Construction and Analysis of Evolving Data Summaries: an Application on Web Usage Data

Alzennyr da Silva, Yves Lechevallier, Fabrice Rossi
Project AxIS, INRIA-Rocquencourt
Domaine de Voluceau, B.P. 105
78153 Le Chesnay cedex, France
{*Alzennyr.Da_Silva, Yves.Lechevallier, Fabrice.Rossi*}*@inria.fr*

Francisco de Carvalho
Centro de Informatica - CIn / UFPE
Av. Prof. Luiz Freire, s/n, CDU
50740-540 Recife, Brazil
fatc@cin.ufpe.br

Abstract

Taking the temporal dimension into account during the analysis of Web usage data has become a necessity since the way a site is visited may well evolve due to modifications in the structure and content of the site, or even due to changes in the behavior of certain user groups. Consequently, the models associated with these behaviors must be continuously updated. One solution to this problem is to update these models using summaries obtained by means of an evolutionary approach based on clustering methods. To do this, we carry out various clustering strategies that are applied on time sub-periods. We compare the results obtained using this method with those reached by traditional global analysis.

1. Introduction

The access patterns to Web pages are indeed of a dynamic nature, due both to the on-going changes in the content and structure of the Web site and to changes in the users' interest. The access patterns can be influenced by certain parameters of a temporal nature, such as the time of the day, the day of the week, recurrent factors (summer/winter vacations, national holidays, Christmas) and non-recurrent global events (epidemics, wars, economics crises, the World Cup).

Web Usage Mining (WUM) [2] [12] consists in extracting interesting information from files which register Web usage traces (log files). Most methods in this domain take into account the entire period during which usage traces were recorded, the results obtained naturally being those which prevail over the total period. Consequently, certain types of behaviors, which take place during short sub-periods are not detected and thus remain unknown by traditional methods. It is however important to study these behaviors and thus carry out an analysis related to significant time sub-periods. It will then be possible to study the temporal evolution of users' profiles by providing descriptions that can integrate the temporal aspect. Furthermore, as the volume of mined data is great, it is important to define summaries to represent user profiles.

WUM has recently started to take account of temporal dependence in usage patterns. In [9], the authors survey the work to date and explore the issues involved and the outstanding problems in temporal data mining by means of a discussion about temporal rules and their semantic. In addition, they investigate the confluence of data mining and temporal semantics. Recently in [7], the authors outline methods for discovering sequential patterns, frequent episodes and partial periodic patterns in temporal data mining. They also discuss techniques for the statistical analysis of such approaches. Notwithstanding these considerations, the majority of methods in the WUM are applied on the entire period that covers all the available data. Consequently, these methods reveal the most predominant behaviors in data, and the interesting short-term behaviors which could occur during short periods of time are not taken into account. For example, when the data analysed is inserted into a dynamic domain related to a potential long period of time (such as in the case of Web log files), it is to be expected that behaviors evolve over time.

These considerations have given rise to many studies in data analysis, especially concerning the adaptation of traditional static data-based methods to the dynamic data framework. In this line of research, we use summaries obtained by an evolutionary clustering approach applied over time sub-periods to carry out a follow-up of the user profile evolution.

This article is organized as follows. Section 2 presents the benchmark data set analysed. Section 3 describes the proposed clustering approach based on time sub-periods and discusses the results analysis. The final section presents the conclusion and suggestion for future work.

2. Usage data

As a case study, we use a benchmark Web site from Brazil [1]. This site contains a set of static pages (details of teaching staff, academic courses, etc.) and dynamic pages (see [4][3][10][11] for an analysis of this part of the site). We studied the accesses to the site from 1^{st} July 2002 to 31^{st} May 2003.

Concerning the Web usage data pre-processing, we adopt the methodology proposed by [13] who defines a *navigation* as a succession of requests not more than 30 minutes apart, coming from the same user. In order to analyse the more representative traces of usage, we selected long navigations (containing at least 10 requests and with a total duration of at least 60 seconds) which are assumed to have originated from human users (the ratio between the duration and number of requests must be at least 4, which means a maximum of 15 requests per minute). This was done in order to extract human navigations and exclude those which may well have come from Web robots. The elimination of short navigations is justified by the search for usage patterns in the site rather than simple accesses (performed by a search engine, for example) which do not generate a trajectory in the site. After filtering and eliminating outliers, we obtained a total of 138,536 navigations.

3. Clustering approach based on time sub-periods

The approach proposed in this article consists initially in dividing the analysed time period into more significant sub-periods (in our case, the months of the year) with the aim of discovering the evolution of old patterns or the emergence of new ones, which would not have been revealed by a global analysis over the whole time period. After that, a clustering method is carried out on data of each sub-period, as well as over the complete period. The results provided for each clustering are then compared.

Concerning the clustering itself, we carried out four types of clustering:

- **Global clustering**: this clustering is performed on all existing navigations. By intersecting the obtained clusters with the temporal partition, global clustering generates a partition in each temporal group.
- **Local independent clustering**: this clustering is performed on the set of navigations occurring in each time sub-period separately. We have one clustering for each month of the period analysed. The final partitions are thus independent.
- **Local "previous" clustering**: this clustering can be performed by starting from another clustering when the algorithm is able to assign new individuals to previous clusters. We thus use the clustering results performed on the preceding time sub-period to obtain a partition on the navigations belonging to the current time sub-period.
- **Local dependent clustering**: this clustering can be performed by an iterative algorithm like the dynamic clustering algorithm, initialised in an adapted way. Here, we initialise the algorithm with the prototypes of the clusters from the previous time sub-period.

3.1. The algorithm and evaluation criteria

Our method uses an adapted version of the dynamic clustering algorithm [1][5] [8] applied on a data table containing the navigations in its rows and real-value variables in its columns (for example: number of successful requests, number of repeated requests, total duration, number of transfered bytes in a navigation, etc). As a distance measure, we adopt the Euclidean distance. For all the experiments, we defined an *a priori* number of clusters equal to 10 with a maximum number of iterations equal to 100. The number of random initialisations is equal to 100, except when the algorithm is initialised with the results obtained from a previous execution.

To analyse the results, we apply two criteria. For a cluster by cluster analysis, we compute the F-measure [14]. To compare two partitions, we look for the best representation of the cluster A in the first partition by a cluster B in the second partition, i.e., we look for the best match between the clusters of two partitions. This gives us as many values as there are clusters in the first partition. For a global analysis, we apply the corrected Rand index [6] to compare two partitions. The F-measure takes a value in the range [0,+1], whereas the corrected Rand index values are in the range [-1,+1]. In both cases, the value 1 indicates a perfect agreement and values near 0 correspond to cluster agreements found by chance.

3.2. Results and discussion

The values of the corrected Rand index reveal that the results from the local independent clustering are very different from those of the global and local dependent clustering (cf. figure 1). These differences are confirmed by the

[1]This web site is available at the following address: `http://www.cin.ufpe.br/`

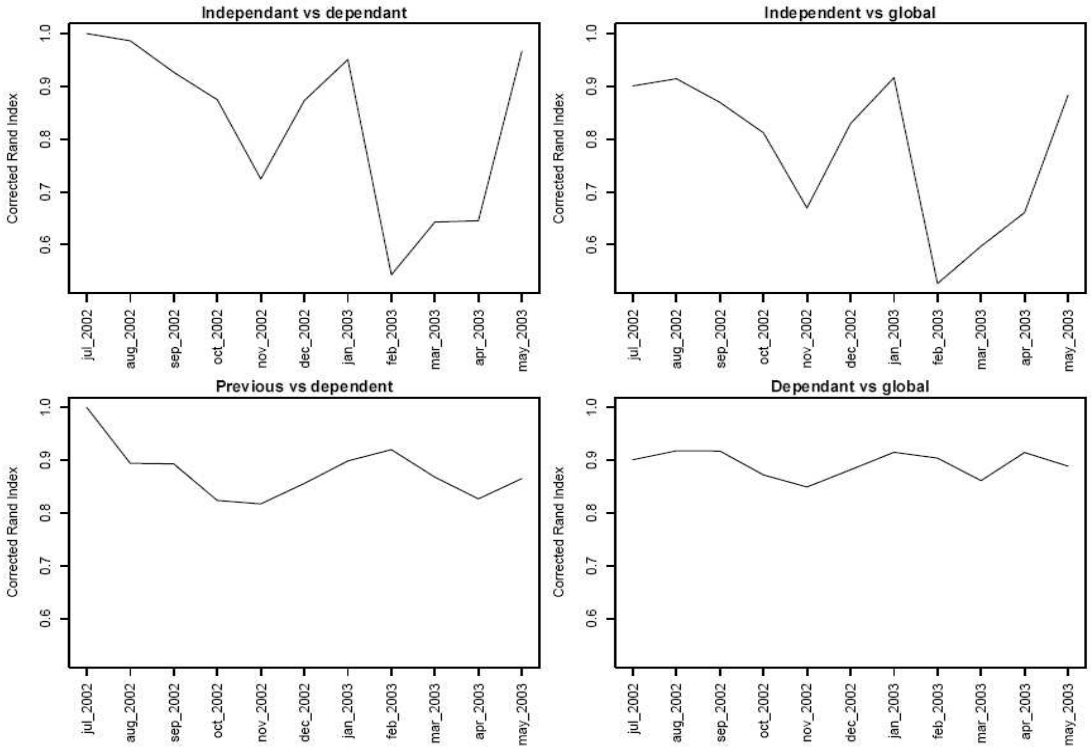

Figure 1. Corrected Rand index values computed partition by partition

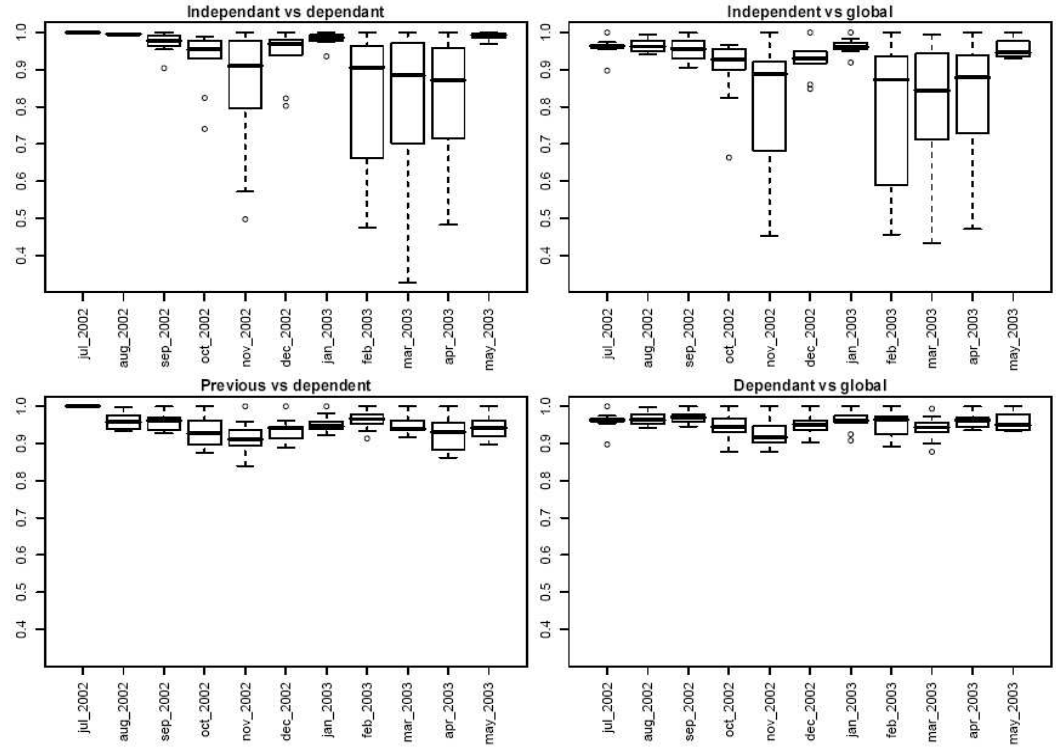

Figure 2. F-measure values computed cluster by cluster

F-measure. As we obtain 10 values (one per cluster) from the F-measure for each month, we trace the corresponding boxplot to summarise these values (cf. figure 2). We can see by the confrontation of the local independent clustering versus the global clustering that there are almost always low values, i.e., certain clusters resulting from the local independent clustering are not found by the global clustering. We can also notice that the local previous clustering does not give very different results from those obtained by the local dependent clustering.

