Diverses approches permettant l'introduction du temps dans la fouille de données d'usage du Web

Alzennyr Da Silva

Projet AxIS, INRIA Rocquencourt
Domaine de Voluceau, Rocquencourt, B.P. 105
78153 Le Chesnay cedex, France
Alzennyr.Da_Silva@inria.fr
http://www-rocq.inria.fr/axis

Résumé. Le volume des transactions commerciales réalisées en ligne sur l'Internet ne cesse de croître. Le Web est donc devenu l'une des plates-formes commerciales les plus importantes et beaucoup d'opérateurs de sites Web sont incités à analyser l'utilisation de leurs sites afin d'améliorer leur rentabilité. Cependant, la manière dont un site Web est visité peut changer en raison de modifications liées à la structure et au contenu du site lui-même, ou bien en raison de l'évolution du comportement de certains groupes d'utilisateurs. Ainsi, les modèles d'usage doivent être mis à jour continuellement afin de refléter le comportement réel des visiteurs. Dans ce contexte, nous présentons une brève vue de l'ensemble des techniques appliquées à la fouille des données temporelles. Une attention particulière est apportée aux méthodes consacrées à la découverte de modèles sur les données d'usage du Web. Nous évoquons également certaines questions en suspens dans ce contexte.

1 Introduction

De nombreuses sociétés souhaitent exploiter les grandes quantités de données accumulées quotidiennement durant leurs opérations commerciales pour mieux comprendre leur activité globale et mieux appréhender le contexte commercial, afin d'adopter de meilleures décisions stratégiques. Il en résulte une utilisation croissante de la fouille de données (Data Mining), qui est définie comme l'application de l'analyse de données et des algorithmes de découverte sur les grandes bases de données ayant comme but la découverte de modèles non triviaux (Fayyad et al., 1996). Plusieurs algorithmes ont été proposés afin de formaliser les nouveaux modèles découverts, de construire des modèles plus efficaces, de traiter de nouveaux types de données et de mesurer les différences entre les ensembles de données. Cependant, la plupart des algorithmes de fouille de données supposent que les modèles sont statiques et ne tiennent pas compte de l'éventuelle évolution de ces modèles au cours du temps.

Concernant les données volumineuses et dynamiques, le Web est devenu l'exemple le plus pertinent grâce à l'augmentation colossale du nombre de documents mis en ligne et des nouvelles informations ajoutées chaque jour. Les modèles d'accès au Web ont une nature dynamique, dû au contenu et à la structure d'un site Web ou bien dû au changement d'intérêt de

ses utilisateurs. Les modèles d'accès à un site Web peuvent être dépendants de certains paramètres, comme par exemple : l'heure et le jour de la semaine, des événements saisonniers (vacances d'été, d'hiver, Noël, etc.), événements externes dans le monde (épidémies, guerres, crises économiques, coupe du monde de football), etc.

Ces considérations ont motivé d'importants efforts dans l'analyse de données temporelles et l'adaptation des méthodes de la fouille de données statiques aux données qui évoluent pendant le temps. Le présent article se place dans ce courant de recherche et présente trois volets principaux : la fouille du Web, la fouille de données temporelles et une combinaison de ces deux dernières, la fouille de données temporelles d'usage du Web. L'objectif de l'article est d'apporter dans un premier temps une exploration initiale (et non exhaustive) de ces trois axes, liés à la recherche et aux applications, puis de susciter une discussion sur les limites et les enjeux rapportés à ce domaine.

L'article est organisé comme suit : dans la section 2 nous présenterons les définitions et les applications liées à la fouille du Web, dans la section 3 nous discuterons des techniques de la fouille de données temporelles et dans la section 4 nous présenterons une nouvelle conception de fouille de données qui trouve son fondement dans la combinaison des deux premières, la fouille de données temporelles d'usage du Web, ainsi que ses applications. La section 5 est centrée sur une discussion portant sur plusieurs exjeux importants dans la fouille de données temporelles d'usage du Web et présente également des propositions pour répondre à quelques uns de ces défis. La dernière section présente les conclusions de cette étude.

2 Fouille du Web

La fouille du Web (Web Mining, WM) (Kosala et Blockeel, 2000) s'est développée à la fin des années 90 et consiste à utiliser l'ensemble des techniques de la fouille de données afin de développer des outils permettant l'extraction d'informations pertinentes à partir de données du Web (documents, traces d'interactions, structure des pages, structure des liens, etc.). Selon l'objectif visé, plusieurs types d'études peuvent être réalisés : le premier se concentre sur l'analyse du contenu des pages Web (Web Content Mining), le deuxième concerne l'analyse des liens entre pages Web (Web Structure Mining) et le dernier traite de l'usage (Web Usage Mining) (Srivastava et al., 2000), c'est-à-dire, des traces laissées sur les sites Web lors des connexions, notamment les clics effectués par un utilisateur sur un ou plusieurs sites. Plus précisément, la fouille de données d'usage du Web désigne l'ensemble de techniques basées sur la fouille de données pour analyser le comportement des utilisateurs d'un site Web (Cooley et al., 1999; Spiliopoulou, 1999) et c'est précisément sur cette partie que le présent article se concentre. Une des motivations de l'analyse de l'usage est la fidélisation des internautes et le recrutement de nouveaux visiteurs. Pour se faire, on analyse le comportement des utilisateurs afin d'extraire des modèles ou schéma d'accès au(x) site(s) étudié(s) en vue d'une amélioration, voire d'une personnalisation, de ce(s) dernier(s).

Lorsqu'un utilisateur consulte un site Web, il navigue en effectuant une suite de "clics" avec sa souris et introduisant du texte avec son clavier. Ces informations déclenchent des requêtes (affichage d'une page du site, téléchargement d'un fichier, identification de l'utilisateur via un mot de passe, etc.) qui sont enregistrées en format texte et stockées de manière standardisée dans un fichier qui s'appelle "log web". Ce fichier est maintenu par le serveur HTTP. Suivant la fréquentation du site web, la taille du fichier "'log web" peut atteindre des proportions im-

portantes, pouvant croître de quelques centaines de mégabytes jusqu'à plusieurs dizaines de gigabytes par mois.

A chaque page visitée, le fichier log enregistre notamment : l'identité du visiteur (via son adresse IP), la date et l'heure de la requête, la page consultée ou les fonctionnalités mises en oeuvre (téléchargement de fichier, clic sur image, etc.), le code de statut attribué à la requête (prend la valeur 200 en cas de réussite), le nombre d'octets transmis, la page précédemment visitée (ou le moteur de recherche utilisé pour rejoindre la page Web suivi des mots clés demandés), la configuration du client, c'est-à-dire, son navigateur Web (Firefox, Internet Explorer, etc.) et son système d'exploitation (Windows, Linux, Mac OS, etc.). Le format d'un fichier log (Common Log Format) a été standardisé par W3C (W3C, 1995). Le tableau 1 illustre un extrait de fichier log d'accès du serveur Web de l'INRIA.

En ce qui concerne l'identification de l'utilisateur, pour les sites Web exigeant un login préalable, le nom de l'utilisateur est enregistrée dans le troisième champ du fichier log Web. Dans ce cas-ci, cette information peut être utilisée pour identifier l'utilisateur. Autrement, l'identification de l'utilisateur à partir des fichiers log Web enregistrés par le serveur devient une tâche assez complexe en raison des serveurs proxy-cache, des adresses dynamiques, des utilisateurs multiples (dans une bibliothèque, un cyber café, etc.) ou encore quand un utilisateur accède à l'Internet à partir de plusieurs ordinateurs. Les points positifs qui justifient l'utilisation d'un serveur proxy-cache sont : (a) la sécurité, une seule machine accède à l'Internet ; et (b) les transferts, une page demandée par plusieurs personnes est téléchargée une seule fois (Abrams et al., 1995). Toutefois, en ce qui concerne l'identification de l'utilisateur, la présence d'un serveur proxy-cache implique des effets secondaires négatifs, comme par exemple le fait que les logs ne refléteront plus exactement l'utilisation d'un site Web, dès lors que l'adresse enregistrée dans les fichiers log sera celle de la machine proxy-cache et non plus celles des machines des clients. De plus, la demande d'une page présente dans le cache du serveur proxy-cache ne sera plus transmise au serveur qui héberge la page en question. Par conséquence, cette demande ne sera pas enregistrée sur le log du site Web, ce qui rend ce fichier incomplet vis-à-vis des vraies demandes des utilisateurs. De plus, le cache ne fonctionne pas avec les sites Web dynamiques car puisqu'une page demandée par deux personnes est générée différemment car elle prend en compte les paramètres propres à chacune de ces personnes.

