Vers une Ingénierie Ontologique à Base du Web Usage Mining

GRABA Abdelmadjid Guessoum, ELBERRICHI Zakaria

Laboratoire EEDIS, Université Djilali Liabes de Sidi Bel Abbes, ALGERIE

{majid.guesse, elberrichi}@gmail.com

Résumé. Récemment, de nouvelles approches ont intégré l'utilisation de techniques de fouille de données dans le processus d'enrichissement d'ontologies. En effet, les deux domaines, fouille de données et méta-données ontologiques sont extrêmement liés : d'une part les techniques de fouille de donnée aident à la construction du Web sémantique, d'autre part le Web sémantique aide à l'extraction de nouvelles connaissances. Ainsi, beaucoup de travaux utilisent les ontologies comme un guide pour l'extraction de règles ou de motifs, permettant de discriminer les données par leur valeur sémantique et donc d'extraire des connaissances plus pertinentes. Il s'avère à l'inverse que peu de travaux visant à mettre à jour l'ontologie s'intéressent aux techniques de fouilles de données. Dans ce papier, nous présentons une approche pour supporter la gestion des ontologies des sites Web basée sur l'utilisation des techniques de Web Usage Mining. L'approche présentée a été expérimentée et évaluée sur une ontologie d'un site Web, que nous avons construit et ensuite enrichie en se basant sur les motifs séquentiels extraits sur le Log.

Mots-clés: ontologies, Web Usage Mining, enrichissement, Web sémantique, fouille de données, motifs séquentiels.

1 Introduction

Le Web sémantique désigne un espace d'échange et de manipulation de grandes sources de données visant à rendre le contenu des pages Web accessibles aux humains et aux agents artificiels. En effet, la recherche d'information dans le Web classique se base essentiellement sur la structure des documents, ce qui rend l'exploitation du contenu quasiment impossible par les machines. A la différence de cela, dans le Web sémantique, les machines accèdent aux ressources grâce à la représentation sémantique du contenu. Cette représentation inclut une formalisation du contenu, permettant d'encoder l'information dans un format lisible par la machine, ainsi que l'ajout de métadonnées sémantiques modélisant l'information disponible. La combinaison des données formalisées et de la couche sémantique donne alors accès à la connaissance et ouvre la voie à un large panel d'applications.

Il est nécessaire d'utiliser un moyen d'échange commun afin de partager l'information entre différentes communautés. Les ontologies sont l'un des modèles de représentation de connaissances les plus avancés. Constituées de concepts liés par des relations, et souvent structurés hiérarchiquement, elles permettent d'organiser des connaissances en fonction du domaine considéré. Au cœur du Web sémantique, elles ajoutent une couche sémantique au Web classique en décrivant les connaissances contenues dans les ressources. Considérées désormais dans ce domaine comme métadonnées de référence, les ontologies, ainsi que leur création et leur développement, font l'objet de nombreux travaux de recherche. En particulier, l'évolution permanente des ressources, nécessite la mise au point de techniques permettant l'évolution des ontologies et leurs mises à jours.

Nous proposons donc dans cet article une approche d'enrichissement d'ontologies basée sur une technique de fouille de données, la recherche de motifs séquentiels à l'aide de l'algorithme Vpsp. Appliquée sur des fichiers logs, cet algorithme permet de mettre en évidence des séquences fréquentes d'accès Web, dans un ordre donné. Notre approche consiste ainsi, à partir de fichiers logs, à extraire des motifs séquentiels qui sont ensuite utilisés afin d'enrichir l'ontologie, en y ajoutant d'une part de nouveaux concepts, d'autre part, les relations sémantiques qui peuvent exister entre eux.

La suite de cet article est organisée de la manière suivante : dans la section 2 nous présentons les étapes de notre approche. La section 3 présente les résultats d'expérimentations conduites sur la mise à jour de l'ontologie du site Web étudié. Nous terminons l'article par une discussion de travaux connexes et une conclusion.

