Classer pour Découvrir : une nouvelle méthode d'analyse du comportement de tous les utilisateurs d'un site Web

Doru Tanasa, Brigitte Trousse, Florent Masseglia

Projet AxIS, INRIA Sophia Antipolis,
2004, Route des Lucioles, BP 93, 06902 Sophia Antipolis
{Doru.Tanasa, Brigitte.Trousse, Florent.Masseglia}@sophia.inria.fr,
http://www-sop.inria.fr/axis/

Résumé. L'analyse du comportement des utilisateurs d'un site Web est un domaine riche et complexe. Le grand nombre de méthodes d'extraction de connaissances appliquées aux logs Web, ainsi que la diversité du type de ces méthodes en est une preuve. Cependant, compte tenu de cette complexité, nous posons dans cet article la question suivante : *Est-il possible de combiner des méthodes existantes pour proposer une analyse qui tire profit des résultats de plusieurs spécialités et extraire par exemple des comportements fréquents minoritaires?* Notre étude a donc porté sur une nouvelle approche hybride (issue de la classification neuronale et de la recherche de motifs séquentiels) visant à classer les navigations des utilisateurs d'un site (à l'aide de leurs résumés sémantiques) puis, pour chaque classe de navigations, d'en extraire les comportements fréquents. Notre objectif est 1) de pallier les limites de l'extraction des motifs fréquents par rapport à la quantité de données à traiter et aussi par rapport à la qualité des résultats et 2) de pallier les limites d'une première méthode d'analyse du comportement appelée "Diviser pour Découvrir", que nous avons proposée en 2003. Nous avons mené des expérimentations sur les logs HTTP des sites INRIA. Les résultats obtenus confirment le bien fondé de notre approche vis à vis de l'état de l'art.

1 Introduction

L'analyse du comportement des utilisateurs d'un site Web, également connue sous le nom de Web Usage Mining (WUM), est un domaine de recherche qui consiste à adapter des techniques de fouille de données sur les enregistrements contenus dans les fichiers access logs. Ces fichiers regroupent des informations sur l'adresse IP de la machine, l'URL demandée, la date, et d'autres renseignements concernant la navigation de l'utilisateur. Les techniques de Web Usage Mining s'intéressent à la recherche de motifs (voire des connaissances sur les comportements des utilisateurs d'un site Web) afin d'extraire des relations entre les données stockées [Cooley *et al.*, 1999] [Masseglia *et al.*, 2000] [Mobasher *et al.*, 2002] [Spiliopoulou *et al.*, 1999]. Parmi les méthodes développées, celles consistant à extraire des motifs séquentiels [Agrawal et Srikant, 1995] s'adaptent particulièrement bien au cas des logs. En théorie, l'extraction de motifs séquentiels sur un fichier access log permet d'extraire le type de relations suivant : "*Sur le site de l'IN-*

RIA, 10% des clients ont visité, dans l'ordre, la page d'accueil, la page des opportunités de travail, la page du recrutement des ITA[1]*, la page des missions des ITA et enfin la page des annales de concours des ITA.*"

Ce type de comportements existe, mais en théorie seulement, car l'extraction de motifs séquentiels à partir d'un fichier de type access log se heurte à de multiples problèmes : la présence du cache, la grande diversité des pages sur le site[2], la représentativité de la partie du site visitée par rapport au site dans sa globalité[3] ou encore la représentativité des utilisateurs de cette partie du site par rapport au nombre total d'utilisateurs sur le site global.

Pour comprendre l'enjeu de ces travaux, reprenons l'exemple de navigation que la théorie prétend obtenir. Bien qu'ils soient les plus représentés, les utilisateurs consultant la partie "opportunité de travail" du site de l'INRIA en Janvier 2003 sont 0,5% des utilisateurs du site global. De la même manière, les utilisateurs consultant la partie "enseignement" du projet de recherche AxIS, ne sont que 0,01% des utilisateurs du site global. Ainsi l'étude du log pour un tel site Web doit passer par la prise en compte de cette représentativité très particulière pour prétendre à des résultats satisfaisants.

Dans [Masseglia *et al.*, 2003] nous avons proposé une approche de division récursive du problème, basée sur une classification des motifs séquentiels extraits du log à chaque étape. Dans cet article nous proposons une méthode de classification capable de travailler directement sur les navigations enregistrées dans le log, afin d'économiser les différentes étapes d'extraction de motifs séquentiels intermédiaires. L'extraction des motifs séquentiels étant alors appliquée à l'étape suivante du processus pour chaque classe construite, nous obtenons plus rapidement des résultats de meilleure qualité.

Cet article est structuré de la manière suivante : la section 2 présente les principales notions ainsi que leur formalisation utilisée dans cet article. Les sections 3 et 4 détaillent notre nouvelle méthodologie "Classer pour Decouvrir" basée sur une classification neuronale des navigations. Avant de conclure en section 6, la section 5 décrit nos experimentations réalisées sur les fichiers logs de l'INRIA.

