Extraction des connaissances à partir des fichiers Logs

Malika Charrad*, Mohamed Ben Ahmed*, Yves Lechevallier**

*Ecole Nationale des Sciences de l'Informatique, Laboratoire RIADI
Université de la Manouba, 1010 La Manouba
{malika.charrad, mohamed.benahmed}@riadi.rnu.tn
** INRIA – Institut National de Recherche en Informatique et en Automatique
Domaine de Voluceau-Rocquencourt, B.P.105, 78153 Le Chesnay Cedex, France
yves.lechevallier@inria.fr

Résumé. L'approche que nous proposons de caractériser les utilisateurs d'un site Web en se basant sur leurs motifs de navigation sur le site comporte trois phases : prétraitement des fichiers Logs, classification des pages et classification des internautes. Dans la phase de prétraitement, les requêtes sont organisées en visites. Dans la phase de classification des pages, des paramètres introduits à partir des statistiques sur les accès aux pages sont utilisés pour la catégorisation des pages Web en pages auxiliaires et pages de contenu. Les requêtes aux pages de contenu servent à la découverte des motifs de navigation. Pour construire des groupes d'utilisateurs, deux méthodes hybrides de classification automatique basées sur l'analyse en composantes principales, l'analyse des correspondances multiples et les cartes de Kohonen sont appliquées aux visites. Une expérience effectuée sur des données réelles prouve l'efficacité de cette méthodologie.

1 Introduction

Au cours de ces dernières années, la croissance exponentielle du nombre des documents en ligne a entraîné une croissance rapide de l'activité sur le Web, et une explosion des données résultant de cette activité. En effet, le nombre des utilisateurs d'Internet dans le monde a atteint 972.8 millions au mois de Novembre 2005[1], et le nombre de sites Web a atteint 74.4 millions au mois d'Octobre 2005[2]. Ces données, en particulier celles relatives à l'usage du Web, sont traitées dans le Web Usage Mining (WUM). Dans cet article, nous nous intéressons à l'analyse des fichiers Logs afin de comprendre le comportement des internautes sur un site Web. L'apport de ce travail réside principalement dans trois points :

- Utiliser plusieurs heuristiques pour l'identification des robots Web et l'identification des images dans la phase du prétraitement des fichiers Logs.
- Associer la classification des pages à la classification des usagers du site Web. En d'autres termes, exploiter les résultats de la classification des pages dans la classification des internautes.

[1] www. Internetworldstats.com
[2] www.netcraft.com

- Intégrer et combiner deux types de méthodes de fouille des données, deux méthodes factorielles et une méthode neuronale pour la classification des utilisateurs.

Cet article est organisé en trois sections distinctes. La première section présente les différentes étapes du prétraitement des fichiers Logs ainsi que les résultats de leur application sur des données réelles. La deuxième section présente une méthodologie de classification des pages Web. La dernière est consacrée à la classification des utilisateurs du site Web étudié.

2 Approche proposée

Récemment, de nombreux travaux en Web Usage Mining ont été menés. Certains de ces travaux se sont intéressés à la phase du prétraitement des données du Web tels que les travaux de Tanasa (2003) et Srivastava (2000) ; d'autres travaux sont concentrés sur la détermination des modèles comportementaux des internautes fréquentant les sites Web. Ce second axe est traité dans les travaux de Pierrakos (2003) et Mobasehr (2002). Notre approche consiste à intégrer la classification des pages dans la classification des utilisateurs. En d'autres termes, exploiter les résultats de la classification des pages dans la classification des utilisateurs.

