World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:6, No:8, 2012

Clustering Methods Applied to the Tracking of user Traces Interacting with an e-Learning System

Larbi Omar, Elberrichi Zakaria

Open Science Index, Computer and Information Engineering Vol:6, No:8, 2012 waset.org/Publication/15037

***Abstract*—**Many research works are carried out on the analysis of traces in a digital learning environment. These studies produce large volumes of usage tracks from the various actions performed by a user. However, to exploit these data, compare and improve performance, several issues are raised. To remedy this, several works deal with this problem seen recently. This research studied a series of questions about format and description of the data to be shared. Our goal is to share thoughts on these issues by presenting our experience in the analysis of trace-based log files, comparing several approaches used in automatic classification applied to e-learning platforms. Finally, the obtained results are discussed.

***Keywords*—**Classification, , e-learning platform, log file, Trace.

I. INTRODUCTION

THE growth in the use of the web and more specifically e-learning platforms were accompanied by a special interest in data analysis in order to serve users and to present customized content. Recently, new approaches have incorporated the use of data mining techniques in the field of e-learning. The field of e-learning has grown considerably in recent years as a result of the increasing number of visitors and consultation to web documents. In this context, this work presents a summary report based on the analysis of various recent works in this area.

The behavior of the user on a website is represented by a series of mouse clicks and typed on a keyboard. This information triggered queries that result in the display of the pages. These queries are recorded in the log file. Analysis of log files allows determining, for example, what are the queries that do not (missing page, broken link ...) or what is the frequency of a specific page. Each line of this file provides information about the user, date and time of the request, the page or the course requirements, the action taken and some information related to the visited page.

The ultimate goal of this work is to classify users into previous unknown classes, this classification allows for different types of profiles that are interested in the services available on this platform. But we still have some questions posed in this gap:

How an individual is assigned a specific class and what is the relevance and usefulness of the classification decision point of view, to provide some answers to these questions, several research projects and start this thread: ([Reffay & Betbeder 09], [SETTOUTI et al. 09], [DYKE, et al. 10], [Pham Thi Ngoc et al. 09], [Ziani, 2007] [Merzoug, 2009], [CHOQUET & Iksal 07], [CHARRAD 10], [Bousbia 11])[2]…

II. TRACES CONCEPT

The notion of trace covers various thoughts such as footprint, memory and writing. In the field of data processing, due to the lack of a precise definition of the notion of trace, we chose the definition of [CHAMPIN et al. 04] which defines the trace as a sequence of states and transitions representing user activity: “the time sequence of objects and operations mobilized by the user when using the system is called trace of use”. Its study has become important, especially with the digitization of traces and the development of specific tools for treatment. In information technology and particularly in communication systems, where all communication processes produce trace, written or not (text, data, fingerprints,) observation of the communication process necessitate the collection of such traces, their classification and even their organization. This is also applicable to interaction between a human and a machine. It is in this context that we focus on traces and their exploitation. In this frame, we chose a standard model of a trace.

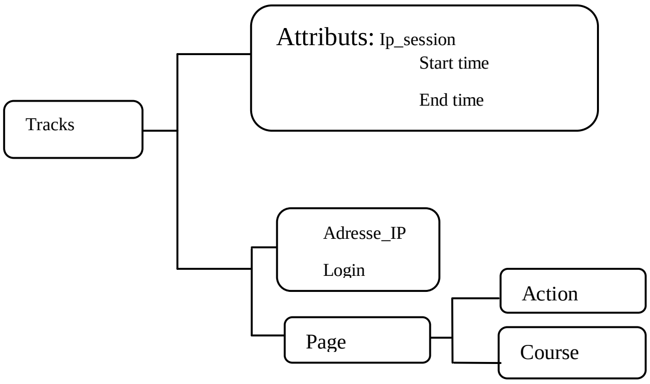
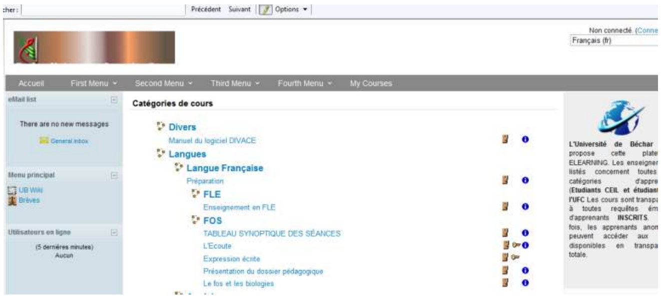