Pour tout cela, une détection stricte des différents utilisateurs connectés au même serveur proxy-cache devient impossible. Malgré tout, un certain nombre de techniques ont été proposées afin de contourner les difficultés imposées sur l'identification des utilisateurs. Pour identifier un utilisateur qui a déjà visité un certain site Web, les techniques les plus communes sont les cookies (petits fichiers textes stockés chez le client), le login d'utilisateur et les navigateurs modifiés. Toutes ces techniques ont l'inconvénient de s'introduire dans le domaine privé de l'utilisateur. Autrement, dans (Berendt et al., 2002), les auteurs rapportent que la combinaison des champs *IP Address* et *User Agent* (le navigateur Web) d'un fichier log Web identifie correctement l'utilisateur dans 92,02% des cas et seul un nombre limité de ces combinaisons (1,32%) sont utilisés par plus de trois utilisateurs (en raison des serveurs proxy-cache).

Applicant cette strategie sur les données du tableau 1, il est possible d'apprendre qu'un certain utilisateur identifié par le couple adresse IP (194.78.232.8) et navigateur (MSIE 5.0b1) a démarré une visite sur le site de l'INRIA à partir de la page d'accueil et a en suite demandé trois pages : *liens.htm*, *Telesc.html* et *Vsurv.html*.

Une fois l'utilisateur (approximativement) identifié, le problème consiste alors en détecter

adresse IP	date/heure	URL	statut	bytes	page précédente	navigateur
194.78.232.8	[10/Sep/2001 :15 :33 :01 +0200]	GET /orion/index.htm HTTP/1.1	200	1502	-	Mozilla/4.0 (compatible ; MSIE 5.0b1 ; Mac_PowerPC)
194.78.232.8	[10/Sep/2001 :15 :33 :43 +0200]	GET /orion/liens.htm HTTP/1.1	200	1893	http ://www-sop.inria.fr/ orion/index.html	Mozilla/4.0 (compatible ; MSIE 5.0b1 ; Mac_PowerPC)
100.64.30.6	[10/Sep/2001 :15 :34 :07 +0200]	GET /stacs2002/home.html HTTP/1.0	200	483	http ://www.google.fr/ search ?hl=fr&q=inria +statistiques&meta=	Microsoft Internet Explorer 5.0
194.78.232.8	[10/Sep/2001 :15 :34 :09 +0200]	GET /orion/Telescope/Telesc.html HTTP/1.1	200	4433	http ://www-sop.inria.fr/ orion/liens.htm	Mozilla/4.0 (compatible ; MSIE 5.0b1 ; Mac_PowerPC)
100.64.30.6	[10/Sep/2001 :15 :34 :09 +0200]	GET /stacs2002/Images/affiche.jpg HTTP/1.0	200	281281	http ://www-sop.inria.fr /stacs2002/home.html	Microsoft Internet Explorer 5.0
194.78.232.8	[10/Sep/2001 :15 :34 :23 +0200]	GET orion/Telescope/Vsurv.html HTTP/1.1	200	2979	http ://wwwsop.inria.fr/ /orion/Telescope/Telesc.html	Mozilla/4.0 (compatible ; MSIE 5.0b1 ; Mac_PowerPC)

TAB. 1 – *Fragment d'un fichier log Web contenant 6 requêtes HTTP (unités élémentaires).*

toutes les requêtes d'un utilisateur. Concernant cette problématique, la méthode la plus simple pour le groupement des requêtes en sous-ensembles de requêtes appartenant au même utilisateur (navigations) est d'utiliser le temps de latence maximum entre deux requêtes successives. Les auteurs de (Catledge et Pitkow, 1995) ont estimé ce temps de manière empirique à 25,5 minutes et la majorité des méthodes d'analyse de l'usage du Web ont adopté le temps de latence de 30 minutes.

Nous pouvons considérer la combinaison adresse IP + navigateur comme étant un critère acceptable pour l'identification d'un utilisateur dans le cadre une activité ponctuelle, c'est-à-dire, elle est capable d'identifier les requêtes en provenance d'un même utilisateur dans le contexte d'une seule navigation. Cependant, on ne peut pas généraliser cette combinaison pour l'identification de plusieurs navigations appartenant à un même utilisateur, puisque nous n'avons aucune garantie que l'utilisateur d'avant aura les mêmes valeurs pour le couple adresse IP + navigateur lors d'une prochaine visite sur le site Web.

Dans les analyses d'usage du Web, on remarque deux grandes catégories : (a) les analyses "orientées serveur", qui s'appuient sur l'analyse de la structure sémantique (plan) du site et des traces de l'usage (fichiers logs) enregistrés sur les serveurs HTTP ; et (b) les analyses "orientées utilisateur" qui s'appuient sur les données capturées par le navigateur installé sur l'ordinateur de l'utilisateur (position et clics de la souris, option de langue, d'affichage, d'ergonomie, etc.). La plupart des analyses menées jusqu'à présent appartiennent à la première catégorie. A ce sujet, les auteurs de (Padmanabhan et al., 2001), en travaillant au niveau intermédiaire d'un fournisseur de contenu disposant de logs relatifs à plusieurs sites commerciaux, ont montré que le point de vue "orienté serveur" est partiel et biaisé. Les conclusions que l'on peut tirer

de ce type d'approche doivent être considérées avec réserve, en particulier lorsqu'elles tendent à dresser des comportements généraux de navigation . Du côté "orienté utilisateur" et dans le cadre du commerce électronique, une étude réalisé par (Licoppe et al., 2002) montre que les consommateurs en ligne ont un comportement très volatile, variant entre l'achat réfléchi et l'achat d'impulsion, et en outre, que l'achat en ligne implique la mobilisation de ressources à l'extérieur des sites de commerce électronique (moteurs de recherche, comparateurs, etc.) et aussi à l'extérieur du Web.

Toutefois, l'analyse des fichiers log Web est particulièrement utile car elle fournit des informations sur la manière dont les utilisateurs naviguent réellement sur le site Web. Après la réalisation d'une telle analyse, il est ainsi possible :

- de mettre en évidence les fonctionnalités les plus et les moins utilisées dans le site ;
- de chercher à comprendre les raisons pour lesquelles les fonctionnalités les moins utilisées sont délaissées par les utilisateurs afin, selon les cas, de les améliorer ou de les supprimer ;
- de mettre en évidence les erreurs les plus fréquemment commises par les utilisateurs, afin d'identifier et de résoudre les problèmes d'organisation ou d'utilisation qui pourraient en être la cause.

Au delà, la fouille de données d'usage du Web peut apporter des avantages à d'autres domaines, comme par exemple, l'ajout dynamique de liens dans des pages Web (Yan et al., 1996), la recommandation d'objets, la caractérisation de groupes d'utilisateurs, la performance d'un réseau d'ordinateurs, etc.

Un système Web de recommandation d'objets est ainsi capable de fournir des suggestions de produits (pages, livres, films, etc.) susceptibles d'intéresser l'utilisateur (Pierrakos et al., 2003; Mobasher, 2004). Les techniques utilisées peuvent être basées sur différents critères : la correspondance du produit au profil du client, la similarité du produit aux autres produits que le client a déjà acheté ou visualisé, ou bien l'appréciation des autres clients ayant des profils similaires au client actuel. Cette stratégie est intéressante dans le cadre du commerce électronique, mais aussi plus généralement dans un contexte dans lequel une évaluation objective de la qualité des objets à recommander est difficile (voire impossible).