2 Definitions

2.1 Motifs séquentiels

Ce paragraphe expose et illustre la problématique liée à l'extraction de motifs séquentiels dans de grandes bases de données. Il reprend les différentes définitions proposées dans [Agrawal *et al.*, 1993] [Agrawal et Srikant, 1995].

Dans [Agrawal *et al.*, 1993], la notion de support est définie de la manière suivante :

Définition 1 Soit $I = \{i_1,i_2,...,i_m\}$, un ensemble de m achats (*items*). Soit $D = \{t_1,t_2,...t_n\}$, un ensemble de n transactions ; chacune possède un unique identificateur appelé TID et porte sur un ensemble d'items (*itemset*) I. I est appelé un *k-itemset* où

1. Postes d'Ingénieurs, Techniciens, Administratifs.

2. Sur un site comme celui de l'INRIA, on peut compter plus de 70000 ressources filtrées (après l'étape de sélection de données) pour le siège, et plus de 82000 ressources filtrées pour le site de l'unité de Sophia Antipolis.

3. Une équipe de recherche peut représenter moins de 0,7% de la globalité du site de l'INRIA.

k représente le nombre d'éléments de I. Une transaction $t \in D$ contient un itemset I si et seulement si $I \subseteq t$. Le *support* d'un itemset I est le pourcentage de transaction dans D contenant I : $supp(I) = \|\{t \in D \mid I \subseteq t\}\|/\|\{t \in D\}\|$.

Définition 2 Une *transaction* constitue, pour un client C, l'ensemble des items achetés par C à une même date. Dans une base de données client, une transaction s'écrit sous forme d'un triplet : <id-client, id-date, itemset>. Un *itemset* est un ensemble non vide d'items noté $(i_1 i_2 \dots i_k)$ où i_j est un *item* (il s'agit de la représentation d'une transaction non datée). Une *séquence* est une liste ordonnée, non vide, d'itemsets notée $< s_1 s_2 \dots s_n >$ où s_j est un itemset (une séquence est donc une suite de transactions avec une relation d'ordre entre les transactions). Une *séquence de données* est une séquence représentant les achats d'un client. Soit T_1, $T_2, \dots,$ T_n les transactions d'un client, ordonnées par date d'achat croissante et soit $itemset(T_i)$ l'ensemble des items correspondants à T_i, alors la séquence de données de ce client est $< itemset(T_1)\ itemset(T_2) \dots\ itemset(T_n) >$.

Exemple 1 *Soit C un client et $S = <$ (3) (4 5) (8) $>$, la séquence de données représentant les achats de ce client. S peut être interprétée par "C a acheté l'item 3, puis en même temps les items 4 et 5 et enfin l'item 8".*

Définition 3 Soit $s_1 =< a_1 a_2 \dots a_n >$ et $s_2 =< b_1 b_2 \dots b_m >$ deux séquences de données. s_1 est *incluse* dans s_2 ($s_1 \prec s_2$) si et seulement si il existe $i_1 < i_2 < \dots < i_n$ des entiers tels que $a_1 \subseteq b_{i_1}$, $a_2 \subseteq b_{i_2}, \dots$ $a_n \subseteq b_{i_n}$.

Exemple 2 *La séquence $s1 =<$ (3) (4 5) (8) $>$ est incluse dans la séquence $s2 =<$ (7) (3 8) (9) (4 5 6) (8) $>$ (i.e $s1 \prec s2$) car* (3) $\subseteq$ (3 8), (4 5) $\subseteq$ (4 5 6) *et* (8) $\subseteq$ (8). *En revanche* $<$ (3) (5) $>\not<$ (3 5) $>$ *(et vice versa).*

Définition 4 Soit s, une séquence. Le *support* de s, noté $supp(s)$, est le pourcentage de toutes les séquences dans D qui supportent (contiennent) s. Si $supp(s) \geq minsupp$, avec une valeur de support minimum $minsupp$ fixée par l'utilisateur, la séquence s est dite *fréquente*.

2.2 Web Usage Mining

Nous rappelons ici les principales définitions issues de [Tanasa et Trousse, 2003].

Définition 5 *Ressource Web* - une ressource[4] accessible par une version du protocole HTTP ou un protocole similaire (ex. HTTP-NG).

Définition 6 *Requête Web* - une requête pour une ressource Web, faite par un client (navigateur Web) à un serveur Web.

Définition 7 *Page Web* - ensemble des informations, consistant en une (ou plusieurs) ressource(s) Web, identifiée(s) par un seul URI. Exemple : un fichier HTML, un fichier image et une applet Java accessibles par un seul URI constituent une page Web.

4. D'apres [Network Working Group, 1998], une ressource est "toute entité ayant une identité", par exemple : un document électronique, une image, un service et même une collection d'autres ressources.

Définition 8 *Session utilisateur* - l'ensemble des requêtes, pour des pages Web, de l'utilisateur sur un (ou plusieurs) serveur(s) Web.

