Pre-Processing and Clustering Complex Data in E-Commerce Domain

Sergiu Chelcea[1], Alzennyr Da Silva[1&2], Yves Lechevallier[2], Doru Tanasa[1], Brigitte Trousse[1]
[1] *AxIS, INRIA Sophia-Antipolis*
2004, Route des Lucioles, B.P. 93
06902 Sophia Antipolis Cedex, France
[2] *AxIS, INRIA Rocquencourt*
Domaine de Voluceau, Rocquencourt, B.P. 105
78153 Le Chesnay Cedex, France
{Sergiu.Chelcea,Doru.Tanasa,Brigitte.Trousse,Alzennyr.Da_Silva,Yves.Lechevallier }@inria.fr
http://www-sop.inria.fr/axis/

Abstract

This paper presents our preprocessing and clustering method on a clickstream dataset issued from e-commerce domain. The main contributions of this article are double. First, after presenting the clickstream dataset, we show how we build a rich data warehouse based an advanced preprocessing method. We take into account the intersite aspects in the given e-commerce domain, which offers an interesting data structuration. A preliminary statistical analysis based on such complex data i.e. time period clickstreams is given, emphasing the importance of intersite user visits in such a context. Secondly, we describe our crossed-clustering method which is applied on data generated from our data warehouse. Our preliminary results are interesting and promising illustrating the benefits of our WUM methods, even if more investigations are needed on the same dataset.

1. Introduction

This article deals with complex data from Web in e-commerce domain. The data complexity results from its characteristics: large, multi-source (intersite logs), heterogeneous (product description tables, logs) and temporal (time period based clickstreams).
Indeed, the daily access of an Internet Web site can today easily rise to a number of access in millions of pages, executed by a large amount of users spread all over the world. Different contents and heterogeneous necessities constitute the reality of the Web. All this complexity results in extremely rich and varied usage patterns. With this explosive growth of data available on the internet, the discovery and analysis of useful information from the Web becomes a necessity. Web usage mining [14] is the application of data mining technologies on large logs files, collected from Web servers accesses. This is also known as Web log mining, and represents the process of knowledge extraction from the Web access log files. Examples of such applications include: improvements of web sites design, system performance analyses as well as network communications, understanding user reaction and motivation, web personalization, and building adaptive web sites [11][12][24][16][17].
E-commerce organizations are especially interested in the insight that web usage mining provides [13][17]. Such insight helps not only to improve their web site, but also their services and marketing strategies (promotions, banners, etc.).
This paper aims at analyzing a chosen clickstream dataset, which is in the e-commerce domain. Also we adopt for our analysis the vendor/webmaster point of view. According to different time periods, we aim at discovering the usage and the server charge in the context of multiple Web sites in the e-commerce domain. The rest of this paper is organized as follows. Section 2 presents the analyzed dataset and our Intersite advanced data preprocessing of structuring data in terms of multishop consumers visits. Then Section 3 gives first results issued from a preliminary statistical analysis of this dataset. Next, before concluding in Section 5, Section 4 describes our two clustering methods, the used data and the respective analyses.

2. Advanced intersite pre-processing

2.1. Used clickstream dataset description

The used clickstream dataset, provided for the first time in the PKDD Challenge 2005, consists in 576 large log files with a total of 3,617,171 requests for page views. The data was multi-source with requests made on seven different e-commerce Web sites from the Czech Republic (see Table 1). Each log file contains all the requests recorded during one hour on the seven e-commerce Web sites, whilst all the files cover a continuous 24 days period starting from 09:00AM 20th January until 08:59AM on the 13th February 2004.

Table 2. Format of page requests

ShopID	Date	IP address	SessionID	Page	Referrer
11	1074585663	213.151.91.186	939dad92c4…84208dca	/	
11	1074585670	213.151.91.186	87ee02ddcff…7655bb9e	/ct/?c=148	http://www.shop2.cz

The log files were in csv format, with each line containing a request for a page on one of the 7 e-commerce Web servers and having the 6 following fields (Table 2):

- *ShopID*: an ID of the e-commerce Web server (also denoted as shop) that received the request;
- *Date*: the Unix time of the request (seconds count since 00:00:00 1st January 1970);
- *IP address*: the computer's IP address of the user making the request;
- *SessionID*: a php session id automatically generated for each new *visit* on each server (unique IDs);
- *Page*: the requested resource (page) on the server;
- *Referrer*: the referrer of the requested page.