La caractérisation de groupes d'utilisateurs consiste à identifier des traits d'usage partagés par un nombre suffisant d'utilisateurs d'un site Web puis inférer un profil à chaque groupe (Chi et al., 2002; Da Silva et al., 2006). De plus, avec la mise en oeuvre des applications nécessaires, l'exploitation des données Web peut également permettre d'examiner la façon dont des navigateurs sont employés et l'interaction de l'utilisateur avec une interface de navigateur (Catledge et Pitkow, 1995).

En ce qui concerne les réseaux d'ordinateurs, l'exploitation d'usage du Web constitue une stratégie importante d'analyse du trafic et donne des bases pour la définition de politiques comme le *caching* et le *prefetching* anticipés (Sow et al., 2003) (quelques pages Web sont stockées directement sur le cache lors d'un premier accès au site, évitant ainsi une deuxième demande au serveur d'origine).

3 Fouille de données temporelles

L'intérêt pour les bases de données temporelles a augmenté ces dernières années et un nombre de plus en plus important de prototypes et de systèmes mis en application tiennent

compte de la dimension temporelle des données de façon explicite, par exemple pour étudier la variabilité au cours du temps des résultats d'analyse.

La fouille de données temporelles (Temporal Data Mining, TDM) (Laxman et Sastry, 2006) est une extension importante de la fouille de données classique car elle se centre plutôt sur l'analyse des activités plus que de se limiter à celle des états. De ce fait, elle permet de rechercher des associations de cause à effet en exploitant conjointement les proximités contextuelles et temporelles. Il s'agit d'exploiter le fait que les causes précèdent les effets, ce qui est difficile quand la dimension temporelle est négligée ou simplement introduite comme un attribut numérique additionnel dans la description des données.

De ce fait, la fouille de données temporelles présente la capacité d'analyser les aspects comportementaux des objets (ou bien d'ensemble d'objets) par opposition aux règles d'extraction simples qui décrivent un état sur un point spécifique du temps : on cherche à obtenir des enchaînements logiques plutôt qu'à simplement associer des évènements. Comme exemple, considérons une règle d'association simple énonçant que *"la dinde et le champagne X sont achetés ensemble pendant la période de Noël"*. Ceci constitue un exemple de règle d'association temporelle, la règle statique équivalente associerait simplement les deux produits. L'aspect temporel *"pendant la période de Noël"* apporte des informations cruciales. Premièrement, l'association entre les deux produits peut être rare pendant tout le reste de l'année, pour pourrait ne pas être détectée si l'analyse était concentrée uniquement sur les règles fréquentes sur l'ensemble des données. Deuxièmement et dans un contexte commercial, puisque l'association n'est pertinente que pendant la période de Noël, les offres promotionnelles combinant les deux produits devraient également avoir lieu pendant cette période spécifique de l'année. Une offre contenant les deux produits pendant une autre période ayant vraisemblablement un intérêt limité aux yeux des consommateurs. Les associations temporelles comme celle ci-dessus peuvent donc être très utiles pour établir une stratégie de vente et plus généralement pour piloter une offre commerciale. La fouille de données temporelles présente donc un grand intérêt pour "l'intelligence des affaires" mais aussi plus généralement dans les domaines dans lesquels le temps joue un rôle important, comme par exemple la finance, certaines applications médicales, etc.

En outre, la fouille de données temporelles est directement liée à la fouille de données sur des grands fichiers séquentiels. Par des données séquentielles, on entend les données qui sont ordonnées selon l'index de la séquence. Les données temporelles sont un cas particulier de données séquentielles dans lequel le temps joue le rôle d'indexation. Les séquences de gènes (DNA) et les séquences de mouvements dans un jeu d'échecs constituent d'autres exemples de données séquentielles. Ici, bien qu'il n'y ait aucune notion de temps en tant que tel, l'ordre des observations est très important et même indispensable pour la description et l'analyse de telles données.

Historiquement, le problème de la prévision de séries temporelles a été l'un des plus étudiés (Box et al., 1994) dans le cadre de la météorologie, des finances et de la bourse, mais aussi en contrôle. La recherche en reconnaissance de la parole a motivé de nombreux travaux sur la comparaison et la classification des séries temporelles (Rabiner et Juang, 1993; O'Shaughnessy, 2003), comme par exemple ceux portant sur les modèles de Markov cachés ou les techniques d'apprentissage automatique comme les réseaux de neurones à temps retardé.

La principale différence entre la fouille de données temporelles et l'analyse de séries temporelles concerne la nature des informations qu'on veut estimer ou mettre en évidence. Le

cadre de la fouille de données temporelles se prolonge au-delà des applications standards de prévision ou d'applications de contrôle pour l'analyse des séries temporelles. Dans l'analyse de séries temporelles la prévision joue un rôle central alors que dans fouille de données temporelles c'est plutôt l'évolution que l'on essaie de modéliser.

4 Fouille de données temporelles du Web

La fouille de données temporelles du Web (Temporal Web Mining, TWM) est à l'intersection de la fouille du Web avec la fouille de données temporelles. Selon (Samia, 2003), le TWM est le processus de découverte, d'extraction, d'analyse et de prévision des données contenant des informations temporelles, découvertes par l'application en temps réel des techniques de la fouille de données temporelles sur le Web.

Selon (Samia et Conrad, 2004), le TWM soutient l'aspect temporel des données du Web en considérant ces données comme des séries temporelles. Son but est de présenter la prévision comme une issue principale de la fouille du Web, spécifiquement de la fouille sur le contenu du Web (Web Content Mining). En d'autres termes, TWM vise à faire des prévisions à partir du contenu des pages Web.

Par contre, la fouille de données temporelles d'usage du Web (Temporal Web Usage Mining, TWUM) s'appuie sur les techniques de la fouille de données d'usage du Web (issues des fichiers log Web) pour découvrir les modèles et schémas temporels qui décrivent les comportements des utilisateurs d'un site Web, même si ceux-ci ne sont significatifs que sur une période temporelle fixée. L'extraction de séquences fréquentes et la construction de classes pour la segmentation d'utilisateurs constituent les méthodes d'analyse les plus importantes en TWUM.

Dans (Roddick et Spiliopoulou, 2002), les auteurs résument les solutions proposées et les problèmes en suspens dans l'exploitation de données temporelles, au travers d'une discussion sur les règles temporelles et leur sémantique, mais aussi par l'investigation de la convergence entre la fouille de données et de la sémantique temporelle. Tout récemment, dans (Laxman et Sastry, 2006) les auteurs discutent en quelques lignes des méthodes pour découvrir les modèles séquentiels, les motifs fréquents et les modèles périodiques partiels dans les flux de données. Ils évoquent également des techniques concernant l'analyse statistique de telles approches.

4.1 Approches d'extraction de motifs séquentiels

L'extraction de motifs a été initialement présentée dans (Agrawal et al., 1993). Pour étendre cette problématique à la prise en compte du temps des transactions, les mêmes auteurs ont proposé dans (Agrawal et Srikant, 1995) la notion de séquence. Considérons une base de données de séquences où chaque séquence est une liste de transactions triée par le temps et chaque transaction est un ensemble d'items. Le problème consiste à découvrir tous les motifs séquentiels avec un support minimum spécifié par l'utilisateur, où le support d'un motif est défini comme le nombre de séquences qui contient le motif sur toutes les séquences de la base. Un exemple de motif séquentiel est "25% de clients achètent le livre A et B dans une transaction, suivi par l'achat d'un livre D dans une transaction postérieure". Cette approche a été également explorée par d'autres auteurs (Spiliopoulou, 1999) (Masseglia et al., 1999). Les problèmes principaux concernant l'extraction de motifs, à savoir le grand nombre de règles découvertes ainsi que

l'identification des règles intéressantes, sont des questions largement discutées (Chen et Petrounias, 1999; Liu et al., 2001).

Cependant, ces approches ont toujours une caractéristique commune : l'analyste peut choisir d'appliquer sa méthode sur la totalité de données disponibles ou sur un bloc de données qui présentent certaines particularités utiles pour l'analyse. Ainsi, les changements intéressants qui pourraient se produire dans la période considérée ne sont pas pris en compte. Par exemple, quand les données proviennent d'une accumulation portant sur une période de temps potentiellement longue (comme dans le cas des fichiers log Web), on s'attend à ce que les motifs fréquents évoluent avec le temps. Les modèles d'usage découverts doivent également être mis à jour continuellement (avec des algorithmes efficaces) afin de suivre fidèlement le changement de comportement des visiteurs. Ceci exige de la méthode une surveillance continue des modèles existants, pré-réquit d'importance essentielle pour les applications centrées sur la dimension temporelle. Une solution possible pour traiter ce problème est définir un schéma approprié pour effectuer le partitionnement du temps. Dans (Marascu et Masseglia, 2006) les auteurs appliquent une découpage du flux de données en *batches* de taille fixe.

Dans (Chakrabarti et al., 1998), les auteurs proposent la découverte de motifs surprenant, c'est-à-dire, de motifs inattendus et donc intéressants dans l'analyse de ventes du marché à partir de l'observation de la variation de corrélation des achats d'articles sur l'échelle du temps. L'inspiration de ce travail est issue de l'analyse de séries temporelles, et l'accent est donc mis sur la construction d'une subdivision du temps en intervalles de façon que les statistiques sur des règles varient nettement entre deux intervalles consécutifs. Dans (Baron et Spiliopoulou, 2001) les auteurs proposent un modèle générique de règle (Generic Rule Model, GRM) qui modélise le contenu et les statistiques d'une règle comme un objet temporel. Dans des travaux ultérieurs (Baron et Spiliopoulou, 2003), les mêmes auteurs proposent un moniteur de modèles automatisé (Pattern Monitor, PAM) basé sur les mêmes principes de GRM et appliqué à l'observation des changements de comportement des visiteurs d'un site Web.

4.2 Approches de classification

En ce qui concerne la classification des données issues des fichiers log Web, nous avons d'un coté les approches qui considèrent les navigations des utilisateurs comme les ensembles de clics non ordonnés, dans ce cas les mesures les plus populaires sont la distance euclidienne, le cosinus et le coefficient de Jaccard. De l'autre coté, nous avons les approches qui prennent en compte l'ordre d'accès aux pages, par exemple la réquisition d'une page A suivi par la réquisition d'une autre page B peut apporter plus d'information que simplement savoir que les pages A et B appartiennent à la même navigation. Les auteurs de (Rossi et al., 2006a,b) ont réalisé une étude comparative des dissimilarités de Jaccard, cosinus et `tf x idf` dans le but d'obtentir une classification des pages d'un site web en fonction des navigations extraites des fichiers log. Leurs résultats donnent un avantage à la dissimilarité de Jaccard appliquée dans ce contexte.

Les approches de classification les plus indicatives peuvent être récapitulées comme suit :

– L'analyse de flux de données (Kothari et al., 2003) évalue les similarités entre deux flux de données. Plus spécifiquement, la similarité entre deux flux de données est basée sur la différence de similarité entre deux vues de pages, certains cas considérant aussi le temps dédié à la visualisation de page. Puisque l'analyse sémantique n'est pas possible, le degré de similarité entre deux pages est proportionnel à leur fréquence relative de co-

occurrence. Dans ce contexte, dans (Banerjee et Ghosh, 2001) les auteurs appliquent le *clustering* des flux de données en utilisant comme critère la longueur de la plus grande sous-séquence commune (longest common subsequence, LCS) entre deux flux de données.

– La méthode d'alignement de séquences (Sequence Alignment Method, SAM) (Sankoff et Kruskal, 1983), où les navigations sont séquences de pages chronologiquement ordonnées. La SAM mesure la similarité entre deux navigations en termes du nombre d'opérations nécessaires pour égaliser les deux navigations. La méthode de programmation dynamique et la distance de Levenshtein (Levenshtein, 1966) ou distance d'édition sont souvent utilisées dans cette méthode.
– La clusterisation basée sur la généralisation (Fu et al., 2000) utilise la structure sémantique du site (hiérarchie des pages) pour généraliser les navigations des utilisateurs. Puis, les pages en chaque navigation d'utilisateur sont remplacées par les pages générales correspondantes et groupées en utilisant l'algorithme BIRCH (Zhangn et al., 1996).

En outre, nous avons aussi les approches de classification basées sur des modèles. Ces approches ont été également utilisées dans beaucoup d'applications concernant les données du Web (Baldi et al., 2003). Dans ces modèles, le nombre de groupes est déterminé à travers des méthodes probabilistes, telles que BIC (critère bayésien de l'information), approximations bayésiennes, ou méthodes de *bootstrap* (Fraley et Raftery, 1998). La structure du modèle peut être déterminée par les techniques de sélection de modèle et l'estimation de paramètres en utilisant des algorithmes de vraisemblance maximale, par exemple comme l'algorithme Expectation-Maximization (Dempster et al., 1977). Les modèles de Markov (cachés ou du premier ordre) (Baldi et al., 2003; Cadez et al., 2003) sont les modèles les plus indicatifs dans ce contexte.

Comme exemple de traitement des données évolutives, (Ester et al., 1998) présente un algorithme de *clustering* incrémental issu de DBSCAN (Ester et al., 1996) qui examine quelles parties des classes actuelles sont affectées par une mise à jour de la base de données sur une fenêtre de temps et ajuste les groupes par conséquence.

4.3 Approches basées sur le raisonnement à partir de cas

Le Raisonnement à Partir de Cas (RàPC) se dit d'une approche de résolution de problèmes basée sur la réutilisation par analogie d'expériences passées appelées "cas". Un cas est généralement indexé pour permettre de le récupérer suivant des caractéristiques pertinentes et discriminantes, appelées "indices". Les indices déterminent dans quelle situation (ou contexte) un cas peut être de nouveau réutilisé (Aamodt et Plaza, 1994; Kolodner, 1993). Le système BROADWAY (**BRO**wsing **AD**visor reusing path**WAY**s) (Jaczynski et Trousse, 1999) est un assistant pour la navigation sur le Web réutilisant les navigations passées d'un groupe d'utilisateurs. Ce système utilise le RàPC pour proposer des pages à visiter en recherchant des navigations similaires à celles effectuées par l'utilisateur. BROADWAY utilise les navigations des utilisateurs pour en extraire des expériences utiles (cas) permettant d'associer à un comportement (succession de pages visitées), un ensemble de pages qui seront proposées à l'utilisateur. Dans ce système, c'est l'utilisateur qui marque le début et la fin de sa session de recherche au moment de sa demande d'aide. BROADWAY est un serveur HTTP utilisé comme proxy : il est inséré entre le navigateur et le reste du Web et il intercepte ainsi toutes les demandes de documents pour le protocole HTTP. Durant une navigation, Broadway peut afficher un ensemble

de documents qu'il conseille suivant l'état courant de la navigation, et permet aux utilisateurs d'évaluer ou d'annoter les documents traversés grâce à une barre d'outils insérée dynamiquement dans les pages HTML visualisées. Le système Broadway s'appuie, entre autre sur la durée d'affichage des pages web pour en déduire l'importance. De plus, la solution proposée par ce système consiste en un ensemble de pages triées par ordre de pertinence obtenue par moyen d'une similitude calculé sur des comportements de navigation passés et ne peut donc être construite que dans le domaine de pages déjà visitées, ce qui limite considérablement les ressources du WEB. Cependant, le système BROADWAY reste indépendant du navigateur ce qui permet son utilisation sur différentes plates-formes.