Fig. 1 Standard model of a trace

L. O. is with University of Bechar, BP: 417 - 08000 Bechar – Algeria (e-mail: omarlarbi@yahoo.fr).

E. Z., was with University of Sidi Bel Abbes - Algeria.

As application, we used two file logs, one is associated to the moodle e-learning platform (Fig. 2) of Bechar University (http://www2.univ-bechar.dz/elearning/), during the period from 28/04/2010 to 01/06/2010 and the other one associated to the efad platform (http://www.efad.ufc.dz). Each of the two files has a different attributes format. To solve this problem, one proposed a common model of trace (Fig. 1).



There are no new messages
General inbox

(5 dernières minutes)
Aucun

Catégories de cours

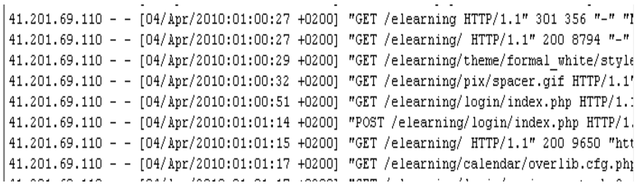
- Divers
  - Manuel du logiciel DIVACE
- Langues
  - Langue Française
    - Préparation
    - FLE
      - Enseignement en FLE
    - FOS
      - TABLEAU SYNOPTIQUE DES SÉANCES
      - L'Ecoute
      - Expression écrite
      - Présentation du dossier pédagogique
      - Le fos et les biologies

L'Université de Béchar propose cette plate ELEARNING. Les enseigner listés concernent toutes catégories d'appre (Etudiants CEIL et étudiant l'UFC Les cours sont transpa à toutes requêtes ém d'apprenants INSCRITS. fois, les apprenants anon peuvent accéder aux disponibles en transpa totale.

Fig. 2 E-learning platform of the university of bechar

In this paper, we are interested in the traces of interaction between user and a computer system. These traces are listed in a log file which is stored in memory supports.

A log file, also named log connections, query history or journal [6], is a file that stores all the actions in a sequence that occurred with a server. It can be transfer error log file, a repository log file or agent log file. In this paper, we are interested in the repository type log file [8].



```
41.201.69.110 - - [04/Apr/2010:01:00:27 +0200] "GET /elearning HTTP/1.1" 301 356 "-" "I
41.201.69.110 - - [04/Apr/2010:01:00:27 +0200] "GET /elearning/ HTTP/1.1" 200 8794 "-"
41.201.69.110 - - [04/Apr/2010:01:00:29 +0200] "GET /elearning/theme/formal_white/style
41.201.69.110 - - [04/Apr/2010:01:00:32 +0200] "GET /elearning/pix/spacer.gif HTTP/1.1'
41.201.69.110 - - [04/Apr/2010:01:00:51 +0200] "GET /elearning/login/index.php HTTP/1.:
41.201.69.110 - - [04/Apr/2010:01:01:14 +0200] "POST /elearning/login/index.php HTTP/1.
41.201.69.110 - - [04/Apr/2010:01:01:15 +0200] "GET /elearning/ HTTP/1.1" 200 9650 "htt
41.201.69.110 - - [04/Apr/2010:01:01:17 +0200] "GET /elearning/calendar/overlib.cfg.php
```

Fig. 3 The form of log file

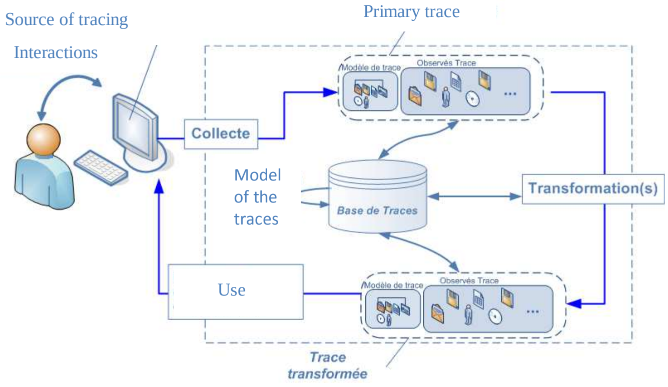


Fig. 4 Principle operation of the system based on traces [1]

