Analyzing the Evolution of Web Usage Data

Alzennyr Da Silva

Projet AxIS, INRIA
Domaine de Voluceau, Rocquencourt, B.P. 105
78153 Le Chesnay cedex, France
Alzennyr.Da_Silva@inria.fr

Abstract. Analyzing Web usage has become a very important strategy for Web site operators as it provides them a better understanding of the users' behavior. This insight can enable the operators to improve their service and thereby attract more visitors. Taking into account the temporal dimension in such analyses has become a necessity since the way a site is visited can indeed 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 in order to reflect the actual behavior of the users. One solution to this problem, proposed in this article, is to update these models using summaries obtained by means of an evolutionary approach based on clustering methods. To do so, we carry out various clustering strategies that are applied on time sub-periods. We compare the results obtained using this method with the results obtained by a traditional global analysis.

Keywords: Clustering, Evolving Data, Web Usage Mining.

1 Introduction

The Web is one of the most relevant examples of an evolving and dynamic data source. This is due to the fact that new information is constantly being added to existing pages while a huge number of new documents are appearing on-line each day. The access patterns to these pages therefore are of a dynamic nature, due both to the on-going changes in the content and structure of the Web site as well as 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, the Christmas period) and non-recurrent global events (epidemics, wars, economics crises, the World Cup), etc.

A usage based analysis consists in studying the traces left by users when they visit a Web site. More precisely, Web Usage Mining (WUM) consists in extracting interesting information from Web server's 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 to carry out an analysis relating to significant time sub-periods. This will make it possible to study, for example, possible shifts in the user's interests concerning the Web site's services over time sub-periods. It will then be possible to study the temporal evolution of users' profiles by providing descriptions that are able to integrate the temporal aspect. Furthermore, as the volume of mined data is very great, it is important to define summaries to represent user profiles.

These considerations have give rise to many studies in data analysis, especially concerning the adaptation of traditional static data based methods to the dynamic data framework. The work presented in this article continues in this line of research and proposes to follow changes in users' behavior. We use summaries obtained by an evolutionary clustering approach applied over time sub-periods to carry out a follow-up of the evolution.

This article is organized as follows. Section 2 presents a short state of the art regarding current challenges in the domain. Section 3 presents the analyzed benchmark data set. Section 4 describes the proposed clustering approach based on time sub-periods and presents the analyses of the results. The final section presents conclusion outlines and future work.

2 State of the art

Web Mining [9] appeared in the end of 90s and consists in using Data Mining techniques in order to develop methods that allow relevant information to be extracted from Web data (such as documents, interaction traces, link structure, etc). A more specialized branch of this domain, named Web Usage Mining (WUM), deals with techniques based on Data Mining that are applied to the analysis of users' behavior in a Web site [3] [16]. The present article is placed is this last axis. In the e-commerce domain, one of the most important motivations for the analysis of usage is the need to build up consumer loyalty and to make the site more appealing to new visitors.

Web Usage Mining has recently started to take account of temporal dependence in usage patterns. In [13], 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. Also, they investigate the confluence of data mining and temporal semantics. Recently in [10], 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 Web Usage Mining 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 analyzed 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.

To deal with this situation, the usage models must be continuously updated (by means of efficient algorithms) in order to follow changes in user's behavior. This requires a continuous monitoring of the discovered models, which is a necessary pre-requisite for applications focusing on temporal dimension. A possible solution to this problem is proposed in this article: to partition time before applying Web usage mining methods and thus integrate a temporal follow-up of behaviors patterns into a clustering algorithm.

3 Usage Data

Data concerning usage of a Web site come primarily from the log files recorded by the Web server. Lines in this file describe requests received by the server concerned. Each line specifies the requested document, the source of the requisition, date and time, etc. There is a range of techniques for preprocessing these data and extracting navigations from Web log files. In this article we use the methodology proposed by [17]. A navigation constitutes the trajectory of a user in the site and is defined as a succession of requests coming from the same user and there are no more than 30 minutes apart.

