Etude de cas en Web Usage Mining : Catégorisation des utilisateurs de la connexion Internet de l'UATL[1]

ZIANI B. *, OUINTEN Y. *
** Département d'informatique, université de Laghouat 03000, Algérie*
Email {bziani, ouinteny}@mail.lagh-univ.dz

Introduction

La fouille de données constitue le cœur du processus de l'ECD (Extraction des Connaissances à partir des Données). Sa motivation principale est la valorisation d'une grande base ou entrepôt de données par la recherche de connaissances nouvelles pour l'aide à la décision[4]. Compte tenu de la croissance exponentielle de l'Internet, les applications des techniques de fouille de données aux fichiers log sont rapidement apparues[8]. Dans ce cas les données analysées proviennent du Web et on parle plutôt de **Web mining**. Selon le type des données analysées, on distingue trois classes d'applications dans le domaine du Web mining:

- **Web Content Mining**: qui concerne l'analyse du contenu des pages Web.
- **Web Structure Mining**: qui s'intéresse à l'analyse de la structure des sites Web.
- **Web Usage Mining**: qui analyse le comportement des utilisateurs des sites Web.

Dans notre travail on s'intéresse à une application du Web Usage Mining qui peut être définit comme étant "*l'application des techniques de data mining pour découvrir des comportements typiques d'utilisation dans le but de mieux comprendre et servir les utilisateurs du Web*"[8].

Ce processus d'analyse comporte trois étapes principales :

1. Le prétraitement: préparation des données à analyser.
2. Extraction des modèles: découverte des comportements typiques à partir des données analysées.
3. Interprétation: analyse et validation des modèles découverts.

[1] Université Amar Télidji de Laghouat

Objectifs

Le présent travail s'intéresse à l'analyse du comportement des utilisateurs de la connexion Internet par application de la technique de clustering. L'objectif étant de comprendre les habitudes de navigations chez les l'ensemble des utilisateurs (**enseignants**, **étudiants** et **personnel administratif)**. Un premier clustering a été réalisé à l'aide de l'algorithme **K-means** implémenté dans la plate forme WEKA[9]. Ayant remarqué un taux d'erreurs (instances mal classées) élevé, nous voulons confirmer ce résultat par l'exploration de l'algorithme **EM**. Les différents clusters qui peuvent être éventuellement identifiés présentent chacun un intérêt pour une catégorie différente d'utilisateurs. Ceci permet une bonne vision de l'ensemble des utilisateurs et, en conséquence, une amélioration des services offerts. On peut procéder en particulier à :

- Une gestion des plages horaires et de partage de la bande passante par profil d'utilisateurs
- Une configuration des serveurs proxy par groupe basés sur le profil des utilisateurs

Le Clustering

Le clustering peut être définie comme étant l'ensemble des méthodes visant à découper un ensemble d'objets en plusieurs groupes (clusters) en fonction des attributs qui les décrivent. L'objectif du clustering est ainsi de regrouper dans le même cluster les observations jugées similaires, selon une certaine métrique, (homogénéité intra-classe) et de placer les observations jugées dissimilaires dans des clusters distincts (hétérogénéité inter-classe). Nous ne prétendons pas fournir une liste exhaustive de l'ensemble des notions et méthodes existant dans le cadre du clustering, mais simplement de présenter le principe de fonctionnement des deux méthodes que nous avons utilisé dans notre expérimentation. Il s'agit de deux méthodes couramment utilisées et disponibles dans la plupart des logiciels de data mining : K-means et EM. Dans le domaine du Web Usage Mining, il y'a généralement, deux types de clusters à découvrir[3]:

- Les clusters **d'utilisateurs** dont l'objectif est de trouver des groupes d'internautes ayant des modèles de navigations similaires.
- Les clusters de **pages Web** regroupant les pages dont les contenus sont sémantiquement proches.

Algorithme k-means

Proposé en 1967 par MacQueen[6], l'algorithme des centres mobiles (K-means) est l'algorithme de clustering le plus connu et le plus utilisé, tout en étant très efficace et simple. Ce succès est dû au fait que cet algorithme présente un rapport coût/efficacité avantageux[2]. Le k-means utilise une mesure de distance pour attribuer un point au cluster le plus proche et fonctionne selon les étapes suivantes:

1. On choisit d'abord K points représentant les K clusters recherchés et appelés "centroides" des clusters (le nombre K est fixé par l'utilisateur).
2. On assigne chaque point non classé au cluster dont le centroide est le plus proche.
3. On réévalue les nouveaux centroides des clusters.
4. On recommence (réassignation des points au cluster dont le centroide est le plus proche, et recalcul des centroides) jusqu'a ce que les centroides ne changent plus significativement.

