & P. Ferreira
Department of Informatics, University of Minho, Portugal

Abstract

Collecting and mining clickstream data from e-commerce sites has become increasingly important for marketing, advertising, and traffic analysis activities. Organizations are promoting many initiatives concerning user's navigation pattern discovery, in order to implement better sites, more functional and close to customers' needs. Basically, the main idea is to provide more quality of attendance in their sites, and, consequently, get more profitability. However, clickstream processing is not a simple task. The sequences of clicks are very difficult to handle using conventional techniques, essentially due to their diversity and nature. They include a lot of aspects that reveal the multidimensional perspective of web data. OLAP technology provides today the means and techniques to represent, store and analyse such kinds of multidimensional data. However, it does not offer discovery driven analysis to support traversal pattern identification processes on web sites. Mining traversal pattern techniques can be applied in conjunction with OLAP as an integrated alternative for understanding those particular sequences of clicks. In this paper we present an integrated OLAP and mining approach specially conceived for exploring user navigation patterns based on clickstreams. We also describe the multidimensional structure provided for modelling click sequences and the OLAP operations and mining techniques that can be pushed over data cubes to bring up navigation patterns.

1 Introduction

The concepts and techniques of data mining and knowledge discovery could be applied efficiently on web sites. Furthermore, this specific application of data mining called Web Mining, has taken much attention of researchers and

Data Mining V, A. Zanasi, N. F. F. Ebecken & C. A. Brebbia (Editors)

companies [1]. Besides, from Web Mining, new research areas were derived to guide the solutions to its specific needs. In short, some researchers have worked on mining the content of a web site (web content mining), others have decided to study the structure of a web site (web structure mining), and finally, some of them have analyzed the usage of a web site (web usage mining).

Web usage mining has attracted much attention recently from research and e-business professionals and it offers many benefits to an e-commerce website, namely: targeting customers based on usage behaviour or profile (personalization); adjusting web content and structure dynamically based on page access pattern of users (adaptive web site); enhancing the service quality and delivery to the end user (cross-selling, up-selling); improving web server system performance based on the web traffic analysis; or identifying hot areas of a web site. Many successful data mining systems can handle very large data files like clickstreams [2][3]. However, we have seen few efforts on a systematic study and development of data warehousing and mining systems for knowledge extraction on clickstreams. Although, among many different paradigms and architectures of data mining systems, Analytical Data Mining/On-Line Analytical Mining – OLAM (also called OLAP Mining), which integrates on-line analytic processing (OLAP) with data mining and mining knowledge in multidimensional databases, is a promising direction [4]. As a research effort on analytical data mining and web usage mining, in this paper we present an integrated OLAP and mining architecture specially conceived for exploring user navigation patterns on clickstreams. The purpose is providing guidelines to build a profile of web users.

2 Clickstreams

As the web grows and organizations move some or all of their business to the web, the opportunity for logging and analyzing user's navigation path grows. Though, before any kind of analysis, it is necessary to pay attention on what kind of actions can be done and which information can be gathered from clickstreams to achieve effective actions. An interesting use of clickstream information is personalization of the site to the individual user's desires [7], e.g., delivering services and advertisements based on user interest and, in that way, improving the quality of user interaction and leading to higher customer loyalty. Customer interests are based on what a user have looked at on the web site and especially which sections of the site a user visits and spends some time in. One of the design goals of most sites is to have high stickiness, meaning that users spend a long time productively on the site. A way to achieve this is by identifying pages that often lead to undesired termination of user sessions - the so-called killer pages.

The data needed to fulfil the mentioned tasks is derived from web server log files, probably supported with cookies. From web server log file in Common Log Format (CLF), for each click on the web site, we know the IP-address of the user, the page (URL) the user requested, and the timestamp of the event. From these single facts, we can obtain extra information. First, we can identify

Data Mining V, A. Zanasi, N. F. F. Ebecken & C. A. Brebbia (Editors)

individual users, through IP-address, and through it the users' origin. We can identify users sessions, including start page, end page, and all pages visited between those two. Basically, an user session is an ordered sequence of clicks made by the same user, starting from the click where the user enters the site until the last click registered in the log for that user during that visit. Nevertheless, extracting user sessions from web server logs, especially in the lack of cookie information, is a difficult task and we will not address it here. We shall assume that the sessions have already been extracted either using cookies or by some reasonable heuristics [8].