Définition 9 *Navigation (visite)* - Un sous-ensemble des requêtes Web appartenant à une session utilisateur. Les requêtes Web d'une session utilisateur peuvent être groupées en plusieurs navigations en calculant la distance temporelle entre deux requêtes HTTP consécutives et si cette distance excède un certain seuil (en général 30 min.), une nouvelle navigation commence.

2.3 Adapter la problématique des motifs séquentiels

Ce paragraphe propose de reprendre les concepts essentiels d'un processus de Web Usage Mining, afin de présenter de façon synthétique, les procédés mis en œuvre lors de l'analyse du comportement des utilisateurs d'un site Web. Les principes généraux sont similaires à ceux d'un processus d'extraction de connaissances (cf. [Fayad *et al.*, 1996]).

Les données brutes sont collectées dans des fichiers access log des serveurs Web. Une entrée dans le fichier access log est automatiquement ajoutée chaque fois qu'une requête pour une ressource atteint le serveur Web (*demon http*). Les fichiers access log peuvent varier selon les systèmes qui hébergent le serveur, mais présentent tous en commun trois champs : l'adresse du demandeur, l'URL demandée et la date à laquelle cette demande a eu lieu. Parmi ces différents types de fichiers, nous avons retenu dans cet article le format ECLF [Luotonen, 1995] qui permet des enregistrements (cf. figure 1) formés de 8 champs séparés par des espaces :

host rfc931 authuser [date:time] "request" status bytes referrer user-agent

```
138.96.69.8 - - [03/Mar/2003:18:42:14 +0100] "GET /axis/logoToile3.swf HTTP/1.1" 200 36214 "-" "Mozilla/4.76 [en] (X11;
U; Linux 2.4.20 i686)"
138.96.69.8 - - [03/Mar/2003:18:48:00 +0100] "GET /aid/personnel/ HTTP/1.1" 404 4856 "-" "Mozilla/4.76 [en] (X11; U;
Linux 2.4.20 i686)"
138.96.69.7 - - [03/Mar/2003:19:03:14 +0100] "GET /axis/cbrtools/ HTTP/1.0" 200 15969 "-" "Mozilla/4.76 [en] (X11; U;
Linux 2.4.20 i686)"
138.96.69.7 - - [03/Mar/2003:19:04:44 +0100] "GET /axis/cbrtools/manual/ HTTP/1.0" 200 907 "-" "Mozilla/4.76 [en] (X11;
U; Linux 2.4.20 i686)"
138.96.69.7 - - [03/Mar/2003:19:12:45 +0100] "GET /axis/broadway/ HTTP/1.0" 200 671 "-" "Mozilla/4.76 [en] (X11; U;
Linux 2.4.20 i686)"
```

FIG. 1 – *Extrait du fichier access log du serveur Web de l'INRIA Sophia Antipolis.*

Deux types de traitements sont effectués sur les entrées du serveur log. Le premier type est le processus de prétraitement décrit en détails dans [Tanasa et Trousse, 2003] et sert à grouper les requêtes en navigations. Le deuxième type est la phase de codage de données entre les deux premières étapes de notre méthode (cf. section 3.2). Afin de rendre plus efficace le traitement de l'extraction de données, les URLs et les clients sont codés sous forme d'entiers. Toutes les dates sont également traduites en temps relatif par rapport à la plus petite date du fichier.

Définition 10 Soit *Log* un ensemble d'entrées dans le fichier access log. Une entrée g, $g \in Log$, est un tuple $g =< ip_g, \{(l_1^g.URL, l_1^g.time), ..., (l_m^g.URL, l_m^g.time)\} >$ tel que pour $1 \leq k \leq m$, $l_k^g.URL$ représente l'objet demandé par le client ip_g à la date $l_k^g.time$, et pour tout $1 \leq j < k$, $l_k^g.time > l_j^g.time$.

La figure 2 illustre un exemple de fichier obtenu après la phase de pré-traitement pour une classe de navigations. A chaque client correspond une suite de "dates" (événements) et la traduction de l'URL demandée par ce client à cette date.

	Date1	Date2	Date3	Date4	Date5
Client1	10	30	40	20	30
Client2	10	30	60	20	50
Client3	10	70	30	20	30

FIG. 2 – *Exemple de fichier résultat issu de la phase de pré-traitement*

L'objectif est alors de déterminer, grâce à une phase d'extraction, les séquences de ce jeu de données, qui peuvent être considérées comme fréquentes selon la définition 4. Les résultats obtenus sont du type `< ( 10 ) ( 30 ) ( 20 ) ( 30 ) >` (ici avec un support minimum de 66% et en appliquant les algorithmes de fouille de données sur le fichier représenté par la figure 2). Ce dernier résultat, une fois re-traduit en termes d'URLs, confirme la découverte d'un comportement commun à *minSup* utilisateurs et fournit l'enchaînement des pages qui constituent ce comportement fréquent. Enfin, l'exploitation par l'utilisateur des résultats obtenus est facilitée par un outil de requête et de visualisation.

