Visual Mining of Web Logs with DataTube2

Florian Sureau[1], Frederic Plantard[1], Fatma Bouali[2,1], and Gilles Venturini[1]

[1] Université François Rabelais de Tours, Laboratoire d'Informatique,
64 Av. Jean portalis,
37200 Tours, France
venturini@univ-tours.fr
[2] Université de Lille2, IUT, Dpt STID,
2527, Rue du Maréchal Foch
59100 Roubaix, France
Fatma.Bouali@univ-lille2.fr

Abstract. We present in this paper a new method for the visual and interactive exploration of Web sites logs. Web usage data is mapped onto a 3D tube which axis represents time and where each facet corresponds to the hits of a given page and for a given time interval. A rearrangement clustering algorithm is used to create groups among pages. Several interactions have been implemented within this visualization such as the possibility to add annotations or the use of a virtual reality equipment. We present results for two Web sites (1148 pages over 491 days, and 107 pages over 625 days). We highlight the actual limits of our system (9463 pages over 153 days) and show that it outperforms similar existing approaches.

1 Introduction

Web usage mining (WUM) [1] is a challenging problem for data mining methods because it consists of analyzing large amount of complex and time-dependent data, and often with the constraint of presenting the results to non-expert in data mining. Among the WUM methods [2], we have concentrated our attention on those that involve the so-called Visual Data Mining (VDM) domain: these methods use data visualizations and interactions to let the user discover useful knowledge in an intuitive and interactive way [3] [4] [5]. They also facilitate the presentation of results to other people.

The remaining of this paper is organized as follows: section 2 describes existing VDM approaches for WUM and positions our work with respect to these approaches. Section 3 presents our approach, called DataTube2, and more precisely the organization of the visualization and the visual encoding of the data, the use of a rearrangement clustering algorithm, and finally, the graphical interactions (selection, annotations, etc) and their implementation in a virtual reality environment. Section 4 presents the results obtained on real Web logs. Section 5 concludes and proposes several perspectives on this work .

G. Vossen, D.D.E. Long, and J.X. Yu (Eds.): WISE 2009, LNCS 5802, pp. 555–562, 2009.

2 Existing Approaches: Visual Mining of Log Data

VDM methods applied to WUM should deal with a large amount of data and should take into account the complexity of the data and their temporal aspects. They should provide the user with easy to use and understand visualizations and interactions. Reaching these objectives is a challenging research task for VDM. One of the pioneers is probably Webviz [6] which displays a graph of web pages where links can be colored according to the visits to these pages. VisVIP [7] uses the same pages display principle but represents the web navigation as a curve, and the time spent on each page with a column. MIR [8] is one of the unique use of a metaphor to represent web logs: the web site is a city where each building represents a web page, and the users' navigation is represented by the moves of an avatar. Among the methods which can deal with the largest amounts of data, one must mention TimeTube [9] where the logs of a 7588 pages Web site have been represented (tree representation). This system has the advantage of representing the Web site structure, but its main drawback is that it gives time a minor role: only a few time instants are visualized. DataJewel [10] is another example which uses a calendar representation in conjunction with a pixel-based visualization (each day of the calendar is filled with pixels that represent access to pages for instance). Calendar representation is easy to understand for the user. However, the filling of the calendar does not help the user to perceive the absence of hits or pages with similar behavior. One must also mention the use of Kohonen's Self-Organizing Maps [11]) where pages are clustered together according to their co-occurrence in users navigation. Our method belongs to this last kind of methods where a priority is given to the amount of data and to knowledge discovery using a clustering algorithm.

Finally, one must mention basic plots and graphs of Web usage data as provided by commercial or industrial tools, like for example Google Analytics. Those tools are very useful for many web sites: they may trace the activity of a page, the global activity of a site, the origin of users, etc. When many pages are considered, such standard tools cannot provide for instance a graph that include the activity of all pages (such as TimeTube for instance). In addition, if one wishes to highlight additional information (pages with similar activities), then such functionalities are not provided by standard tools because the use of a clustering algorithm would be time-consuming.