3 An Architecture for mining clickstream data

As shown in Figure 1, our architecture was designed around three main issues on web usage mining: filtering, discovering and exploring process [5]. The filtering process takes place in the beginning of the whole process. First, clickstreams are gathered from web servers and then put into a staging area (Relational Database). Next, some heuristics are applied for identification of crawlers. We don't get further on crawler's detection in this paper, so additional information can be obtained in [6]. Clickstream analysis should demand multidimensional analysis. How can we get interesting insights over the clickstream by analysing its attributes (session, sequence, time, user agent, etc.) in an isolated manner? Thinking multidimensional we set up a data cube for getting better comprehension of the data. The data cube is the basic structure to allow exploration analysis (OLAP mechanisms), and also for discovering activities (mining techniques).

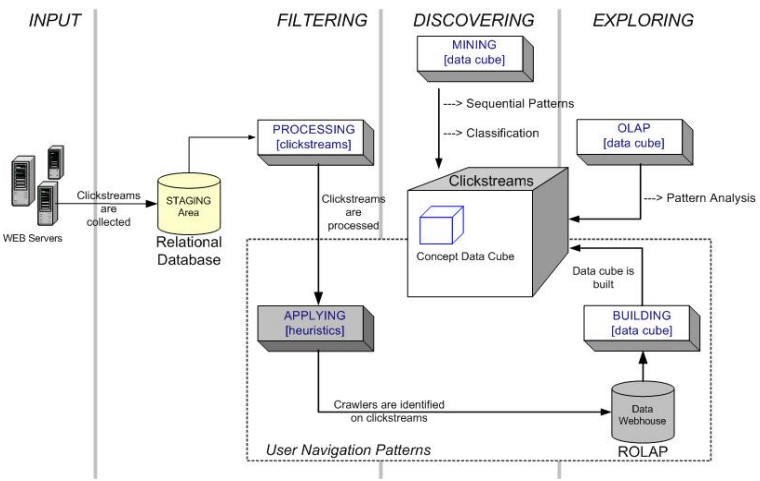

Figure 1: An overall perspective of the architecture.

Data Mining V, A. Zanasi, N. F. F. Ebecken & C. A. Brebbia (Editors)

The data cube structure is defined adding dimensions with its specific hierarchies to aggregate and compute the measures (such as visit count). This modelling facilitates the navigation inside the clickstreams. Besides, a sliced data cube (concept data cube) is available for classification and sequence analysis. The concept data cube plays an interesting role making available the concept sequence path of each session, the first concept visited, the last concept, the most longer, and so far. If paths are considered at the webpage level of resolution, paths tend to have very little similarity with one another. This is because, at such a high resolution, there are very few exact webpage matches between the paths. Thus, webpages can be first grouped into categories (that we call concepts) based on suitable analytics (OLAP mechanisms) and metadata information. When we convert the raw paths to concept-based paths, the average size of the path reduces, and we get paths which can be easily understood. Also, those concepts are aggregated by merging successive concepts (Table 1).

Table 1: Original paths converted to concept-based paths.

original path	concept-based path
main/home/.jhtml	Main
main/leg_news/.jhtml	-
products/productDetailLegcare/.jhtml	products
articles/new_shipping/.jhtml	Articles
main/home/.jhtml	Main
main/shopping_cart/.jhtml	-
account/credit_info/.jhtml	account
checkout/confirm_order/.jhtml	checkout
checkout/expressCheckout/.jhtml	-
main/registration/.jhtml	Main

From the concept data cube structure (Figure 2) it is also possible to keep a track of users distinguishing characteristics such as user agent, referrer and session duration. Additionally, it was introduced the notion of "concept", which is intended to group pages based on their template, generating information such as the start concept, the end concept, the sequence length, and the number of unique concepts associated to a given session.

4 User navigation pattern analysis

The case study selected is about a small dot-com company called Gazelle.com, a legwear and legcare retailer, which agreed to volunteer their data to the KDD-Cup 2000 [9]. In terms of background information there is to say that the home page contained more than 70 images (making modem-based downloads extremely slow), there were many promotions (affecting both the traffic to the site and the type of users it received), a TV advertisement was ran at prime-time and their registration form significantly changed at a given moment in time. In Table 2, some statistics about data volume, session length and duration, among others, are provided.

After taking the firsts steps into the architecture to achieve the data cube structure, it is possible to apply OLAP analysis for getting better comprehension

Data Mining V, A. Zanasi, N. F. F. Ebecken & C. A. Brebbia (Editors)

of the information available. Further, mining techniques can be pushed over the data cube for classification and sequence analysis. From these mining techniques, we can characterize the user' sessions by getting the order of the concepts visited (concept path), the frequent concepts accessed and associates these results with a classification (for instance, a decision tree technique) in order to obtain patterns (rules) for personalization tasks.