2 Approche proposée

Cette approche est divisée en quatre étapes :

1- Construction d'une ontologie pour le site Web étudié.
2- Prétraitement des fichiers Logs.
3- Application de l'algorithme Vpsp sur le fichier Log.
4- Enrichissement de l'ontologie de base à l'égard des visites des utilisateurs.

2.1 Construction de l'ontologie

Les ontologies sont l'un des modèles de représentation de connaissances les plus avancés. Constituées de concepts liés par des relations, et souvent structurés hiérarchiquement, elles permettent d'organiser des connaissances en fonction du domaine considéré.

La construction de l'ontologie peut être effectuée manuellement ou semi-automatiquement. Dans le premier cas, cette tâche est difficile et prend du temps. C'est la raison pour laquelle de nombreuses méthodes et méthodologies ont été conçues pour semi-automatiser ce processus. Les sources de données peuvent être du texte, des données semi-structurées, les données relationnelles, etc. Dans la suite, nous décrivons certaines méthodes dédiées à l'extraction des connaissances à partir des pages Web.

L'approche proposée par Navigli et Velardi [9] tente de réduire la confusion conceptuelle et terminologique entre les membres d'une communauté virtuelle. Les concepts et les relations sont tirés d'une série de sites Web en utilisant l'outil Ontolearn. Les principales étapes sont: l'extraction de terminologie à partir des sites Web, interprétation sémantique des termes, et l'identification des relations taxonomiques.

Certaines approches transforment les pages html en une hiérarchie sémantique structurée codées en XML, en tenant compte des régularités HTML [3].

Enfin, on peut aussi remarquer que certaines approches consacrées à la construction de l'ontologie à partir des pages Web sans l'utilisation d'aucune connaissance a priori. L'approche décrite dans [10] est basé sur les étapes suivantes: (1) extraire des mots clés représentatifs du domaine, (2) trouver une collection de sites Web liés aux anciens mots clé (en utilisant par exemple Google), (3) analyse exhaustive de chaque site, (4) l'analyseur recherche les mots clés initiales dans un site Web et trouve les mots suivants et précédents; ces mots sont candidats à être des concepts, (5) pour chaque concept sélectionné, une analyse statistique est effectuée sur la base du nombre d'occurrences de ce mot dans les sites Web et, finalement, (6) pour chaque concept extrait en utilisant une fenêtre autour du mot clé initial, un nouveau mot clé est défini et l'algorithme réitère récursivement.

Dans [5] une méthode est proposée pour extraire l'ontologie de domaine à partir des sites Web sans l'utilisation de connaissances a priori. Cette approche prend la structure des pages Web en considération et définit une hiérarchie contextuelle. Le prétraitement des données est une étape importante pour définir les termes les plus pertinents pour pouvoir être classés. Des poids sont associés aux termes en fonction de leur position dans cette hiérarchie conceptuelle. Ensuite, ces termes sont automatiquement classés et les concepts sont extraits.

Dans [2] les auteurs définissent une architecture ontologique basée sur un triplet sémantique, à savoir: sémantique du contenu, la structure et les services d'un domaine.

2.2 Prétraitement des fichiers Logs

La première étape d'un processus du WUM consiste en un prétraitement des fichiers Log. En effet, le format des fichiers log web est impropre à une analyse directe par les diverses techniques de fouille des données. Leur nettoyage et leur structuration sont donc nécessaires avant toute analyse.

La première étape d'un processus WUM se compose principalement de deux types de taches :

- Tâches classiques de prétraitement : fusion des fichiers logs web, nettoyage et structuration de données.
- Tâches avancées de prétraitement : stockage des données structurées dans une base de données, généralisation et agrégation des données.