3 "Classer pour Découvrir" : méthodologie

Après un bref rappel de nos motivations et d'un premier résultat, la méthode "Diviser pour Découvrir" [Masseglia *et al.*, 2003], nous présentons une nouvelle méthode d'analyse de tous les comportements appelée "Classer pour Découvrir" basée principalement sur une classification des navigations suivi d'une extraction des motifs fréquents.

3.1 Motivations

Considérons des sites Web comme celui du siège de l'INRIA ou celui de l'unité de Sophia Antipolis. Il s'agit de sites riches, dont les thèmes peuvent varier de l'emploi à l'INRIA, au plan stratégique en passant par les pages des membres d'une unité (pages relatives aux activités de recherche, d'enseignement, etc. de chacun). Les leçons que l'on peut tirer d'une activité d'analyse du log correspondant à ces sites sont les suivantes :

1) Généralement, les motifs séquentiels issus du fichier log d'un site de cette ampleur sont assez décevants. En effet leur significativité est assez faible et leur évidence les rend peu utiles (par ex. "0,1% des utilisateurs sont passés par la page d'accueil puis la page du sommaire").

2) Les comportements intéressants sont contingentés à une partie très précise du log. Par exemple, la partie du log correspondant au colloque EPIQUE'03 sera consultée par **0,01%** des utilisateurs enregistrés dans le log. Les utilisateurs ayant visité la page des actualités de l'INRIA, eux, représentent 0,5% des accès sur le site.

3) Si l'on veut extraire des motifs séquentiels intéressants sur ce log, il faut donc spécifier un support global (par rapport à tout le fichier log) extrêmement bas.

Pour découvrir les motifs avec un faible support global, nous avons proposé une méthode itérative – "Diviser pour Découvrir" (D&D) [Masseglia *et al.*, 2003] – qui repose sur une division récursive du log à traiter : celle-ci est basée sur une étape d'extraction de motifs séquentiels et sur une étape de classification des motifs extraits.

Pour pallier certains limites de la méthode D&D, dans cet article, nous proposons une nouvelle méthode, "Classer pour Découvrir" (C&D), qui est également une méthode hybride d'analyse, mais qui évite les nombreuses itérations de la méthode D&D.

3.2 Principe général

Dans les grandes lignes, notre objectif est de découvrir des classes d'utilisateurs (regroupés en fonction de leur comportement sur le site) et d'analyser, ensuite, leurs navigations grâce à une extraction de motifs séquentiels. Le principe général de notre méthode (cf. figure 3) est donc le suivant :

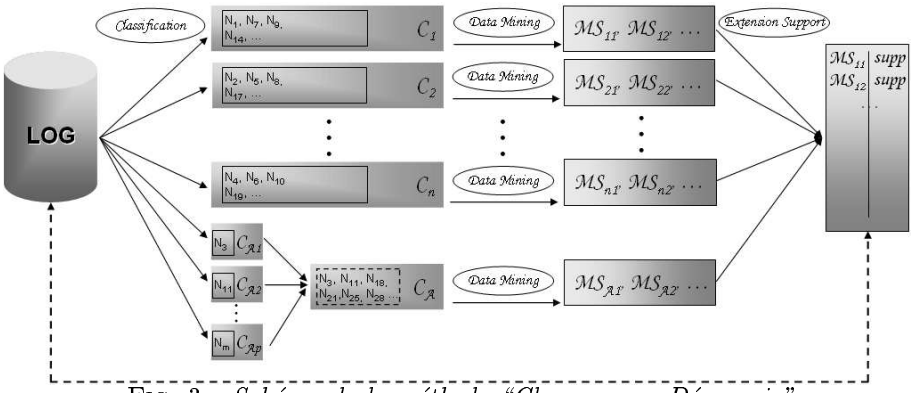

FIG. 3 – *Schéma de la méthode "Classer pour Découvrir"*

`Etape 1 - Classification des navigations.` Classer les navigations du fichier log (structuré en navigations) en classes (sous-logs) en appliquant un algorithme de classification neuronale sur les navigations. Nous utilisons l'algorithme de classification neuronale détaillé en section 4. Une notion de seuil minimum de représentativité pour chaque prototype obtenu est utilisée et permet de définir si un prototype est atypique ou non. Tous les prototypes atypiques sont réunis dans une même classe (sous-log) C_a. Suite à la classification, le fichier sera divisé en classes distinctes (sous-logs) qui regrouperont les comportements proches.

`Etape 2. - Fouille de données.` Extraire les motifs fréquents dans les sous-logs, en appliquant l'algorithme PSP [Masseglia *et al.*, 2000], d'extraction de motifs séquentiels. Pour automatiser tout le processus, nous proposons d'utiliser une même valeur du *support local* (i.e. support relatif à un sous-log) au départ pour tous les sous-logs créés.

`Etape 3. - Calcul du support global.` Etendre le support local des motifs vis-à-vis du log d'origine en recherchant toutes les navigations qui vérifient les motifs fréquents trouvés intéressants dans le log d'origine. Le nouveau support ainsi obtenu est appellé le *support global*.

