WSEAS TRANSACTIONS on COMPUTERS

Issue 2, Volume 3, April 2004

ISSN 1109-2750 **http://www.wseas.org**

[14] C. Diaz, S. Seys, J. Claessens, B. Preneel Towards measuring Anonymity, *Proceedings of PET 2002, In Hannes Federath (Ed.), Designing Privacy Enhancing Technologies*, LNCS 2482, pp. 54-68, 2002.
[15] A. Serjantov, G. Danezis, Towards an Information Theoretic Metric for Anonymity, Hannes Federath (ed), *Designing Privacy Enhancing Technologies*, LNCS 2482, pp. 41-53, 2002.
[16] The Anonymizer. *http://anonymizer.com.*
[17] The Lucent Personalized Web Assistant. *http:// www.ProxyMate.com.*
[18] D. Chaum, Untraceable electronic mail, return addresses, and digital pseudonyms *Communication. ACM,* Vol. 24, No. 2, pp. 84–88.
[19] D.M. Goldschlag, M.G. Reed, P.F. Syverson, Onion Routing for Anonymous and Private Internet Connections. *Communications of the ACM*, Vol. 42, No. 2, pp. 39-41, 1999.
[20] P.F. Syverson, D.M. Goldschlag, M.G. Reed (May 1997) Anonymous Connections and Onion Routing. *Proceedings of the 18th Annual Symposium on Security and Privacy*, IEEE CS Press, Oakland, CA, pp. 44-54, May 1997.
[21] M.G. Reed, P.F. Syverston, D.M. Goldschlag, Anonymous Connection and Onion Routing. *IEEE Journal on Selected Areas in Communication*, Vol. 16, No. 4, pp.482-494, 1998.
[22] D.M. Goldschlag, M.G. Reed, P.F. Syverson, Hiding Routing Information, *Information Hiding*, R. Anderson (editor), Springer-Verlag LNCS 1174, pp. 137-150, 1996.
[23] S. Dolev, R. Ostrovsky, Xor-Trees for Efficient Anonymous Multicast and Reception, *ACM Transactions on Information and System Security*, Vol. 3, No. 2, pp. 63–84, 2000.
[24] M.J. Freedman, R. Morris, Tarzan: A Peer-to-Peer Anonymizing Network Layer, *Proceedings of the ACM Conference on Computer and Communications Security (CCS 9)*, pp. 193 - 206, 2002.
[25] M.K. Reiter, A.D. Rubin, Anonymous Web Transactions with Crowds, *Communications of the ACM*, Vol. 42, No. 2, pp. 66-92, 1999.

Web Access Data Visualization with Data Analysis Methods

NIKOS KOUTSOUPIAS
Department of Balkan Studies
Aristotle University of Thessaloniki
3rd km Florina-Niki, 83100 Florina
GREECE
nickk@auth.gr

Abstract: - In the course of a travel around the internet, users visit various host servers located all over the world and, inadvertently, place digital traces of their surfing habits. These hosts gather and register information about all users and their client computers for various reasons, such as security violation detection, frequency of page access and browser types. This paper shows how Correspondence Analysis and Hierarchical Classification, two popular data analysis methods can be utilized with web log statistics for the examination of internet user activities and access patterns. We plotted a web access data set from a university department on the first factorial plane and employed a clustering dendrogram to clarify group patterns. We believe that these two data analysis methods, when applied on web access data can aid both web designers and design theorists to attain a better picture on client habits and assess the usability and effectiveness of their design choices. Findings of such a methodology can also be applied in marking, risk management and customer relationship management.

Key-words: **Clustering, Factor Analysis, Web Access, Web Log Analysis, Web Log Visualization**

1 Introduction

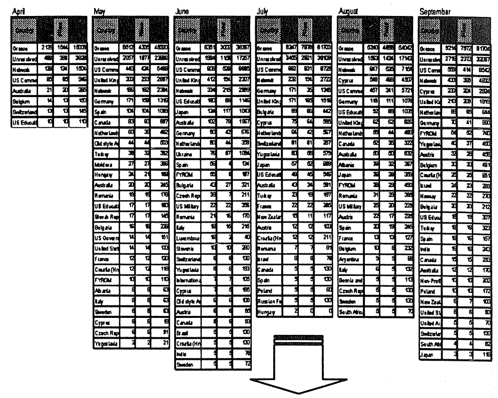
During the recent years, the fast increase of connections to the Internet has led to the availability of large quantities of web access log data.