2.2.1 Prétraitement classique

Fusionner tous les fichiers log dans un seul fichier log afin de pouvoir reconstruire les sessions réalisées à travers plusieurs serveurs Web ;

Nettoyer le fichier log :

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

- Les requêtes non valides. Ce sont les requêtes dont le statut est inférieur à 200 ou supérieur à 399. En effet, le code d'état (statut), entier codé sur trois chiffres, a un sens propre dont la catégorie dépend du premier chiffre:
  - 1xx indique uniquement un message informel,
  - 2xx indique un succès,
  - 3xx redirige le client sur un autre URL,
  - 4xx indique une erreur côté client,
  - 5xx indique une erreur côté serveur.

- Requêtes provenant des robots Web. Il est presque impossible aujourd'hui d'identifier tous les robots Web puisque chaque jour apparaissent des nouveaux. Pour les robots dont l'adresse IP et le User-Agent sont inconnus, nous procédons à un examen de leurs comportements sachant que les robots Web procèdent à une visite relativement exhaustive (nombre de pages visitées par un robot est supérieur au nombre de pages visitées par un utilisateur normal) et rapide et qu'ils cherchent généralement un fichier nommé «\robot.txt». Ainsi, pour identifier les requêtes provenant des robots ou leurs visites nous avons utilisé des heuristiques en considérant qu'il suffit de vérifier une d'entre elles pour considérer la requête correspondante comme étant générée par un robot Web :
  a. Identifier les adresses IP connus comme étant des robots Web. Ces informations sont fournies généralement par les moteurs de recherche.
  b. Identifier les IP ayant fait une requête à la page « \robots.txt».
  c. Utiliser un seuil pour la vitesse de navigation BS «Browsing Speed» égale au nombre de pages visitées par seconde. Le calcul du Browsing Speed n'est possible qu'après détermination des sessions et des visites.

- Requêtes aux images. Cette étape de nettoyage consiste à supprimer les fichiers dont les extensions sont : .jpg, .gif, .png, etc... et les fichiers multimédia dont l'extension est : .wav, .wma, .wmv, etc.

- Requêtes dont la méthode est différente de «GET». Les méthodes généralement utilisées sont: GET, HEAD, PUT, POST, TRACE et OPTIONS :
  - La méthode GET est une requête d'information. Le serveur traite la demande et renvoie le contenu de l'objet.
  - La méthode HEAD est très similaire à la méthode GET. Cependant le serveur ne retourne que l'en-tête de la ressource demandée sans les données. Il n'y a donc pas de corps de message.
  - La méthode PUT permet de télécharger un document, dont le nom est précisé dans l'URI, ou d'effacer un document, toujours si le serveur l'autorise.
  - La méthode POST est utilisée pour envoyer des données au serveur.

- La méthode TRACE est employée pour le débogage. Le serveur renvoie, dans le corps de la réponse, le contenu exact qu'il a reçu du client. Ceci permet de comprendre, en particulier, ce qui se passe lorsque la requête transite par plusieurs serveurs intermédiaires.
- La méthode OPTIONS permet de demander au serveur les méthodes autorisées pour le document référencé.

Vu que le WUM s'intéresse à l'étude du comportement de l'internaute sur le Web et par conséquent aux ressources qu'il demande, il faut garder seulement les requêtes dont la méthode utilisée est GET.

- Les Scripts. Généralement, le téléchargement d'une page demandée par un utilisateur est accompagné automatiquement par le téléchargement des scripts tels que les scripts Java (fichiers .js), des feuilles de style (fichiers .css), des animations flash (fichier .swf), etc. Ces éléments doivent être supprimés du fichier Log étant donné que leur apparition ne reflète pas le comportement de l'internaute.

Structurer le fichier log :

La structuration des données consiste à identifier les utilisateurs, les sessions et les visites :

Identification des utilisateurs et des sessions : Une session est composée de l'ensemble de pages visitées par le même utilisateur durant la période d'analyse. Plusieurs moyens d'indentification des utilisateurs ont été proposés dans la littérature (login et mot de passe, cookie, IP), cependant, tous ces moyens présentent des défaillances à cause des systèmes de cache, des firewalls et des serveurs proxy. Dans notre cas, nous considérons que deux requêtes provenant de deux adresse IP différents, appartiennent à deux sessions différentes donc elles sont effectuées par deux utilisateurs différents. Toutefois, nous ne pouvons nier la limite inhérente à cette méthode. En effet, une confusion entre deux utilisateurs différents utilisant la même adresse IP est toujours possible surtout en cas d'utilisation d'un serveur Proxy ou d'un firewall.

Identification des visites : Une visite est composée d'une série de requêtes séquentiellement ordonnées, effectuées pendant la même session et ne présentant pas de rupture de séquence de plus de 30 minutes. L'identification des visites sur le site, est effectuée selon la démarche suivante:

Déterminer la durée de consultation des pages. La durée de consultation d'une page est le temps séparant deux requêtes successives. Si la durée de consultation d'une page dépasse 30 minutes alors la page suivante dans la même session est attribuée à une nouvelle visite.

Une fois les visites identifiées, la durée de consultation de la dernière page de chaque visite est obtenue à partir de la moyenne des temps de consultation des pages précédentes appartenant à la même visite.

Les visites composées d'une seule requête ne sont pas considérées car, d'une part, il n'est pas possible d'estimer la durée d'une seule requête, d'autre part, elles sont éliminées dans la phase de retraitement puisqu'elles ne présentent aucun intérêt pour notre analyse.

2.2.2 Prétraitement avancé

Dans cette étape, les données structurées sont enregistrées sous une forme persistante, généralement, dans une base de données. On produit souvent des variables dérivées à partir des premières (requête, utilisateur, session ou visite).

Afin de pouvoir traiter l'information contenue dans la base le plus simplement et le plus efficacement possible, il faut restructurer la base selon le schéma relationnel.

2.3 Application de l'algorithme Vpsp sur le fichier Log

Introduits dans [1] et largement étudiés dans [7], les motifs séquentiels peuvent être vus comme une extension de la notion de règles d'association intégrant diverses contraintes temporelles. La recherche de tels motifs consiste ainsi à extraire des enchaînements d'ensembles d'items, couramment associés sur une période de temps bien spécifiée.

2.3.1 L'approche Vpsp (Vertical Prefix-Tree for Sequential Pattern)

L'algorithme Vpsp [4] est un algorithme de type "générer-élaguer" et combine les avantages principaux des algorithmes PSP [6] et SPADE [13] grâce à une nouvelle structure de données en forme d'arbre préfixée (héritée de PSP) couplée à un chargement de la base de données en représentation verticale. Cette union de structure d'arbre et de transformation de bases de données permet :

1. d'optimiser l'espace mémoire utilisé. La différence principale, distinguant notre structure de celle utilisée dans PSP, provient du fait que la base de données n'est parcourue qu'une seule fois tout au long de l'algorithme permettant l'extraction. Lors de cette unique lecture de la base, le premier étage de l'arbre préfixé est construit. Chaque nœud de profondeur 1, représentant un item de la base de données, garde une trace, pour chaque apparition dans la base, du visiteur et de la date d'apparition correspondante. Ce processus reprend donc la transformation de la base de données opérée dans SPADE, pour la projeter dans l'arbre préfixé.

2. de simplifier l'opération de générations (inutilité des classes d'équivalences). Cet algorithme reprend les fondements de la génération des candidats effectuée dans PSP. Cependant, dans Vpsp lorsqu'un candidat est généré, celui-ci se voit attribué un vecteur d'apparition issu de la base de données qui permet de vérifier plus efficacement le comptage du support.

3. de simplifier le comptage du support. Le comptage des séquences candidates tire profit de la structure de données utilisée par Vpsp. En effet, grâce à la liste d'apparition dont dispose chaque candidat, nous pouvons déterminer efficacement le nombre de visiteurs qui ont participés à l'incrémentation du support.