Along with these logs files, information on the data structure was given in different description tables. A table containing the seven shops names and their ID was provided (see Table 1, first two columns). For confidentiality reasons, the names of the seven e-commerce Web sites have been anonymized in this table as well as in the *Referrer* field when present.

Table 1. Number of requests per shop

ShopID	Site name (shop)	#Requests
10	www.shop1.cz	509,688
11	www.shop2.cz	400,045
12	www.shop3.cz	645,724
14	www.shop4.cz	1,290,870
15	www.shop5.cz	308,367
16	www.shop6.cz	298,030
17	www.shop7.cz	164,447

The Web pages on these servers are interconnected, meaning users can navigate from one *shop* to another using only the links in the pages (clicks). However, since the *SessionID* is generated when first entering a page of a web shop, a user will get a new *SessionID* when he/she switches to a yet unvisited shop.

The *Page* field contains the path on the server corresponding to the *ShopID* field to the requested page. There are 21 "page types" corresponding to the distinct 21 first level syntactic topics of all pages (see Table 3 below).

Table 3. Page types

ID	Page type	Description	#Requests	%
1	/ct	Product category	228,991	6.33
2	/ls	Product sheet	**1,363,187**	**37.68**
3	/dt	Detail of product	1,233,570	34.1
4	/znacka	List of brand names or brand detail	88,189	2.43
5	/akce	Actual offers	26,260	0.72
6	/df	Comparing product parame-ters	57,939	1.60
7	/findf	Fulltext search for products and accessories	55,139	1.52
8	/findp	Parameters based search	93,455	2.58
9	/setp	Setting displayed parameters	11,752	0.32
10	/poradna	On-line advice	107,711	2.97
11	/kosik	Shopping cart, details of contract, submitting order	35,487	0.98
12	/	Main page	219,218	6.06
13	/obchody -elektro	List of shops with electronics	10,926	0.30
14	/kontakt	Contact info	6,104	0.16
15	/faq	Frequently-asked questions	861	0.02
16	/onakupu	Info about shopping	6,659	0.18
17	/splatky	Variants of hire-purchase	2,846	0.07
18	/mailc	Availability of products	6,680	0.18
19	/mailp	Send this page	6,905	0.19
20	/mailf	Send feedback	1,855	0.05
21	/mailr	Complaint form	494	0.01
		Total	**3,564,228**	**98,45**

Using these page types and their provided descriptions, we can thus find out if the user has made a request for a specific product, category or theme, if some filters were applied, etc. For example, in the second entry presented in Table 1, the user has requested the products from the Earphones category (code 148). To describe the variables present in the requested pages, another four tables were made available:

1. *kategorie*: containing 60 descriptions of product categories, table hereafter denoted category table;
2. *list*: containing 157 descriptions of products, hereafter denoted product table;
3. *znacka*: containing 197 descriptions of product brands, hereafter denoted brand table;
4. *tema*: contains 36 descriptions of product themes, denoted theme table.

The *Referrer* field represents the URL of the Web page containing the link that the user followed to get to the current page. This field sometimes can be empty (address entered manually, blocked referrer, etc).

2.2. Intersite pre-processing method

In order to prepare the dataset for our analyses, we used a recently proposed methodology for multisites logs data preprocessing [19][20], which extends the Cooley's previous work [5].

Unfortunately, the provided raw data was not formatted in the CLF (Common Log Format [21] and some fields were not available (i.e. the status code, the user agent, logname). Thus, we rely only on the *SessionID* to identify a user's visit.

The data preprocessing was done in four steps: *data fusion*, *data cleaning*, *data structuration*, *data summarization*.

Generally, in the *data fusion* step, the log files from different Web servers are merged into a single log file. This was already done, so we only merged the 579 log files. We also changed the *Date* format into Gregorian time in order to facilitate our analyses interpretation, and merged the *Page* and the *ShopID* fields into the *URL* field (see Table 4) in order to have same format as the *Referrer* field.

Table 4. Transformed logs lines

Datetime	IP	SessionID	URL	Referrer
2004-01-20 09:01:03	213.151.91.186	939dad92c4…84208dca	http://www.shop2.cz/	-
2004-01-20 09:01:10	213.151.91.186	87ee02ddcff…7655bb9e	http://www.shop2.cz/ct/?c=148	http://www.shop2.cz/

During the *data cleaning* step, the non-relevant resources are eliminated (e.g. jpg, js files). Here, this was also already done but since the status code field is missing, we assume that all requests have succeeded (code 200).