Une des critiques pour l'approche K-means est qu'elle est paramétrée par les centres de gravité des clusters sans tenir compte de leur dispersion. Elle s'appuie en fait sur l'hypothèse selon laquelle les nuages de points ont la même forme sphérique, ce qui n'est pas toujours le cas. Ainsi, à la vue des limitations imposées par la notion de distance, un intérêt s'est développé autour de l'approche probabiliste, avec la notion de modèle de mélange[5].

Algorithme EM

Introduit par Dempster, Laird et Rubin en 1978[7], l'algorithme **EM** utilise une approche dite **probabiliste**. Dans ce contexte, on considère que les données ont été générées par un mélange de modèles statistiques dont on cherche à déterminer les paramètres. Dès lors, il est possible d'associer les paramètres (moyenne, variance,...etc) de chaque modèle à un cluster. Ainsi, il est possible d'estimer la probabilité qu'un point ait été généré par un cluster[1]. La vraisemblance des données en entrée correspond alors à la probabilité qu'elles aient été générées par le mélange de modèles.
Partant d'une valeur initiale arbitraire θ^0, la $q^{ième}$ itération de l'algorithme EM consiste à effectuer les deux étapes **E** et **M** suivantes:

- L'étape E(**Expectation**) : calcul des probabilités conditionnelles $t_{ik}^{q}=t_k(x_i\ ;\ \theta^{q-1})$ que x_i appartienne à la $K^{ième}$ classe en utilisant la valeur courante θ^{q-1} du paramètre.
- L'étape M(**Maximisation**): L'estimateur θ^q de θ est actualisé en utilisant les probabilités conditionnelles □□□□.

L'approche la plus connu est le modèle de mélange gaussien où les densités élémentaires sont des lois normales. Dans cette technique, chaque cluster (groupe) est décrit par une loi de distribution normale, paramétrée par son centre de gravité et sa matrice de variance covariance. L'objectif est de maximiser la log-vraisemblance de l'échantillon de données compte tenu d'un nombre de clusters défini au préalable. Le calcul avec EM tient compte des centres de clusters (comme pour K-means), mais également de la forme des nuages de points à travers la matrice de variance co-variance. En pratique, cet algorithme fournit de bons résultats même s'il n'échappe pas aux principaux problèmes des algorithmes de clustering, en particulier sa convergence peut être très lente[7].

Etude expérimentale

Description

Les serveurs proxy de l'UATL enregistrent l'ensemble de l'activité des utilisateurs de la connexion Internet. On y trouve la trace de toutes les requêtes effectuées par tous les postes clients. On peut ainsi savoir comment chacun utilise le Web dans l'établissement. L'exploration des données brutes du trafic, nous a permis dans un premier temps de décrire les utilisateurs par différents attributs tels que le temps de navigation, le temps de téléchargement et le volume des données échangées[9]. Ces attributs sont ensuite utilisés pour réaliser un regroupement (clustering) des utilisateurs. Notre analyse s'est déroulée en trois étapes:

1. Préparation des données : Récupération et concaténation des fichiers log des différents serveurs proxy de l'UATL sur une période d'une semaine.
2. Génération d'un fichier de données en format .arff, format reconnu par la plate forme d'expérimentation Weka.
3. Regroupement (clustering) des utilisateurs.

Préparation des données

En collaboration avec les responsables du centre de calcul, nous avons pu obtenir un jeu de données sur la période du 02 au 08 Mai 2006. Les différents fichiers log sont concaténés en un seul fichier qui a constitué notre source de données. Le tableau 1 indique la taille des différents fichiers log récupérés sur la période étudiée.

Jour	Taille du log (en Mo)
02/05 (Mardi)	227.2
03/05 (Mercredi)	193.2
04/05 (Jeudi)	73.2
05/05 (vendredi)	8.6
06/05 (Samedi)	186.5
07/05 (Dimanche)	199.9
Moyenne	155.9

Tableau 1 : Taille des fichiers log analysés

Pour réaliser notre expérience de clustering, nous avons effectué une sélection des attributs jugés pertinents pour décrire les utilisateurs. En deuxième lieu nous avons généré un fichier, au format **arff** (format reconnu par Weka), contenant les données nécessaires à l'expérimentation. Le jeu de données utilisé concerne un ensemble de 458 utilisateurs décrits par 6 attributs. Le tableau 2 présente les attributs utilisés pour l'opération de clustering.