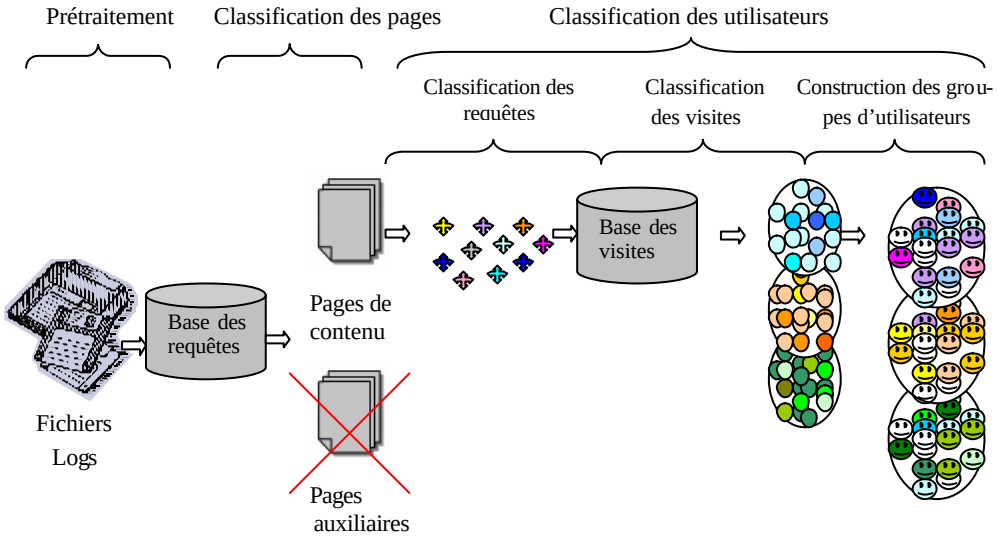

FIG. 1 - *Etapes de classification des utilisateurs.*

2.1 Prétraitement des données

2.1.1 Nettoyage des données

Le nettoyage des données consiste à supprimer les requêtes inutiles des fichiers logs à savoir :

- Les requêtes non valides. Ce sont les requêtes dont le statut est inférieur à 200 ou supérieur à 399 ;

- Les requêtes provenant des robots Web. Il est presque impossible aujourd'hui d'identifier tous les robots Web puisque chaque jour apparaissent des nouveaux. Pour identifier les requêtes provenant des robots, nous avons utilisé cinq heuristiques, dont les trois premières ont été utilisées par Tanasa et al. (2003), en considérant qu'il suffit de vérifier l'une de ces heuristiques pour considérer la requête correspondante comme étant générée par un robot Web :
  - Identifier les adresses IP et les « User-Agents » connus comme étant des robots Web. Ces informations sont fournies généralement par les moteurs de recherche[3],
  - Identifier les adresses IP ayant fait une requête à la page « \robots.txt »,
  - Utiliser un seuil pour la vitesse de navigation BS « Browsing Speed » égale au nombre de pages visitées par seconde. Le calcul du Browsing Speed n'est possible qu'après détermination des sessions et des visites.
  - Identifier les « User-Agents » comportant l'un des mots clés suivants : « crawler », « spider » ou encore « bot »,
  - Identifier les requêtes effectuées par des aspirateurs de sites Web (HTTrack par exemple), ou par des modules de certains navigateurs permettant la consultation de pages hors ligne tels que DigExt d'Internet Explorer. L'identité de ces aspirateurs ou de ces modules est trahie par la mention de leurs noms au niveau de leurs User-Agents. Pour les aspirateurs qui cachent leurs User-Agents, leur identification est effectuée en se basant sur la durée de leurs requêtes généralement nulle.
- Les requêtes aux images. Cette étape de nettoyage consiste à supprimer les fichiers dont les extensions sont : .jpg, .gif, .png, etc... et les fichiers multimédia dont l'extension est : .wav, .wma, .wmv, etc… Deux méthodes ont été utilisées pour supprimer les requêtes aux images. La première consiste à utiliser la carte du site afin d'identifier les URLs des images nécessitant de cliquer sur un lien pour être affichées. Les images inclues dans les fichiers HTML sont supprimées car elles ne reflètent pas le comportement de l'internaute. Cependant, ce n'est pas toujours possible d'identifier toutes les images inintéressantes quand le site est volumineux. Dans ce cas, nous proposons une autre méthode dont l'application nécessite tout d'abord l'identification des sessions. Cette méthode consiste à supposer qu'un utilisateur ne peut cliquer à la fois (au même instant) sur plusieurs images pour les visualiser ; Tenant compte de cette hypothèse, nous déterminons pour chaque utilisateur l'ensemble des requêtes effectuées au même instant. Les requêtes correspondantes à des fichiers images sont éliminées ;
- Les requêtes dont la méthode est différente de « GET » ;
- Les scripts. Généralement, le téléchargement d'une page demandée par un utilisateur est accompagné par le téléchargement automatique des scripts tels que les scripts Java (fichiers .js), des feuilles de style (fichiers .css), des animations flash (fichier .swf),…etc. Ces éléments doivent être supprimés du fichier Log étant donné que leur apparition ne reflète pas le comportement de l'internaute ;

[3] Informations obtenues au site http://www.searchturtle.com/search/Computers_Internet/Robots/ , http://www.robotstxt.org/wc/active/html/index.html et http://www.iplists.com/

- Les requêtes spécifiques à l'activité sur le site. Ce sont les requêtes relatives au trafic sur le site objet de l'analyse. Cette étape montre que la méthodologie d'analyse du comportement des internautes sur le Web n'est pas unique et qu'elle dépend de plusieurs facteurs, en particulier du site analysé. Par exemple, en considérant le site du CCK, cette étape consiste à supprimer :
  - Les requêtes pour les pages « proxy.pac »,
  - Les requêtes pour les annonces (les popups). En effet, les annonces apparaissent toutes seules dès que l'utilisateur se connecte sur le site du CCK. De ce fait, les requêtes correspondantes ne reflètent pas son comportement.

2.1.2 Transformation des fichiers logs

Identification des utilisateurs et des sessions : Une session est composée de l'ensemble de pages visitées par le même utilisateur durant la période d'analyse. Afin d'identifier les sessions, nous considérons que deux requêtes provenant de la même adresse IP mais de deux user-agents différents appartiennent à deux sessions différentes donc elles sont effectuées par deux utilisateurs différents. Ainsi, le couple (IP, User-Agent) représente un identifiant des utilisateurs. Toutefois, nous ne pouvons nier la limite inhérente à cette méthode. En effet, une confusion entre deux utilisateurs différents utilisant la même adresse IP et le même User-Agent est toujours possible surtout en cas d'utilisation d'un serveur Proxy ou d'un firewall.

```
Tant qu'il y'a des enregistrements dans la base faire
   Lire l'enregistrement i
   Récupérer l'adresse IP_i et le User Agent UA_i
      Si le couple (IP_i, UA_i) = (IP_(i-1), UA_(i-1))
         Alors ajouter l'enregistrement i à la session S_(i-1)
         Sinon recommencer une nouvelle session S_i
      Fin Si
Fin Tant Que
```

FIG. 2 – *Algorithme d'identification des utilisateurs et des sessions.*

Identification des visites : Une visite est composée d'une suite de requêtes séquentiellement ordonnées, effectuées pendant la même session et ne présentant pas de rupture de séquence de plus de 30 minutes (d'après les critères empiriques de Kimball (2000)). L'identification des visites sur le site, est effectuée en suivant cette démarche :

1. Déterminer la durée de consultation des pages. La durée de consultation d'une page est le temps séparant deux requêtes successives. Si la durée de consultation d'une page dépasse 30 minutes alors la page suivante dans la même session est attribuée à une nouvelle visite.
2. Une fois les visites identifiées, la durée de consultation de la dernière page de chaque visite est obtenue à partir de la moyenne des temps de consultation des pages précédentes appartenant à la même visite.

2.1.3 Résultats de l'analyse des fichiers Log du CCK

Le tableau suivant présente les résultats du prétraitement des fichiers Logs du site du Centre de Calcul elKhawarizmi[4] collectées pendant la période allant du 17 Septembre au 14 Octobre 2004.

	Nombre de requêtes	**Pourcentage**
Total de requêtes	**279879**	**100 %**
Requêtes non valides	*13028*	4.6 %
Requêtes provenant des WRs	*7187*	2.7 %
Identification par IP ou UA	2651	
Requêtes à « /robots.txt »	0	
Identification par mots clés	3689	
Requêtes effectuées par des aspirateurs	847	
Identification par BS	0	
Requêtes aux images et fichiers multimédia	*144025*	55.4 %
Requêtes dont méthode<>GET	*158*	0.13 %
Scripts et feuilles de style	*7426*	6.4 %
Requêtes spécifiques au site du CCK	*83163*	76.9 %
Requêtes à « /proxy.pac »	*78439*	
Annonces	4724	
Total	**254987**	**91.1 %**
Nombre de requêtes après nettoyage et retraitement	35353	3.7 %
Nombre des sessions	1770	
Nombre des visites destinées à l'analyse	2700	

TAB. 1 - *Tableau récapitulatif des résultats.*

2.1.4 Création de nouvelles variables

A partir des variables préexistantes, des nouvelles variables sont crées pour faciliter l'analyse envisagée. D'autres variables peuvent être créées suivant la nature de l'analyse envisagée.

Variable Originale	**Variables créées**	**Type**	**Valeurs variables**
Time	Période-journée	Discrète	Matin, Midi, Après midi, Soir, Nuit
Statut	Statut-200	Discrète	1, 0 (1 si le statut est 200, 0 sinon)
URL	Extension	Discrète	PDF, html, rtf, asp, doc, dot, jpg, ppt,
	Niveau 1	Discrète	
	Niveau 2	Discrète	
	⋮		
User-Agent	Navigateur	Discrète	MSIE, Netscape, Autres navigateurs
	Plateforme	Discrète	Windows, Unix/Linux, MacOS

TAB. 2 - *Création de nouvelles variables.*

[4] www.cck.rnu.tn

URL	Extension	Niveau 1	Niveau 2
/français/espace_chercheur.htm	htm	français	espace_chercheur

TAB. 3 - *Transformation de la variable URL.*

2.2 Classification des pages

La classification des pages a pour objectif de distinguer les pages de contenu présentant l'information recherchée par l'internaute des pages de navigation utilisée pour faciliter la navigation de l'utilisateur sur le site de manière à ne garder dans la base que les requêtes aux pages présentant un contenu intéressant aux visiteurs. Notre approche consiste à définir des variables servant à la caractérisation des pages et les utiliser pour la classification des pages.

2.2.1 Collecte des informations sur les accès

Afin de caractériser les pages visitées par les internautes, les variables suivantes sont définies pour chaque page :

- Nombre de Visites (NV) effectuées à chaque page ;
- Nombre des Inlinks (NI) : nombre d'hyperliens qui mènent à la page en question à partir des autres pages ;
- Nombre des Outlinks (NO) : nombre d'hyperliens dans la page qui mènent vers d'autres pages ;
- Durée Moyenne par page : temps moyen de visite de chaque page (DM) ;
- Taille du Fichier (TF) ;
- Type du Fichier (.html, .doc, .pdf, .rtf, …etc) (TYF).

Ainsi, chaque page peut être représentée par un vecteur :

Page = {NV, NI, NO, DM, TF, TYF}

Les variables TF et DM sont obtenues directement à partir de la base des requêtes après nettoyage et transformation. Pour la variable TYF, nous supposons que les pages dont l'extension est «.doc, .pdf, .rtf…» sont des pages de contenu. Par conséquent, nous ne considérons dans la suite que les pages «.asp» et «.html» auxquelles nous appliquons l'ACP.

La détermination des variables NV, NI et NO nécessite tout d'abord l'indexation des pages du site Web pour faciliter leur manipulation et la construction de deux matrices : matrice d'hyperliens et matrice d'accès.

URL	Referrer
/arabe/index.htm	/haut.htm
/arabe/main.htm	/arabe/index.htm
/arabe/espace_chercheur/cher_ar.htm	/arabe/main.htm
/arabe/services_en_ligne/serv_ligne_ar.htm	/arabe/espace_chercheur/cher_ar.htm
/arabe/haut.htm	/arabe/index.htm
/arabe/services_internet/serv_internet_ar.htm	/arabe/haut.htm

TAB. 4 - *Exemple de visite.*

URL	Index
/haut.htm	A
/arabe/index.htm	B
/arabe/main.htm	C
/arabe/haut.htm	D
/arabe/espace_chercheur/cher_ar.htm	E
/arabe/services_en_ligne/serv_ligne_ar.htm	F
/arabe/services_internet/serv_internet_ar.htm	G

TAB. 5 - *Indexation des pages de la visite.*

Matrice d'accès. Cette matrice est utilisée pour déterminer le nombre de visites effectuées par les internautes à chaque page. Chaque entrée (i,j) de la matrice représente le nombre de visites effectuées de la page i à la page j. Si cette entrée est égale à zéro alors la page j n'a jamais été visitée à partir de la page i.

	A	B	C	D	E	F	G	**Total**
A	0	0	0	0	0	0	0	**0**
B	5	0	0	0	0	0	0	**5**
C	0	24	0	0	0	0	0	**24**
D	0	9	0	0	0	0	0	**9**
E	0	0	2	0	0	0	0	**2**
F	0	0	0	0	1	0	0	**1**
G	0	0	0	1	0	0	0	1

TAB. 6 - *Matrice d'accès.*

Matrice d'hyperliens. Cette matrice est utilisée pour calculer le nombre d'inlinks et le nombre d'outlinks. En effet, le nombre d'inlinks est le total sur les lignes alors que le nombre d'outlinks est le total sur les colonnes.

	A	B	C	D	E	F	G	**Inlinks**
A	0	0	0	0	0	0	0	**0**
B	1	0	0	0	0	0	0	**1**
C	0	1	0	0	0	0	0	**1**
D	0	1	0	0	0	0	0	**1**
E	0	0	1	0	0	0	0	**1**
F	0	0	0	0	1	0	0	**1**
G	0	0	0	1	0	0	0	**1**
Outlinks	**1**	**2**	**1**	**1**	**1**	**0**	**0**	**6**

TAB. 7 - *Matrice d'hyperliens.*

Chaque ligne de la matrice correspond à une page du site. Il en est de même pour chaque colonne. Ainsi, s'il existe N pages différentes visitées par les internautes, la matrice d'hyperliens sera de dimension (N, N). Chaque entrée (i,j) de la matrice prend la valeur 1 si l'utilisateur a visité la page j à partir de la page i (présence d'un lien direct entre les deux pages) et la valeur 0 sinon. Toutefois, il ne faut pas oublier que certaines pages du site ne sont pas visitées par les internautes et que certains liens dans les pages visitées ne sont pas utilisés. Ces pages et hyperliens ne sont pas considérés dans cette représentation matricielle qui ne prend que les accès enregistrés dans les fichiers Logs.

2.2.2 Application de l'analyse en composantes principales

En considérant les variables présentées ci-dessus, nous avons appliqué l'ACP au tableau (pages × variables).

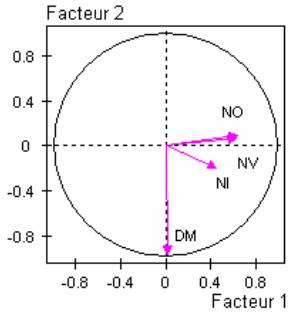

FIG. 4 - *Projection des variables sur les axes factoriels.*

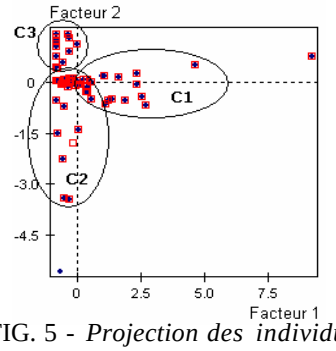

FIG. 5 - *Projection des individus sur les axes factoriels.*

D'après Fig.10, le premier axe factoriel est expliqué par les trois variables NV, NI et NO. Il oppose les pages les plus fréquentées et ayant un nombre important d'inlinks et d'outlinks aux pages les moins fréquentées et caractérisées par un faible nombre d'inlinks et d'outlinks. Le second axe factoriel est celui de la durée moyenne de consultation de pages. Les pages projetées définissent quatre classes.

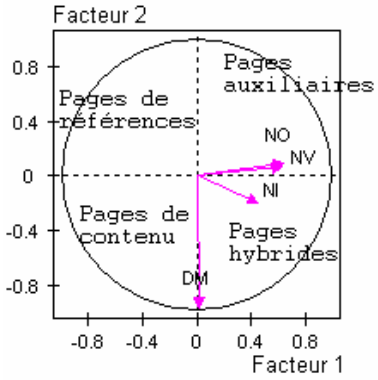

FIG. 6 - *Classification des pages.*

Type de page	NV	NI	NO	DM
Page de contenu	-	-	-	+
Page auxiliaire	+	+	+	-
Page hybride	+	+	+	+
Page de référence	-	-	-	-

TAB. 8 - *Caractérisation des pages.*

La première classe (C1) est composée de pages visitées fréquemment et caractérisées par un nombre important d'inlinks et d'outlinks. Elle correspond à la classe de pages auxiliaires ou de navigation. La deuxième classe (C2) est celle de pages de contenu caractérisées par une durée de consultation assez élevée. L'intersection de ces deux classes est composée de pages présentant à la fois les caractéristiques des pages de contenu et des pages auxiliaires. C'est la classe de pages hybrides (C4). La dernière classe (C3) est celle des pages visitées rarement, qui ne pointent nulle part et vers lesquelles pointent peu de pages. La durée moyenne de consultation de ces pages est faible. Nous considérons que ces pages correspondent à ce que Rao (1996) appelle « pages de références » utilisées pour définir un concept ou expliquer des acronymes. Cependant, nous considérons que ces pages sont, dans une certaine mesure, des pages de contenu.

2.3 Classification des utilisateurs

Cette phase est réalisée en trois étapes. La première consiste à classifier les requêtes effectuées par les internautes afin de découvrir des motifs de navigation. Les résultats de cette première classification sont injectés dans la base des visites. La clusterisation des visites permet de construire des groupes d'utilisateurs.

2.3.1 Découverte de motifs de navigation

La découverte de motifs de navigation est effectuée à deux niveaux en combinant deux méthodes de classification. La première est l'analyse des correspondances multiples appliquées aux variables présentées dans TAB. 1. La seconde est la carte topologique de Kohonen utilisée pour déterminer des groupes de requêtes. Les axes factoriels résultant de l'application de l'ACM servent de variables d'entrée (inputs) pour la carte de Kohonen.

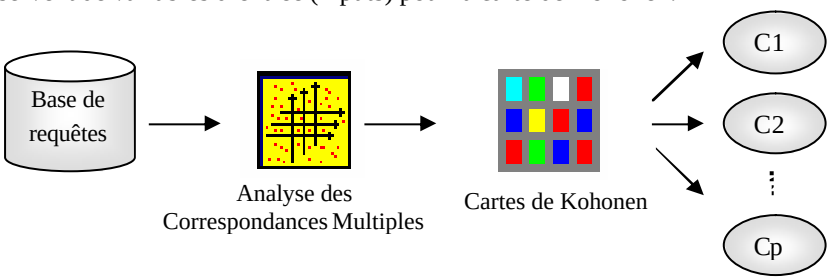

FIG. 7 - *Etapes de la classification des requêtes.*

La caractérisation des classes obtenues par la projection des variables « niveau 1» et « niveau 2 » déterminées à partir de la variable « URL » a donné le résultat présenté par la figure 14. En examinant les éléments caractérisant chaque classe, il est possible d'attribuer un label à chaque classe. La classe 1-1 par exemple comporte des requêtes aux institutions universitaires, la classe 3-4 comporte des requêtes dont l'objectif est le téléchargement des cours. la carte de Kohonen, après division en aires logiques et labellisation, met en évidence cinq aires logiques correspondant à cinq motifs de navigation : visites aux institutions universitaires, activités de recherche, demandes des informations sur les congrès, services CCK et téléchargement des cours.