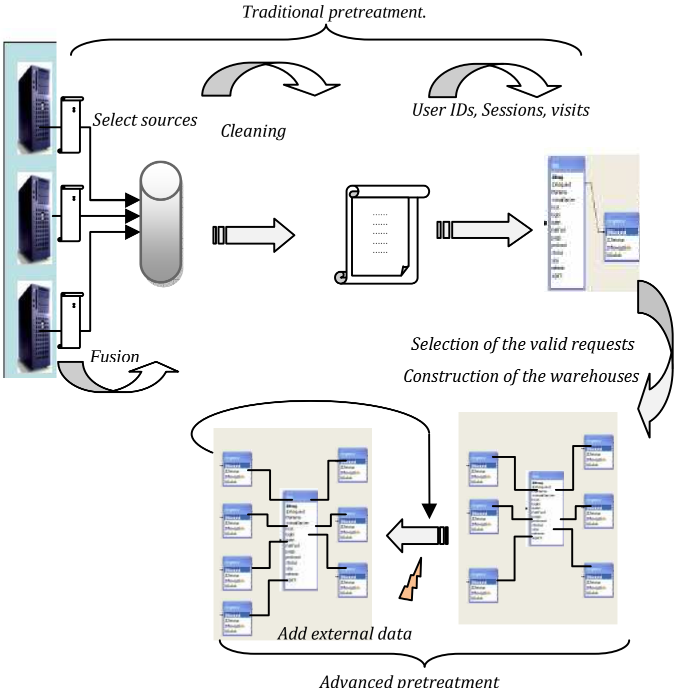
III. Process of Web Mining

There are three phases in the process of web mining: preprocessing, the discovery of navigation patterns (pattern discovery) and the analysis of these models (pattern analysis). This generic process is adapted to each axis of Web mining according to the nature of the used data (text, log, links ...).

* Pretreatment *Discovery models * Analysis of models

A. Pretreatment Process Data

The pretreatment process is an important initial phase. It usually consists of five phases:
Cleaning, structuring, transformation, reprocessing, extension.



Traditional pretreatment.
Select sources
Cleaning
User IDs, Sessions, visits
Fusion
Selection of the valid requests
Construction of the warehouses
Add external data
Advanced pretreatment

Fig. 5 Process for pretreatment of log files

Indeed, the format of log files that we have chosen is different than CLF format as these files contains user traces on a platform e-learning. Their cleaning and structuring are needed before any analysis. First, we present a methodology for pretreatment and the results of its application on the log files.

B. Identification of users

The identification of users from the log file is not a simple task. Indeed, the IP address and even the name of the user do not identify a user.

IP Address: The same IP address can be assigned to multiple users.

Username: The existence of users with similar names.

In this case, the pair (username, password) is an identifier of a user.

C. Data cleaning

Data cleaning means the remove of unnecessary queries log files such as requests from anonymous visitors, from teachers or the webmaster, The failed connection requests or invalid requests, requests for connection and disconnection of users and request for images.

D. Limits the use of log files

These files are disadvantaged by their heterogeneity, being full of useless details and lack of legibility.

Open Science Index, Computer and Information Engineering Vol:6, No:8, 2012 waset.org/Publication/15037

IV. Experiments and Results

A. Introduction

We will carry out experiments on some algorithms of classification, in order to test and compare the performance of these algorithms. In our experiment, we chose three algorithms which belong to the two big families of classification (hierarchical, partitioning): CLARA and BIRCH for the first family and K-means for the other one. These three algorithms proved their effectiveness in solving problems similar to ours. Indeed, CLARA algorithm can treat the bulky data sources, which corresponds to the case considered in this paper. On the other hand BIRCH does not treat that important extracts of the considered data base.

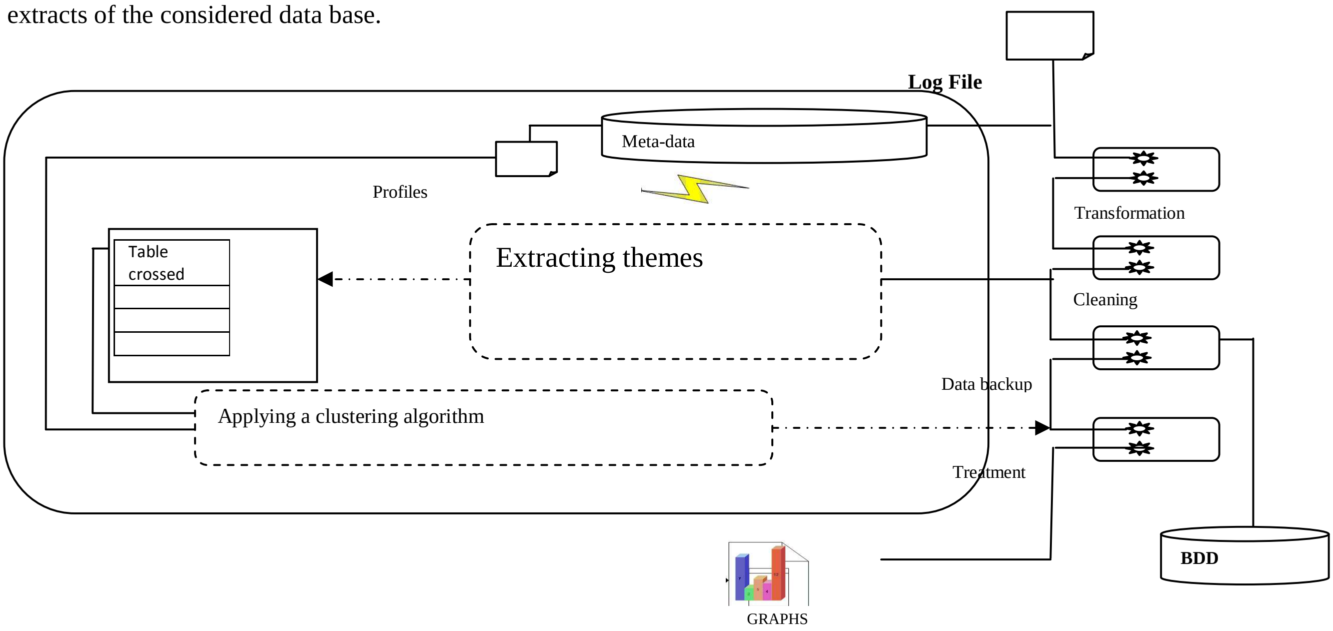


Fig. 6 Model of suggested analysis

The figure above shows the general model that includes the steps of web usage mining process applied to a log file.

We will test, initially, the algorithms of clustering such as: K-means, Clara and Birch. Series of individual tests will be carried out with the aim of determining the best parameters for each algorithm. The performances (execution time, quality of the results…) will be then measured for all the algorithms then compared to each other. All tests are performed on a Toshiba machine with an Intel I3 of 2 GB of RAM running under Microsoft Windows © XP Professional.

B. Description of algorithms

1) The algorithm k-Means

The algorithm k-Means is considered as standard algorithm in the field of clustering due to its simplicity of implementation and ability to handle large population.

2) The algorithm CLARA

CLARA is an algorithm that applies hierarchical algorithms on several random samples.

3) The algorithm BIRCH

BIRCH is an agglomerative algorithm of clustering hierarchical. It was introduced by Zhang, Ramakrishan and Livny. The community of the classification filed finds that BIRCH is one of the best algorithms being able to treat large data files [6]. The principal idea is that classification is carried out on a whole of compacted objects and organized in a structure of balanced tree called CF_Tree (Clustering Feature Tree) of limited size, where each level presents a clustering phase.

C. Application of clustering methods

The quality of the results of clustering algorithms and their performance depends eventually on their parameters. We will, test each of our algorithms with different parameters values.