3 DataTube2

3.1 Definition of the Visualization

From a general point of view, we consider that the temporal data to analyze is described with n attributes over k time steps (or intervals). These n attributes (denoted by $A_1, .., A_n$) are supposed to be numeric. Several scales can be used for the $t_1, ..., t_k$ time steps (e.g. hours, days, months, etc). The input data of our method can thus be represented as a $n \times k$ matrix where $A_i(t_j)$ denotes the value of the i-th attribute at time-step (or interval) t_j. As far as WUM is concerned,

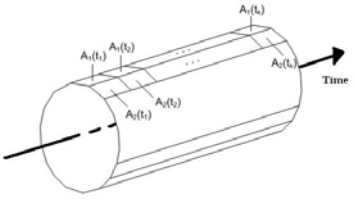

Fig. 1. Definition of the temporal tube

each attribute A_i represents for instance the number of hits of a page P_i of the considered Web site.

DataTube2 is initially based on Mihael Ankerst's DataTube visualization [12] that we have significantly extended in order to fit WUM requirements. This includes the addition of new visual elements and new interactions, the integration of a clustering algorithm, and the ability to deal with much larger amount of data (the initial version was apparently limited to $n = 50$ and $k < 100$). The visualization in DataTube2 uses a 3D temporal tube as shown in figure 1 where the tube axis represents the time flow and where each attribute value $A_i(t_j)$ is represented by a rectangular facet. A time-step t_j is thus represented as a ring in the tube, while the evolution of a given attribute over time is represented as a line which is parallel to the tube axis. We have defined several ways to visually encode the Web usage data: with colors (three values are defined by the user: minimum, intermediate, maximum), with the width of the facets, or with the height of the facets (w.r.t. the tube axis). These different modes can be combined together. Missing values (which may reflect the lack of event in the considered time interval) are represented in a default color (black, for example, in all of our visualizations).

We have added in DataTube2 an explicit time axis in order to give a scale for the time flow and to help the user in locating events over time. This axis takes the form of a "path" consisting of slabs, where each slab corresponds to a time-step t_i. This path is placed inside the tube, below the tube axis. The slabs are transparent in order to enable the perception of data located below them. In addition, a text label periodically indicates to which time-step a slab corresponds to. To highlight a given time-step, one can select it by clicking on the corresponding slab. Finally, the user can add annotations on the time axis.

3.2 Clustering Algorithm

Since many pages may have a similar activity, it is very important to help the user to visually detect those groups of pages, as well as other information about them like their size or the reason why they form a group. This greatly improves the clarity and the usefulness of the visualization. For this purpose, we have used a simple rearrangement clustering algorithm where similar attributes (i.e. with similar temporal behavior) are displayed next to each other in the tube. Many methods of such reorganization exist, including recent work like [13]. Here we

use a classical and popular method for matrices reorganization, the Bond Energy Algorithm [14]. Its complexity is polynomial and we show in section 4 that it performs well.

3.3 Interactions

Regarding the navigation in the visualization, the user is initially placed at the tip of the tube axis and he faces the inside of the tube (see figure 2 for instance): he obtains a global view of all data, exhibiting for example, major trends, groups of similar pages (see previous section) and missing data (pages which did not exist yet, or pages which disappeared). Obtaining a zoom is achieved by "perspective effects" (data located far away from the user appears smaller and with less details than data close to the user) and by the user's moves: the sides of the tube are close to the central axis, which allows the user to quickly reach them and to locally observe attributes with more details.

Regarding the interactive selection of data, each facet is clickable. A left click triggers the display of the attribute name (page name for instance), the considered time-step and the value $A_i(t_j)$, in the upper right corner of the screen. A right click also enables to dynamically add annotations on a facet. Thus the user can store notes or "landmarks" related to the discovered knowledge and mark a special event in order to better observe its causes or consequences (for instance, an "advertising campaign" and its successive effects on the pages activities). The user can also use these annotations for the interactive presentations of data and extracted knowledge. These annotations can be either a graphical element such as an image chosen by the user or a visual marker (sphere). One can associate with these annotations a link (URL), a sound or a text displayed in a separate window.