Attribut	Type	Signification
Nom	String	Nom désignant un utilisateur et remplaçant son adresse IP (pour confidentialité)
Tdown	Real	Temps épuisé pour les téléchargements
Tnav	Real	Temps épuisé pour les navigations
In	Real	Volume de données entrantes
Out	Real	Volume de données sortantes
Profil	Enumerate	Le profil de l'utilisateur (Ens=Enseignant, Etd=Etudiant, Adm=personnel administratif)

Tableau 2 : Les attributs retenus pour le clustering

La préparation des données au format **arff** a nécessité un travail laborieux. Dans cette phase nous avons procédé principalement à :

1. La codification des adresses IP des utilisateurs (pour confidentialité)
2. La codification des profils visités par les utilisateurs pour faciliter leur manipulation en type énuméré.
3. La conversion des attributs Tnav et Tdown, de la forme hh:mm:ss, en seconde pour faciliter leur manipulation sous forme numérique.

Clustering avec Weka

Le clustering, par K-Means et EM, sous Weka nécessite principalement la détermination de deux paramètres qui sont le nombre de clusters recherchés et une "graine". Cette notion de graine correspond à un point de départ aléatoire pour calculer les différents clusters (initialisation). Selon la graine de départ choisie, les clusters vont évoluer jusqu'à un état de stabilité qui conduira l'arrêt du processus de clustering. L'objectif de l'expérience de clustering étant de confirmer (ou d'infirmer) l'existence de comportement typique selon le profil d'utilisateur, le nombre de clusters est ainsi fixé à trois (03). En sortie, les deux méthodes nous fournissent essentiellement:

1. Le nombre d'instances assignées à chacun des clusters
2. La distribution des instances dans les clusters
3. Une estimation des instances mal classées

Notre intérêt est d'obtenir des clusters les plus significatifs possibles. Le "meilleur" résultat sera évalué en nous basant sur le taux d'erreurs (pourcentage des instances mal classées).

Résultats

Les deux méthodes utilisées sont sensibles au choix aléatoire à l'initialisation. C'est pourquoi nous avons répété l'expérience 20 fois, en faisant varier la valeur de la graine (Seed) de 1 à 20. Nous avons gardé enfin les résultats obtenus qui correspondent au taux d'erreurs (les instances mal classées) le plus faible. Ces résultats sont résumés ce qui suit :

- **Pourcentage des instances dans les clusters :**

Cluster	Nombre des instances affectées aux clusters	
	K-means	EM
0	82 (18\%)	196 (43\%)
1	185 (40\%)	222 (48\%)
2	191 (42\%)	40 (9\%)

- **Composition des clusters en fonction du profil des utilisateurs**

	K-means			EM		
Profil	**Clusters**			**Clusters**		
	0	**1**	**2**	**0**	**1**	**2**
Ens.	**12 (08.27%)**	**82 (56.55%)**	**51 (35.17%)**	**81(41.33%)**	**45(20.27%)**	**19(47.50%)**
Etd.	**55 (31.43%)**	**40(22.86%)**	**80 (45.71%)**	**81(41.33%)**	**123(55.41%)**	**12(30.00%)**
Adm.	**15 (10.87%)**	**63 (45.65%)**	**60 (43.48%)**	**75(38.26%)**	**54(24.32%)**	**9(22.50%)**

- **Nombre des instances mal classées**

K-means	EM
261 (56.98%)	**241 (52.62%)**

Discussion

Le taux élevé des instances mal classées ainsi que la composition des clusters en fonction des profils des utilisateurs montrent bien que les classes trouvées ne se détachent pas visiblement les unes des autres. Ce résultat reflète un comportement assez homogène chez l'ensemble des utilisateurs, indépendamment de leurs profils.

Conclusions et perspectives

Les deux méthodes de clustering que nous avons expérimenté sont basées sur des techniques différentes. En effet, K-means utilise uniquement un centroide pour représenter un cluster alors que EM définit une probabilité d'appartenance de chaque objet à un cluster en utilisant un modèle gaussien. Le taux d'erreurs, qui représente les instances mal classées dans les différents clusters, est important dans les deux expérimentations. Celui obtenu avec l'algorithme EM est légèrement inférieur. Globalement, les résultats dans les deux expérimentations montrent qu'il n'y a pas de groupes d'utilisateurs qui se détachent d'une manière apparente. Ainsi l'ensemble des utilisateurs peut être vu comme un seul groupe homogène. Il serait intéressant de valider ce résultat par d'autres expérimentations. Une piste possible est de réaliser des expérimentations avec différentes descriptions des utilisateurs, par élimination de certains attributs par exemple. Notons, enfin que les méthodes expérimentées nécessitent la connaissance du nombre de clusters K. Une amélioration possible consisterait à identifier automatiquement ce paramètre et utiliser le résultat pour confirmer (ou infirmer) l'hypothèse émise dans la problématique.

Références

1. B. Devèze B. Bargeton. