Afrique SCIENCE *02(3) (2006) 335 - 351*
ISSN 1813-548X

Traitement et exploration du fichier Log du serveur web, pour l'extraction des connaissances : Web usage mining

Mostafa Hanoune* et Faouzia Benabbou

Laboratoire des technologies de l'information et modélisation (TIM), Faculté des sciences Ben M'Sik, Université Hassan II, Mohammedia, Casablanca, Maroc

(Reçu le 10 Juillet 2006, accepté le 19 Septembre 2006)

* Correspondance, courriel : *m_hanoune@yahoo.fr*

Résumé

L'objectif de ce travail est la conception et la réalisation d'un outil logiciel, en utilisant les concepts du « Web usage mining », qui permettra au « Webmaster » d'avoir l'ensemble des connaissances sur le site web qu'il gère, en vue d'une amélioration et personnalisation.
Il s'agit en fait, d'extraire de l'information à partir du fichier Log du serveur Web, hébergeant le site Web, et prendre les décisions pour découvrir les habitudes des internautes, et de répondre à leurs besoins en adaptant le contenu, la forme et l'agencement des pages web.

Mots-clés : *« Web usage mining », extraction de connaissances, fichier Log, méthode « A Priori », algorithme, théorie de décision.*

Abstract

Log file treatment and exploration of the web site, for extracting and mining the knowledges : Web usage mining

The aim of this work is to design and produce one tool and implement, base on KDD (Knowledge discoverer on data base), by using the concept of Web usage mining, to give to the Webmasters, collections of Knowledge, including the statistics on site, in order to take the favourable decisions.
We will use the "Log file" on the Web server, accommodating the site Web, to discover the habits of users and respond to there requirements and needs, the shape and layout of Web pages and the contain generally speaking, with waiting of users.

Keywords : *Web usage mining, extraction of knowledge, log file, «A Priori» method, decision theory.*

1. Introduction

L'activité sur le Web et les données résultantes ont connu une croissance très rapide, vu la croissance exponentielle du nombre des documents mis en ligne. D'après des statiques sur des sites spécialisés, le nombre des utilisateurs d'Internet dans le monde a dépassé le milliard (1 076 203 987), au mois de novembre 2006 [1], et le nombre de sites Web a atteint (105 244 649), au mois de décembre 2006 [2]. Ces données, en particulier celles relatives à l'usage du Web, sont traitées dans le Web Usage Mining (WUM) [3]. Dans cet article, nous nous intéressons à l'analyse des fichiers « Log » [4-5], afin de comprendre le comportement des internautes sur un site Web (*Site de l'université Hassan II-Mohammedia du Maroc* : *www.univh2m.ac.ma*).
L'apport de ce travail réside principalement dans les points suivants :

1-1. Connaissances sur les visiteurs

Détenir le pourcentage des visiteurs par semaine, par mois et par an,
Avoir une visibilité internationale : d'où proviennent les visiteurs ?

1-2. Connaissances sur les pages

Reconnaître les pages Web les plus et les moins consultées (pages populaires et pages impopulaires),
Connaître les combinaisons des pages consultées,
Savoir quels sont les liens qui référencent le mieux.

1-3. Connaissances sur les navigateurs et les « OS »

Connaître le pourcentage des navigateurs les plus utilisés,
Connaître également le pourcentage des systèmes d'exploitations les plus utilisés,

Le présent article est donc subdivisé en trois sections distinctes :
- La première section présente la conception, par la méthode UML, de la solution mise en place.
- Dans la deuxième section, on trouve les différentes étapes du prétraitement et nettoyage du *fichier Log.*
- La dernière est consacrée à l'exploration et l'analyse du *fichier Log.*

2. Analyse du problème et conception de la solution méthode *UML*

2-1. Diagramme de cas d'utilisation

Dans cette étape, il s'agira de structurer les besoins des utilisateurs et les objectifs correspondants.

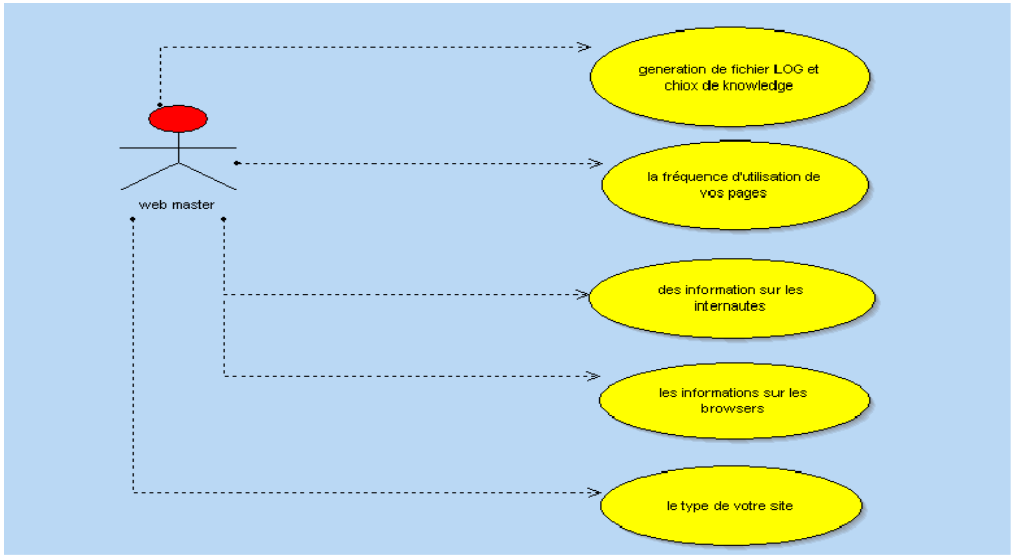


Figure 1 : *Cas d'utilisation*

2-2. Diagramme de classe

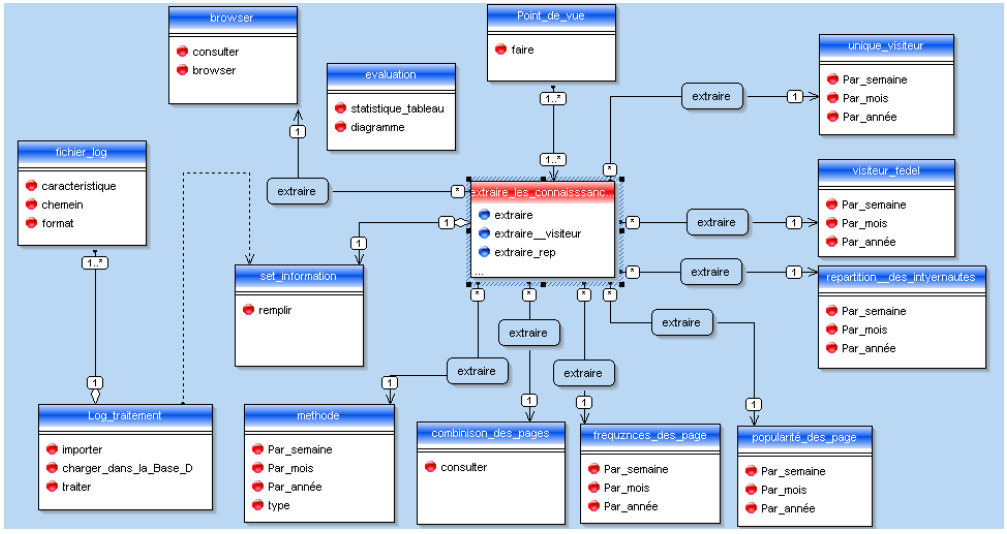


Figure 2 : *Diagramme des classes*

A la ***Figure 1***, le cercle en rouge signifie que c'est une fonction. On utilisera le diagramme de classe pour montrer la structuration du modèle à adopter, en mettant en évidence les différentes associations et lien entre les classes.

2-3. Diagramme d'état de transition

2-3-1. DET de traitement de fichier LOG

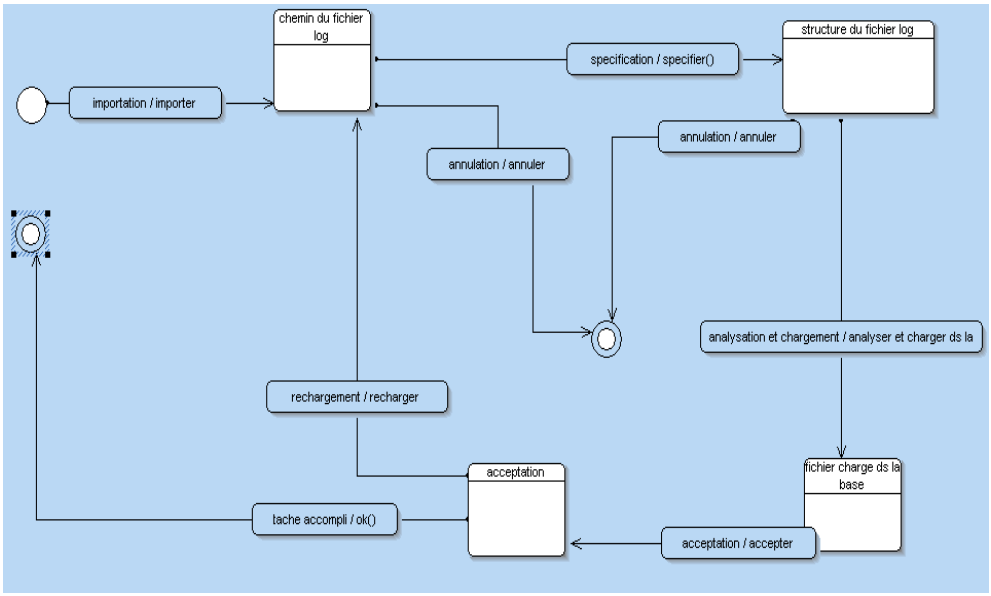


Figure 3 : *Etats de transition*

Le diagramme d'état de transitions nous permettra de décrire les changements d'états d'un objet ou d'un composant, en réponse aux interactions avec d'autres objets/composants ou avec des acteurs.

2-3-2. DET extraction connaissance

On montrera le processus suivi pour aboutir aux connaissances souhaitées.

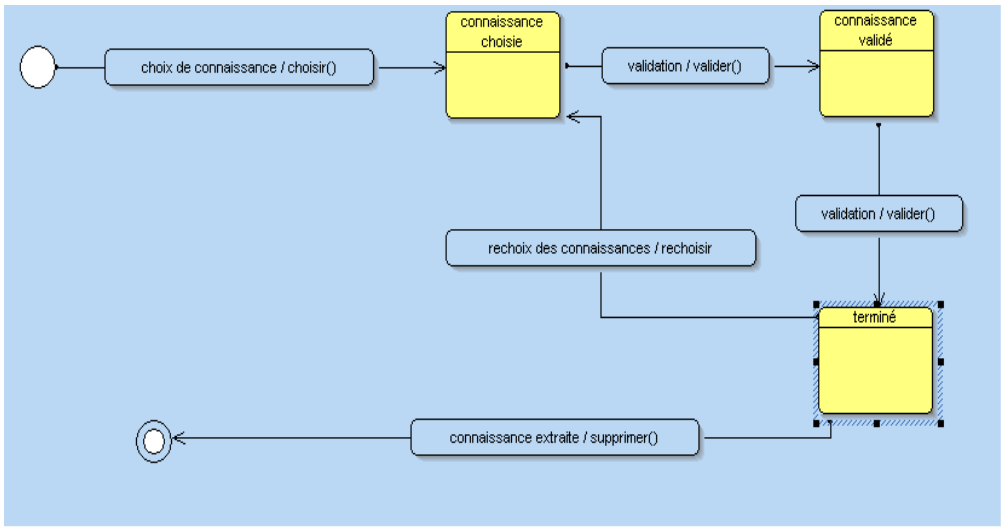


Figure 4 : *Diagramme d'extraction de la connaissance*

2-4. Diagramme de séquences

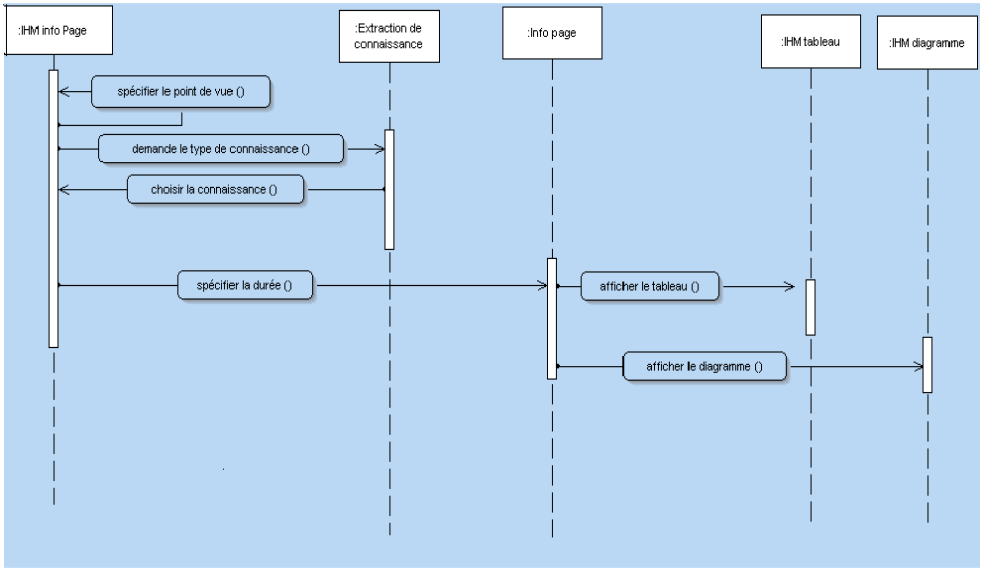


Figure 5 : *Séquences d'utilisation*

Le diagramme de séquences permet de représenter des collaborations entre objets selon un point de vue temporel, on y met l'accent sur la chronologie des envois de messages.

3. Prétraitement et nettoyage du *fichier Log*

3-1. Chargement du *fichier Log* et transformation en une table d'une BDD

Le *fichier LOG* est un fichier Texte, appelé aussi journal des connexions, qui conserve les traces des requêtes et des opérations traitées par le serveur. Généralement il est de la forme suivante, pour laquelle les différents champs de ce fichier seront importés dans une base de données définie comme suit :

- Le *fichier Log* se transforme en une table composée de plusieurs colonnes. Chaque colonne correspond à un champ spécifique du *fichier Log* ,
- La colonne « hote_client » correspond aux adresses IP des visiteurs,
- La colonne « login_client » correspond au Nom du serveur utilisé par le visiteur,
- La colonne « utilisateur_client » correspond au Nom de l'utilisateur (en cas d'accès par mot de passe),
- La colonne « date_et_heure » correspond à la date d'accès,
- La colonne « methode » correspond à la méthode utilisée (GET/POST),
- La colonne « url_des_pages » correspond au URL demandé,

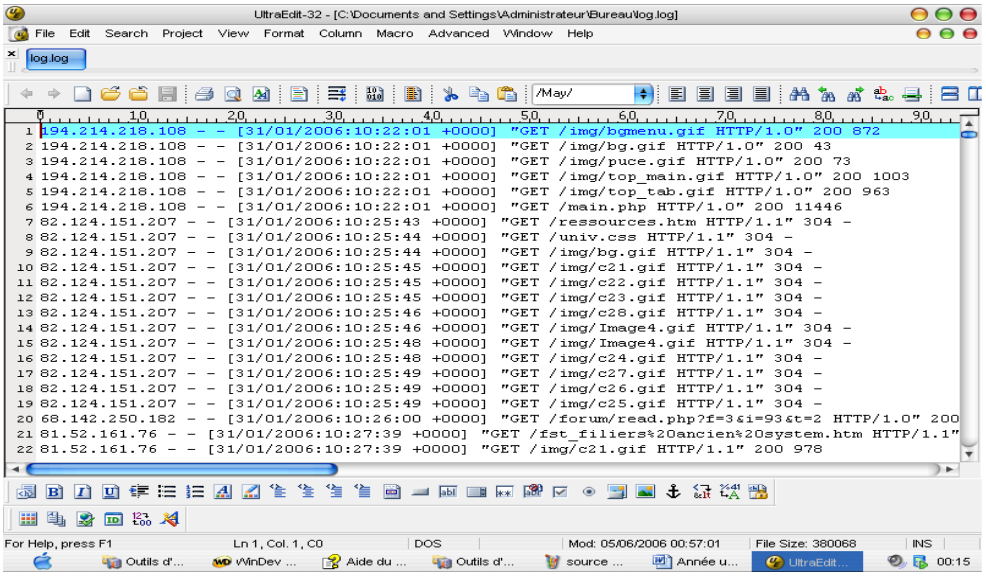


Figure 6 : *Fichier Log brut*

- La colonne « protocole » correspond au protocole utilisé,
- La colonne « code_de_retour »,
- La colonne « taille_chargé » correspond à la taille chargée, ce fichier doit impérativement être pré-traité et nettoyé [6].

hote_client	login_client	utilisateur_clien	date_et_heure	methode	url_des_pages	protocole	code_de_retour	c
127.0.0.1	-	-	l/2005 01:25:09	GET	/uh2m/	HTTP/1.1"	200	3
127.0.0.1	-	-	l/2005 01:25:23	GET	/	HTTP/1.1"	200	2
127.0.0.1	-	-	l/2005 01:25:23	GET	/images_easyp	HTTP/1.1"	304	-
127.0.0.1	-	-	l/2005 01:25:23	GET	/images_easyp	HTTP/1.1"	304	-
127.0.0.1	-	-	l/2005 01:25:23	GET	/images_easyp	HTTP/1.1"	304	-
127.0.0.1	-	-	l/2005 01:25:23	GET	/images_easyp	HTTP/1.1"	304	-
127.0.0.1	-	-	l/2005 01:25:23	GET	/images_easyp	HTTP/1.1"	304	-
127.0.0.1	-	-	l/2005 01:32:25	GET	/	HTTP/1.1"	200	2
127.0.0.1	-	-	l/2005 01:32:25	GET	/images_easyp	HTTP/1.1"	304	-
127.0.0.1	-	-	l/2005 01:32:25	GET	/images_easyp	HTTP/1.1"	304	-
127.0.0.1	-	-	l/2005 01:32:25	GET	/images_easyp	HTTP/1.1"	304	-
127.0.0.1	-	-	l/2005 01:32:25	GET	/images_easyp	HTTP/1.1"	304	-
127.0.0.1	-	-	l/2005 01:32:25	GET	/images_easyp	HTTP/1.1"	304	-
127.0.0.1	-	-	l/2005 01:34:20	GET	/	HTTP/1.1"	200	2
127.0.0.1	-	-	l/2005 01:34:20	GET	/images_easyp	HTTP/1.1"	304	-
127.0.0.1	-	-	l/2005 01:34:20	GET	/images_easyp	HTTP/1.1"	304	-
127.0.0.1	-	-	l/2005 01:34:20	GET	/images_easyp	HTTP/1.1"	304	-
127.0.0.1	-	-	l/2005 01:34:20	GET	/images_easyp	HTTP/1.1"	304	-
127.0.0.1	-	-	l/2005 01:34:20	GET	/images_easyp	HTTP/1.1"	304	-
127.0.0.1	-	-	l/2005 01:35:18	GET	/uh2m/	HTTP/1.1"	200	3
127.0.0.1	-	-	l/2005 01:36:16	GET	/uh2m/	HTTP/1.1"	200	3
127.0.0.1	-	-	l/2005 01:36:29	GET	/uh2m/	HTTP/1.1"	200	3
127.0.0.1	-	-	l/2005 01:36:30	GET	/uh2m/	HTTP/1.1"	200	3
127.0.0.1	-	-	l/2005 01:36:36	GET	/uh2m/	HTTP/1.1"	200	3
127.0.0.1	-	-	l/2005 01:36:54	GET	/uh2m/	HTTP/1.1"	200	3

Figure 7 *: Base de données après import*

3-2. Nettoyage des données

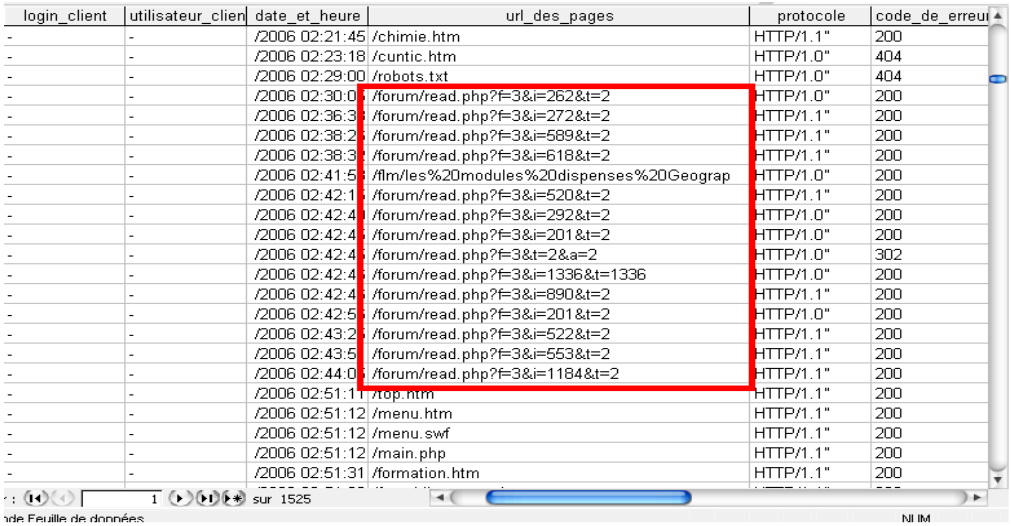
3-2-1. Nettoyage des graphiques, images ou scripts

Les données concernant les pages possédant des graphiques, des images ou des scripts, n'apportent rien à l'analyse. Elles seront donc filtrées et éliminées.
Pour cela on est amené à supprimer de notre base de données les URLs qui ont les formes suivantes :

Les urls correspondant aux images d'extension «.gif» par la requête
("delete from LOGUNIV where url_des_pages like '%.gif%")
Les urls correspondant aux images d'extension «.jpg» par la requête
("delete from LOGUNIV where url_des_pages like '%.jpg%")
Les urls correspondant aux images d'extension «.ico» par la requête
("delete from LOGUNIV where url_des_pages like '%.ico")
Les urls correspondant aux feuilles de styles d'extension «.css» par la requête
("delete from LOGUNIV where url_des_pages like '%.css")
Les urls correspondant aux images d'extension «.png» par la requête
("delete from LOGUNIV where url_des_pages like '%.png")

3-2-2. Résolution des pages uniques interprétées différemment par l'interpréteur du serveur

Pour les pages qui ont des extensions «.php», on doit résoudre le problème de « la page unique qui est interprétée différemment » par l'interpréteur du serveur.



login_client	utilisateur_clien	date_et_heure	url_des_pages	protocole	code_de_erreu
-	-	/2006 02:21:45	/chimie.htm	HTTP/1.1"	200
-	-	/2006 02:23:18	/cuntic.htm	HTTP/1.0"	404
-	-	/2006 02:29:00	/robots.txt	HTTP/1.0"	404
-	-	/2006 02:30:0	/forum/read.php?f=3&i=262&t=2	HTTP/1.0"	200
-	-	/2006 02:36:3	/forum/read.php?f=3&i=272&t=2	HTTP/1.1"	200
-	-	/2006 02:38:2	/forum/read.php?f=3&i=589&t=2	HTTP/1.1"	200
-	-	/2006 02:38:3	/forum/read.php?f=3&i=618&t=2	HTTP/1.1"	200
-	-	/2006 02:41:5	/flm/les%20modules%20dispenses%20Geograp	HTTP/1.0"	200
-	-	/2006 02:42:1	/forum/read.php?f=3&i=520&t=2	HTTP/1.1"	200
-	-	/2006 02:42:4	/forum/read.php?f=3&i=292&t=2	HTTP/1.0"	200
-	-	/2006 02:42:4	/forum/read.php?f=3&i=201&t=2	HTTP/1.0"	200
-	-	/2006 02:42:4	/forum/read.php?f=3&t=2&a=2	HTTP/1.0"	302
-	-	/2006 02:42:4	/forum/read.php?f=3&i=1336&t=1336	HTTP/1.0"	200
-	-	/2006 02:42:4	/forum/read.php?f=3&i=890&t=2	HTTP/1.1"	200
-	-	/2006 02:42:5	/forum/read.php?f=3&i=201&t=2	HTTP/1.0"	200
-	-	/2006 02:43:2	/forum/read.php?f=3&i=522&t=2	HTTP/1.1"	200
-	-	/2006 02:43:5	/forum/read.php?f=3&i=553&t=2	HTTP/1.1"	200
-	-	/2006 02:44:0	/forum/read.php?f=3&i=1184&t=2	HTTP/1.1"	200
-	-	/2006 02:51:11	/top.htm	HTTP/1.1"	200
-	-	/2006 02:51:12	/menu.htm	HTTP/1.1"	200
-	-	/2006 02:51:12	/menu.swf	HTTP/1.1"	200
-	-	/2006 02:51:12	/main.php	HTTP/1.1"	200
-	-	/2006 02:51:31	/formation.htm	HTTP/1.1"	200

1 sur 1525

Figure 8 : *Pages uniques interprétées différemment par le serveur*

Alors l'idée est de modifier la colonne « url_des_pages » de telle sorte à supprimer pour une même page, la partie de son adresse url, qui commence par «?» jusqu'à la fin de l'url.

Par exemple, couper la partie de l'url, qui suit le «?» jusqu'à la fin de la ligne, et ensuite la supprimer :

/forum/read.php|||?f=3&i=940&t=625|||

L'algorithme utilisé pour cette modification est le suivant :

```
ch est une chaîne
m est un entier
y est un entier
POUR m=1 A Table2..Occurrence
    ch est une chaîne
    ch est un entier
    ch=Table2.Pages[m]
    y=Position(ch,"?",DepuisDébut)
    SI y<>0 ALORS
        ch=ExtraitChaîne(ch,1,"?")
        Table2.Pages[m]=ch
    FIN
```

3-2-3. Résolution du Problème du Format Date et Heure

L'un des problèmes à résoudre aussi, est le format des champs « date_et_heure » du *fichier LOG* qu'on utilise pour garder la trace de l'instant d'entrée de l'internaute. Ce format est incompatible avec celui utilisé dans la base de données ***(Figure 8)***.
Le format adéquat pour ce champ est « **date_et_heure** ».

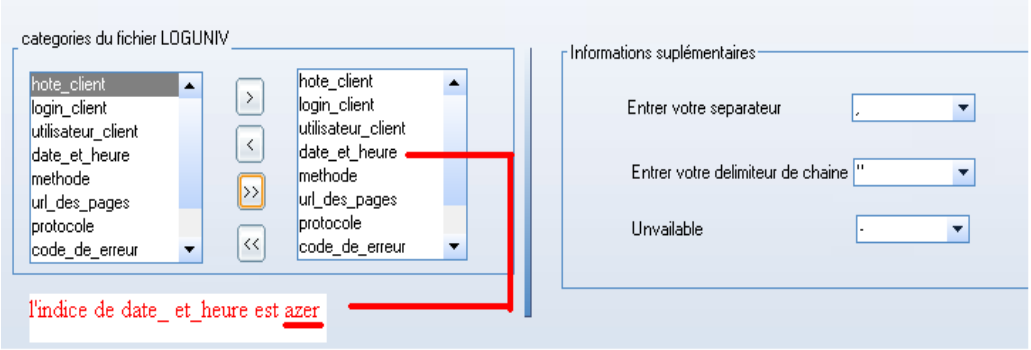


Figure 9 : *Format date et heure*

L'algorithme utilisé pour rendre la date et l'heure du *fichier LOG* reconnaissable par la colonne « date_et_heure » du SGBD est comme suit :
On commence par repérer la position courante de la colonne « date_et_heure » dans la liste de catégorie du *fichier LOG*, qui nous renseigne sur son emplacement dans le *fichier LOG* ***(Figure 9).***

Algorithme

```
BOUCLE
    Ch est chaine
    ch=fLitLigne(id)
    SI ch = EOT ALORS SORTIR
    ch1 est une chaîne=""
    ch11 est une chaîne=""
    ch12 est une chaîne
    ch13 est une chaîne
    h est un entier
    h=azer+1
    //azer est l'indice de la colonne date_et_heure dans la liste
    POUR i=1 A nbre
        ch11=ExtraitChaîne(ch,azer,sep)
        ch12=ExtraitChaîne(ch11,1,":")
```

```
            ch12=Droite(ch12,10)
            ch13=Droite(ch11,7)
            ch13=ch12+" "+ch13
            SI i<>azer ET i<>h  ALORS
                ch1=ch1+"""+ExtraitChaîne(ch,i,sep)+"","
            FIN
            SI i=azer ALORS
                ch1=ch1+"""+ch13+"","
            FIN
        FIN
        ch1=ch1+"""+ExtraitChaîne(ch,nbre,"sep")+"""
        SQLExec("insert into LOGUNIV values("+ch1+")", "REQ1")
    FIN
    SQLExec("alter table LOGUNIV ALTER COLUMN date_et_heure DATETIME",
"REQ1")
```

Après ces quatre grandes étapes de pré-traitement et nettoyage, le *fichier Log* est enfin prêt pour l'exploration et l'analyse.

4. Exploration et analyse du *fichier Log*

Pour l'exploration et l'analyse du *fichier Log*, un outil logiciel a été conçu et réalisé : « *LOG ANALYZER* », dont l'interface est comme suit :

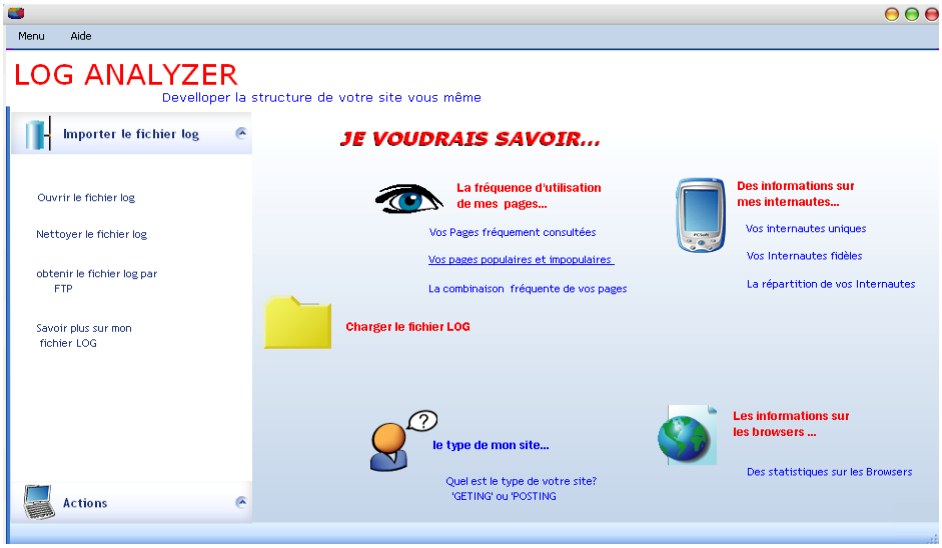


Figure 10 : *Interface du logiciel « LOG ANALYZER »*

4-1. Les combinaisons fréquentes des pages

Pour déterminer la combinaison des pages les plus utilisées, on a utilisé la méthode ***« A Priori »***.
On commencera tout d'abord par définir quelques notions utilisées dans cette méthode :
- *Nombre de pages* : c'est le nombre total des pages dans le *fichier Log* (dans la base de données)
- *Une session* : peut être soit une adresse IP, soit une demi-heure si l'adresse IP se répète pour une durée supérieure à une demi-heure.
- *Support d'une page* : c'est la somme du nombre d'apparition de la page dans toutes les sessions, divisée par le nombre total des sessions.
- *Support minimum* : seuil de support.
- *Page fréquente* : c'est la page dont le support est supérieur au support minimum.
- *Les combinaisons des pages* : elles sont déduites en utilisant l'algorithme *« A Priori »*

L'algorithme A Priori – 2

Fk : ensemble des itemsets fréquents de taille k
Ck : ensemble des itemsets candidats de taille k
K<-1
C1<- items
Tant que Ck‡Ø faire
Fk<- candidats de Ck dont le support >=σ
Ck+1<- candidats sont générés à partir de Fk
K<-k + 1
Fin
<-UFk
Pour chaque X ϵ UkFk
Pour chaque Y ⊂ X
Tester la confiance de la règle X\Y->Y

4-2. Les pages fréquemment consultées

On doit spécifier, tout d'abord, la période pour laquelle on aimerait faire la consultation ***(Figure 11)***, puis on calcule, pour chaque page, la fréquence de consultation.