DataTube2 can be run on a standard computer with a 2D screen. However, the perception of the third dimension is important in this tubular representation because it greatly helps the user to perceive the time flow (tube axis). So we have therefore develop the possibility to run DataTube2 in a virtual reality environment called VRMiner [15].

4 Results

We have applied DataTube2 to several logs obtained from real Web sites as shown in table 1. Our tests were performed on a MacBook Pro (2.4GHz Intel Core Duo, 4GB of RAM).

Table 1. Logs from real Web sites used for our experimental tests

Web site	# values	n	k	Vis. time	Clust. time	Clust. efficiency
Polytech	563 668	1148	491 days	1.4s	22.5s	1.19
Polytech-Init	1 447 839	9 463	153 days	34s	8 min.	1.03
Antsearch	66 875	107	625 days	407ms	234ms	1.39

In table 1, we have evaluated the execution times of both the visualization and clustering algorithms. In this table, "Vis. time" represents the time needed to build the visualization and "Clust. time" represents the time needed to reorganize the visualization with BEA. The construction of the visualization is linear and thus fast compared to BEA which requires more time especially when the number of pages increases (quadratic complexity). However the global execution time is quite acceptable for the user. In this table we have also measured the efficiency of the clustering algorithm both in a quantitative and visual way. "Clust. efficiency" is thus the ratio between the sum of similarities of adjacent attributes in the final reorganized visualization and the same sum but for the initial visualization. Obviously the clustering step reorganizes the visualization by placing pages with similar behavior next to each others.

We present now typical results and knowledge which can be visually extracted with our tool. In all visualizations, we have visually encoded the number of hits per day with a color ranging from green (low number of hits) to red (high number). First, the user initially has a global view of all log data, such as those represented in figure 2. On top of this figure, pages are ordered around the tube according to their date of creation, which results in a spiral-like shape. One immediately perceives the pages which, after their creation, do not receive much attention like the pages labeled "1" in the figure. It is also possible to detect the periods where no pages have been added to the site (see the area labeled "2" in the figure) or where many pages have been added (see area "3"). The area "4" corresponds to pages which, during a given time interval, received more attention than the others.

In the bottom of figure 2, the clustering algorithm has been used to group together pages with similar activities. One notices the many differences between the top (sorted by date of creation) and bottom pictures. The clustering algorithm groups together pages which were 1) created at the same time (period with no activity) and 2) which, once created, behave similarly. Many such groups can be detected (see for instance the groups labeled "A"). Then, more specific groups can be found. Group "B" corresponds to pages which, after a period of high activity, where not viewed anymore, and then viewed again. The webmaster may detect in such a way pages which are not accessible for some time. Finally, we highlight another group "C": this group can be divided into two subgroups which share important similarities and which were placed next to each others in the visualization.

We have let the Webmasters test DataTube2 and we report here the obtained comments. They were not aware of such a visualization tool before and they quickly learn its use. The time needed to explain the characteristics of the visualization and the interactions was short thanks to the tubular shape which can be easily explained and understood. These people were easily able to highlight the above-mentioned information and they especially appreciated being able to have a global view of the data. They were able to give name to groups of pages (i.e. "News", "Press", "Gallery", "Courses", etc) by recognizing specific clusters of pages with similar behavior, and were able to understand the shown behavior