Using a cluster-by-cluster confrontation via the F-measure between the global clustering and the local dependent and independent clustering, we refine the analysis. What appears quite clearly is that the clusters are very stable over time if we apply the local dependent clustering method. In fact, no value is lower than 0.877, which represents a very good score. On the other hand, in the case of local independent clustering, we detect clusters that are very different from those obtained by the global clustering (some values are lower than 0.5).

In sum, we can say that the local dependent clustering method shows that the clusters obtained change very little or do not change at all, whereas the local independent clustering method is more sensitive to changes which occur from one time sub-period to another.

4. Conclusion

In this article, we addressed the problems of processing dynamic data in the WUM domain. The questions discussed highlight the need to define or adapt methods to extract knowledge and to follow the evolution of this type of data.

Although many powerful knowledge discovery methods have been proposed for WUM, very little work has been devoted to handling problems related to data that can evolve over time.

Through our experiments, we showed that the analysis of dynamic data by time sub-periods offers a certain number of advantages such as making the method sensitive to cluster changes over time. Analysing changes in clusters over time can provide important clues about the changing nature of how a web site is used, as well as any changing loyalties of its users. Furthermore, as our approach splits the data and concentrates the analysis on fewer sub-sets, some constraints regarding hardware limitations could be overcome.

Possible future work could involve applying other clustering methods and implementating techniques that enable the automatic discovery of the number of clusters as well as identifying fusions and splits over time.

5. Acknowledgements

The authors are grateful to the collaboration project between INRIA and FACEPE (France/Brazil) and CAPES (Brazil) for their support for this research.

References

[1] M. R. Anderberg. *Cluster analysis for applications*. Probability and Mathematical Statistics, New York: Academic Press, 1973, 1973.

[2] R. Cooley, B. Mobasher, and J. Srivastava. Data preparation for mining world wide web browsing patterns. *Journal of Knowledge and Information Systems*, 1(1):5–32, 1999.

[3] A. Da Silva, F. De Carvalho, Y. Lechevallier, and B. Trousse. Characterizing visitor groups from web data streams. *Proceedings of the 2nd IEEE International Conference on Granular Computing (GrC 2006)*, pages 389–392, May 10-12 2006.

[4] A. Da Silva, F. De Carvalho, Y. Lechevallier, and B. Trousse. Mining web usage data for discovering navigation clusters. *11th IEEE Symposium on Computers and Communications (ISCC 2006)*, pages 910–915, 2006.

[5] E. Diday and J. C. Simon. Clustering analysis. In K. Fu, editor, *Digital Pattern Classification*, pages 47–94. Springer Verlag, 1976.

[6] L. Hubert and P. Arabie. Comparing partitions. *Journal of Classification*, 2:193–218, 1985.

[7] S. Laxman and P. S. Sastry. A survey of temporal data mining. *SADHANA - Academy Proceedings in Engineering Sciences, Indian Academy of Sciences*, 31(2):173–198, 2006.

[8] J. MacQueen. Some methods for classification and analysis of multivariate observations. In *5th Berkley Symposium on Mathematics and Probability*, volume 1, pages 281–297, 1967.

[9] J. F. Roddick and M. Spiliopoulou. A survey of temporal knowledge discovery paradigms and methods. *IEEE Transactions on Knowledge and Data Engineering*, 14(4):750–767, 2002.

[10] F. Rossi, F. De Carvalho, Y. Lechevallier, and A. Da Silva. Comparaison de dissimilarits pour l'analyse de l'usage d'un site web. *Actes des 6me journes Extraction et Gestion des Connaissances (EGC 2006), Revue des Nouvelles Technologies de l'Information (RNTI-E-6)*, II:409–414, January 2006.

[11] F. Rossi, F. De Carvalho, Y. Lechevallier, and A. Da Silva. Dissimilarities for web usage mining. *Actes des 10me Confrence de la Fdration Internationale des Socits de Classification (IFCS 2006)*, July 2006.

[12] M. Spiliopoulou. Data mining for the web. *Workshop on Machine Learning in User Modelling of the ACAI99*, pages 588–589, 1999.

[13] D. Tanasa and B. Trousse. Advanced data preprocessing for intersites web usage mining. *IEEE Intelligent Systems*, 19(2):59–65, March-April 2004.

[14] C. J. van Rijsbergen. *Information Retrieval*. Butterworths, London, second edition, 1979.