Web logs are becoming an important area for determining various characteristics of user behavior and an increasingly popular approach in the direction of discovering such information through such web access files is that of Web mining. Overviews on research, demands and challenges of Web mining is given by various sources [24][25][28][29]. Web log detail and potential is beginning to draw the attention of data analysts and miners and finds supporters in fields such as e-commerce [2][21][27], marketing [5][6], library science [31], the media [23] and, of course, web design and engineering [4][8]. Also, interesting conclusions about web usage sequences and patterns can be drawn from various papers [22][26][30]. Further on, various real-world e-service applications require analyzing large volumes of transaction data to extract web access information. Such systems include WAV [14], WebQuilt [15], the system described in "The Scent of a Site" [8], i-Miner [1] and PVA [7]. The idea presented in this paper is to utilize two data analysis and visualization methods – correspondence analysis and hierarchical classification- on web log data files to present useful results for evaluation and let web administrators and designers identify access trends, take a comprehensive view of how users are accessing a site and answer significant questions with regard to site organization and content.

2 Methodology and Application

In a previous paper [17] Correspondence analysis (a factor analysis variation) was presented as a method for analyzing and interpreting access logs. This method along with clustering has also been utilized in various fields of computing [16][20]. We view that Correspondence Analysis [3][12 can be augmented by the Hierarchical Classification method [13](a variation of Clustering) -two multidimensional scaling methods that utilize factorial diagrams (or factor score plots) and dendrograms- in order to aid the interpretation of the analyzed phenomenon [3], the internet surfer's viewing preferences. In the outputs of these methods are included plots, absolute and relative contribution tables along with factorial axes, factorial panes and dendrograms coupled with their corresponding interpretation tables. We used a relatively new software implementation of the two methods as described in [18]. Methodologically, Hierarchical Classification and Correspondence Analysis are not based on a canonical distribution or any other theoretical distribution, while their output is characterized by the absence of restrictive technical prerequisites, which in the case of classical statistical methods lead in distorted results.

Both methods can be applied in any table of positive numbers driven from by web site access log. For the beginning, we will see the steps followed when applying Correspondence analysis, as demonstrated in [17]. In the case an administrator chooses to view the picture of visits per foreign country each month, we propose s/he uses log data relating to Country-Months with respect to Page Hits, number of Files and page size in Kilobytes (i.e. as drawn from the Webalizer, a log file statistical preprocessing package installed in the same internet server with the site in question). This means that, our data table will have N rows (of Country-Months) over a period of six months and three columns (Hits, Files, Kbytes). In our case, the time period under examination commences April 2001 and concludes September 2001, excluding Greek and Unresolved address visitor URLs, since, they comprise about 60% of the site's traffic. The resulting table will be a 138X3 data set (fig 1).



	Country-Month		Files	
1	Network-April	136	134	1504
2	US Commercial-April	85	85	946
3	Australia-April	21	20	265
4	Belgium-April	14	13	150
5	Switzerland-April	13	13	145
6	US Educational-April	10	10	110
:	:		:	
132	Poland-September	10	10	172
133	New Zealand (Aotearoa)-September	8	7	103
134	United States-September	6	6	93
135	United Arab Emirates-September	5	5	70
136	Switzerland-September	5	5	130
137	South Africa-September	4	4	82
138	Japan-September	3	3	118

fig 1: the process of the data table creation

The next step would be the categorization of the values of each variable (column) into three categories, following the convention that High values correspond to the upper 25% of the observations, Mid values to 50% and Low values to the rest 25% of the values in the entire data set. This process will allow for the immediate distinction of extreme values when comes to the interpretation of the factorial planes produced by the analysis. Further on, in order to achieve homogeneity in the data examined, each variable column will collapse into three (High-Mid-Low) and the values corresponding to each Country-Month will transform into bits of two 0's and a 1 for each variable triplet. Thus the dimensions of the final table for processing will be 138X9.

Based on the results produced by the application of Correspondence Analysis using the resulting table, three factorial axes incorporate 93,4% of the total information presented in the initial data table, which is a very satisfactory rate.