In our approach, as a case study, we use a benchmark Web site (see figure 1) from Brazil [1]. It is the site of the Computer Science Department of UFPE (Federal University of Pernambuco, Brazil), the laboratory of one of the authors. This contains a set of static pages (details of teaching staff, academic courses, etc.) and dynamic pages (see figure 2). The dynamic pages form the site front-end and consist of 91 pages organized in a well defined semantic structure (see [4][5][14][15] 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 which corresponds approximately to 2 Gigabytes of raw data.

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

Fig. 1. Screenshot from the CIn's Web site.

In order to analyze 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 be 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 does not generate a trajectory in the site. After filtering and eliminating of outliers, we obtained a total of 138,536 navigations.

4 Clustering Approach based on Time Sub-periods

Characterizing user groups consists in identifying features shared by a sufficiently large number of users and which thus provide elements that allow a profile for each user group to be inferred [2] [4] [5].

The approach proposed in this article consists initially in dividing the analyzed 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.

It is important to notice that the partitioning obtained by the clustering method concerns the total set of navigations, which are then broken down into groups. In the final partition, each

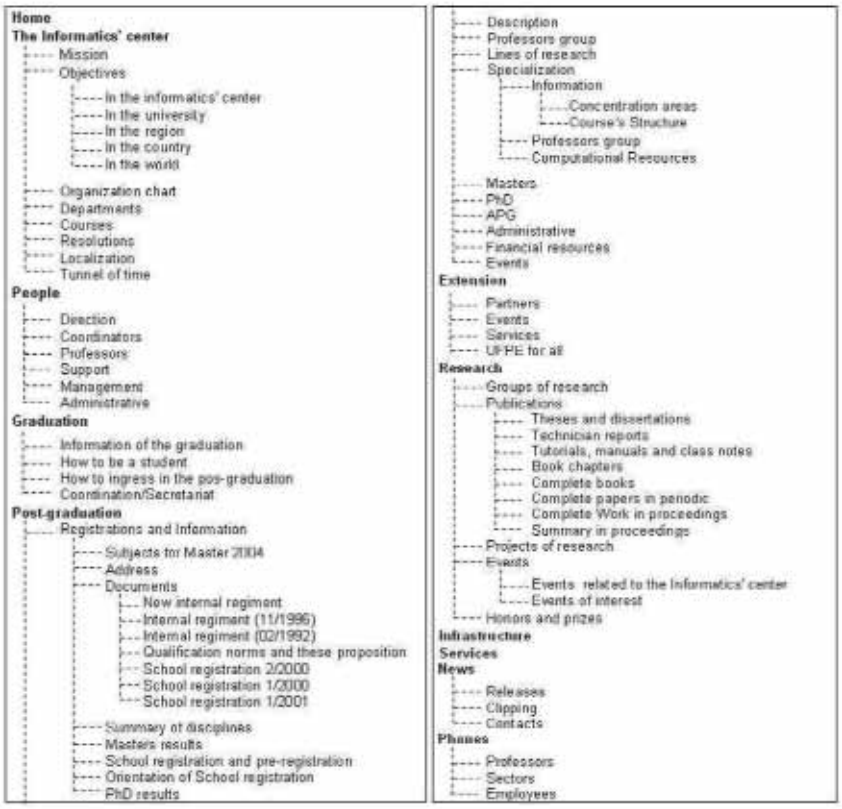

Fig. 2. Semantic structure of the dynamic pages from the CIn's Web site.

navigation is assigned to a specific cluster, which represents a rather different approach from that taken by elementary statistical analyse where page hits occurring during a specified time slot are simply counted, without taking into account the navigation involved.

In our approach, the time partitioning is formulated according to the calendar months. However, other possibilities for time partitioning (such as 15-day intervals, national holidays, times of the day: morning, afternoon, evening, etc.) could also be applied.

As concerns the clustering itself, we carried out four types of clustering:

Global clustering: this clustering is performed on all existing navigations. By intersecting of 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 analyzed. 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 (see the next section). 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 (see the next section) initialized in an adapted way. Here, we initialize the algorithm with the prototypes of the clusters from the previous time sub-period.

4.1 The algorithm and evaluation criteria

To deal with clustering of navigations, our method uses an adapted version of the dynamic clustering algorithm [1][6][11] applied on a data table containing the navigations in its rows and real-valued variables in its columns (see table 1 for variables description). Other clustering algorithms could of course be used, however they must be compatible with the clustering strategies described in the previous section. In particular the algorithm must be able to:

1. assign new observations to an existing partition;
2. initialize the algorithm with the cluster prototypes of a previous clustering that it has carried out.

It would thus be possible to use an algorithm such as Kohonen's self-organizing maps [8].

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 initializations is equal to 100, except when the algorithm is initialized with the results obtained from a previous execution.

Table 1. Description of the Navigation Variables

N°	Field	Signification
1	IDNavigation	Navigation code
2	NbRequests_OK	Number of successful requests (status = 200) into the navigation
3	NbRequests_BAD	Number of failed requests (status $\neq$ 200) into the navigation
4	PRequests_OK	Percentage of successful requests (= NbRequests_OK/ NbRequests)
5	NbRepetitions	Number of repeated requests into the navigation
6	PRepetitions	Percentage of repetitions (= NbRepetitions / NbRequests)
7	TotalDuration	Total duration of the navigation (in seconds)
8	AvDuration	Average of duration (= TotalDuration / NbRequests)
9	AvDuration_OK	Average of duration among successful requests (= TotalDuration_OK/NbRequests_OK)
10	NbRequests_SEM	Number of requests related to pages in the site's semantic structure
11	PRequests_SEM	Percentage of requests related to pages in the site's semantic structure (=NbRequests_Sem/ NbRequests)
12	TotalSize	Total size of transfered bytes in the navigation
13	AvTotalSize	Average of transfered bytes (= TotalSize / NbRequests_OK)
14	MaxDuration_OK	Duration of the longest request in the navigation

To analyze the results, we apply two criteria. For a cluster-by-cluster analysis, we compute the F-measure [18]. 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. This measure allows a detailed analysis, but it does not take into account the relative size of the clusters.

For a global analysis, we apply the corrected Rand index [7] 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. In fact, an analysis made by [12] confirmed corrected Rand index values near 0 when presented to clusters generated from random data, and showed that values lower than 0.05 indicate clusters achieved by chance.

4.2 Results and discussion

To better understand the evolution of clusters compared over the time sub-periods, we carried out a follow-up of the cluster prototypes (month by month) for the local independent clustering and the

local dependent clustering. We projected these prototypes in the factorial plan which is computed on the total population (cf. figure 3). In this representation, each disc represents a prototype. In the dependent clustering (on the left), the ten clusters are represented by different colors and the lines represent the trajectory of these prototypes. It is possible to notice a certain stability notwithstanding the diversity of the months analyzed. In the case of the local independent clustering (on the right), the temporal trajectory is simply illustrated by the lines which join a prototype to its nearest neighbor in the previous time sub-period. This does not give clear trajectories because, in some cases, certain prototypes share the same predecessor. In fact, we notice that only four clusters are perfectly identified, the others ones undergo fusions and splits over time.

If we project on the factorial plan the prototypes of the global clustering (G1, G2..., G10), then the prototypes obtained by the local independent and by the local dependent clustering (cf. figure 4), it becomes clear that the local independent clustering is able to identify new clusters that are not found by the two other clusterings. On the other hand, the global and local dependent clustering present very close prototypes.