2.4 Recommandations pour enrichir l'ontologie

Pour soutenir la gestion de l'ontologie, nous fournissons des recommandations pour la mise à jour de l'ontologie par les techniques du Web Mining, principalement par le Web Usage Mining. Les mises à jour concernent principalement l'extension de l'ontologie qui ne modifie pas complètement l'ontologie initiale:

1. L'ajout d'une feuille concept dans une hiérarchie.
2. L'ajout d'un sous-arbre de concepts dans la hiérarchie.
3. L'ajout d'une relation entre deux concepts.

Notre démarche consiste à fouiller le fichier Log afin d'en extraire des séquences de pages apparaissant fréquemment. Ces motifs séquentiels sont ensuite eux-mêmes analysés afin d'identifier les items représentant de nouveaux concepts.

De la même façon, une nouvelle relation entre deux concepts peut être identifiée grâce à l'extraction des patterns séquentiels. L'exemple choisi dans l'expérimentation décrite dans la section suivante nous permet d'illustrer le premier cas.

3 Expérimentations

Étape 1 : Construction de l'ontologie

Dans cette expérimentation, nous avons utilisé une ontologie qui existe déjà et nous l'avons modifié en fonction de la structure du site Web (UNIVERSITE DJILALI LIABES www.uni-sba.dz) qui a été étudiée.

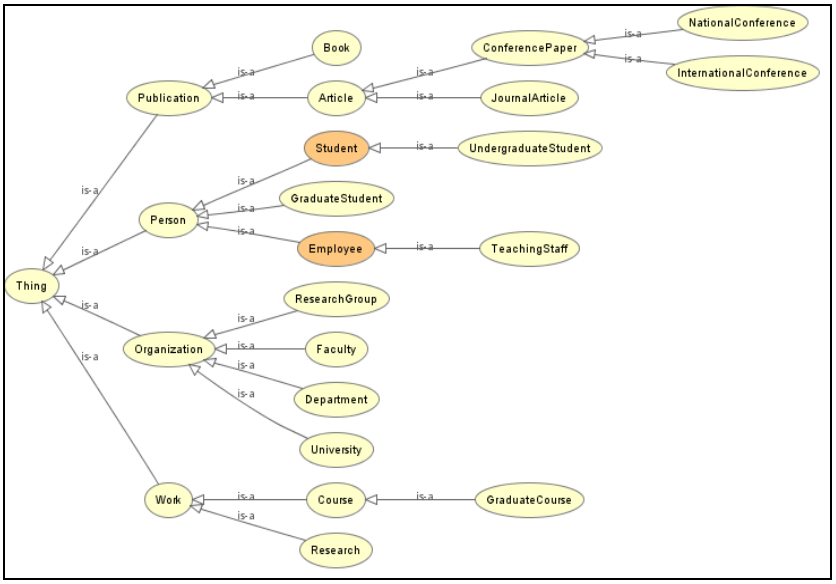

Fig. 1. L'ontologie du site Web.

Étape 2 : Résultats de l'analyse du fichier Log

Corpus expérimental

Il est constitué de l'ensemble de requêtes adressées au site (UNIVERSITE DJILALI LIABES www.uni-sba.dz) pendant la période allant du 24 Mars au 21 Avril 2008. Le fichier est composé de 115703 requêtes enregistrées suivant la norme CLF (Common Log Format). Pour chaque requête, nous disposons des champs suivants: la date de réalisation de la requête et l'heure à laquelle elle s'est produite (timestamp), l'adresse IP du client ayant accédé au serveur (client), la méthode i.e. l'action que tentait de réaliser le client (method), la requête que le client a essayé d'effectuer (url), la réponse du serveur (status), la longueur du contenu du document transféré (size).

Résultats

Le tableau suivant présente les résultats du prétraitement du fichier Log du site collectées pendant la période allant du 24 Mars au 21 Avril 2008.