La méthode C&D n'implique ainsi que trois étapes à l'inverse des diverses itérations nécessaires de la méthode D&D. De cette façon les motifs fréquents sont découverts dès les deux premières étapes successives et le support global est calculé lors de la troisième étape.

4 Classification neuronale des navigations

Cette section décrit la classification neuronale que nous avons appliqué aux navigations lors de l'étape 1.

4.1 Méthode

Nous avons adapté au cas des navigations Web une méthode de classification issue d'un travail réalisé en 2000 par [Benedek et Trousse, 2002] et intégré dans une plateforme objet CBR*Tools[5] [Jaczynski, 1998] pour les besoins d'un contrat France Telecom-INRIA (1998-2000). Cette méthode (utilisée également dans D&D) s'appuie sur un modèle hybride de mémoire composé d'une partie connexionniste inspirée du modèle ARN2 [Azcarraza et Giacometti, 1991] et constituée d'un réseau neuronal à base de prototypes avec une structure évolutive[6] et d'une partie de mémoire plate qui contient les différents groupes de navigations.

Architecture d'un réseau à base de prototypes : celle-ci contient trois couches [Giacometti, 1992] comme le montre la figure 4 :

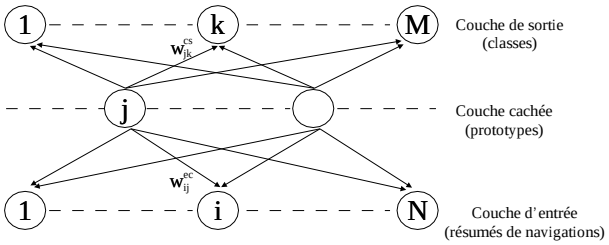

FIG. 4 – *Architecture générale d'un réseau à base de prototypes*

1. Une couche d'entrée qui comprend N unités, N correspond à la dimension de l'espace de description de la navigation (cf. section 4.2).
2. Une couche cachée qui comprend les prototypes décrits dans le même espace de description que les navigations : par exemple, le prototype j sera représenté par le vecteur de pondération : $W^{ec}_j = (w^{ec}_{1j}...w^{ec}_{Nj})$, poids des connexions qui lient cette unité aux unités de la couche d'entrée.
3. Une couche de sortie qui comprend M unités correspondant aux M classes ; $w^{cs}_{jk} = 1$ si le prototype j appartient à la classe k; $w^{cs}_{jk} = 0$ sinon

Un prototype est caractérisé par son vecteur de référence dans l'espace de description, d'une région d'influence et d'un ensemble de navigations qu'il représente. A chaque prototype, on associe un seuil s_i qui sera modifié pendant l'apprentissage, celui-ci déterminant une région dans l'espace des entrées appelée *région d'influence*. Cette région est définie par l'ensemble des vecteurs d'entrée ayant une mesure de distance inférieure à un seuil donné Se.

5. CBR*Tools URL=http://www-sop.inria.fr/axis/software.html#tools

6. i.e. dans le sens où le nombre d'unités cachées (prototypes) dans la couche cachée n'est pas fixé au départ et il peut croître au cours de l'apprentissage.

Dynamique d'activation Soit n une navigation à classer et U l'ensemble des prototypes contenant n dans leur région d'influence. L'activation des prototypes est donnée par l'équation :

$A_c = 1$ *si* $D(n,W_c^{ec}) = \min_{j \in U} D(n,W_j^{ec})$: $A_c = 0$ *sinon*

où c est le prototype gagnant, $W_c^{ec} = (w_{1c}^{ec}...w_{Nc}^{ec})$ et D une distance euclidienne (avec normalisation). Les poids w_{jk}^{cs} entre la couche cachée et la couche de sortie sont donnés par les formules :

$w_{jk}^{cs} = 1$ *si* le prototype j appartient à la classe k; $w_{jk}^{cs} = 0$ *sinon.*

L'activation de la classe k (cas où O_k=1) est donnée par : $O_k = \sum_{j \in U} A_j \times w_{jk}^{cs}$

Mode d'apprentissage et de mémorisation Pendant le mode d'apprentissage, une navigation n ayant le vecteur de description $(n_1,..,n_N)$ est présentée au réseau :

1. La navigation tombe dans *une*, *plusieurs* ou *aucune* région(s) d'influence, le prototype le plus proche à n est activé, appelons le prototype gagnant C[7]. Par conséquent, une seule classe *au plus* sera activée.
2. Si n n'est tombée dans aucune région d'influence, un nouveau prototype représentant cette navigation est ajouté à la couche cachée avec un seuil d'influence.
3. Les poids du prototype gagnant sont modifiés de la façon suivante : $W_c^{ec} = W_c^{ec} + \alpha(t) \times (m - W_c^{ec})$ où $\alpha(t)$ est une suite décroissante avec le temps, elle est donnée par la relation : $\alpha(0) = 1; \alpha(1) = c_1; \alpha(t+1) = \frac{\alpha(t)}{1 + c_2 \times \alpha(t)}$ où c_1 et c_2 sont deux constantes. Cette opération permet d'obtenir des prototypes représentatifs pour chaque classe.