RADIX (**R**echerche **A**ssistée de **D**ocuments **I**ndexés sur l'e**X**périence) (Corvaisier et al., 1997) est un système d'aide à la recherche d'information sur le Web et se base sur une description plus détaillée de navigation que le système BROADWAY. L'observation des actions des utilisateurs (telles que la sélection de pages pour le *bookmark*, l'édition de l'adresse de page, les actions de retour ou avance de pages, etc.) sont utilisées dans la représentation de la navigation de l'utilisateur et de ses composants. Pour cela, RADIX utilise un navigateur spécifique avec des fonctions adaptées à la surveillance des actions de l'utilisateur. Par conséquence, l'utilisation de ce système se limite à une plate-forme spécifique.

Le système CASEP (**CA**se-based reasoning system for **SE**quence **P**rediction) (Zehraoui et al., 2003a) (Zehraoui et al., 2003b) est utilisé pour la prédiction à partir de séquences d'actions d'un utilisateur sur un site Web. Ce système présente quelques limitations, comme par exemple la définition d'un nombre fixe d'états de la séquence courante représentant la partie problème du cas cible. Celle-ci et autres limitations ont motivé la définition du système hybride CASEP2 (Zehraoui et al., 2004) par les mêmes auteurs. Ce nouveau système a eu pour but d'apporter des améliorations au précédent système CASEP tout en prenant en compte l'aspect temporel des données ainsi que l'utilisation du système à long terme. Dans CASEP2, l'aspect temporel des données est pris en compte par la modélisation de toute séquence sous la forme d'une matrice de covariance dynamique (Zehraoui et Bennani, 2004b,a). Cette matrice prend en compte la distribution des états de la séquence dans l'espace ainsi que leur ordre temporel. La base de cas du système de RàPC est partitionnée en utilisant un réseau de neurones M-SOM-ART ayant les propriétés de stabilité et de plasticité (Zehraoui et Bennani, 2004a), importantes pour une utilisation à long terme du système. Dans ce contexte, dans (Corchado et Lees, 2001; Fdez-Riverola et Corchado, 2003), les auteurs utilisent les réseaux de neurones dans les différentes phases du cycle RàPC pour traiter des données temporelles, mais les systèmes proposés prennent en compte l'aspect temporel des données juste en définissant des fenêtres temporelles de longueurs fixées pour représenter les cas. Dans CASEP2, il n'y a pas de restriction sur la longueur de la séquence dans la représentation de cas. Le système CASEP2 consiste à prédire le comportement d'un internaute à partir de traces de navigations dans le but d'adapter de façon dynamique un site d'utilisateurs à partir de séquences de navigation sur le Web.

Le système COBRA (CBR-based **CO**llaborative **BR**owsing **A**dvisor) (Malek et Kanawati, 2001) effectue la prédiction des actions des utilisateurs sur un site Web. COBRA permet de guider un utilisateur à naviguer dans un site afin de prédire ses futures pages à visiter en réutilisant des traces de navigations passées qui sont similaires à la navigation courante. La stratégie consiste soit de modifier les liens qui relient ses différentes pages, soit de modifier ses contenus sémantiques en fonction de leur utilisation. Pour chaque page demandée le serveur Web insère

un *applet Java* invisible qui envoie des événements au serveur quand la page est visualisée par un client, même si la page est chargée à partir du cache du client. Cette technique simple permet de tracer les navigations des utilisateurs malgré l'existence d'un cache. En outre, elle permet également calculer le temps de visualisation de page sans prendre en compte le temps de transfert du fichier sur le réseau. Un des avantages de COBRA est la structure de cas et la phase de réutilisation proposées qui rendent possible la prévision d'accès aux pages qui n'ayant jamais été visitées auparavant.

Le système Letizia (Lieberman, 1995) aide l'utilisateur en déterminant ses centres d'intérêt par l'analyse des pages parcourues et en explorant, pendant le temps de lecture des pages, les pages liées qui semblent le mieux correspondre aux attentes de l'utilisateur. Letizia est un agent installé du côté client qui recherche sur le Web des pages semblables à celles que l'utilisateur a déjà visité ou mis dans le *bookmark*. La principale faiblesse de Letizia est de ne pas garder trace des sessions de recherche effectuées. Letizia n'apprend pas à partir de ses expériences et établit simplement une description d'intérêt de l'utilisateur pour la session courante en enregistrant ses actions.

Le système HYPERCASE (Micarelli et Sciarrone, 1996) s'appuie sur des recherches prototypiques effectuées par des experts du domaine pour guider la navigation des utilisateurs. Ce système se référant au RàPC ne peut pas apprendre à partir des navigations réelles d'un groupe d'utilisateurs puisque son raisonnement est uniquement basé sur des cas préenregistrés et construits par des experts. Il ne peut pas y avoir d'évolution des propositions dans le temps. L'expérience accumulée par les utilisateurs reste donc inexploitée.

4.4 D'autres approches

La problématique de l'analyse temporelle a été largement explorée sur les fichiers log Web sous l'optique de différentes techniques : logique floue (Zhou et al., 2005; Suryavanshi et al., 2005), abstractions temporelles (Moskovitch et Shahar, 2005), cartes auto-organisatrices de Kohonen (Hogo et al., 2003), arbre de décision (Hulten et al., 2001) et arbre de concepts (Acharyya et Ghosh, 2003) et graphes (Desikan et Srivastava, 2004).

Dans (Zhou et al., 2005), les auteurs proposent une approche basée sur un modèle de treillis d'usage du Web (Web Usage Lattice model) qui utilise une hiérarchie des activités d'accès du Web décrites par le moyen de la logique floue pour représenter des concepts temporels réels tels que le matin, l'après-midi, la soirée et les catégories significatives de pages Web. Dans (Suryavanshi et al., 2005) les auteurs présentent un schéma de maintenance du profil de l'utilisateur Web basé sur un algorithme de clusterisation flou (Relational Fuzzy Subtractive Clustering algorithm, RFSC) capable d'ajouter de nouvelles données d'usage à un modèle préexistant sans les coûts d'une remodélisation complète. Les auteurs définissent une mesure quantitative qui indique quand la remodélisation complète doit être exécutée afin d'éviter une dégradation du modèle. Leurs résultats montrent que la technique adoptée est presque aussi bonne que celle utilisant la remodélisation complète.

Dans (Moskovitch et Shahar, 2005), les auteurs proposent l'utilisation des abstractions temporelles définies sur des intervalles de temps. Puis, ils définissent une ontologie pour représenter la connaissance temporelle extraite à partir des données. Ils proposent d'exploiter les résultats obtenus de l'abstraction temporelle dans l'étape de prétraitement des techniques de la fouille de données afin de développer un genre d'analyse intelligente du domaine médical. La problématique de cette approche est celle de la fouille d'ordre supérieur (higher order mi-

ning). Les auteurs arguent que l'abstraction temporelle permet une réduction potentielle sur la quantité de données et de bruit.

Dans (Hogo et al., 2003), les auteurs proposent une étude de cas de la fouille de données temporelles d'usage sur un site Web éducatif à travers l'application des cartes auto-organisatrices de Kohonen adaptées et basées sur les propriétés d'ensembles rudes (rough set).

Dans (Hulten et al., 2001), les auteurs présentent un algorithme basé sur VFDT (Very Fast Decision Tree learner) pour la construction incrémentale d'un arbre de décision pour la fouille sur les flux de données à travers un apprentissage défini sur des fenêtres glissantes. Les auteurs de (Acharyya et Ghosh, 2003) ajoutent à l'analyse statistique classique des logs une information de "concept" rattachée à chaque page. L'objectif est de prendre en compte un "changement de centre d'intérêt" de l'utilisateur au cours de la session, de pré-segmenter les sessions sur la base des contenus visités et de mieux prédire les liens qui seront suivis en fonction de la position dans un arbre de concepts. Les auteurs attestent une augmentation significative du taux de prédiction en ayant recours à cette méthode.

Concernant l'aspect temporel d'évolution dans la fouille d'usage du Web, les auteurs de (Desikan et Srivastava, 2004) construisent des graphes d'utilisation d'un site Web à partir de l'analyse des fichiers log du centre d'informatique de l'université de Minnesota afin d'étudier leur évolution sur le temps à travers de la fouille des sous-graphes séquentiels.

5 Problèmes ouverts et propositions

La croissance explosive de l'utilisation d'Internet a induit une croissance correspondante des fichiers log Web. De plus, les sites Web deviennent de plus en plus grands parce que plus de pages Web sont ajoutées que supprimées. Les propriétaires de sites Web optent pour garder les vieilles pages Web parce que le stockage de l'information est relativement peu coûteux de nos jours et aussi parce qu'ils savent que leurs clients marquent les pages Web dans leurs *bookmarks* et comptent toujours les trouver en ligne, ce qui est tout à fait compréhensible dans le cadre de concurrence commerciale. Le stockage des données requiert alors une quantité considérable de mémoire. Cependant, il existe des stratégies qui peuvent résoudre ce problème, par exemple, il y a des moyens de faire un changement automatique de pages lors qu'un client demande une adresse obsolète. Ceci pourrait d'une certaine façon éviter les informations redondantes et minimiser le volume de pages hébergées par les sites Web. Contradictoirement à la quantité colossale des données mises en ligne sur l'Internet, l'une des difficultés le plus importantes liées à la fouille d'usage du Web est la pénurie (voir inexistence) de *benchmarks* de données d'usage du Web pour l'application et comparaison de différentes techniques. Cela est dû au fait que les données d'usage contiennent des informations privées souvent maintenues sous la propriété d'entreprises du commerce électronique.

Le défi que tente de relever la fouille des données temporelles est de pouvoir prendre en compte la totalité des informations disponibles concernant les entités qui constituent les données de nature temporelle et les données de nature non temporelle. Cela suppose d'être capable d'intégrer des informations de nature différente et de pouvoir ainsi les rattacher à une même catégorie sémantique. Sur le plan méthodologique, il s'agit donc de définir une mesure de ressemblance ou de dissemblance entre deux objets dont la description est fournie par un ensemble complexe d'informations. Jusqu'à présent, les approches proposées ont surtout été fondées sur la juxtaposition de similarités ou de dissimilarités partielles. Il serait donc intéressant d'avoir

une mesure de distance, à la fois capable de prendre en compte l'aspect temporel et de traiter les données classiques (non temporelles).

Un problème qui n'a pas beaucoup été abordé dans la fouille de données temporelles est le traitement différencié pour des événements ayant des durées différentes dans une séquence. Quand différents événements ont différentes durées, il est souhaitable de prolonger le cadre de base des modèles temporels pour définir des structures qui rendent possible des traitements spéciaux à ces évènements et pas seulement la pondération en fonction de leur durée. Cependant, en ce qui concerne l'analyse de données temporelles d'usage du Web, le temps consacré par chaque utilisateur à visiter une page Web ne constitue pas spécialement un facteur décisif pour la caractérisation des pages d'un site Web, car chaque utilisateur a son propre rythme de navigation. On peut cependant imaginer la prise en compte des durées relatives des consultations des pages Web.

Concernant à la fouille des flux de données, l'extraction de motifs fréquents se conçoit sur un ensemble fini de résultats possibles (l'ensemble de combinaisons entre les items enregistrés dans les données) et ceci n'est pas le cas pour les motifs séquentiels où l'ensemble de résultats est infini. En fait, en raison de l'aspect temporel des motifs séquentiels, un item peut être répété sans limites menant à un nombre très grand de séquences potentiellement fréquentes. Cependant, la fouille de motifs séquentiels se concentre sur les motifs qui semblent avoir des trajectoires communes. Dans ce contexte, la fouille d'ordre supérieur (Higher Order Mining) dans laquelle la fouille est appliquée aux règles précédemment extraites, est un secteur prometteur car il permet de réduire les coûts généraux de la fouille de données (Roddick et Spiliopoulou, 2002).

La plupart des méthodes consacrées à la fouille de données d'usage du Web prennent en compte, dans leur analyse des données, toute la période qui enregistre les traces d'usage d'un site Web : les résultats obtenus sont donc naturellement ceux qui prédominent sur la totalité de la période. En conséquence, les modèles de comportements qui ont lieu pendant de courtes sous-périodes ne sont pas pris en compte et restent donc ignorés par les méthodes classiques. Touchant à l'extraction de motifs séquentiels, les auteurs de (Masseglia et al., 2004) proposent une méthode de division récursive pour la découverte de motifs séquentiels capables de mettre en évidence l'existence des comportements parfois minoritaires mais pourtant intéressants qui sont présents dans les fichiers log Web. Cette technique repose sur des résumés des motifs et les méthodes neuronales.

Ainsi, il est désirable d'avoir une sous-division de la période totale analysée en sous périodes plus significatives, soit divisée en intervalles de temps réguliers et uniformes, soit divisée de façon adaptée par la méthode en sous-périodes qui mettent en évidence l'occurrence de comportements significatifs. Dans ce contexte, une étude réalisé par les auteurs de (Da Silva et al., 2007) met en évidence la nécessité de partitionnement de la période de temps en sous-périodes. Leur analyse montrent que les résultats obtenus sont vraisemblablement différents si on effectue des analyses sur des sous-périodes de temps au lieu de considérer la période toute entière. Il est donc nécessaire d'analyser le changement d'intérêt des utilisateurs d'un site Web sur ces sous-périodes de temps et mettre à jour les profils de ses utilisateurs afin de pouvoir caractériser le vrai intérêt de ceux-ci en définissant une méthodologie de modélisation d'évolution de ces profils sur le temps. En outre, dû au volume des données existantes sur les fichiers log, il semble intéressant d'utiliser une structure représentative de données afin de résumer l'information.

Cette remarque reste valable pour le domaine des systèmes de recommandation si on postule que le goût d'un utilisateur peut varier selon le temps. Supposons par exemple, un groupe d'utilisateurs qui possèdent une préférence pour les films d'horreur en général mais qui, pendant l'hiver, optent pour des comédies. Le système doit donc être capable de détecter le changement d'intérêt d'un utilisateur et être suffisamment habile pour ajuster les recommandations aux besoins de ses utilisateurs. Dans le même contexte, la vulnérabilité d'un système de recommandation constitue une des issues débutantes dans ce secteur. Quelques commerçants sans scrupules et toujours cherchant à pénétrer dans le marché profitent de cette faiblesse pour "tromper" le système afin d'avoir leurs produits recommandés plus souvent que ceux de leurs concurrents. Cette problématique est traitée sous une nouvelle nomenclature : *shilling attacks* (Lam et Riedl, 2004).

6 Conclusions

Dans cette étude, nous avons abordé la problématique du traitement des données temporelles dans le contexte de l'analyse de l'usage du Web. Les questions abordées ont montré la nécessité de définition et/ou d'adaptation de méthodes capables d'extraire des connaissances et de suivre l'évolution de ce type de données. Bien qu'il existe de nombreuses méthodes performantes d'extraction de connaissances, peu de travaux ont été consacrés à la problématique de données temporaires et évolutives. Afin d'explorer cette problématique, une étude bibliographique a été consacrée aux concepts et techniques utilisés pour traiter les données temporaires, plus précisément dans le contexte des données qui représentent les traces d'usage du Web.

Finalement, malgré toutes adaptations et améliorations possibles pour les méthodes d'analyse de données, il est important de noter que les analyses appliquées dans un contexte dynamique doivent être toujours mises à jour. Sur les moyen et long termes, la nature dynamique du Web induit nécessairement une évolution des pratiques. De ce fait, les résultats obtenus sur des données d'usage deviennent progressivement obsolètes au fur et à mesure qu'on s'éloigne de la période d'enregistrement des données.

Remerciements

Je tiens à remercier l'INRIA-France et la CAPES-Brésil pour leur soutien. Mes remerciements se tournent aussi vers Monsieur Yves Lechevallier et Monsieur Fabrice Rossi pour leur aide qui a précieusement contribué à l'aboutissement de ce travail de recherche.

Références

Aamodt, A. et E. Plaza (1994). Case-based reasoning : foundational issues, methodological variations, and system approaches. *AI Communications 7*(1), 39–59.