Table 1. Tableau récapitulatif des résultats.

	Nombre de requêtes	**Pourcentage**
Total de requêtes	**115703**	**100 %**
Requêtes non valides	8250	7.13 %
Requêtes provenant des WRs	3243	2.8 %
Identification par IP	0	
Requêtes à « /robots.txt »	3243	
Identification par BS	0	
Requêtes aux images et fichiers multimédia	79074	68.34 %
Requêtes dont méthodes <> GET	118	0.1 %
Scripts et feuilles de style	1541	1.33 %
Total	92226	79.71 %
Nombre de requêtes après nettoyage et retraitement	9578	8.28 %
Nombre des sessions	**1544**	
Nombre des visites destinées à l'analyse	**2103**	

La nouvelle taille de la base (8.28 % de la taille initiale) montre bien l'importance de l'étape du prétraitement des fichiers Log, en particulier la phase du nettoyage. Cette étape a abouti à des fichiers nettoyés et structurés, prêts à l'analyse par l'application des méthodes de fouille des données.

Afin de pouvoir traiter l'information contenue dans la base le plus simplement et le plus efficacement possible, il faut restructurer la base selon le schéma relationnel.

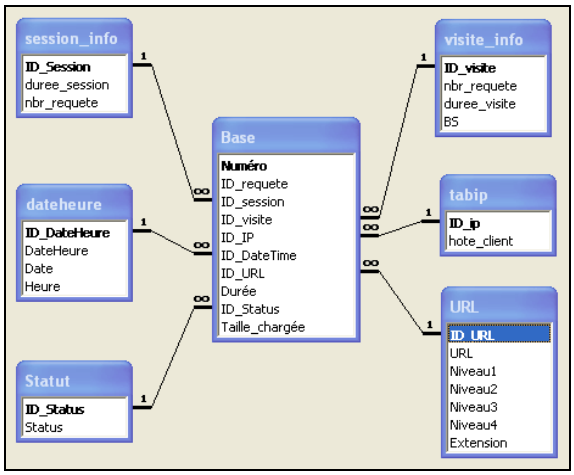

Fig. 2. Schéma relationnel.

Étape 3 : Résultats de l'extraction des motifs séquentiels

Après l'étape de prétraitement, le fichier log du site web (UNIVERSITE DJILALI LIABES www.uni-sba.dz) contient 241 URLs et 2103 séquences de visite. Les séquences extraites reflètent les comportements fréquents des internautes connectés sur le site.

Nous rapportons dans cette section quelques motifs séquentiels extraits du fichier Log :

Pattern 1:

http://www.univ-sba.dz/
http://www.univ-sba.dz/fsi/
http://www.univ-sba.dz/fsi/lmd/progr_LMD/S1-S2.htm

Support : 0.009715

Ce comportement a un support de 0,9715%. Cela signifie qu'il correspond à 15 utilisateurs du site Web. Ces utilisateurs sont susceptibles d'être intéressées par le programme du nouveau système LMD.

Pattern 2:

http://www.univ-sba.dz/
http://www.univ-sba.dz/laboratoires.php

Support : 0.027202

Ce comportement a un support de 2,7202%. Cela signifie qu'il correspond à 42 utilisateurs du site web. Ces utilisateurs sont susceptibles d'être intéressées par les laboratoires de l'université DJILALI LIABES.

Étape 4 : Mise à jour de l'ontologie

L'interprétation des résultats de l'étape précédente nous permet de faire des suggestions en vue d'appuyer la gestion de l'ontologie, et plus précisément d'étendre l'ontologie de base.
Après l'étape précédente (extraction des motifs séquentiels), on a trouvé que les internautes ont consultés la page de vice-rectorat (http://www.univ-sba.dz/vrpg) après la page principale du site Web (http://www.univ-sba.dz/). Le concept "Rectorate" n'existe pas dans notre ontologie et pourrait être ajoutée comme le montre la Figure 3. Nous avons également trouvé que plusieurs internaute ont consultés la page du programme LMD (http://www.univ-sba.dz/fsi/lmd/progr_LMD/S1-S2.htm) et aussi la page des cours LMD (http://www.univ-sba.dz/fsi/lmd/courtd.htm). Nous proposons d'ajouter le concept "LMDStudent" sous-concept du concept "UndergraduateStudent".