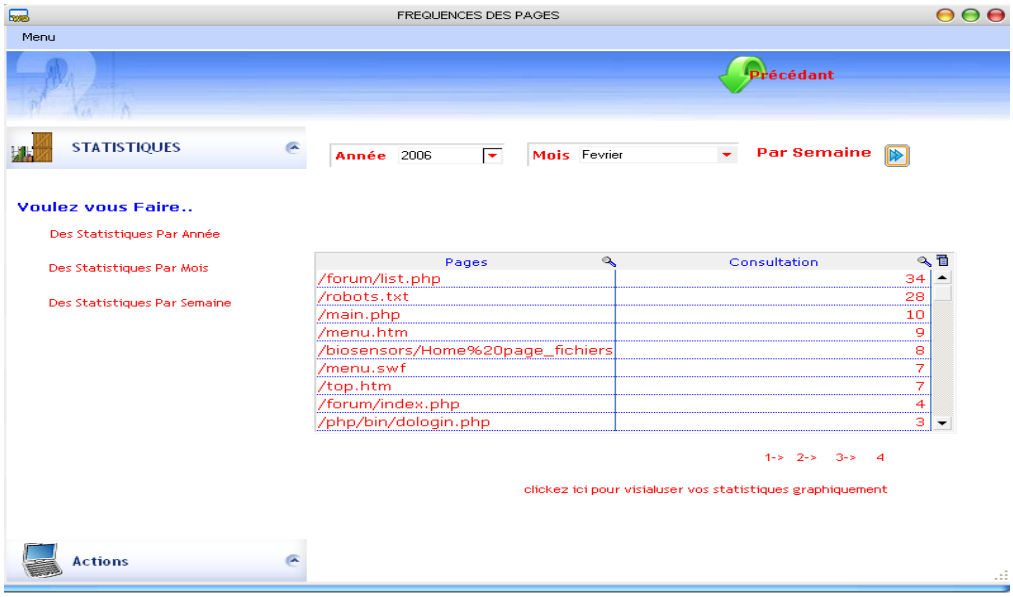


Figure 11 : *Pages fréquemment consultées?*

On peut l'illustrer graphiquement comme suit :

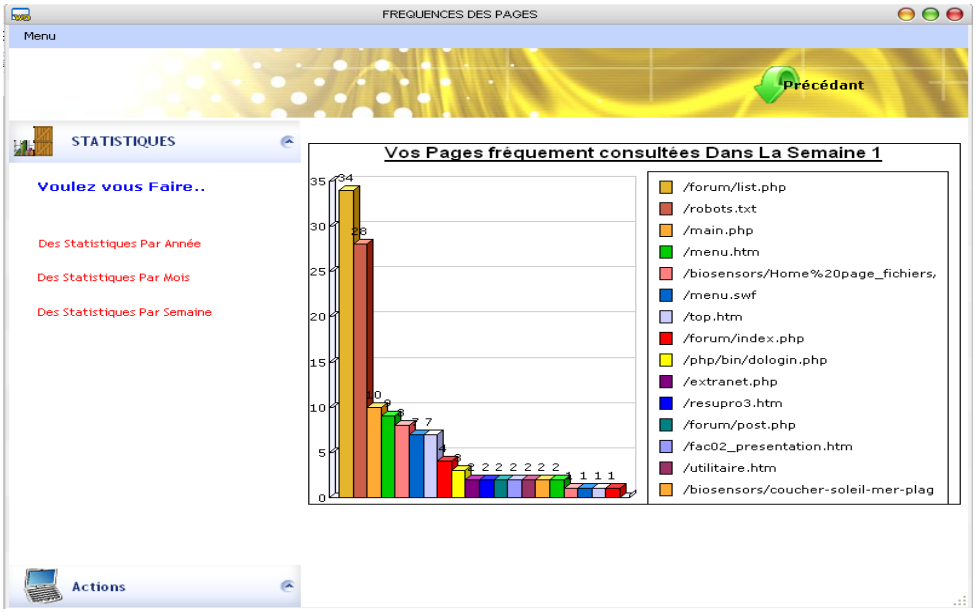


Figure 12 : *Graphique des pages fréquemment consultées*

4-3. Pages populaires et impopulaires

On doit choisir la période pour laquelle on voudrait faire la consultation. Un tableau de statistiques sur les pages populaires et impopulaires selon la période spécifiée sera affiché ***(Figure 13)***.

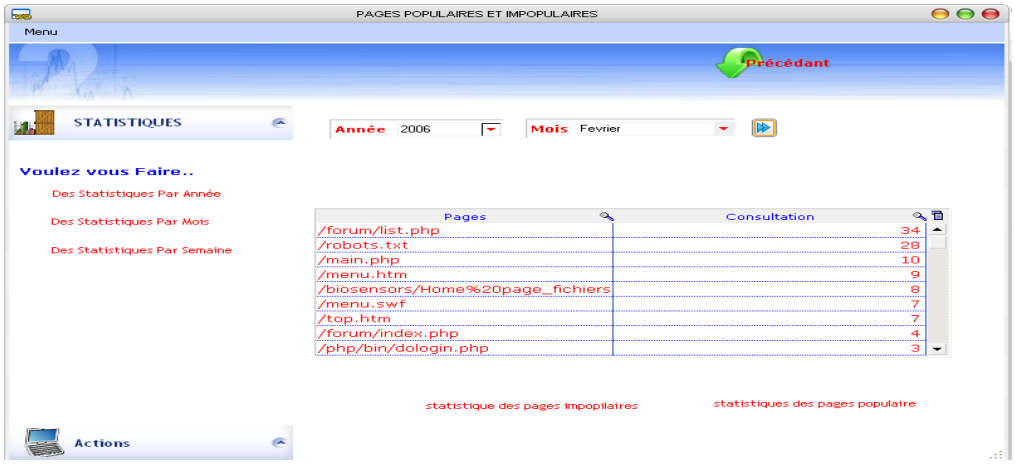


Figure 13 : *Pages populaires et impopulaires*

Ici l'idée est de donner au Webmaster toutes les pages populaires et impopulaires selon leurs degrés de popularité, en se basant sur le nombre de fois qu'une page a été visitée dans une période spécifiée [7,8].