TABLE I
COMPARATIVE RESULTS TABLE

Algorithm	number of the documents classified	number of the classes.	Execution time	inertia Intra/Inter
K-means	1775	6	313	36.57
Clara	1775	6	997	70.20
Birch	1775	6	2871	23.37

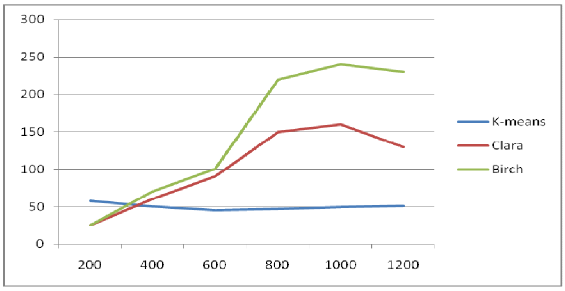


Fig. 7 The contribution of execution time according to number of iterations

D. Discussion

We carried out a series of tests on data mining algorithms using real data. These tests allowed us to determine the best settings for each algorithm. First, the tests of K-means algorithm have shown that this algorithm gives good results either in time or in accuracy. We also concluded that the size of the chosen test population should not be too big or too small to achieve a good compromise between execution time and quality of the result. We have shown, through comparative tests that the algorithm CLARA has the best execution time dealing with large populations. Indeed, we were confronted with various difficulties in the implementation of the algorithm Birch. Finally, we deduced that the number of classes must be fixed by the user according to his needs.

V. CONCLUSION

Teaching by using the platforms e-Learning is in full expansion because of the increasing number of platform visitors. The lack of adaptability of educational content is the main difficulty facing e-learners. In this context, we proposed to study traces of interaction of learners with these e-learning systems.

In this paper, a comparative study between platform e-learning user's automatic classification algorithms is developed. We started, at first, to define the various concepts related to data mining process and the KDD. We have shown that data mining is a partial process of the process KDD. In a second step, we have performed some data mining algorithms. We have seen that there is a wide variety. Each algorithm is adapted to a particular context, so it can succeed in one context and fail in another. It can be concluded that choosing the right algorithm is based on both data and needs. Tests on classification algorithms have shown the high sensitivity of the k-means algorithm which shows a good effectiveness when its parameters are properly chosen.

As perspective, we propose carrying out a detailed study on the different approaches currently used in classification especially new approaches such as: the mixture model and methods of bi-partitioning. Besides that, at the level of different approaches performance comparison, concrete results of these works can be used as comparison criterion.

REFERENCES

[1] N. Bousbia, "Analyse des traces de navigation des apprenants dans un environnement de formation dans une perspective de détection automatique des styles d'apprentissage", PhD thesis in Computer Science, University Pierre and Marie Curie (France) and Higher National School of Computer Science, ESI, Algeria, 2011.

[2] M. Charrad, "Une approche générique pour l'analyse croisant contenu et usage des sites web par des méthodes de bipartitionnement", Presented for obtaining a doctorate in Computer Science from the CNAM, Paris and ENSI, University of Manouba, 2010.

[3] W. Hengshan and al., "Design and implementation of a web usage mining model based on fpgrowth and prefixspan". In: Communications of the IIMA, 2006.

[4] L. Settouti, Y. Prié, A. Mille, J-C. Marty, "Système à base de traces pour l'apprentissage humain", In: ICTE International Symposium, Information Technology and Communication in Higher Education and Enterprise, Toulouse, 2006.

[5] B. Arnaud, “Personnalisation et prise en compte du contexte dans les modèles conceptuels pour la conception des si”, Prise Accounting for the User Information Systems, Proceedings pecus, Toulouse, 2009.

[6] BENATCHBA, “Application de techniques de data mining pour la classification automatique des données”. Thesis studies to obtain the engineering degree in Computer Science, 2010.

[7] M. Charrad, “Techniques d’extraction de connaissances appliquées aux données du web”, Master Thesis in Computer Science, National School of Computer Science, University of Manouba, Tunisia, 2005.

[8] H. BENSEFIA, “Fichiers logs: preuves judiciaires et composant vital pour forensics”, RIST Vol.15 No. 01-02, 2005.

[9] F. Marius, “Data mining, fouille de données: concepts et techniques”. Faculty of Medicine, Marseille, 2006.

[10] P. Giacomini, “Un environnement collaboratif d’enseignement à distance adapté au profil de l’apprenant”. International Journal of Information Sciences for Decision Making, TICE Mediterranean Sfax Tunisia, 2008.