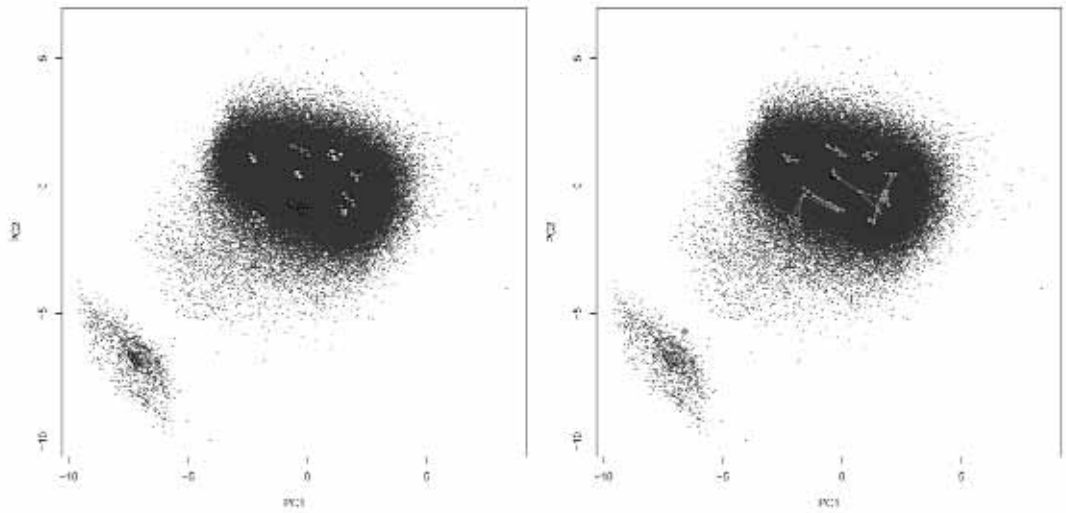

Fig. 3. Projection and follow-up of cluster prototypes for the local dependent clustering (left) and the local independent clustering (right).

We have also computed (month-by-month for the resulting partitions) the intra-cluster variance which represents the sum of the distance between each individual and the prototype of its assigned cluster. As expected, the best score is for the local independent clustering, then for the local dependent clustering and finally for global clustering (cf. figure 5). In other words, this result shows that the partition discovered by the local independent clustering is the most cohesive one.

Given the small difference between theses scores, we might then expect the clusters to be somewhat close. However, the representation in the factorial plan raises an initial doubt, as the prototypes in the local independent clustering seem sometimes rather different from those of the global and the local dependent clustering. In fact, 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 6). These differences are confirmed by the F-measure. As we obtain 10 values (one per cluster) from the F-measure for each month, we trace the corresponding boxplot to summarize these values (cf. figure 7). 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

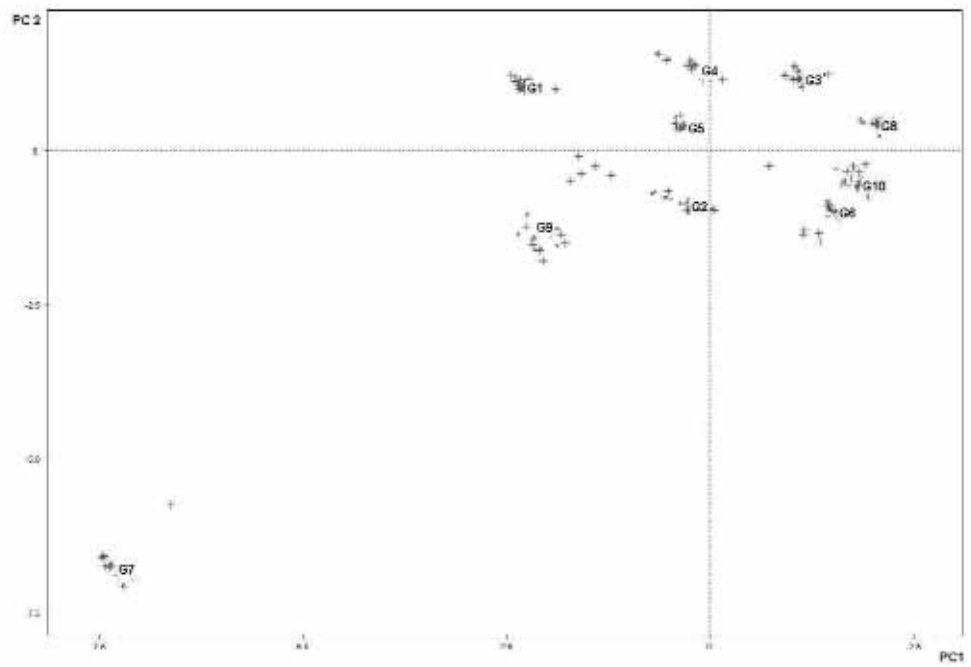

Fig. 4. Projection of cluster prototypes for the clusterings: global (G1, G2..., G10), local independent (+) and local dependent (o).