In the *data structuration* step, requests are grouped by user, user session, and visit. On this dataset, as mentioned previously, a user changing shops can have during a single visit multiple *SessionIDs*, one on each shop. For this reason, we decided to group such *SessionIDs* that belong to a single user (same IP) into a group of sessions, corresponding to the user's actual visit. This was done by comparing the *Referrer* with the *URLs* previously accessed (in a reasonable time window), each time the user moves to another shop. If the *Referrer* (actually a page on another shop) was previously accessed, we group the two *SessionIDs* together (the actual one and the one on the previous shop). We thus grouped the existing 522,410 *SessionIDs* into 397,629 groups, equivalent to a 23.88% reduction in the user visits number. For example, in Table 4 we grouped into the same session group the two *SessionIDs* because the *Referrer* of the second line was recently requested from the same IP address and recorded in the URL field of the first line. Thus, we obtained cross-server user visits which can be used to perform global analyses on all the shops.

But these days, still for privacy reasons, a large number of security software (firewalls, anti spyware, etc) are blocking user's cookies (used for *SessionID*) and/or the URL's *Referrer* on each request. In this case, blocking the cookies and the *Referrer* will result in a different SessionID for each request that a user makes even on the same shop. This also happens in the case of a Web robot [1] used to index the pages on these Web sites or in the case of a download manager. We thus found that 2,54% IPs (2,020 out of 79,526) have 141,976 distinct requests (3,92%) corresponding to 141,976 distinct *SessionIDs* (27,18%). Knowing this, we could furthermore reduce the total number of user visits to achieve a more accurate dataset.

Finally, after identifying each variable in the accessed URL and their corresponding descriptions, we defined a relational database model to use (see Fig. 1).

Next, the preprocessed data was stored in a relational DB, during the *data summarization* step. As in [19][20], this model can be further extended by adding new tables or new attributes to the existing tables. When analyzing this data, we can select only the information which interests us. Table LOG is the main table in this model and contains on each line, information about one request for page view from the provided log files.

During the pre-processing step and the statistical analysis (see Tables 1 and 3), we found several inconsistency problems with the dataset. For example we found missing and redundant entries in the given description tables, missing variables information, malformed SessionIDs, etc. Also to have a better data

analysis, more information like the *User Agent*, the *Status Code*, the *Login Name* (if available) and the general PHP session configuration should be provided.

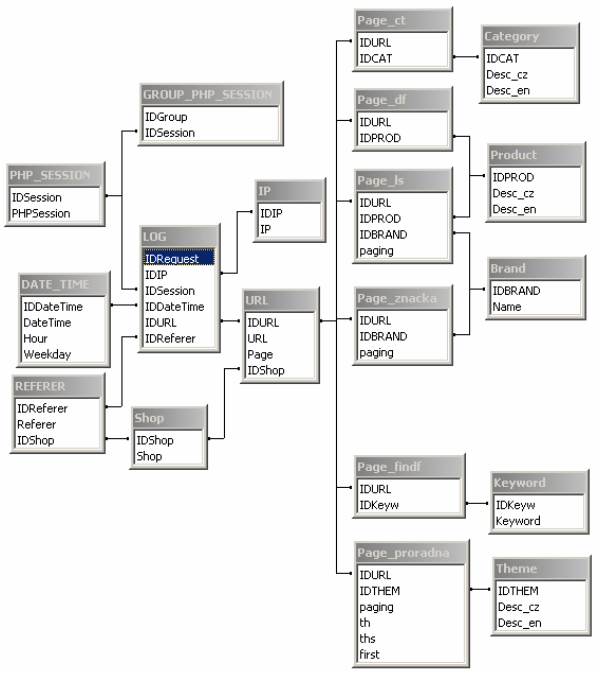

Fig. 1. Relational model of the clickStream database

In the two next sections, we illustrate the usefulness of the structured pre-processed data and the flexibility of our DB model via a statistical analysis and two clustering methods. In order to analyze the traffic charge in the seven shop sites, we have decided to use time units. For the first statistical analysis, we used the classical day and hour time unit, whilst for the two clustering methods we used slices of date and hour, which we call *Time Periods*.

3. Statistical data analysis based on time periods

The objective of this section is to show the kind of results we obtained based on an Intersite analysis and on the previously added notion of "Group of SessionsIDs", which actually represents the user *visits*.

Table 5 shows that Wednesdays are the most important days in terms of new visits (visits are grouped by their start date). More, we observe that on Wednesdays and Sundays the reduction rates of SessionsIDs to visits are the highest (more than 35%). We believe this is caused by frequent users re-connexions, fact confirmed also by the high ratio of number of SessionIDs per visit. We note that the multi-shop analysis was not significant (multi-shop visits percentage varied from 2,72 to 4,49% per day).

Table 5. Days analysis

Day	#Requests	#SesIDs	#Visits	Red. %	#MS Visits	#Ses/MS Visits
Mon	551,138	73,700	53,373	27,58	2,903	7
Tues	675,984	80,649	66,565	17,46	3,753	3,75
Wed	677,243	112,580	***73,126***	***35,04***	3,576	***11,03***
Thu	612,158	76,211	64,338	15,57	3,410	3,48
Fri	461,607	64,737	57,065	11,85	2,430	3,15
Sat	296,334	51,018	44,706	12,37	1,601	3,94
Sun	342,707	63,515	38,456	***39,45***	1,859	***13,47***
Total	**3,617,171**	**522,410**	**397,629**		**19,532**	

As for the hours analysis, Table 6 shows that four time periods (7-8, 12-14 and 20-21) are very important in terms of the server charge (re-connexions), while Fig. 2 presents the global visits distribution per hour.

Table 6. Hour analysis

Hour	#Req	#SesDs	#Visits	Red. %	#MS Visits	#Sess/ Visits
0-1	59,205	12,804	9,407	26.53	274	12,39
1-2	32,110	9,352	8,309	11.15	165	6,32
2-3	19,183	6,628	6,376	3.80	90	2,8
3-4	13,302	5,937	5,815	2.05	58	2,1
4-5	14,082	6,999	6,743	3.65	51	5,01
5-6	15,691	7,772	7,265	6.52	65	7,8
6-7	43,459	11,178	10,161	9.09	258	3,94
7-8	103,445	22,827	14,589	***36.08***	521	***15,81***
8-9	156,642	24,805	17,913	27.78	899	7,66
9-10	200,170	24,746	21,006	15.11	1,152	3,24
10-11	228,906	26,685	23,112	13.38	1,286	2,77
11-12	246,296	29,661	22,967	22.56	1,332	5,02
12-13	264,805	43,493	24,598	***43.44***	1,424	***13,26***
13-14	275,854	36,981	24,767	***33.02***	1,454	8,4
14-15	264,876	31,040	24,788	20.14	1,433	4,36
15-16	242,962	29,220	23,092	20.97	1,304	4,69
16-17	206,331	23,841	20,531	13.88	1,179	2,8
17-18	179,601	21,241	18,259	14.03	957	3,11
18-19	185,730	23,714	20,086	15.29	1,068	3,39
19-20	187,077	22,933	19,158	16.46	1,051	3,59
20-21	219,793	37,663	20,599	***45.30***	1,174	***14,53***
21-22	200,343	27,394	19,612	28.40	1,048	7,42
22-23	154,582	20,645	15,462	25.10	762	6,8
23-0	102,726	14,851	13,014	12.36	527	3,48
		Total	**397,629**		**19,532**	

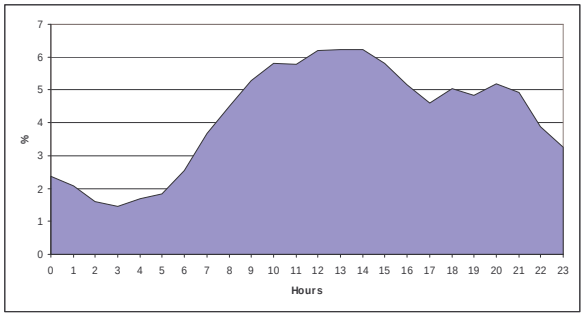

Fig. 2. Visits per hour

Fig. 3a (global visits) and 3b (multi-shop visits) show clearly the low number of customers new visits on Saturdays and Sundays during the lunch time and the high number on Tuesdays and Wednesdays.