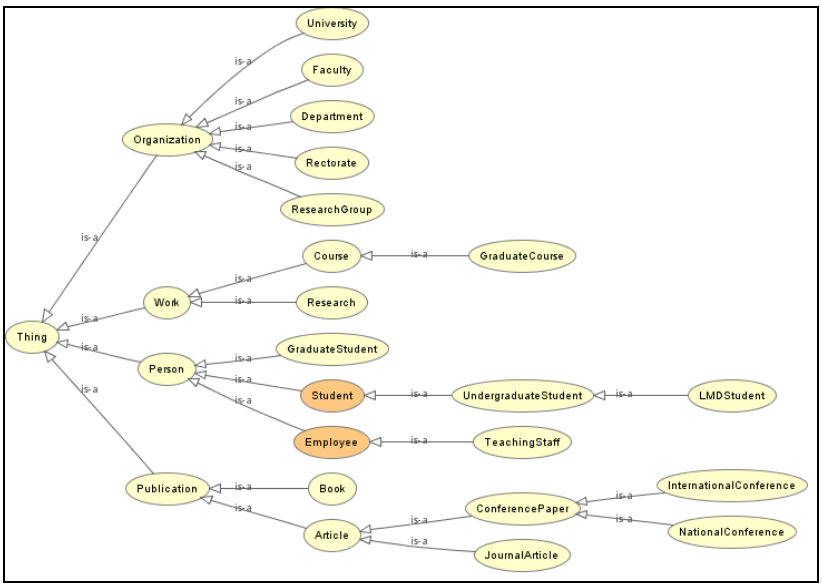

Fig. 3. Mise à jour de l'ontologie.

4 Travaux connexes

Peu de travaux visant à mettre à jour l'ontologie s'intéressent aux techniques de Web Usage Mining. Parmi ses travaux :
Mikroyannidis & Theodoulidis [8] Ici, les auteurs ont proposé une architecture pour l'adaptation des sites Web. Après l'étape de prétraitement des fichiers Log, les algorithmes pour l'extraction des itemsets fréquents sont appliqués afin de produire l'ensemble des pages (pagesets) qui sont souvent accessibles en même temps tout au long de la même session. Les pagesets extraits sont classées en fonction de deux critères. Le premier critère est basé sur la relation entre les pages de chaque pageset,

selon la topologie du site. Le deuxième critère de classification est basé sur le contenu des pages contenues dans chaque pageset. Une fois que les modifications proposées ont été révisées par le webmaster, elles peuvent être utilisées pour mettre à jour l'ontologie et modifier la structure du site Web.

Trousse & all. [12] Les auteurs présentent une approche basée sur l'analyse des usages du site Web. Ensuite, ils appliquent deux techniques de fouille de données sur les fichiers Log : extraction des motifs séquentiels et classification des pages Web dans le but de proposer de nouvelles relations entre les concepts de l'ontologie. L'approche a été illustrée sur un site Web de tourisme.

5 Conclusion et perspectives

Dans ce travail, nous avons tenté de démontrer l'impact potentiel du Web Usage Mining sur la mise à jour de l'ontologie. Nous avons illustré un tel impact dans le domaine de l'université en examinant le site Web de l'université DJILALI LIABES (http://www.univ-sba.dz/), nous partons d'une ontologie de domaine obtenu grâce à l'adaptation d'une ontologie existante à la structure actuelle du site Web. Ensuite, l'algorithme Vpsp a été appliqué sur le fichier Log généré à partir de ce site. Web Usage Mining fournit des informations pertinentes aux utilisateurs et il est donc un outil très puissant pour la recherche d'information. Le Web Usage Mining peut également être utilisés pour appuyer la modification de la structure du site Web et donner quelques recommandations aux visiteurs.