4.2 Construction d'un résumé pour chaque navigation

Pour caractériser une navigation, nous choisissons, comme dans la méthode D&D, d'utiliser quatre attributs basés sur une *généralisation des pages Web* le constituant, d'une part sur l'aspect "multi-sites" et d'autre part sur l'aspect "rubriques sémantiques visitées de premier niveau tous sites confondus". Les *rubriques sémantiques* correspondent à des regroupements faits à partir des rubriques de premier niveau obtenues syntaxiquement : par exemple les rubriques de premier niveau ACACIA, AXIS, ICARE, etc. du site *www-sop.inria.fr* sont regroupées dans une rubrique sémantique 'PROJETS de recherche de l'Inria, celles de *cma* et *w3c* du site *www-sop.inria.fr* dans une rubrique sémantique *partenaires*. Les quatre attributs sont :1) le nombre des documents de la navigation par site (DOCSPARSITE), 2) le nombre de rubriques de deuxième niveau explorées par site à partir des différentes pages de la navigation (RUBRIQUE2PARSITE), 3) le nombre de documents de la navigation par rubrique sémantique de premier niveau (tous sites confondus) (DOCSPARRUBRIQUESEM1) et enfin 4) le nombre de rubriques de deuxième niveau explorées par rubrique sémantique de premier niveau tous sites confondus (RUBRIQUE2PARRUBRIQUESEM1) à partir des différentes pages de la navigation. Chaque attribut fait l'objet de l'affectation d'un poids d'importance relativement au contexte choisi.

7. Un traitement particulier proposé dans [Giacometti, 1992] et basé sur la notion de région d'incertitude a été implanté pour les navigations frontières, évitant la création d'un nombre considérable d'unités cachées dans le réseau.

Cette modélisation des navigations correspond à des vecteurs dans un espace de description de dimension égale à (2 x nb sites considérés + 2 x nb rubriques sémantiques de premier niveau intervenant dans les navigations, tous sites confondus i.e. union des rubriques 1) soit le vecteur (DocsParSite Rubrique2ParSite DocsParRubriqueSem1 Rubrique2ParRubriqueSem1).

L'utilisation des rubriques sémantiques de premier niveau (à définir sur les sites considérés) permettent de réduire l'espace de description d'une navigation davantage qu'en utilisant les rubriques de premier niveau obtenues par analyse syntaxique de l'URL de chaque requête.

`Illustration sur une navigation` : supposons que nous fusionnions les logs de deux sites *www.inria.fr* et *www-sop.inria.fr* pour le log d'origine, que nous les structurions en navigations. Soit la navigation :

(a) http://www.inria.fr/ actualites/index.fr.html
(b) http://www.inria.fr/actualites/colloques/index.fr.html
(c) http://www-sop.inria.fr/axis/jft-2004/.

Les pages Web intervenant dans cette navigation sont : deux du site *www.inria.fr* et une du site *www-sop.inria.fr* : DocsParSite = (2 1).

Les rubriques sémantiques mises en jeu au premier niveau sont : *actualités* pour les deux premières pages et *projets* pour la dernière. Supposons ici que les logs considérés ne mettent en jeu que six rubriques sémantiques de niveau un (dont *actualités* et *projets*) tous sites confondus : DocsParRubriqueSem1 = (0 2 0 0 1 0).

La page *(a)* n'a pas de rubrique de deuxième niveau, pendant que les pages *(b)* et *(c)* ont *colloques* et respectivement *jft-2004* comme rubrique de deuxieme niveau. Enfin la page *(a)* sera considérée comme un document du site *www.inria.fr* et de la rubrique de premier niveau *actualites*. Cette page n'a pas de rubrique de deuxième niveau : Rubrique2ParSite = (1 1), Rubrique2ParRubriqueSem1 = (0 1 0 0 1 0).

Ainsi, cette navigation sera résumée de la manière suivante à l'aide du vecteur de description (2 1 1 1 0 2 0 0 1 0 0 1 0 0 1 0). Le réseau de prototypes ainsi constitué comprend une couche d'entrée de dimension 16 (chaque navigation donnant lieu à un vecteur de description de dimension 16).

5 Expérimentations

Les logs sur lesquels nous avons effectué nos expérimentations portent sur une période d'un mois (Février 2003) pour le site du siège (1 607 Mo, 9 139 391 lignes) et pour le site de Sophia Antipolis (837 Mo, 4 193 695 lignes) et représentent au total environ 2,4 Go.

Après l'étape de prétraitement (cf. [Tanasa et Trousse, 2003]), les nouvelles caractéristiques du fichier log initial et du log structuré en navigations sont présentées dans le tableau 1. Nos analyses portent sur les navigations comportant au moins deux pages soit sur 178 086 navigations. Nos programmes d'analyse des logs sont réalisés en Perl pour le prétraitement, en Java pour l'algorithme de classification et en C++ pour d'extraction de motifs séquentiels[8].