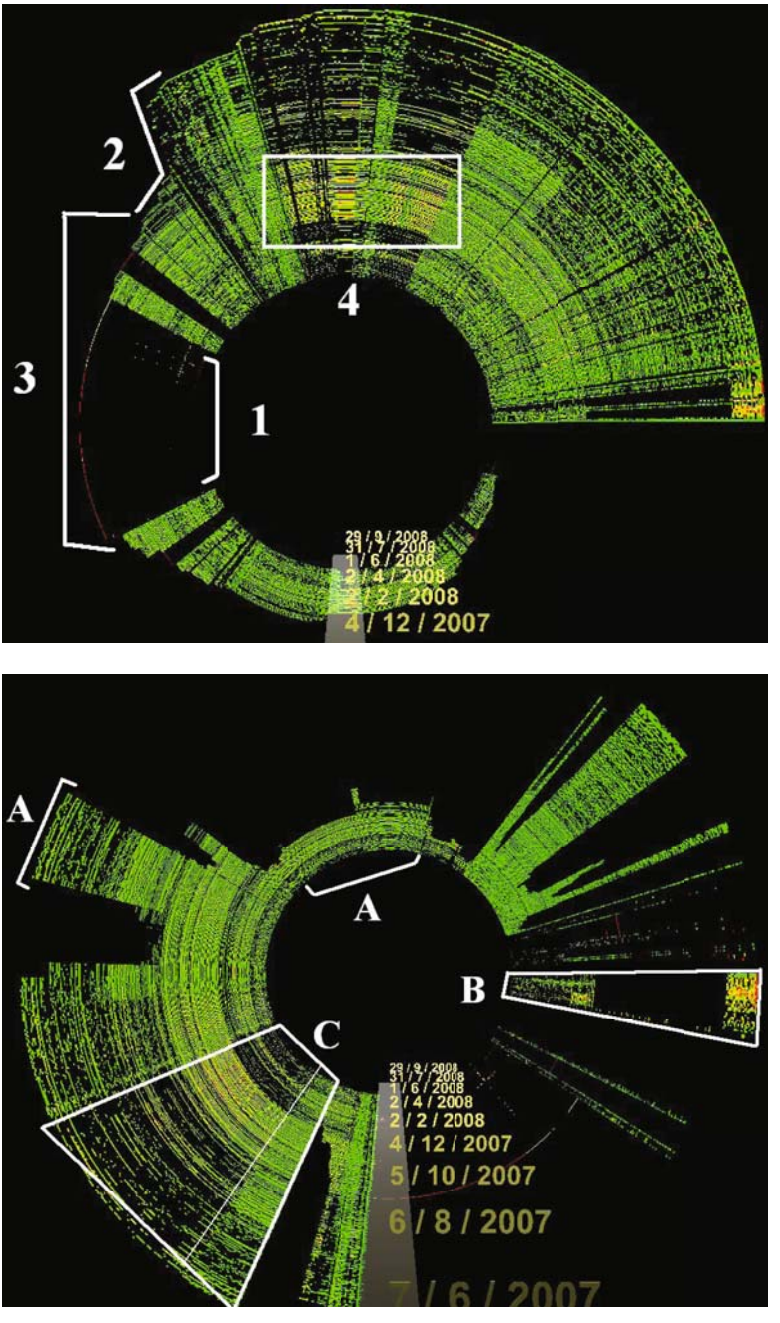

Fig. 2. DataTube2 typical visualizations: the first visualization (top) corresponds to the ordering of pages by date of creation, while the second one (bottom) is obtained after running the clustering algorithm (see text for explanation)

(like for instance pages dealing with "Courses" which often change). For the Polytech logs, the Webmaster has studied the influence of the end of the school period, the holidays and the beginning of the school on the site visits.

Finally, we have tested the actual maximum capacity of DataTube2 with a large amount of pages. In this case, pages have been ordered according to their date of creation, which results in a spiral that represents the history of the Web site since its creation. The activity of all pages is globally represented. We have a total amount of 9463 pages over 153 days. However, the interactions are limited because the frame rate of the display is too slow (about one per second). So this visualization should be considered as static, i.e. the user cannot easily move around. As mentioned in the next section, DataTube2 outperforms previous visualization of Web logs, both in terms of number of pages or time steps.

5 Discussion and Conclusion

We have presented here a new VDM method and its application to WUM. We have tested it on different web sites with hundreds of pages. Webmasters easily understood its functionalities. The clustering algorithm improves the visualization by showing similarities on pages access. So webmaster have a complete vision of the activity over time and are able to perceive at least the following information: the number of visits to web pages, the areas of a site which are difficult to reach and which receive a small number of hits, important times of the year, the week or the day (when an hour scale is selected). The user is also informed on groups of pages with similar activities. He may obtain a global view but also details on any part of the data. Moreover, we have implemented DataTube2 in a virtual reality in order to allow the user to efficiently explore and analyze the temporal aspect of the data.