Les résultats de ce travail nous ouvrent des opportunités futures, notamment la possibilité de combiner Usage et Content Mining pour confirmer les mises à jour proposées. Cette combinaison pourrait nous permettre de construire des ontologies selon le contenu des pages Web et de les affiner avec les comportements extraits à partir des fichiers Log.

Références

1. Agrawal R., & Srikant R. (1995). Mining Sequential Patterns, Proceedings of the 11th International Conference on Data Engineering (ICDE'95), Tapei, Taiwan.
2. Ben Mustapha N., Aufaure M-A., & Baazhaoui-Zghal H. (2006). Towards an architecture of ontological components for the semantic web. In Proceedings of Wism (Web Information Systems Modeling) Workshop, CAiSE 2006, Luxembourg (pp. 22-35).
3. Davulcu H., Vadrevu S., & Nagarajan S. (2003). OntoMiner: Bootstrapping and populating ontologies from domain specific websites. In Proceedings of the First International Workshop on Semantic Web and Databases (SWDB 2003), Berlin.
4. Di-Jorio L., Jouve D., Kraemer D., Serra A., Raissi C., Laurent A., Teisseire M., & Poncelet P. (2006). VPSP : extraction de motifs séquentiels dans weka. In Démonstrations dans les 22émes journées "Bases de Données Avancées" (BDA'06).
5. Karoui L., Aufaure M.-A., & Bennacer N. (2004). Ontology discovery from web pages: Application to tourism. Workshop on Knowledge Discovery and Ontologies (KDO), co-located with ECML/PKDD, Pisa, Italy, pp. 115-120.

6. Masseglia F., Cathala F., & Poncelet P. (1998). The PSP Approach for Mining Sequential Patterns , Proceedings of the 2nd European Symposium on Principles of Data Mining and Knowledge Discovery (PKDD'98), LNAI, Vol. 1510, Nantes, France, September 1998, p. 176-184.
7. Masseglia F. (2002). Algorithmes et applications pour l'extraction de motifs séquentiels dans le domaine de la fouille de données : de l'incrémental au temps réel, PHD Dissertation, Université de versailles St Quentin - France.
8. Mikroyannidis A., & Theodoulidis B. (2004). A Theoretical Framework and an Implementation Architecture for Self Adaptive Web Sites, in Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence (WI'04), Beijing, China, pp. 558-561.
9. Navigli R., & Velardi P. (2004). Learning domain ontologies from document warehouses and dedicated web sites. Computational Linguistics, 30(2), 151-179.
10. Sanchez D., & Moreno A. (2004). Automatic generation of taxonomies from the WWW. In Proceedings of the 5th International Conference on Practical Aspects of Knowledge Management (PAKM 2004). LNAI, Vol. 3336 (pp. 208-219). Vienna, Austria.
11. Stumme G., Hotho A. & Berendt B. (2006). Semantic web mining : State of the art and future directions. Web Semantics : Science, Services and Agents on the World Wide Web, 4(2), 124–143.
12. Trousse B., Aufaure M.-A., Le Grand B., Lechevallier Y., & Masseglia F. (2007). Web Usage Mining for Ontology Management, in: Data Mining with Ontologies: Implementations, Findings and Frameworks. N. Hèctor Oscar, Gonzalez Cisaro. Sandra Elisabeth G, X. Daniel Hugo (editors), Information Science Reference, 2007, chap. 3, p. 37-64.
13. Zaki D. (2001). SPADE: An Efficient Algorithm for Mining Frequent Sequences, Machine Learning, vol. 42, 2001, p. 31-60, Kluwer Academic Publishers.