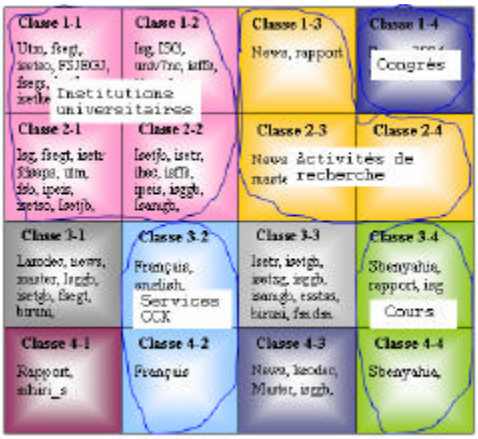

FIG. 9 - *Carte de Kohonen après division en aires logiques et labellisation.*

2.3.2 Construction de groupes d'utilisateurs

Pour construire des groupes d'utilisateurs, il faut tout d'abord attribuer à chaque visite un ou plusieurs motifs de navigation, caractériser les visites par un ensemble de variables et les regrouper en classes en suivant le schéma suivant. La seconde étape consiste à construire des groupes d'utilisateurs et les caractériser.

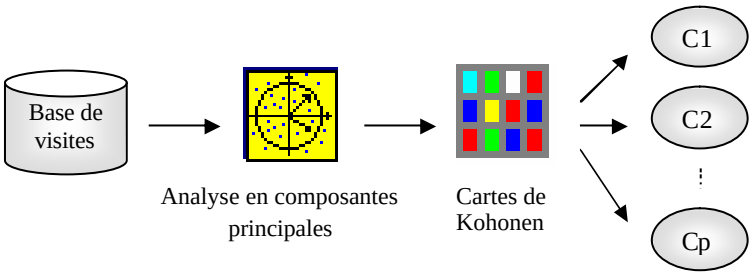

FIG. 10 - *Etapes de la clusterisation des visites*

Analyse en composantes principales : L'analyse en composantes principales s'applique à un tableau (individus x variables). Dans notre cas, les individus sont les visites et les variables considérées sont présentées dans Tab.5.

Nom de la variable	Description de la variable
Durée_visite	La durée de chaque visite
DuréeMoyPage	Durée moyenne de consultation des pages dans chaque visite
NbReqVisite	Nombre de requêtes dans chaque visite
PourcReqDiff	Pourcentage de requêtes différentes dans chaque visite
PourcReqOk	Pourcentage de requêtes réussies dans chaque visite

TAB.9 - *Variables utilisées dans l'ACP.*

Cartes de Kohonen : L'application des cartes de Kohonen met en évidence trois classes de visites. La première classe est composée de visites dont la durée moyenne, le nombre moyen de requêtes par visite et la durée moyenne de consultation des pages sont assez élevées en comparaison avec les deux autres classes. Ceci s'explique par le fait que ces visites sont effectuées principalement dans le but de télécharger des cours ou visiter des institutions universitaires. La deuxième classe est caractérisée par le pourcentage le plus élevé de requêtes réussies (95%) et de requêtes différentes (98%). Ces visites sont effectuées afin de profiter des services fournis par le CCK tels que les services Internet, les services de calcul et le compte Internet. La dernière classe comporte des visites dont l'objectif est d'avoir des informations sur les congrès, les colloques, ..etc.