Graphiquement, on obtient ceci à la ***Figure 14*** *:*

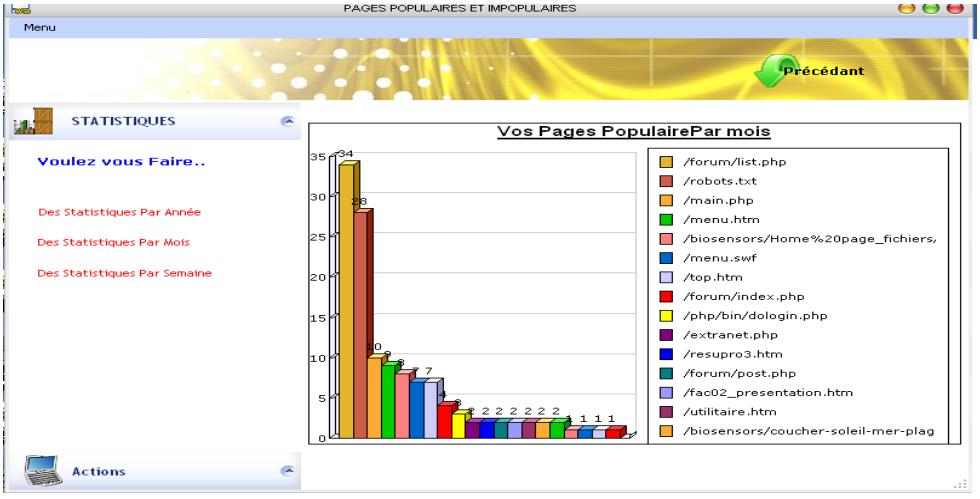


Figure 14 : *Graphique des pages populaires et impopulaire*

4-4. Informations sur les internautes

Cette rubrique renseigne le Webmaster sur la fidélité de ses visiteurs ***(Figure15).***

NB : *chaque adresse IP est équivalente à un utilisateur ou internaute.*

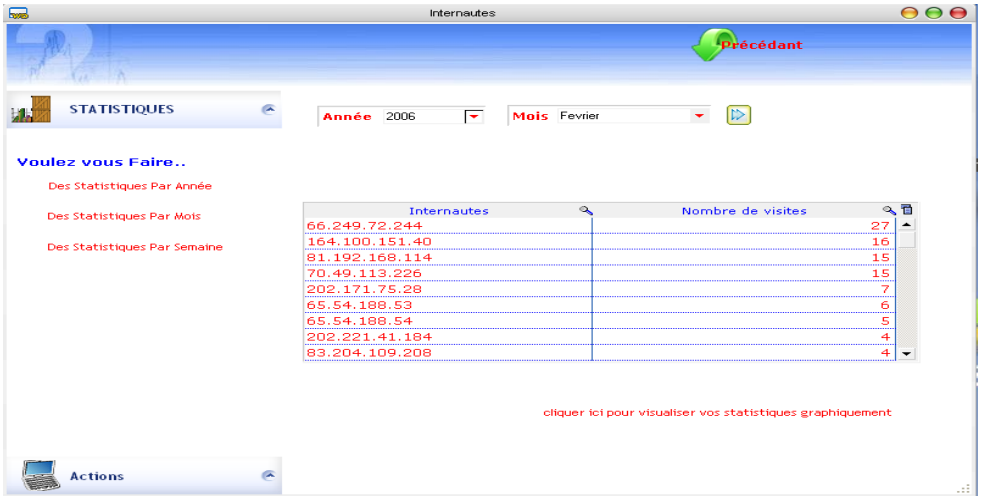


Figure 15 : *Informations sur les internautes.*

4-5. Catégorie du site

Dans cette rubrique on déterminera si notre site joue toujours son rôle préalablement défini (consultatif ou de téléchargement), en utilisant les méthodes sur le site *(GET OU POST).*

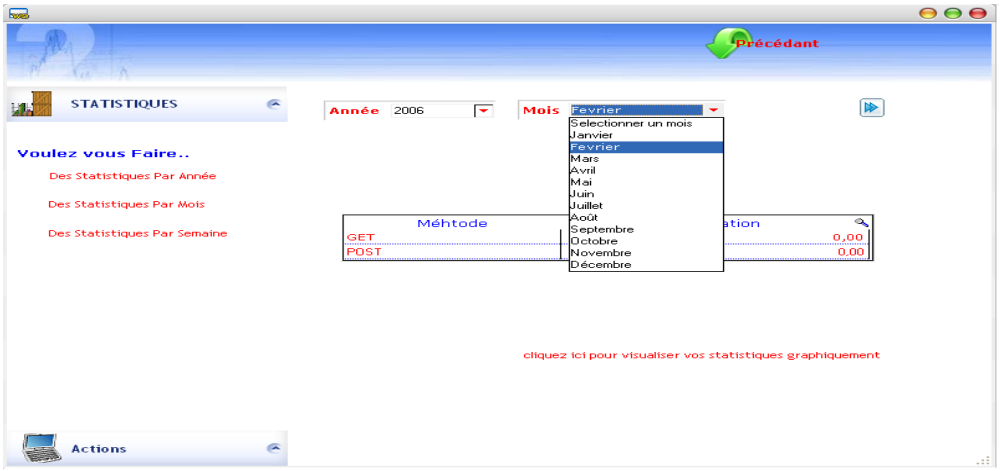


Figure 16 : *Catégorie du site*

Parfois, un site défini préalablement comme consultatif avec quelques fichiers à télécharger se transforme en un site de téléchargement, car les utilisateurs ne le consultent que pour les téléchargements. Par conséquent, ce site ne satisfait pas ses utilisateurs comme son Webmaster l'avait prévu. On peut l'illustrer graphiquement par la ***Figure 17***:

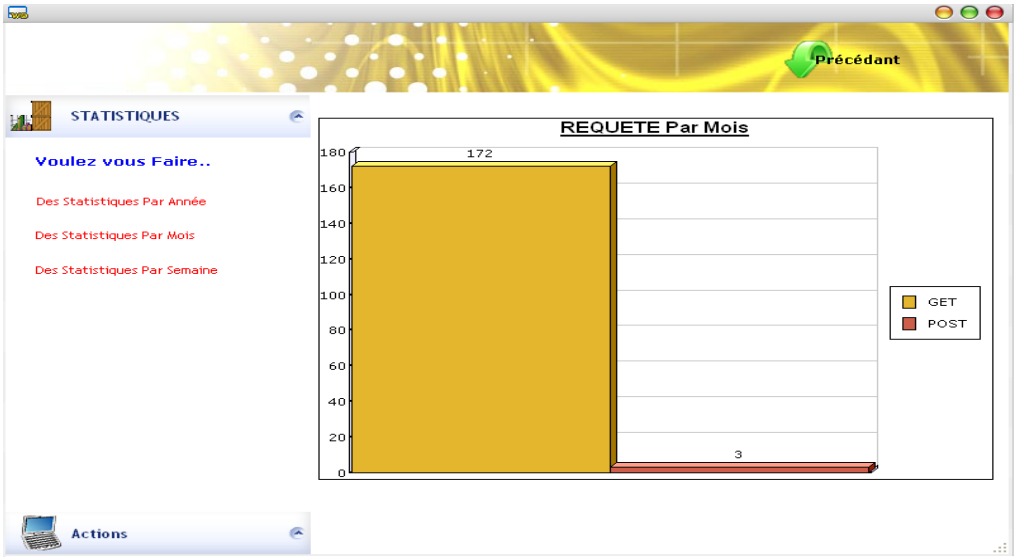


Figure 17 : *Graphique sur la catégorie du site*

4-6. Types de navigateurs (Browsers)

Si on clique sur le lien « *statistiques sur les browsers* », une page nous demande de choisir la période pour laquelle on aimerait afficher un diagramme résumant les pourcentages d'utilisation des browsers pour accéder au site :

NB : *seuls les navigateurs "célèbres" sont pris en compte*

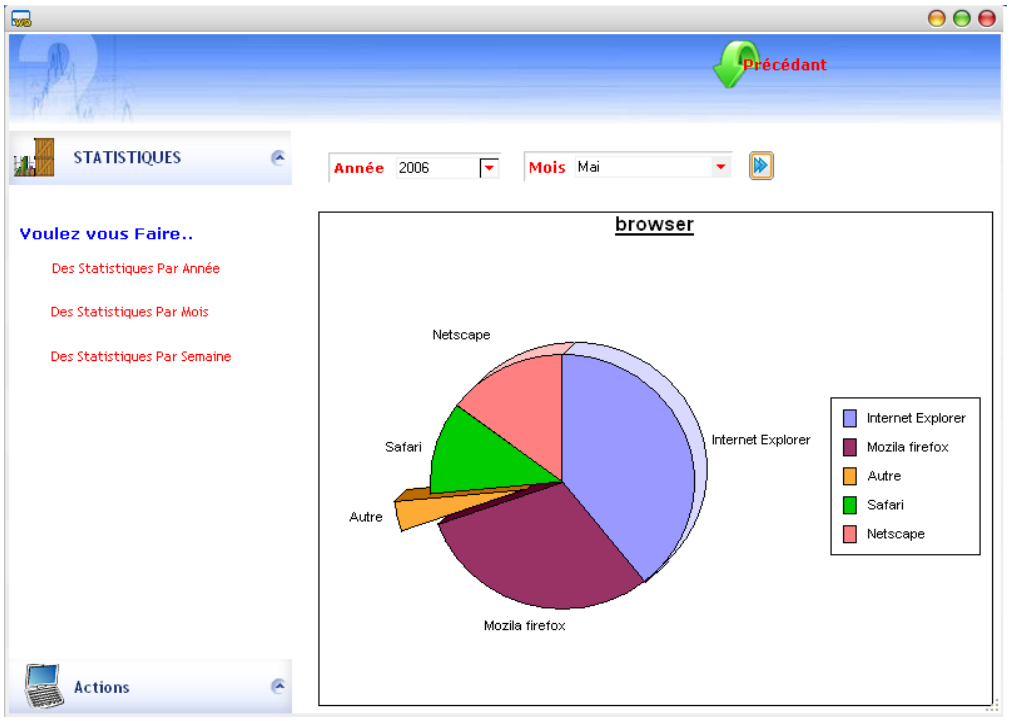


Figure 18 : *Types de navigateurs*

4. Conclusion et perspectives

Dans ce travail, nous avons montré comment les techniques du Web Usage Mining peuvent contribuer à l'amélioration et au diagnostic de site Web, sur un cas concret : le site de l'Université Hassan II-Mohammedia : *www.univh2m.ac.ma* , en explorant le *fichier Log* du serveur Web de ladite Université.

Ce travail est toujours en cours d'amélioration, en ce qui concerne les thèmes abordés et les fonctionnalités. On focalisera notre intérêt sur le *« profil du visiteur »* et de *« groupe de visiteurs »*, et on enrichira le travail par l'association d'une *« cartographie »* des visiteurs.

Références

[1] - Site *http://www.internetworldstats.com/stats.htm*, "world internet usage and population statistics", *http://www.netcraft.com*, December 2006, Web Server Survey.

[2] - R. Rao, P. Pirolli, J. Pitkow. "Silk from a sow's ear : extracting usable structures from the web". Human Factors in Computing Systems", CHI. (1996). in *Proceeding of ACM Conference on Human Factors in Computing Systems*, Vancouver, British Columbia , Canada, April 13-18 (1996). pages 118-125. *http://www.sigchi.org/chi96/proceedings/papers/Pirolli_2/pp2.html*

[3] - B. Mobasher, H. Dai, T. Lou, et M. Nakagawa "Discovery and evaluation of aggregate usage profiles for web personalization", *Data Mining and Knowledge Discovery*, Vol. 6, N°1 (2002), Jan., pages 61-82.

[4] - H. Azzag, C. Guinot, G. Venturini, « Classification hiérarchique et visualisation de pages Web », *Actes de l'atelier Fouille du Web des 6èmes journées francophones, « Extraction et Gestion des Connaissances EGC 2006 »*. ENIC Telecom Lille1, Cité Scientifique, Lille, France. 17-20 Janvier (2006) pages 5-16.

[5] - R. Cooley, B. Mobasher, and J. Srivastava, « Data Preparation for Mining World Wide Web Browsing Patterns », *J. Knowledge and Information Systems,* Vol. 1, N°1 (1999). pages 5-32.

[6] - M. Charrad, M. Ben Ahmed et Y. Lechevallier « Web Usage Mining: WWW pages classification from log files ». *Actes de l'atelier Fouille du Web des 6èmes journées francophones, « Extraction et Gestion des Connaissances EGC 2006 »* à ENIC Telecom Lille1, Cité Scientifique Lille, France. 17-20 Janvier (2006) pages 41-52.

[7] - Y. Lechevallier, D. Tonasa, B. Trousse, R. Verde. « Classification automatique : Applications au Web Mining ». in *Proceeding of SFC 2003*, Neuchatel, Swiss. Septembre (2003) pages 10-12.

[8] - M. Charrad, « *Techniques d'extraction des connaissances appliquées aux données du Web* ». Mémoire de Mastère présenté en vue de l'obtention du diplôme de Mastère en Informatique, Ecole Nationale des Sciences de l'Informatique de Tunis, Laboratoire RIADI (2005).

[9] - R. Kimball et R. Merz « *Le data webhouse. Analyse des comportements clients sur le Web* ». Editions Eyrolles, Paris (2000).