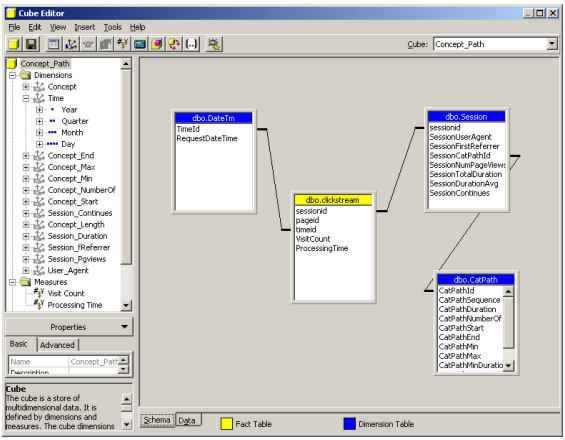

Figure 2: The data cube modelled with Microsoft Analysis Services Manager.

Table 2: Some statistics about the dataset.

Number of Records	742.338
Number of Sessions	234.954
Number of Distinct Referrers	36.737
Number of Distinct User Agent	8.241
Number of Unique concepts	6
Minimum Sequence Length	1
Maximum Sequence Length	137
Minimum Number of Session Page Views	1
Maximum Number of Session Page Views	5.487
Minimum Session Duration (ms)	0
Maximum Session Duration (ms)	90.236

4.1 OLAP mining

OLAP Mining provides such mechanisms in order to allow interchangeably data mining techniques and OLAP operations [4] over multidimensional databases. The architecture discussed in section 3 provides such facilities providing data

mining over a specific dimension or over a data cube (Figure 3), and integrates the results achieved in a virtual cube (a multidimensional view).

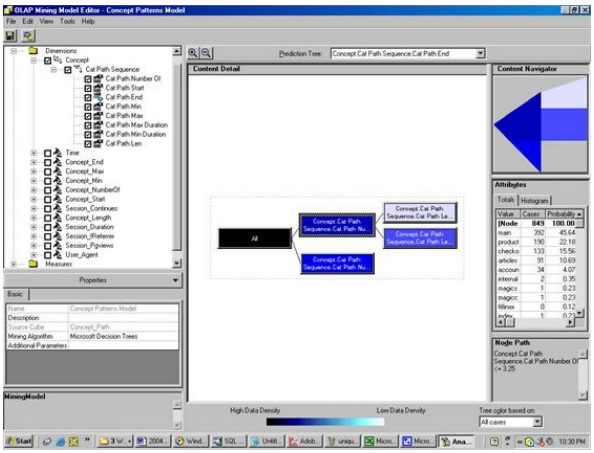

Figure 3: The OLAP mining model editor in Microsoft Analysis Manager.

As we are working with concepts, it is necessary to obtain information (metadata) related to each concept with the aim of the enrichment of that kind of analysis.

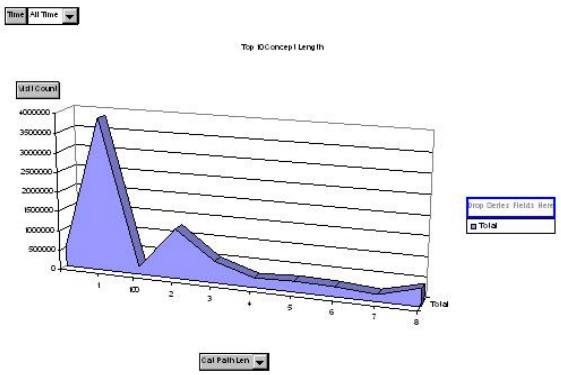

Figure 4: Top 10 most frequent sequences (concept length).

In Figure 4 we can view the distribution of the top 10 frequent concept length. Like it is presented, there are a lot of sessions with length equals to one, and the remains distributions share lengths between 2 and 5. Perhaps those sequence with length equal to one point out crawlers or uninterested user. But, for e-commerce purposes it is very hard to deal with sequence like that. So we can not

pushed any forward analysis on these sequences. On the other hand, it is necessary to inspect how much unique concepts are on each sequence. For this specific case study, there are around 5 most significant concepts which are: main, articles, product, account and checkout. However, the number of concepts in the root level of the webpage ranges from 1 to 6. It is important to say that only the main concept has around 30 sub categories. Besides almost web sessions start on the main, it doesn't mean that it is start at main/home. Further, sequence analysis will be confirmed this assumption.