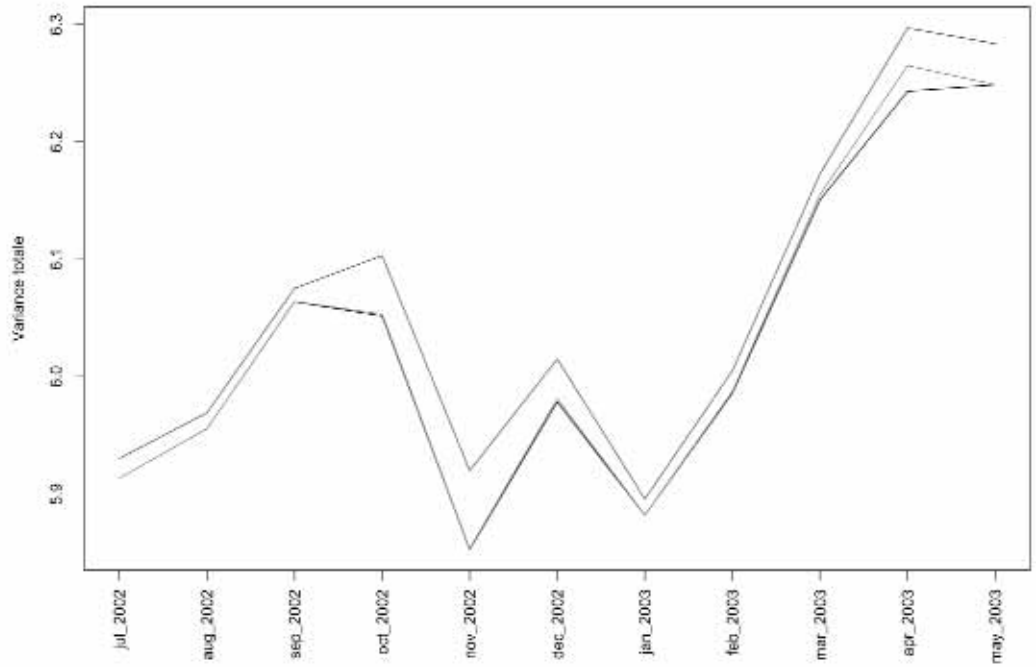

Fig. 5. Intra-cluster variance for the global (blue line), local dependent (red line) and local independent (black line) clusterings.

the local dependent clustering, which confirms the intuitive assumption gained by observing the prototypes in the factorial plan: these prototypes "move" very little over time.

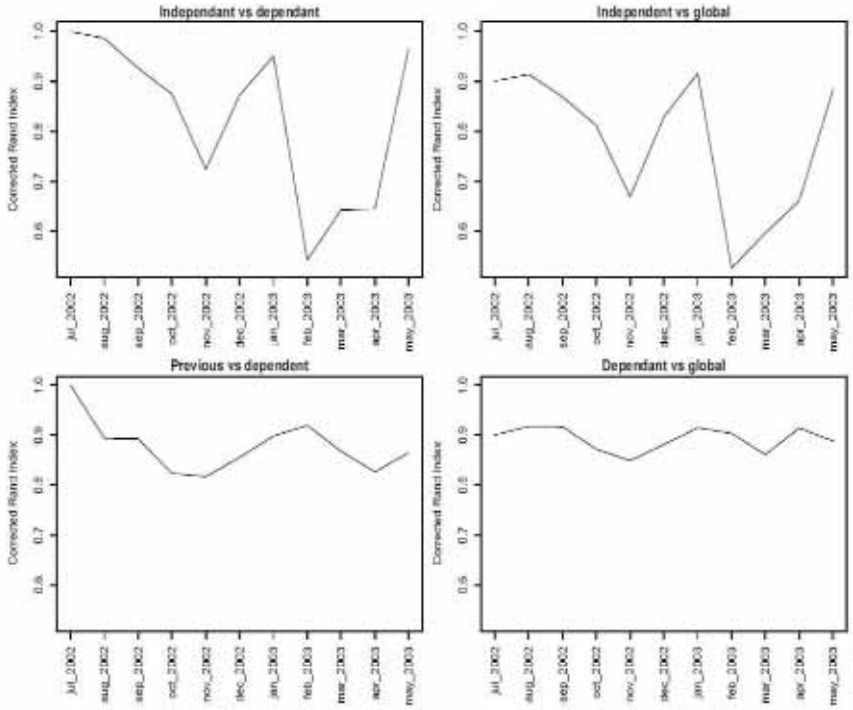

Fig. 6. Corrected Rand index values computed cluster-by-cluster.

By 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 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).

It is quite surprising that the partitions defined by the local dependent clustering are very similar to those from the global clustering. We could thus speculate that an analysis carried out on time sub-periods would able to obtain results supposed to be revealed by a global analysis is carried out on the entire time period. Moreover, the local dependent clustering can be considered as a divide-to-conquer-like approach and, moreover, one that can deal with certain constraints such as processing time and hardware limitations (memory size, microprocessor speed, etc).

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

5 Conclusions

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

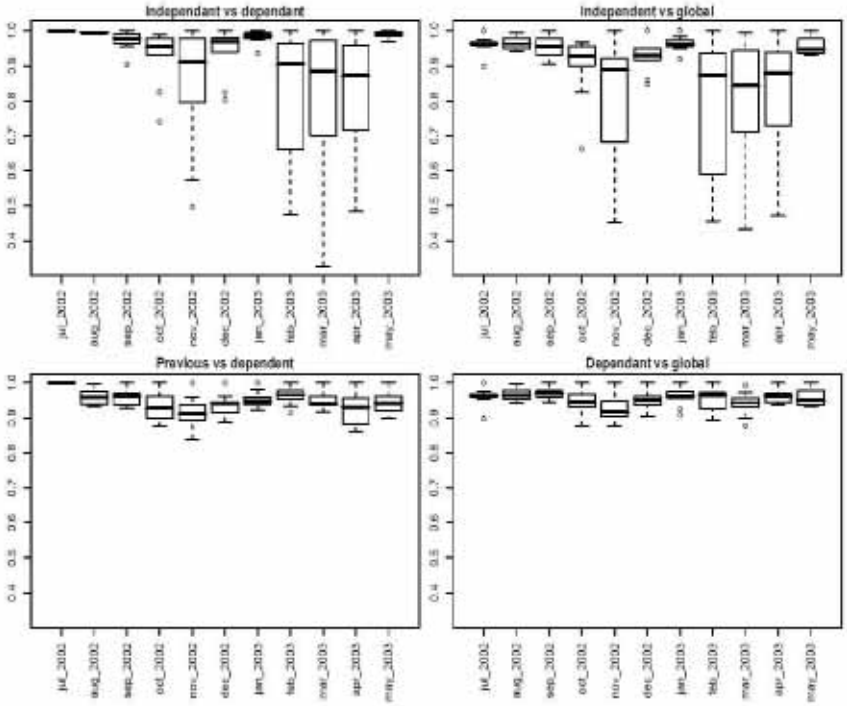

Fig. 7. F-measure values computed cluster-by-cluster.

Although many powerful knowledge discovery methods have been proposed for Web Usage Mining, very little work has been devoted to handling problems related to data that can evolve over time. Moreover, the duration of each event in a sequence is a problem that has not yet attracted sufficient attention in temporal data mining. When events are of different durations, it is desirable to extend the basic temporal model framework to define structures which take these variations into account and not only their date of occurrence.