Compared to industrial tools such as AWStats or Google Analytics, our visualization allows the user to perceive at once the activities of many more pages, as well as groups of pages with similar activities. We have shown that our implementation of DataTube may visualize about 1.500.000 values with a log of 9463 pages over 153 days (these numbers are higher than the values mentioned in the state of art of visual methods, but also for other non visual approaches [16]). If ones compares our results to those obtained in the VDM literature (see section 2) which, as far as we know, are the only one to visually handle large volumes of log data, one must notice that Datatube2 visualizes at least as many pages and many more time-steps than TimeTube [9] (limited to a few time steps), and many more pages than DataJewel [10] (limited to a few pages).

We are currently adding more interactions and graphical requests in the visualization, and we are preparing the visualization of other information such as users' sessions. In this case attributes would be the users and "time" steps would be the pages or groups of pages. We are also studying how to represent the structure of the Web site within the tubular representation.

References

1. Srivastava, J., Cooley, R., Deshpande, M., Tan, P.N.: Web usage mining: Discovery and applications of usage patterns from web data. SIGKDD Explorations 1(2), 12–23 (2000)
2. Facca, F.M., Lanzi, P.L.: Mining interesting knowledge from weblogs: a survey. Data Knowl. Eng. 53(3), 225–241 (2005)
3. Cleveland, W.S.: Visualizing Data. Hobart Press, New Jersey (1993)
4. Shneiderman, B.: The eyes have it: A task by data type taxonomy for information visualizations. In: IEEE Visual Languages. Number UMCP-CSD CS-TR-3665, College Park, Maryland 20742, U.S.A., pp. 336–343 (1996)
5. Wong, P.C., Bergeron, R.D.: 30 years of multidimensional multivariate visualization. In: Scientific Visualization — Overviews, Methodologies and Techniques, pp. 3–33. IEEE Computer Society Press, Los Alamitos (1997)
6. Pitkow, J., Bharat, K.: WEBVIZ: A Tool for World-Wide Web Access Log Visualization. In: Proceedings of the First International World Wide Web Conference, May 1994, pp. 271–277 (1994)
7. Cugini, J., Scholtz, J.: VISVIP: 3D visualization of paths through web sites. In: Proceedings of the Tenth International Workshop on Database and Expert Systems Applications, pp. 259–263 (1999)
8. Kizhakke, V.: MIR: A tool for visual presentation of web access behavior. Master's thesis. University of Florida (2000)
9. Chi, E., Pitkow, J., Mackinlay, J., Pirolli, P., Gossweiler, R., Card, S.: Visualizing the evolution of web ecologies. In: Proceedings of the Human Factors in Computing Systems, pp. 400–407 (1998)
10. Ankerst, M., Jones, D., Kao, A., Wang, C.: Datajewel: Tightly integrating visualization with temporal data mining. In: ICDM Workshop on Visual Data Mining (1996)
11. Benabdeslem, K., Bennani, Y., Janvier, E.: Visualization and analysis of web navigation data. In: Dorronsoro, J.R. (ed.) ICANN 2002. LNCS, vol. 2415, pp. 486–491. Springer, Heidelberg (2002)
12. Ankerst, M.: Visual Data Mining. PhD thesis, Faculty of Mathematics and Computer Science. University of Munich (2000) ISBN 3-89825-201-9
13. Climer, S., Zhang, W.: Rearrangement Clustering: Pitfalls, Remedies, and Applications. The Journal of Machine Learning Research 7, 919–943 (2006)
14. McCormick, W., Schweitzer, P., White, T.: Problem decomposition and data reorganization by a clustering technique. Operations Research 20(5), 993–1009 (1972)
15. Azzag, H., Picarougne, F., Guinot, C., Venturini, G.: Vrminer: A tool for multimedia database mining with virtual reality. In: Processing and Managing Complex Data for Decision Support, pp. 318–339 (2005)
16. Jin, X., Zhou, Y., Mobasher, B.: Web Usage Mining Based on Probabilistic Latent Semantic Analysis. In: Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2004), pp. 197–205 (2004)