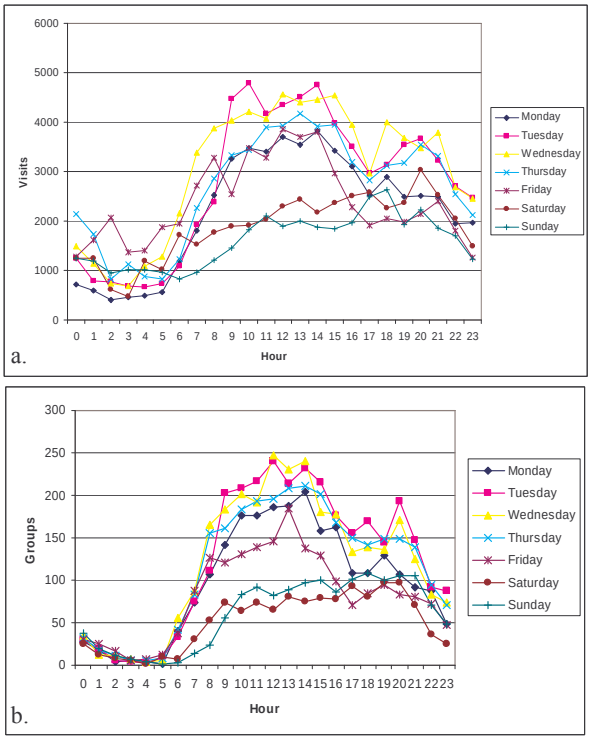

Fig. 3. Visits per day and hours: (a) globally, (b) multi-shop

4. Crossed clustering approach

4.1. Dimension problem

Appropriate use of a clustering algorithm is often a useful first step in extracting knowledge from a database. Clustering, in fact, leads to a classification, *i.e.* the identification of homogeneous and distinct subgroups in data [7] [4], where the definition of homogeneous and distinct depends on the particular algorithm used: this is indeed a simple structure, which, in the absence of a priori knowledge about the multidimensional shape of the data, may be a reasonable starting point towards the discovery of richer, more complex structures.

In spite of the great wealth of clustering algorithms, the rapid accumulation of large databases of increasing complexity poses a number of new problems that traditional algorithms are not equipped to address. One important feature of modern data collection is the ever increasing size of a typical database: it is not so unusual to work with databases containing from a few thousands to a few millions of individuals and hundreds or thousands of variables. Now, most clustering algorithms of the traditional type are severely limited as to the number of individuals they can comfortably handle (from a few hundred to a few thousands).

To analyze the traffic charge in the seven shop sites, we grouped the requests in terms of *Time Periods*. We also limited our clustering analysis to the 1,363,187 requests registered on the page type *ls* for the first analysis and to the 490,883 requests registered on the shop 4 and same page type *ls* for the second analysis.

4.2. Method

According to our aim to obtain rows partition, a classification of the complex descriptors is accomplished. Some authors [8][9] proposed the maximization of the chi-squared criterion between rows and columns of a contingency table. The main advantage to use our algorithm is to get a tool for comparing and clustering aggregated and structured data. Our approach preserves the flexibility and the generality of the dynamic algorithm in the Knowledge Discovery. The cluster characterization is based on the distributions or multi-valued descriptors and the interactions between descriptors and individuals.

As in the classical clustering algorithm the criterion optimized is based on the best fitting between classes of objects and their representation. In our context of analyze the relations between time periods and products, we propose to represent the classes by prototypes which summarize the whole information of the time periods belonging to each of them. Each prototype is even modeling as an object described by multi-categories variables with associated distributions. In this context, several distances and dissimilarity functions could be proposed as assignment. In particular, if the objects and the prototypes are described by multi-categories variables, the

dissimilarity measure can be chosen as a classical distance between distributions (*e.g.* chi-squared).
The convergence of the algorithm to a stationary value of the criterion is guaranteed by the best fitting between the type representation of the classes and the properties of the allocation function. Different algorithms, even referred the same scheme, has been proposed according to the type of descriptors and to the choice of the allocation function.
The generalized dynamic algorithm [15] [26] on objects has been proposed in different contexts of analysis, for example: to cluster archaeological data, described by multi-categorical variables [25]; to compare social-economics characteristics in different geographical areas with respect to the distributions of some variables (e.g.: economics activities; income distributions; worked hours; etc).

4.3. Analysis

4.3.1. Time periods/product data generation from the DB

Here, we have considered a crossed table where each line describes a individual that covers the couple weekday and hour of requests for the page /ls in the shop 4 (the more visited one according to Table 1), and the column describes one multi-categorical variable witch represents the number of products requested by users into a specific time slice (see Table 7, where we have 7 x 24 individuals). We have limited our clustering analysis to the requests registered on the shop 4 although the same analysis can be made for all the other shops.

Table 7. Product clustering

Product_5 Cardinal: 1
/product/Free standing combi refrigerators

4.3.2 Results

We return now to the subset of the time period dataset describing the products accessed on shop 4 (see Table 8). The 168 periods of time summarize 490,883 requests on all products from shop 4. Table 9 presents the results after applying the crossed clustering method [15] specifying 7 classes of periods and 5 classes of products.

Table 8. Quantity of products requested by weekday x hour and registered on shop 4

Weekday x Hour	Product (number of requests)
Monday_0	Built-in electric hobs (10), Built-in dish washers 60cm (64), Corner single sinks (50), ...
Monday_1	Free standing combi refrigerators (44), Corner single sinks (50), Built-in hoods (60), ...
...	…
Sunday_22	Built-in microwave ovens (27), Built-in dish washers 45cm (38), Built-in dish washers 60cm (85), ...
Sunday_23	Built-in freezers (56), Kitchen taps with shower (45), Garbage disposers (32), ...

Table 9. Confusion table

	Prod_1	Prod_2	Prod_3	Prod_4	Prod_5	Total
Per_1	2847	5084	3284	2265	2471	15951
Per_2	11305	31492	12951	1895	9610	67253
Per_3	33107	55652	36699	5345	20370	151173
Per_4	22682	46322	30200	5165	27659	132028
Per_5	9576	20477	19721	2339	7551	59664
Per_6	1783	3515	2549	392	*11240*	**19479**
Per_7	15019	14297	8608	1397	6014	45335
Total	96319	176839	114012	18798	**84915**	490883

Surprisingly the product class 5 was defined only by one product, namely *Free standing combi refrigerators* witch was consulted predominantly on Fridays between the hours 17:00 and 20:00. It is important to note that although this product belongs to the a class product (see Table 7) which is responsible for only 17.3 % of the total requests on shop 4, the requests occurring on its period (see Table 10) is 57.7 % based on this product. In other words it means that product *Free standing combi refrigerators* is the more requested on Fridays from 17:00 to 20:00. Such information could be used on marketing strategies, like cross selling involving this product.

Table 10. Time period clustering

Period_6 Cardinal: 8
Friday_2, Friday_6, Friday_17, Friday_18, Friday_19, Friday_20, Saturday_5, Tuesday_4

5. Conclusions and future works

In this paper, we first proposed a pre-processing method in the context of a multi-shop e-commerce domain. Our analysis on the proposed PKDD Web logs showed the great flexibility offered by the built data warehouse and the benefits of an enriched data structuration in terms of customer visits. Secondly we presented two original clustering analysis based on

efficient clustering methods applied on Web time period-based clickstreams: our first results on the used dataset in a short time due to the PKDD challenge are promising. Such analyses allow us to identify best hours for marketing strategies, like fast promotions, on-line advices and publish banners, etc. Others analyses could be planned (cf. Table 3) in the future, exploiting for example the link between the consumer activities and the time periods by shop or focusing on multi-shop user visits, etc.