Bearing that in mind, we then use the factorial axes for the interpretation of the method's results.

The first axis separates extreme (High-Low) values while the second axis in the analysis deals with the average. When these axes are combined into on (factorial) plane, the analyst can draw useful conclusions just by looking at possible trend formations and groupings in the plotted data points. Each point corresponds to a Country-Month or a High, Mid or Low Value of each one of the parameters. In this way, both row and column points are depicted in a concise manner on a 2-dismesional space, allowing for quick estimation of general access patterns in the data.

In the first factorial plane (formed by the two first factorial axes) the Country-Month and Hits/Files/KB points are shown below (fig 3).

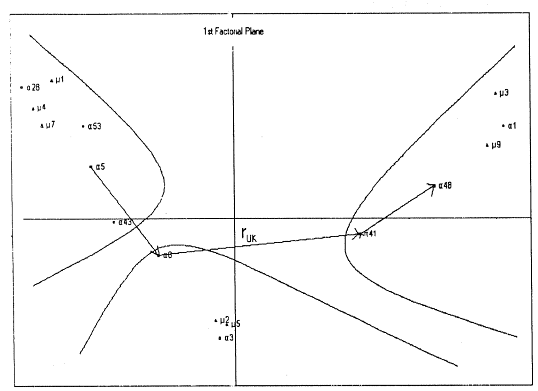


figure 2 : The first factorial plane

In the plot above, α1 to α138 corresponds to a separate country-month and μ1 to μ9 to a different variable category (μ1-μ3: HitsLow-HitsHigh, μ4-μ6: FilesLow-FielsHigh and μ7-μ9 KBLow-KBHigh). From this plot (along with a table containing points below the ones already depicted) it is easy to determine the general groups with regard to the access of each separate country in the period of 6 month. Furthermore, the 1st factorial plane allows for an estimation of the possible repositioning (i.e of hits coming from UK – line r_{UK}) of each Country among the High-Mid-Low monthly groupings.

Moreover, the method provides the means for a definition of the level of importance of contribution (COR) and quality of representation (CTR) of Country-Months and Hits/Files/Kb Variables, for all factorial axes. We choose the desirable (minimum) values of the above COR/CTR indices and so the factorial axes are reproduced carrying only the Country-Month and/or Hits/Files/KB points that satisfy the criteria (minimum values). The resulting axis (for rows and columns), using the example data and setting the value 100 as a minimum for both COR and CTR appears as follows:

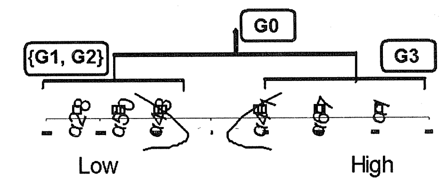


fig 3 : First Axis (criterion minCOR>=100)

Here, as in the factorial plots described above, the method provides information about the number of possible hidden (or overlapped) County-Month/Parameter points, as well as a detailed description of the visible points that cover the hidden ones. The produced factorial axes provide an easy way to distinguish the main trends in the data set.

At this point, we suggest the use of Hiererchical Classification that will help extensively to clarify the formation of groupings and determine exactly which points (Country- Month) are encompassed in each group. The process produces, along with the corresponding tables for the interpretation, a dentrogram of hierarchically produced break ups of the initial data set into distinct groups. Member points (Country-Months) in each one of these groups exhibit common characteristics and traits unique to each group. These characteristics can be brought forward through the study of the accompanying interpretation tables.

Consequently, in our example, the produced tree can be superimposed initially over the fisrt factorial axis (fig 4) and next over the first factorial plane, thus enhancing and clarifying the general view of the plot picture (see fig. 5 below).

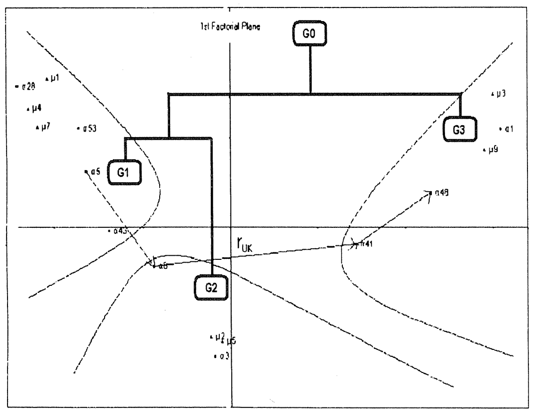


fig 5 : Superinposing the dendrogram over the first factorial plane.

As seen above, group G0 includes the population of all examined CoutryMonths. During the tree formation two groups emerge, G3 and the group comprised of {G1,G2} (fig 4), which in turn produces two separate ones, G1 and G2. With the aid of Classification interpretation tables, the analyst can then attach a descriptive title in each one of the produced groups in an effort to further refine the visual representation of the analyzed web access data set.

Following, the next step in Correspondence Analysis is the combination of any two axes formed during the previous phase. That is, the specification of the characteristics of factorial plots with points that satisfy the axis criteria indicated earlier. In consequence, the plot of the first factorial plane (for axes 1 and 2) encloses only those points that characterize that plane. Thus, it is made obvious which Country-Months form the main trends and relationships in the data set examined. In conjunction to this phase of the analysis, the classification method can isolate specific sub-groupings that contribute most in the interpretation of each formation, providing further feedback to the analyst.

3 Conclusion

The employment of hierarchical classification method as a supplement to correspondence analysis can be a very effective tool when exploring web access logs. One of the major advantages of using these methods is that they help to gain a deeper and crispier insight of the internet user browsing pattern.

A closer look on the merged visualization outputs of these analyses can shed light on the structure and grouping of resource providers irrespective to the type of web pages. Applying these techniques to access logs can unveil interesting access patterns that can be used to restructure sites more effectively, pinpoint advertising better, and target ads to specific internet users.

As further research we suggest the use of these data analysis tools to examine information patterns on preformatted data from discussion groups and online chats. In fact, online discussions could be a good way to discover knowledge, while internet users who are active in online chatting present an interesting target population for many organizations and researchers.

References:

[1] Abraham A., Wang X., i-Miner: A Web Usage Mining Framework Using Hierarchical Intelligent Systems, *The IEEE International Conference on Fuzzy Systems FUZZ-IEEE'03, pp. 1129-1134, 2003.*

[2] Ansari S., Kohavi R., Mason L., Zheng Z., Integrating E-Commerce and Data Mining: Architecture and Challenges, in *Proceedings of the 2001 IEEE International Conference on Data Mining (ICDM 2001)*, San Jose, California, USA Nick Cercone and Tsau Young Lin and Xindong Wu (eds), IEEE Computer Society, pp. 27-34, 2001.

[3] Benzecri, J.-P. *Pratique ed la Analyse des Donnees (T.1: Analyse des Correspondances, expose elementaire)* , Dunod, Paris, 1980.

[4] Berendt B., Web Usage Mining, site semantics, and the support of navigation, *WEBKDD 2000 Papers: Workshop on Web Mining for e-commerce - Challenges and Opportunities* , Boston, MA, Aug 20, pp. 83-93, 2000.

[5] Buchner A.G., Anand S.S., Mulvenna M.D., Hughes J.G. Discovering Internet Marketing Intelligence through Web Log Mining, *Proceedings of Unicom99 Data Mining & Datawarehousing: Realizing the full Value of Business Data* , pp. 127-138, 1999.

[6] Chang W., Yuan S., A Synthesized Learning Approach for Web-Based CRM , *in Proceedings of WEBKDD 2000 - Web Mining for E-Commerce -- Challenges and Opportunities* , Second International Workshop, Boston, MA, USA, 2000.

[7] Chen C., Chen M., Su Y., PVA: A Self-Adaptive Personal View Agent, *Journal of Intelligent Information Systems*, vol. 18, v.2-3, pp.172-194, 2002.

[8] Chi E.H., Pirolli P., Pitkow J., The Scent of a Site: A System for Analyzing and Predicting Information Scent, Usang, and Usability of a Web Site, *CHI Letters*, vol. 2, n.1, pp.161-168, 2000.

[9] Cockburn A., McEnzie B., What do Web users do? An Empirical Characterization of Web Use, *Journal of Human-Computer Studie*, n. 54, pp.903-922, 2001.

[10] Cooley R., Mobasher B., Srivastava J., Data Preparation for Mining World Wide Web Browsing Patterns, *Knowledge and Information Systems* , vol. 1, n. 1, 1999.

[11] Ed H. Chi, Rosien A., Heer J., LumberJack: Intelligent Discovery and Analysis of Web User Traffic Composition, *Proceedings of the 4th WebKDD 2002 Workshop*, at the ACM-SIGKDD Conference on Knowledge Discovery in Databases, 2002.

[12] Greenacre, M.. *Correspondance Analysis in Practice* , Academic Press, London, 1993.

[13] Hair J.F., Anderson R.E., Tatham R.L., Black W.C., *Mutlivariate Data Analysis* , Prentice Hall, N.J. 1998.

[14] Hao M., Garg P., Dayal U., Machiraju V., Cotting D.Visualization of Large Web Access Data Sets, *Proceedings of the symposium on Data Visualisation*, Barcelona, Spain , pp.201-ff, 2002 .

[15] Hong J., Heer J., Waterson S., Landay J., WebQuilt: A proxy-based approach to remote web usability testing, *ACM Transactions on Information Systems* (TOIS) , ACM Press, , Vol.19, n.3, pp.263-285, 2001.

[16] Huang J., Ng M., Ching W., Ng J., Cheung D., A Cube Model and Cluster Analysis for Web Access Sessions, WEBKDD 2001 - *Mining Web Log Data Across All Customers Touch Points* , Third International Workshop, San Francisco, CA, USA, Lecture Notes in Computer Science,R. Kohavi et. Al (eds), , vol. 2356, pp. 48-67, 2002.

[17] Koutsoupias N., Exploring Web Access Logs with Correspondence Analysis, *Proceedings of the 2nd Hellenic Conference on Artificial Intelligence* , pp. 229-236 Thessaloniki, 2002.

[18] Koutsoupias N., S-Pro: An Implementation of Data Analysis Methods, *Tetradia Analysis Dedomenon* , v.1, pp.118-128, 2002 (in Greek).

[19] Madria S., Bwomich S., Ng W.-K., Lim P., Research Issues in Web Data Mining, *Data Warehousing and Knowledge Discovery* , pp. 303-312, 1999.

[20] Mathews G.J., Towheed S. S., NSSDC OMNIWeb: The first space physics WWW-based data browsing and retrieval system, *Computer Networks and ISDN Systems* , n. 27, pp. 801-808, 1995.

[21] Mobasher B., Dai H., Luo T., Sun Y., Zhu J. Integrating Web Usage and Content Mining for More Effective Personalization, in *Proceedings of the Int'l Conf. on E-Commerce and Web Technologies (ECWeb2000)*, Greenwich, UK, 2000.

[22] Nanopoulos A., Manolopoulos Y., Mining patterns from graph traversals, *Data and Knowledge Engineering* , No. 37, pp.243-266, 2001.

[23] Nicholas D., Huntington P., Lievesley N., Withey R., Cracking the Code: Web Log Analysis, *Online & CD-ROM Review* , vol. 23, n. 5, pp. 263-269, 1999.

[24] Pabarskaite, Z. Implementing Advanced Cleaning and End-User Interpretability Technologies in Web Log Mining, *Proceedings of the 24th International Conference on Information Technology Interfaces ITI 2002* , Cavtat, Croatia, pp.109-114, 2002.

[25] Pal S., Talwar V., Mitra P., Web Mining in Soft Computing Framework: Relevance, State of the Art and Future Directions, *IEEE Transactions on Neural Networks*, pp.1163-1177, vol 13, n.5, 2002.

[26] Pei J., Han J., Mortazavi-asl B., Zhu H., Mining Access Patterns Efficiently from Web Logs, *ProcEEDINGS. of the 4th Pacific Asia Conf. on Knowledge Discovery and Data Mining* , pp. 396-407, 2000.

[27] Shahabi C. a Banaei-Kashani F., A Framework for Efficient and Anonymous Web Usage Mining Based on Client-Side Tracking, *Proceedings of WEBKDD 2001* -, Third International Workshop, San Francisco, CA, USA, Kohavi R. et al (eds), Lecture Notes in Computer Science,Springer. vol. 2356, pp. 113-144, 2002.

[28] Spiliopoulou M., The laborious way from data mining to web mining, *Int. Journal of Comp. Sys., Sci. & Eng* ., Special Issue on Semantics of the Web, vol.14, pp.113-125, 1999.

[29] Srivastava J., Cooley R., Deshpande M., Tan P.-N., Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data, vol. 1, no.2, 2000.

[30] Xiao Y., Dunham M.H., Efficiency mining of traversal patterns, *Data& Knowledge Engineering* , n. 39, pp. 191-214, 2001.

[31] Zaïane O.R., Xin M., Han J., Discovering Web Access Patterns and Trends by Applying OLAP and Data Mining Technology on Web Logs, *Proceedings of Advances in Digital Libraries Conference (ADL'98)* , Sanda Barbara, CA, pp. 19-29, 1998.

Performance Study of Hierarchical Video-on-Demand (VoD) Systems for Layered Videos in Broadcast Environment

W.F. Poon and K.T. Lo

Department of Electronic and Information Engineering
The Hong Kong Polytechnic University
Hung Hom, Hong Kong
Email: enwil@polyu.edu.hk, enktlo@polyu.edu.hk

Abstract: - Streaming layered video is one of the possible solutions in providing large-scale video-on-demand (VoD) services in a heterogeneous network environment. In this paper, we analyze the performance of a hierarchical VoD system using layered videos in broadcast environment, in which we not only store some of the popular videos in the proxy server but also broadcast some of the video layers to the customers so as to minimize the blocking probability of the system. By deriving a performance model, we analyze the impacts of different parameters such as the proxy size, the number of broadcasting video layers and the bandwidth reserved for broadcasting on the system performance.

Key-Words: - broadcasting, video streaming, network performance model, video-on-demand

1 Introduction

With the advances in digital video technology, video-on-demand (VoD) has come into practice in recent years. However, VoD has not been a commercial success because its technology is still very expensive. Thus, many researchers have paid enormous efforts on the design of a VoD system in order to reduce the operational cost. Currently, there are mainly two directions to provide a cost-effective VoD service. The first is to use multicasting/broadcasting techniques to share the system resources. The second is to use a proxy server to minimize the backbone network transmission cost.

For the last few years, different multicasting /broadcasting protocols have been developed to improve the efficiency of a VoD system. Staggered broadcasting [1] is the simplest broadcasting protocol proposed in the early days. Since staggered broadcasting scheme suffered from the long start-up latency, some efficient broadcasting protocols such as skyscraper [2], fast data [3] and harmonic [4] were then proposed. In skyscraper and fast data broadcasting, a video is divided into an increasing size of segments that are transmitted into logical channels of same bandwidth so as to reduce the start-up delay. In harmonic broadcasting, instead of using channels with the same capacity, the video is divided into equal-size segments and the system broadcasts the video segments into logical channels of decreasing bandwidth. In these broadcasting schemes, customers are required to receive data from several channels simultaneously and a buffer should be installed in each receiver. A review of different data broadcasting techniques for VoD services can be found in [5]. The results showed that efficient broadcasting protocols can support a nearly true VoD service and the waiting time can be reduced to less than 15 seconds. In order to support true (zero-delay) VoD service in multicast/broadcast environment, patching [6] and hierarchical stream merging (HSM) [7] protocols were also developed to serve each customer immediately as the request is made. The idea of patching [6] is that a client first downloads data on two channels simultaneously. One is called a regular channel that is used to serve a batch of customers. The other is called a patching channel to provide the leading portion of the video so that the customer can be served without waiting for the next regular channel. For the HSM scheme [7], the customers hierarchically merge with the existing multicast groups so that the bandwidth requirement can be further reduced comparing with the patching protocol. The basic rule of the HSM scheme is that a customer always listens to the closest earlier stream that is still active. Once he/she has caught up the closest stream, he/she merges into it and his/her own stream is terminated.

In addition to data sharing, hierarchical network architectures [8-10] have also exploited to provide cost saving, as well as increased quality of service to end users in a VoD system. In [8], Li et al. developed a queuing model with the two-tier architecture to decide which video and how many copies have to maintain at each distributed server. In this architecture, there are a couple of metropolitan