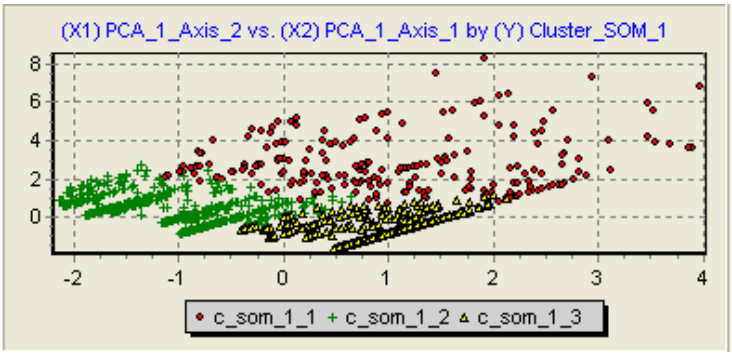

FIG. 11 - *Résultat de la classification des visites.*

Pour chaque groupe de visites, un groupe d'utilisateurs est construit contenant tous les utilisateurs possédant au moins une visite dans ce groupe de visites. Ainsi, nous obtenons des groupes d'utilisateurs ayant le même motif de navigation. Le premier groupe est celui des universitaires dont l'objectif de la navigation sur le site et le téléchargement des cours, l'inscription dans les établissements universitaires et la visite des bibliothèques universitaires. Le deuxième groupe est celui des chercheurs qui demandent des informations sur les congrès, les colloques, les mastères et les thèses et visitent les laboratoires de recherche. Le troisième groupe est celui des visiteurs du site du CCK afin de profiter des services qu'il fournit. Un dernier groupe est déjà défini lors du prétraitement des fichiers Logs et dont les requêtes ont été supprimées pendant la phase du prétraitement. Il s'agit des agents et robots utilisés par les moteurs de recherche pour mettre à jour leurs index de recherche.

3 Conclusion

Dans ce travail, nous avons développé une méthodologie de prétraitement des fichiers Logs permettant de transformer l'ensemble de requêtes enregistrées dans les fichiers Logs à des données structurées et exploitables. L'hybridation des méthodes de classification nous a permis de surmonter l'obstacle de la quantité des données et de tirer profit du pouvoir classificatoire de certaines d'entre elles, à savoir les cartes topologiques de Kohonen.

Références

Charrad, M. (2005) *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.

Charrad, M., M. Ben Ahmed et Y. Lechevallier (2005). Web Usage Mining : WWW pages classification from log files. *In Proceeding of International Conference on Machine Intelligence*, Tozeur, Tunisia, 5-7 Novembre.

Cooley, R., B. Mobasher, et J. Srivastava (1999). Data Preparation for Mining World Wide Web Browsing Patterns. *Journal of Knowledge and Information Systems*.

Kimball, R. et R. Merz (2000). *Le data webhouse. Analyser des comportements clients sur le Web*. Editions Eyrolles, Paris.

Lechevallier, Y., D. Tonasa, B. Trousse, R. Verde (2003). Classification automatique : Applications au Web Mining. *In Proceeding of SFC2003, Neuchatel*, 10-12 Septembre.

Mobasher, B., H. Dai, T. Lou, et M. Nakagawa (2002). Discovery and evaluation of aggregate usage profiles for web personalization. *Data Mining and Knowledge Discovery*, 6: 61-82.

Pierrakos, D., G. Paliouras, C. Papatheodorou, et C.D. Spyropoulos (2003). Web Usage Mining as a tool for personalization: A survey. *User Modeling and User-Adapted Interaction*, 13:311-372.

Rao, R., P. Pirolli, J.Pitkow (1996). Silk from a sow's ear : Extracting usable structures from the web. *In proc. ACM Conf. Human Factors in Computing Systems, CHI.*

Srivastava, J., R. Cooley, M. Deshpande et P.-N. Tan (2000). Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data. *SIGKKD Explorations.*

Tanasa, D. et B. Trousse (2003), Le prétraitement des fichiers Logs Web dans le Web Usage Mining Multi-sites. *In Journées Francophones de la Toile*.

Summary

The approach we proposed to characterize users of a Web site consists of three steps: preprocessing of log files, Web site pages classification and users clustering. In Preprocessing, requests are processed to be organized into sessions. In page classification, parameters are introduced from pages access statistics to help classify pages into auxiliary pages and content pages. Requests to content pages are used to discover browsing patterns. Two hybrid clustering methods based on Principle Component Analysis, Multiple Correspondences Analysis and Self Organizing maps are applied to sessions to construct users' groups. An experiment on real log files shows that the approach is efficient and practical.