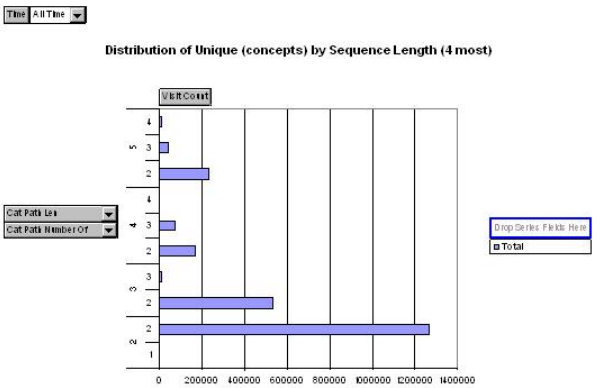

Figure 5: The distribution of the unique concepts by the four most relevant sequence paths.

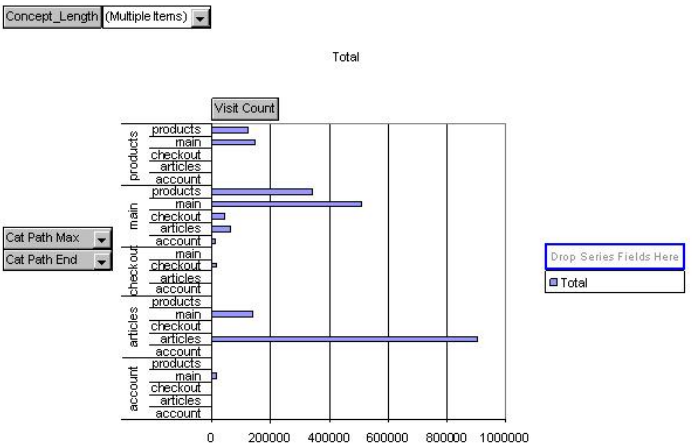

Figure 6: The max concepts duration (ms) and associated end concepts.

We can reach an insight about the unique concepts inside sequences with length between 2 and 5 (Figure 5) if we slice the data cube. This makes clear that

the most web sessions point to two concepts in their sessions which means even the concept length equals to five, the user navigates only in two unique concepts of the website.

In Figure 6, we can get an interesting picture by associating the last concept visited with the concept that the user spend more time in. If the user spends more time on the concept articles he also ends its session on that concept. This points out another insight of users who just wondering about the website.

4.2 Application of decision trees on concept data cube

In order to validate the importance of the concepts for profiling user sessions, it was selected some interesting rules by applying decision trees on the concept data cube [10] (Figure 3). The class used for that analysis was the *session continues* attribute which is available in the KDD Cup dataset for classification purposes. The idea behind that classification is to predict whether a user leaves or not the website. From this point we can analyse the importance of using a concept approach for profiling user sessions, in sense that its possible to see the importance of its attributes on the decision tree results. For instance:

- (sessions with concept length >13 concepts) and (unique concepts between 2 and 3 concepts) and (the last concept not equal to checkout), bend to leave the website. (89,89%);
- (sessions with concept length <=13 concepts) and (unique concepts between 2 and 3 concepts), bend to keep on the website. (47,01%);
- (sessions with concept length >=10 concepts) and (unique concepts between 3 and 4 concepts), bend to leave the website. (91,26%).

4.3 Sequence mining

Sequence mining is the task of finding temporal patterns over a database of sequences, in this case a data base of click streams. Sequence mining is considered to be an extension of association mining that only finds non-temporal patterns. In this context a sequence pattern is considered to be frequent if it appears in a number of database sequences greater or equal than a user defined threshold value, called minimum support. In this work we have used an AprioriAll [11] based implementation, that reports maximal sequences, i.e. sequences that are not contained in any other sequence. We have started by mining the concept paths (sequences) dimension at the root level, but it resulted in very generic frequent sequences which gave us a small insight. Then we used the webpage level (sub-categories) and analyzed it in a concept level, which resulted in more specific, interesting and surprising patterns. Different minimum support values were tried out in a decreasing order. Significant patterns started to be outputted only for values less or equal to 2%. When analyzing the reported maximal sequences for 1% (832 seq.) and 2% (1664 seq.) values of minimum support, we came to the following evidences:

- Only a very few sequences will not start in the Main concept level;
- Smaller sequences, typically 2 to 4 page visits, correspond to 1 (Main) or at most 2 (Main & Products) navigation concepts. This may indicate a relatively interested user that wonders in the website;

- We discover a curious set of small sequences (length = 2), ending with page Main/freegift;
- Longer sequences correspond to effective buys navigations. The concept path for these sequences can be described as: Main -> Products -> Main\shoppcart -> Checkout -> Main\registration -> Main. The buy is confirmed by the action of adding products to the shopping cart, followed by an order confirmation in the Checkout concept pages, than the user authentication/registration and finally logging out and returning to the Main pages;
- A set of small sequences in the form of: Main -> Account -> Main indicates the presence of a potential buyer that doesn't buy but creates an account in the site.

Taking into account the firsts impression in the Section 4.1, and adding the previous evidences, we can conclude that the users of this website can be divided into three main categories: *effective buyers* (supported by evidence 4), *potential buyers* (evidence 5) and *wonderers* (by evidence 2 and 3).

5 Conclusions

The application of OLAP mining to clickstream data provides us the basis to explore it in several dimensions. By applying OLAP operations it was easy to summarize and aggregate the concept data cube for better comprehension of the clickstream data. Selecting some interesting rules by applying decision tree on the concept data cube, we could confirm the importance of using a concept approach to build an users' profile. For instance, it was possible to predict killer sessions according to some characteristics of the concepts, like the concept length, the unique concepts, the max concept duration, etc. The fact that the site has a relative deep navigational depth and that the pages are dynamically generated makes the application of sequence mining very appropriate. Some differentiated and surprising sequence patterns were obtained which helped us to conclude that the website is visited by three classes of users: effective buyers, potential buyers and wonderers. For example, the effective buyers could be achieved by analysing the long concept paths. This information is essential when implementing recommendation systems in the websites. As future work, we plan to extract rules from the previous results in order to make websites more adaptive to users' intentions.

References

[1] Schmitt, E., Manning, H., Paul. Y., & Roshan, S., *Commerce Soft-ware Takes Off,* Forrester Report, 2000.

[2] Fayyad, U. M., Piatetsky-Shapiro, G. , Smyth, P. & Uthurusamy, R., *Advances in Knowledge Discovery and Data Mining*. AAAI/MIT Press, 1998.

Data Mining V, A. Zanasi, N. F. F. Ebecken & C. A. Brebbia (Editors)

[3] Chen, M. S., Han, J. & Yu, P. S., Data Mining: An overview from database perspective. *IEEE Trans. Knowledge and Data Engineering*, 8:866-8883, 1996.

[4] Han, J., Towards on-line analytical mining in large databases. *ACM SIGMOD Record*, 27:97-107, 1998.

[5] Srivastava, J., Cooley, R. , Deshpande, M. & Tan, P-T., Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data. *To appear in SIGKDD Explorations*, (1) 2, 2000.

[6] Pang-Nin, T. & Kumar, V., Discovery of web robot sessions based on their navigational patterns. *Data Mining and Knowledge Discovery*, 6:9--35, 2002.

[7] Baglioni, M. , Ferrara, U., Romei, A., Ruggieri, S. & Turini, F. , Preprocessing and Mining Web Log Data for Web Personalization, *Proc. of the 8th Italian Conf. on Artificial Intelligence* : 237-249. Vol. 2829 of LNCS, September 2003.

[8] Berendt, B. , Mobasher, B., Spiliopoulou, M. & Nakagawa, M., A Framework for the Evaluation of Session Reconstruction Heuristics in Web Usage Analysis, *INFORMS Journal of Computing, Special Issue on Mining Web-Based Data for E-Business Applications* Vol. 15, No. 2, 2003.

[9] Kohavi, R., Brodley, C. E., Frasca, B., Mason, L. & Zheng, Z., KDD-Cup 2000 organizers' report: peeling the onion, *ACM SIGKDD Explorations*, Volume 2, Issue 2, Special issue on “Scalable data mining algorithms”, Pages: 86 - 93, December 2000.

[10] Seidman, C., *Data Mining with Microsoft SQL Server 2000*, Technical Reference: Microsoft Press, 2001.

[11] Agrawal, R., & Srikant, R., Mining Sequential Patterns. *Proc. of the 11th Int'l Conference on Data Engineering*, Taipei, Taiwan, 1995.

Data Mining V, A. Zanasi, N. F. F. Ebecken & C. A. Brebbia (Editors)