Abrams, M., C. R. Standridge, G. Abdulla, S. Williams, et E. A. Fox (1995). Caching proxies : Limitations and potentials. Technical report, Blacksburg, VA, USA.

Acharyya, S. et J. Ghosh (2003). Context-sensitive modeling of web-surfing behaviour using concept trees. In *Proceedings of the Fifth WEBKDD workshop : Webmining as a Premise to Effective and Intelligent Web Applications (WEBKDD 2003)*, Washington, USA.

Agrawal, R., T. Imielinski, et A. Swami (1993). Mining association rules between sets of items in large databases. In P. Buneman et S. Jajodia (Eds.), *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, Volume 22, Washington, D.C., pp. 207–216. ACM Press.

Agrawal, R. et R. Srikant (1995). Mining sequential patterns. *11th International Conference on Data Engineering (ICDE'95)*, 3–14.

Baldi, P., P. Frasconi, et P. Smyth (2003). *Modeling the Internet and the Web : Probabilistic Methods and Algorithms*. Wiley.

Banerjee, A. et J. Ghosh (2001). Clickstream clustering using weighted longest common subsequences. In *Workshop on Web Mining : 1st SIAM Conference on Data Mining*, pp. 33–40.

Baron, S. et M. Spiliopoulou (2001). Monitoring change in mining results. In Y. Kambayashi, W. Winiwarter, et M. Arikawa (Eds.), *Data Warehousing and Knowledge Discovery (DaWaK)*, Volume 2114 of *Lecture Notes in Computer Science*, pp. 51–60. Springer.

Baron, S. et M. Spiliopoulou (2003). Monitoring the evolution of web usage patterns. In B. Berendt, A. Hotho, D. Mladenic, M. van Someren, M. Spiliopoulou, et G. Stumme (Eds.), *European Web Mining Forum (EWMF)*, Volume 3209 of *Lecture Notes in Computer Science*, pp. 181–200. Springer.

Berendt, B., B. Mobasher, M. Spiliopoulou, et M. Nakagawa (2002). The impact of site structure and user environment on session reconstruction in web usage analysis. In *Proceedings of the 4th WebKDD 2002 Workshop, at the ACM-SIGKDD Conference on Knowledge Discovery in Databases (KDD'2002)*, Edmonton, Alberta, Canada.

Box, G. E. P., G. M. Jenkins, et G. C. Reinsel (1994). *Time Series Analysis : Forecasting and Control* (Third ed.). Prentice Hall.

Cadez, I., D. Heckerman, C. Meek, P. Smyth, et S. White (2003). Model-based clustering and visualization of navigation patterns on a web site. *Data Mining and Knowledge Discovery 7*(4), 399–424.

Catledge, L. D. et J. E. Pitkow (1995). Characterizing browsing strategies in the world-wide web. *Computer Networks and ISDN Systems 27*(6), 1065–1073.

Chakrabarti, S., S. Sarawagi, et B. Dom (1998). Mining surprising patterns using temporal description length. In A. Gupta, O. Shmueli, et J. Widom (Eds.), *24th International Conference on Very Large databases (VLDB'98)*, pp. 606–617. Morgan Kaufmann.

Chen, X. et I. Petrounias (1999). Mining temporal features in association rules. *PKDD'99 : Proceedings of the Third European Conference on Principles of Data Mining and Knowledge Discovery*, 295–300.

Chi, E. H., A. Rosien, et J. Heer (2002). Lumberjack : Intelligent discovery and analysis of web user traffic composition. In *Proceedings of the 4th WebKDD Workshop, at the ACM-SIGKDD Conference on Knowledge Discovery in Databases (KDD'2002)*, pp. 1–16.

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

Corchado, J. M. et B. Lees (2001). Adaptation of cases for case based forecasting with neural network support. In *Soft computing in case based reasoning*, London, UK, pp. 293–319. Springer-Verlag.

Corvaisier, F., A. Mille, et J. M. Pinon (1997). Information retrieval on the world wide web using a decision making system. In *Proceedings of the Computer-Assisted Searching on the Internet (RIAO 97)*, Montreal, Canada, pp. 284–295.

Da Silva, A., F. D. Carvalho, Y. Lechevallier, et B. Trousse (2006). Mining web usage data for discovering navigation clusters. *11th IEEE Symposium on Computers and Communications (ISCC 2006)*, 910–915.

Da Silva, A., Y. Lechevallier, F. Rossi, et F. De Carvalho (2007). Analyse de résumés des données évolutives dans le web usage mining. *Actes des 7ème journées d'Extraction et Gestion des Connaissances (EGC 2007) 2*, 539–544.

Dempster, A. P., N. M. Laird, et D. Rubin (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society 1*(39), 1–38.

Desikan, P. et J. Srivastava (2004). Mining temporally evolving graphs. *In Proceedings of sixth WebKDD workshop : Web Mining and Web Usage Analys, in conjunction with the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2004)*, 13–22.

Ester, M., H.-P. Kriegel, J. Sander, M. Wimmer, et X. Xu (1998). Incremental clustering for mining in a data warehousing environment. *Proc. 24th Int. Conf. Very Large Data Bases (VLDB)*, 323–333.

Ester, M., H.-P. Kriegel, J. Sander, et X. Xu (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In E. Simoudis, J. Han, et U. Fayyad (Eds.), *2ndInternational Conference on Knowledge Discovery and Data Mining (KDD-96)*, Portland, Oregon, pp. 226–231. AAAI Press.

Fayyad, U. M., G. Piatetsky-Shapiro, P. Smyth, et R. Uthurusamy (1996). *Advances in Knowledge Discovery and Data Mining*. AAAI Press / The MIT Press.

Fdez-Riverola, F. et J. M. Corchado (2003). Using instance-based reasoning systems for changing environments forecasting. In *Workshop on Applying case-based reasoning to time series prediction in conjunction with ICCBR 2003*, Trondheim, Norway, pp. 219–228.

Fraley, C. et A. Raftery (1998). How many clusters ? which clustering method ? answers via model based cluster analysis. *Computer Journal 41*, 578–588.

Fu, Y., K. Sandhu, et M. Shih (2000). Clustering of web users based on access patterns. In B. Masand et M. Spiliopoulou (Eds.), *Web Usage Analysis and User Profiling*, Berlin, pp. 21–38. Springer-Verlag.

Hogo, M., M. Snorek, et P. Lingras (2003). Temporal web usage mining. In *Proceedings of the IEEE/WIC International Conference on Web Intelligence*, Washington, DC, USA, pp. 450–453. IEEE Computer Society.

Hulten, G., L. Spencer, et P. Domingos (2001). Mining time-changing data streams. In *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, CA, pp. 97–106. ACM Press.

Jaczynski, M. et B. Trousse (1999). Broadway : A case-based system for cooperative infor-

mation browsing on the world-wide-web. In *Collaboration between Human and Artificial Societies, Coordination and Agent-Based Distributed Computing*, London, UK, pp. 264–283. Springer-Verlag.

Kolodner, J. (1993). *Case-Based Reasoning*. 2929 Campus Drive, suite260, SanMateo, CA, USA : Morgan Kaufmann Publishers, Inc.

Kosala, R. et H. Blockeel (2000). Web mining research: A survey. *ACM SIGKDD Explorations: Newsletter of the Special Interest Group on Knowledge Discovery and Data Mining 2*, 1–15.

Kothari, R., P. A. Mittal, V. Jain, et M. K. Mohania (2003). On using page cooccurrences for computing clickstream similarity. In D. Barbará et C. Kamath (Eds.), *SDM*. San Francisco, USA: SIAM (Society for Industrial and Applied Mathematics).

Lam, S. K. et J. Riedl (2004). Shilling recommender systems for fun and profit. In S. I. Feldman, M. Uretsky, M. Najork, et C. E. Wills (Eds.), *WWW*, pp. 393–402. ACM.

Laxman, S. et P. S. Sastry (2006). A survey of temporal data mining. *SADHANA - Academy Proceedings in Engineering Sciences, Indian Academy of Sciences 31*(2), 173–198.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady 10*(8), 707–710.

Licoppe, C., A. S. Pharabod, et H. Assadi (2002). Contribution à une sociologie des échanges marchands sur internet. In *Réseaux*, Volume 20, pp. 97–140.

Lieberman, H. (1995). Letizia: An agent that assists web browsing. In C. S. Mellish (Ed.), *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence (IJCAI-95)*, Montreal, Quebec, Canada, pp. 924–929. Morgan Kaufmann publishers Inc.: San Mateo, CA, USA.

Liu, B., Y. Ma, et R. Lee (2001). Analyzing the interestingness of association rules from the temporal dimension. *ICDM '01: Proceedings of the 2001 IEEE International Conference on Data Mining*, 377–384.

Malek, M. et R. Kanawati (2001). Cobra: A cbr-based approach for predicting users actions in a web site. *Case-Based Reasoning Research and Development : 4th International Conference on Case-Based Reasoning (ICCBR 2001)*, 336–346.

Marascu, A. et F. Masseglia (2006). Extraction de motifs séquentiels dans les flots de données d'usage du web. In G. Ritschard et C. Djeraba (Eds.), *Actes des sixièmes journées Extraction et Gestion des Connaissances (EGC 2006)*, Revue des Nouvelles Technologies de l'Information (RNTI-E-6), Lille, France, pp. 627–638. Cépaduès-Éditions.

Masseglia, F., P. Poncelet, et R. Cicchetti (1999). An efficient algorithm for web usage mining. *Networking and Information Systems Journal (NIS) 2*, 571–603.

Masseglia, F., D. Tanasa, et B. Trousse (2004). Diviser pour découvrir. une méthode d'analyse du comportement de tous les utilisateurs d'un site web. *Ingénierie des Systèmes d'Information 9*(1), 61–83.

Micarelli, A. et F. Sciarrone (1996). A case-based system for adaptive hypermedia navigation. *Advances in Case-Based Reasoning, Proc. of the 3rd European Workshop on Case-Based Reasoning (EWCBRŠ96) 1168*, 266–279.

Mobasher, B. (2004). Web usage mining and personalization. In *The Practical Handbook of Internet Computing*, Florida, USA. Chapman Hall & CRC Press.

Moskovitch, R. et Y. Shahar (2005). Temporal data mining based on temporal abstractions. *Proceedings of the 2nd Workshop on Temporal Data Mining (TDM 2005), held in conjunction with the 5th IEEE International Conference on Data Mining (ICDM'05)*.

O'Shaughnessy, D. (2003). *Speech communications: Human and machine*. Piscataway, NJ: IEEE Press.

Padmanabhan, B., Z. Zheng, et S. O. Kimbrough (2001). Personalization from incomplete data: what you don't know can hurt. In *Knowledge Discovery and Data Mining*, pp. 154–163.

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*(4), 311–372.

Rabiner, L. et B. Juang (1993). *Fundamentals of Speech Recognition*. Prentice Hall Signal Processing Series.

Roddick, J. F. et M. Spiliopoulou (2002). A survey of temporal knowledge discovery paradigms and methods. *IEEE Transactions on Knowledge and Data Engineering 14*(4), 750–767.

Rossi, F., F. De Carvalho, Y. Lechevallier, et A. Da Silva (2006a). Comparaison de dissimilarités pour l'analyse de l'usage d'un site web. *Actes des 6ème journées Extraction et Gestion des Connaissances (EGC 2006), Revue des Nouvelles Technologies de l'Information (RNTI-E-6) II*, 409–414.

Rossi, F., F. De Carvalho, Y. Lechevallier, et A. Da Silva (2006b). Dissimilarities for web usage mining. *Actes des 10ème Conférence de la Fédération Internationale des Sociétés de Classification (IFCS 2006)*, 39–46.

Samia, M. (2003). Temporal web mining. In H. Höpfner, G. Saake, et E. Schallehn (Eds.), *Grundlagen von Datenbanken*, pp. 27–31. Fakultät für Informatik, Universität Magdeburg.

Samia, M. et S. Conrad (2004). From temporal data mining and web mining to temporal web mining. *Sixth International Baltic Conference on Databases and Information Systems (BalticDB&IS'2004)*, 232–245.

Sankoff, D. et J. B. Kruskal (1983). *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison*. Addison-Wesley.

Sow, D. M., D. P. Olshefski, M. Beigi, et G. Banavar (2003). Prefetching based on web usage mining. In M. Endler et D. C. Schmidt (Eds.), *Middleware*, Volume 2672 of *Lecture Notes in Computer Science*, pp. 262–281. Springer.

Spiliopoulou, M. (1999). Data mining for the web. *Workshop on Machine Learning in User Modelling of the ACAI99*, 588–589.

Srivastava, J., R. Cooley, M. Deshpande, et P.-N. Tan (2000). Web usage mining: Discovery and applications of usage patterns from web data. *SIGKDD Explorations 1*(2), 12–23.

Suryavanshi, B. S., N. Shiri, et S. P. Mudur (2005). Incremental relational fuzzy subtractive clustering for dynamic web usage profiling. *In Proceedings of WebKDD Workshop on Taming Evolving, Expanding and Multi-faceted Web Clickstreams held in conjunction with the*

11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2005).

W3C (July 1995). Logging control in w3c httpd. *http://www.w3.org/Daemon/User/Config/Logging.html*.

Yan, T. W., M. Jacobsen, H. Garcia-Molina, et U. Dayal (1996). From user access patterns to dynamic hypertext linking. In *Proceedings of the fifth international World Wide Web conference on Computer networks and ISDN systems*, Amsterdam, The Netherlands, The Netherlands, pp. 1007–1014. Elsevier Science Publishers B. V.

Zehraoui, F. et Y. Bennani (2004a). M-som: matricial self organizing map for sequence clustering and classification. *IEEE International Joint Conference on Neural Networks (IJCNN 2004) 1*, 763–768.

Zehraoui, F. et Y. Bennani (2004b). Som-art : Incorporation des propriétés de plasticité et de stabilité dans une carte auto-organisatrice. *Atelier: Fouille de données complexes dans un processus d'extraction des connaissances (FDC) des 4ème journées Extraction et Gestion des Connaissances (EGC 2004)*, 169–180.

Zehraoui, F., R. Kanawati, et S. Salotti (2003a). Case base maintenance for improving prediction quality. *Case-Based Reasoning Research and Development 2689*, 703–717.

Zehraoui, F., R. Kanawati, et S. Salotti (2003b). Cbr system for sequence prediction: Casep. In *Proceedings of the International workshop on applying case-based reasoning on time series prediction*, Trondhiem, Norway, pp. 260–269.

Zehraoui, F., R. Kanawati, et S. Salotti (2004). Casep2: Hybrid case-based reasoning system for sequence processing. *Advances in Case-Based Reasoning 3155*, 449–463.

Zhangn, T., R. Ramakrishnan, et M. Livny (1996). Birch: An efficient data clustering method for very large databases. In H. V. Jagadish et I. S. Mumick (Eds.), *SIGMOD Conference*, pp. 103–114. Montreal, Quebec, Canada: ACM Press.

Zhou, B., S. C. Hui, et A. C. M. Fong (2005). Discovering and visualizing temporal-based web access behavior. In *Proceedings of the the 2005 IEEE/WIC/ACM International Conference on Web Intelligence (WI'05)*, pp. 297–300. Washington, DC, USA: IEEE Computer Society.

Summary

The number of businesses negotiations put on-line in the Internet is always increasing. The Web has become one of the most important businesses platforms and many Web site operators are incited to analyze the use of their sites to improve their profitability. However, the way in which a website is visited can indeed change, either due to modifications concerning the structure and the content of the site itself or due to the evolution of the behaviour of certain user groups. For consequence, the usage patterns must be updated continuously in order to reflect the actual behaviour of the visitors. In this context, we present a no-exhaustive view of the techniques applied on data mining with emphasis on methods for discovering patterns in web usage data. In addition, we also discuss some challenges tasks regarding pattern discovery in web usage evolving data.