2nd Hellenic Conf. on AI, SETN-2002, 11-12 April 2002, Thessaloniki, Greece,
Proceedings, Companion Volume, pp. 229-236

Exploring Web Access Logs with Correspondence Analysis

Nikos Koutsoupias[1]

[1] Department of Balkan Studies
Aristotle University of Thessaloniki
3rd Km Florina-Niki / 53100 Florina
nickk@auth.gr

Abstract. During the interaction of Internet users with a website, users provide information about themselves and how they respond to the site's content: where they come from, which links they click, or even where they spend most of their time. All this information is stored in a log file or a database. In this paper we will demonstrate the capabilities offered by a data analysis method (Correspondence Analysis) on web log statistics for the examination of user behavior and preferences. Specifically, we observed log statistics of a university department web site an a monthly basis, by plotting each data set on the same factorial plane. We view that this process may produce valuable results both for web-content designers and institutions with Internet presence.

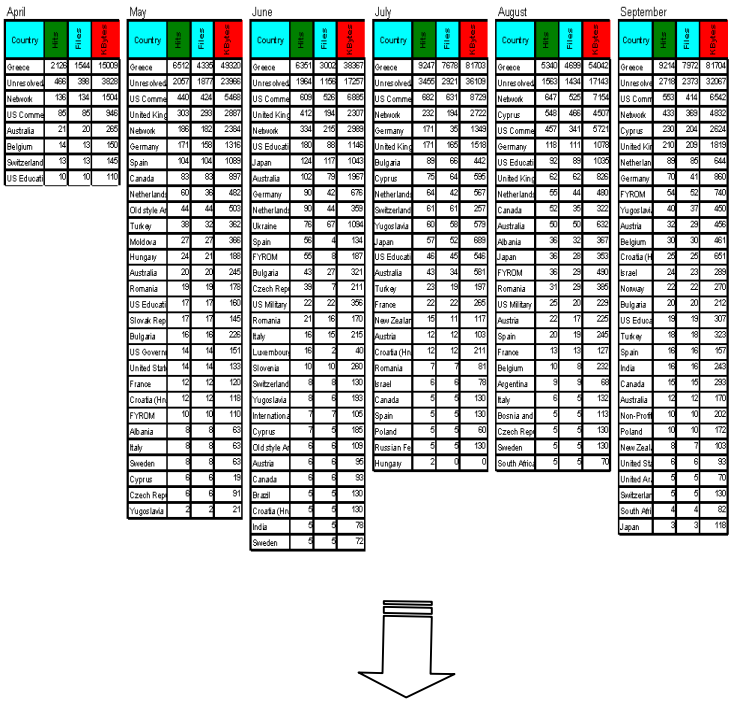
1 Introduction

For the past few years, Internet and the World Wide Web have been expanding at remarkable speed both in size and complexity. Access logs are becoming a valuable field for discovering various characteristics of user behavior. The most popular approach in the direction of discovering such knowledge through log files is that of Web mining. Overviews on methodologies, demands and challenges of Web mining is given by various sources [16][23][24][13][22]. 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 [11][15], marketing [4], library science [26], the media [20] and, of course, web design and engineering [6][3][5]. Also, interesting conclusions about web usage sequences and patterns can be drawn from various papers [7][18][19][21][25]. The idea presented in this paper is to utilize a data analysis/visualization method –correspondence analysis- on web log data files so as to 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 Method and Application

Correspondence analysis, along with other data analysis methods has been utilized in various fields of computing [17][14][8]. Correspondence Analysis like other multidimensional scaling methods [10] utilizes factorial diagrams (or factor score plots) in order to aid the interpretation of the analyzed phenomenon [1][2]. The results of the method include, along with the plots, absolute and relative contribution tables [9]. The fact that Correspondence Analysis is not based on a canonical distribution method or any other theoretical distribution, results in the absence of restrictive technical prerequisites, which in the case of classical statistical methods lead in distorted results.

The method can be applied in any table of positive numbers driven from a site's log. If we assume that we choose to view the picture of visits per foreign country each month, we would prefer to use 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-Month) over a period of six months and three columns (Hits, Files, Kbytes). In our experiment, the period examined starts from April 2001 and ends September 2001, excluding Greek and Unresolved address visitors, since, they comprise about 60% of the site's traffic. The resulting table will be a 138X3 data set (see Fig. 1).



	Country-Month	Hits	Files	KB
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 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 dimension of the final table for processing will be 138X9.

We used the software implementation of AFC97 [12] for Correspondence Analysis. Based on the results produced by the application of Correspondence Analysis using the resulting table, three factorial axes incorporate 93,4% of the information see Fig. 2).

Factor	Eigenvalues	%	Cumulative
1	0,917435	45,859629	45,859629
2	0,821579	41,068071	86,927700
3	0,129721	6,484315	93,412015
4	0,061583	3,078335	96,490350
5	0,046106	2,304691	98,795041
6	0,023574	1,178412	99,973452
7	0,000298	0,014880	99,988333
8	0,000233	0,011667	100,000000

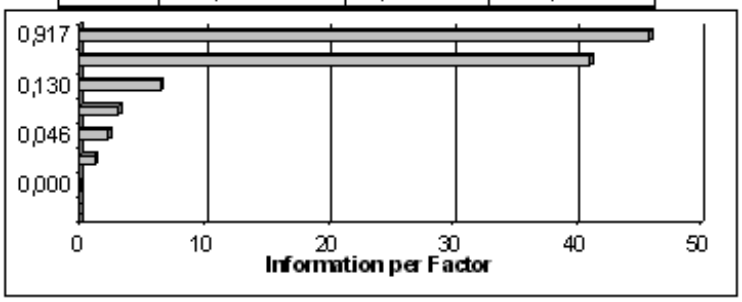


Fig. 2. Factorial axes

Having that in mind, we can 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 (see Fig. 3).

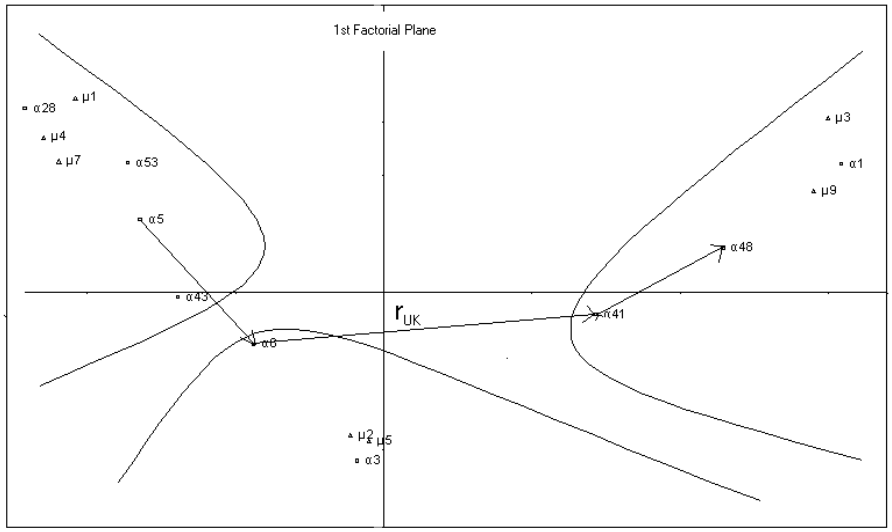


Fig. 3. The first factorial plane

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 in Fig 4.

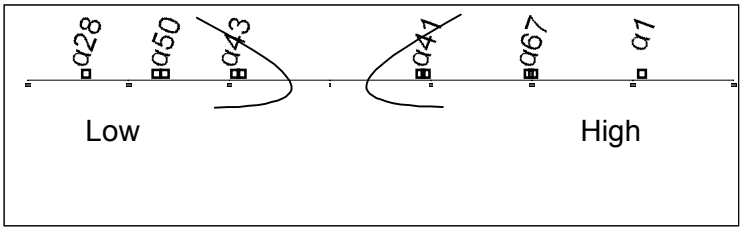


Fig. 4. First Axis (criterion minCOR>=100)

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.

The next step in Correspondence Analysis is the combination of any two axes produced during the previous phase. That is, we specify the characteristics of factorial plots with points that satisfy the axis criteria given earlier. In this way, the plot of the first factorial plane (for axes 1 and 2) contains only those points that characterize that plane. In this way, it is clear to understand which CountryMonths form the main trends and relationships in the data set examined.