8. Nous avons intégré les programmes de fouille de données dans une plateforme appelée "Cluster & Discover".

Caractéristique	Valeur
Taille du fichier log structuré	332 Mo
Nombre de navigations	382 625
Nombre de navigations longues (longueur $\geq$ 2 pages)	178 086
Longueur moyenne des navigations	7.86

TAB. 1 – *Caractéristiques du fichier log structuré en navigations*

Pour nos expérimentations, nous avons comparé notre méthode C&D avec notre première méthode D&D proposée dans [Masseglia *et al.*, 2003] et une extraction des motifs par l'algorithme PSP directement sur le log d'origine. Les tests ont été réalisés sur une machine de type PC équipée de deux processeurs pentium IV à 2,8 Ghz avec 2 GO de mémoire vive et exploitée par un système Linux (2.4).

Nous présentons quelques-uns des comportements fréquents que nous avons mis en évidence (la plupart internes à un site) avec les deux méthodes C&D et D&D :

C1 : avec le préfixe **http://www-sop.inria.fr/mascotte/personnel/Sebastien.Choplin/cours/iut-infocom/excel/** : <(exercices.html) (exo1.xls) (exo2.xls) (exo3.xls) (exo4.xls) (exo5.xls)>. Ce motif reflète le comportement des utilisateurs qui se sont intéressés aux activités d'enseignement de S. Choplin.

C2 : avec le préfixe **http://www-sop.inria.fr/reves/publications/data/2002/SD02/** : < (?LANG=gb) (.psm_divx.avi.gz) (PerspectiveShadowMaps.pdf) (PerspectiveShadowMaps.pdf) (PerspectiveShadowMaps.pdf)>.

C3 : avec le préfixe **http://www-sop.inria.fr/mefisto/java/tutorial1/** : <(tutorial1.html) (node3.html) (node4.html) (node5.html) (node7.html)>.

C4 : avec le préfixe **http://www.inria.fr/** : < (MSADC/root.exe?/c+dir) (scripts/%2e%2e%255c%2e%2e/winnt/system32/cmd.exe?/c+dir) (_mem_bin/%2e%2e%255c%2e%2e/%2e%2e%255c%2e%2e/%2e%2e%255c%2e%2e/winnt/system32/cmd.exe?/c+dir) (scripts/%2e%2e%c1%1c%2e%2e/winnt/system32/cmd.exe?/c+dir) (scripts/%2e%2e%c1%9c%2e%2e/winnt/system32/cmd.exe?/c+dir) (scripts/%2e%2e%25%35%63%2e%2e/winnt/system32/cmd.exe?/c+dir)>. Ce comportement enchaînant des appels à des scripts est typique d'une attaque pirate, d'après le responsable de la sécurité du réseau de l'unité de Sophia Antipolis.

	PSP	**D&D**			**C&D**	
Id	Supp. global	Supp. local	Itérations	Supp. global	Supp. relatif	Supp. global
C1	-	30.7%	3	0.068%	8.5%	0.069%
C2	-	12.8%	2	0.03%	17.1%	0.03%
C3	-	5.53%	4	0.0079%	3.46%	0.009%
C4	-	85.7%	9	0.0033%	2.56%	0.0033%

TAB. 2 – *Comparaison des résultats de deux méthodes*

Nous avons fait notre comparaison en utilisant le même algorithme de classification i.e le même résumé que ce soit pour un motif ou une navigation basé sur des comptages et des rubriques sémantiques (et non des rubriques syntaxiques comme dans [Masseglia *et al.*, 2003]). Il faut noter que, pour la méthode D&D, nous avons eu besoin de plusieurs itérations (cf. tab. 2) pour trouver certains motifs (parfois en changeant les paramètres de configuration des algorithmes), alors que, dans C&D, nous trouvons les motifs dès la deuxième étape. Notons que ces quatre motifs fréquents n'ont pas pu être trouvés avec une approche d'extraction classique (PSP), le support global mis en

jeu étant beaucoup trop faible ($< 0.069\%$). Comme on le remarque dans le tableau 2, le support global d'un motif avec C&D est supérieur ou égal à celui obtenu avec la méthode D&D. La division des logs est aussi différente pour les deux méthodes : dans D&D, nous pouvons avoir une navigation qui appartient à plusieurs classes alors que dans C&D les classes sont disjointes.

6 Conclusions et perspectives

Dans cet article nous avons présenté une nouvelle méthodologie pour l'analyse de tous les comportements des utilisateurs de sites Web à partir d'un fichier access log de grande taille. Celle-ci est composée de seulement trois étapes : une de classification neuronale sur le fichier log structuré en navigations, une étape d'extraction de motifs séquentiels dans les classes obtenues à l'étape 1 et une étape de calcul du support global des motifs.