References

1. ABCInteactive.com: Spiders and Robots. http://www.abcinteractiveaudits.com/abci_iab_spidersandrobots/.
2. Ambroise, C., Séze, G., Badran, F., Thiria, S.: Hierarchical clustering of Self-Organizing Maps for cloud classification. *Neurocomputing*, 30 , pp 47—52, 2000.
3. Arnoux, M., Lechevallier, Y., Tanasa, D., Trousse, B., Verde, R.: Automatic Clustering for the Web Usage Mining. In Dana Petcu and Daniela Zaharie and Viorel Negru and Tudor Jebeleanu, editor, Proceedings of the 5th Intl. Workshop on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC03), Pages 54 -- 66, Editura Mirton, Timisoara, 1-4 October 2003.
4. Bock, H. H.: Classification and clustering~: Problems for the future. In: E. Diday, Y. Lechevallier, M. Schader, P. Bertrand, B. Burtschy (eds.): *New Approaches in Classification and Data Analysis*. Springer, Heidelberg 3—24, 1993.
5. Ciampi A., Lechevallier Y.: Clustering large: An approach based on Kohonen Self Organizing Maps, Proceedings of *PKKD 2000*, Zighed, D. A., Komorowski J., Zytkow J. (Eds), Springer-Verlag, Heidelberg pp 353-358, 2000.
6. Cooley, R.: Web Usage Mining: Discovery and Application of Interesting Patterns From Web Data. PhD Thesis, Dept of Computer Science, Univ. of Minnesota, 2000.
7. Gordon, A. D.: *Classification: Methods for the Exploratory Analysis of Multivariate Data*. Chapman& Hall, London, 1981.
8. Govaert, G.: Algorithme de classification d'un tableau de contingence. In Proc. of *first international symposium on Data Analysis and Informatics*, INRIA, Versailles, pp 487—500, 1977.
9. Govaert, G., Nadif M.: Clustering with block mixture models. *Pattern Recognition*, Elservier Science Publishers, 36, pp 463-473, 2003.
10. Hébrail, G., Debregeas, A.: Interactive interpretation of Kohonen maps applied to curves. Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining. AAAI press, Menlo Park pp 179—183, 1998.
11. Jaczynski, M.: Scheme and Object-Oriented Framework for case Indexing By Behavioural Situations : Application in Assisted Web Browsing. Doctorat Thesis of the University of Sophia-Antipolis (in french), december, 1998.
12. Jaczynski, M., Trousse, B.: WWW Assisted Browsing by Reusing Past Navigations of a Group of Users. In Advanced in Case-based Reasoning, 4th European Workshop on Case-Based Reasoning , Lecture Notes in Artificial Intelligence, 1488, pages 160-171, 1998.
13. Kohavi, R.: Mining E-Commerce Data. KDD 01, San Francisco CA, USA
14. Kohonen, T.*: Self-Organizing Maps*. Springer, New York, 1997.
15. Lechevallier, Y., Verde, R.: Crossed Clustering method: An efficient Clustering Method for Web Usage Mining. In: Complex Data Analysis, Pekin, Chine, October 2004.
16. Mobasher, B.: Mining Web Usage Data for Automatic Site Personalization. In: Proc. 24th Annual Conference of the Gesellschaft Fur Klassifikation E.V., University of Passau, March (2000) 15–17
17. Perkowitz, M., Etzioni, O.: Adaptive sites: Automatically learning from user access patterns. In: Proc. 6th Int'l World Wide Web Conf., Santa Clara, California, April (1997)
18. Srivastava, J. and al.: Web usage mining: Discovery and applications of usage patterns from Web data. ACM SIGKDD Explorations, Vol.1, N.2, January (2000)
19. Tanasa, D., Trousse, B.: Advanced Data Preprocessing for Intersites Web Usage Mining. IEEE Intelligent Systems, Vol. 19(2):59--65, April 2004.
20. Tanasa, D.: Web Usage Mining : Contributions to Intersites Logs Preprocessing and Sequential Pattern Extraction with Low Support, PhD Thesis, University of Sophia Antipolis, June 2005.
21. W3C. Logging Control In W3C httpd. http://www.w3.org/Daemon/User/Config/Logging.html#common-logfile-format, July(1995)
22. Sauberlich, F., Huber K.-P. :A Framework for Web Usage Mining on Anonymous Logfile Data. In : Schwaiger M. and Opitz O.(Eds.): *Exploratory Data Analysis in Empirical Research*, Springer-Verlag, Heidelberg, 309—318, 2001.
23. Srinivasa Raghavan, NR.: Data Mining in e-commerce: A survey. Sadhana (Academy "Proceedings in Engineering Sciences"), Vol. 30, parts 2 & 3, April/June 2005, Indian Academy of Sciences (publisher), pp 275-289.
24. Trousse, B., Jaczynski, M., Kanawati, K.: Using User Behavior Similarity for Recommandation Computation : The Broadway Approach, In proceedings of 8th international conference on human computer interaction (HCI'99), Munich, August, 1999.
25. Verde, R., De Carvalho, F.A.T., Lechevallier, Y. :A Dynamical Clustering Algorithm for Multi-Nominal Data. In : H.A.L. Kiers, J.-P. Rasson, P.J.F. Groenen and M. Schader (Eds.): *Data Analysis, Classification, and Related Methods*, Springer-Verlag, Heidelberg, pp 387—394, 2000.
26. Verde, R., Lechevallier, Y.: Crossed Clustering method on Symbolic Data tables. In :M. Vichi, P. Monari, S. Migneni, A. Montanari (Eds.): *New developments in Classification, and Data Analysis*, Springer-Verlag, Heidelberg, pp 87—96, 2003.