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. Furthermore, as our approach splits the data and concentrates the analysis on fewer sub-sets, some constraints regarding hardware limitations could overcome.

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

Acknowledgements

The author would like to thank the collaboration project INRIA/FACEPE (France/Brazil) and CAPES (Brazil) for their support to this research work. The author deeply thanks Dr. Yves Lechevallier, Dr. Fabrice Rossi and Dr. Francisco de Carvalho for their precious contributions.

References

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

2. Chi, E. H., Rosien, A., Heer, J.: Lumberjack: Intelligent Discovery and Analysis of Web User Traffic Composition. ACM SIGKDD Workshop on Web Mining for Usage Patterns and User Profiles (WEBKDD), Canada: ACM Press (2002) 1 16
3. Cooley, R., Mobasher, B., Srivastava, J.: Data Preparation for Mining World Wide Web Browsing Patterns. Journal of Knowledge and Information Systems, 1 (1), (1999) 5–32
4. Da Silva, A., De Carvalho, F., Lechevallier, Y., Trousse, B.: Mining Web Usage Data for Discovering Navigation Clusters. 11th IEEE Symposium on Computers and Communications (ISCC 2006), Pula-Cagliari, Italy (2006) 910 915
5. Da Silva, A., De Carvalho, F., Lechevallier, Y., Trousse, B.: Characterizing Visitor Groups from Web Data Streams. Proceedings of the 2nd IEEE International Conference on Granular Computing (GrC 2006), Atlanta, USA (2006) 389–392
6. Diday, E., Simon, J. C.: Clustering analysis. Digital Pattern Classification, Fu, K.S. (Eds.), Springer Verlag (1976)47 94
7. Hubert, L., Arabie, P.: Comparing Partitions. Journal of Classification (1985) 193–218
8. Kohonen, T.: Self-Organizing Maps. Springer Series in Information Sciences, v. 30, (1995)
9. Kosala, R., Blockeel, H.: Web Mining Research: A Survey. ACM SIGKDD Explorations: Newsletter of the Special Interest Group on Knowledge Discovery and Data Mining, v. 2, (2000) 1 15
10. Laxman, S., Sastry, P. S.: A survey of temporal data mining. SADHANA - Academy Proceedings in Engineering Sciences, Indian Academy of Sciences, (31) 2, (2006) 173–198
11. MacQueen, J.: Some methods for classification and analysis of multivariate observations. 5th Berkley Symposium on Mathematics and Probability, v. 1, (1967) 281 297
12. Milligan, G. W., Cooper, M. C.: A Study of the Comparability of External Criteria for Hierarchical Cluster Analysis. Multivariate Behavioral Research, (21) 8, (1986) 441–458
13. Roddick, J. F., Spiliopoulou, M.: A Survey of Temporal Knowledge Discovery Paradigms and Methods. IEEE Transactions on Knowledge and Data Engineering, (14) 4 (2002) 750 767
14. Rossi, F., De Carvalho, F., Lechevallier, Y., Da Silva, A.: Comparaison de dissimilarits pour l'analyse de l'usage d'un site web. Actes des 6me journes Extraction et Gestion des Connaissances (EGC 2006), RNTI-E-6, Ritschard, Gilbert and Djeraba, Chabane (Eds.) v. 2, (2006) 409–414
15. Rossi, F., De Carvalho, F., Lechevallier, Y., Da Silva, A.: Dissimilarities for Web Usage Mining. Actes des 10me Confrence de la Fdration Internationale des Socits de Classification (IFCS 2006), Vladimir Batagelj, Anuska Ferligoj, and Ales Ziberna (Eds.) (2006) 39 46
16. Spiliopoulou, M.: Data Mining for the Web. Workshop on Machine Learning in User Modelling of the ACAI99; (1999) 588–589
17. Tanasa, D., Trousse, B.: Advanced Data Preprocessing for Intersites Web Usage Mining. IEEE Intelligent Systems, (19) 2, (2004) 59 65
18. van Rijsbergen, C. J.: Information Retrieval. Butterworths (1979)