Nous avons validé notre méthode "Classer pour Découvrir" en nous comparant à notre première proposition hybride "D&D", décrit dans [Masseglia *et al.*, 2003]. Les expérimentations ont été effectuées sur les mêmes fichiers log de deux sites d'INRIA : *www.inria.fr* et *www-sop.inria.fr*. Celles-ci ont permis de retrouver les mêmes avantages que la méthode D&D i.e. l'extraction de comportements minoritaires très fréquents à un moment donné qu'il aurait été difficile voire impossible d'obtenir à partir du log d'origine avec des algorithmes classiques. De plus la méthode "Classer pour Découvrir" évite les nombreuses itérations de la méthode D&D, facilite davantage l'extraction des comportements de support global extrêmement faible et propose le calcul du support global des motifs extraits vis-à-vis du log d'origine.

Les perspectives envisagées de ce travail concernent l'amélioration de la classification automatique des navigations avec des nouvelles définitions d'un résumé d'une navigation intégrant la notion d'*utilité* des motifs extraits voire des nouvelles méthodes de classification.

Références

[Agrawal *et al.*, 1993] R. Agrawal, T. Imielinski, et A. Swami. Mining Association Rules between Sets of Items in Large Databases. In *Proceedings of the 1993 ACM SIGMOD Conference*, pages 207–216, Washington DC, USA, May 1993.

[Agrawal et Srikant, 1995] R. Agrawal et R. Srikant. Mining sequential patterns. In *Eleventh International Conference on Data Engineering*, Taipei, Taiwan, 1995.

[Azcarraza et Giacometti, 1991] A. Azcarraza et A. Giacometti. A prototype-based incremental network model for classification task. In *Fourth International Conference on Neural Networks and Their Applications*, Nimes, France, 1991.

[Benedek et Trousse, 2002] A. Benedek et B. Trousse. Adaptation of Self-Organizing Maps for CBR case indexing. In *Proceeding of the 4th International Workshopon Symbolic and Numeric Algorithms for Scientific Computing*, pages 31–45, Timisoara, Romania, October 2002.

[Cooley *et al.*, 1999] R. Cooley, B. Mobasher, et J. Srivastava. Data preparation for mining world wide web browsing patterns. *Knowledge and Information Systems*, 1(1):5–32, 1999.

[Fayad *et al.*, 1996] U.M. Fayad, G. Piatetsky-Shapiro, P. Smyth, et R. Uthurusamy, editors. *Advances in Knowledge Discovery and Data Mining*. AAAI Press, Menlo Park, CA, 1996.

[Giacometti, 1992] A. Giacometti. *Modèles Hybrides de l'Expertise*. PhD thesis, Telecom – Paris, 1992.

[Jaczynski, 1998] M. Jaczynski. *Modèle et plate-forme à objets pour l'indexation des cas par situation comportementales: application à l'assistance à la navigation sur le Web*. PhD thesis, Université de Nice Sophia-Antipolis, 1998.

[Luotonen, 1995] A. Luotonen. The commom log file format. http://www.w3.org/pub/WWW/, 1995.

[Masseglia *et al.*, 2000] F. Masseglia, P. Poncelet, et R. Cicchetti. An efficient algorithm for web usage mining. *Networking and Information Systems Journal (NIS)*, April 2000.

[Masseglia *et al.*, 2003] F. Masseglia, D. Tanasa, et B. Trousse. Diviser pour découvrir : une méthode d'analyse du comportement de tous les utilisateurs d'un site web. In *Les 19èmes Journées de Bases de Données Avancées*, Lyon, France, 2003.

[Mobasher *et al.*, 2002] B. Mobasher, H. Dai, T. Luo, et M. Nakagawa. Discovery and evaluation of aggregate usage profiles for web personalization. *Data Mining and Knowledge Discovery*, 6(1):61–82, January 2002.

[Network Working Group, 1998] Network Working Group. RFC 2396, Uniform Resource Identifiers (URI): Generic Syntax. http://rfc-2936.rfc-index.org/, 1998.

[Spiliopoulou *et al.*, 1999] M. Spiliopoulou, L.C. Faulstich, et K. Winkler. A data miner analyzing the navigational behaviour of web users. In *Proc. of the Workshop on Machine Learning in User Modelling of the ACAI'99 Int. Conf.*, Creta, Greece, 1999.

[Tanasa et Trousse, 2003] D. Tanasa et B. Trousse. Le prétraitement des fichiers logs web dans le "web usage mining" multi-sites. In *Journées Francophones de la Toile (JFT'2003)*, Tours, 2003.

Summary

The Web Usage Mining is a new, promising and complex domain of Data Mining. This is shown by the increasing number of data mining techniques applied to the Web logs. However, based on the complexity of these new data, in this paper, we studied the following question: *Is it possible to combine existing methods in order to propose an analysis which benefits from the results of several techniques areas?*. Our study was oriented to a new hybrid approach emerged from neuronal clustering and frequent pattern mining areas and based on semantical summaries of users' navigations.
Our objectives are: 1) to go beyond the limits of these two data mining techniques by increasing the quantity of the data analysed and the quality of the results obtained and 2) to overpass the limitations of a first analysis method for WUM, "Divide & Discover", that we presented in 2003. Our experiments were conducted on the Web logs of two of the INRIA's Web sites. The results confirm that our approach is well-founded compared with the state of art.