3 Concluding Remarks

The statistical data analysis method described above, is a tool for a better understanding and interpretation of the preprocessed statistical results produced from web log data. It can be applied in any log data set and, we believe, it can boost the log data visualization process aiding web administrators to receive a total view of what are the main access trends for their site. Furthermore, when applying correspondence analysis analysts are able to define criteria for both axes and plane points. In this way, the method promotes the interpretation process, since the plots include only significant information in the form of points that characterize each factorial axis and/or plane.

References

1. Benzecri, J.-P. Pratique ed l' Analyse des Donnees (T.1: Analyse des Correspondances, expose elementaire), Dunod, Paris (1980)
2. Benzecri, J.-P., Analyse des Donnees (T. 2: Correspondances), Dunod, Paris (1973)

3. 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 (2000)
4. 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 (1999) 127-138
5. Chi E.H., Pirolli P., Pitkow J., The Scent of a Site: A System for Analyzing and Predicting Information Scent, Usage, and Usability of a Web Site, CHI Letters, Vol. 2, No. 1 (2000) 161-168
6. Cockburn A., McEnzie B., What do Web users do? An Empirical Characterization of Web Use, Journal of Human-Computer Studies, No. 54 (2001) 903-922
7. Cooley R., Mobasher B., Srivastava J., Data Preparation for Mining World Wide Web Browsing Patterns, Knowledge and Information Systems, Vol. 1, No. 1 (1999)
8. Gahegan M., Scatterplots and Scenes: Visualization Techniques for Exploratory Spatial Analysis, Comput. Environ and Urban Systems, Vol. 22, No. 1 (1998) 43-56
9. Greenacre, M., Correspondance Analysis in Practice, Academic Press, London, (1993)
10. Hair J.F., Anderson R.E., Tatham R.L., Black W.C., Mutlivariate Data Analysis, Prentice Hall, N.J. (1998)
11. Kohavi R., Mining E-Commerce Data: The Good, the Bad, and the Ugly, KDD' 2001 Industrial Track, San Francisco, Ca. (2001)
12. Koutsoupias N., AFC97: A New Software Implementation for Correspondence Analysis. In: Signal Processing, Communications and Computer Science, Mastorakis, N. (ed.): Engineering International Reference Book Series. World Scientific & Engineering Society Press (2000) 278-281
13. Madria S., Bwomich S., Ng W.-K., Lim P., Research Issues in Web Data Mining, Data Warehousing and Knowledge Discovery (1999) 303-312
14. Mathews G.J., Towheed S. S., NSSDC OMNIWeb: The first space physics WWW-based data browsing and retrieval system, Computer Networks and ISDN Systems, No. 27 (1995) 801-808
15. Mobasher B., Dai H., Luo T., Sun Y., Zhu J.. Integrating Web Usage and Content Mining for More Effective Personalization, Proceedings of the Int'l Conf. on E-Commerce and Web Technologies (ECWeb2000), Greenwich, UK (2000)
16. Mobasher B., Jain N., Han E., Srivastava J., Web Mining: Information and Pattern Discovery on the World Wide Web, Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'97) (1997)
17. Multivariate Visualization in Observation-Based Testing, Proceedings of ICSE 2000, Limeric, Ireland (2000).
18. Nanopoulos A., Manolopoulos A., Finding Generalized Path Patterns for Web Log Data Mining, Proceedings of East-European Conference on Advanced Databases and Information Systems (ADBIS'00) (2000) 215-228
19. Nanopoulos A., Manolopoulos Y., Mining patterns from graph traversals, Data and Knowledge Engineering, No. 37 (2001) 243-266
20. Nicholas D., Huntington P., Lievesley N., Withey R., Cracking the Code: Web Log Analysis, Online & CD-ROM Review, Vol. 23, No. 5 (1999) 263-269
21. Pei J., Han J., Mortazavi-asl B., Zhu H., Mining Access Patterns Efficiently from Web Logs, Proc. of the 4th Pacific Asia Conf. on Knowledge Discovery and Data Mining (2000) 396-407
22. Spiliopoulou M., Faulstich L.C., Winkler K., A Data Miner analyzing the Navigational Behavior of Web Users, Proceedings of the Workshop on Machine Learning in User Modeling of ACA'99 International Conference, Creta, Greece (1999)
23. 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 (1999)

24. 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)
25. Xiao Y., Dunham M.H., Efficiency mining of traversal patterns, Data& Knowledge Engineering, No. 39 (2001) 191-214
26. 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 (1998) 19-29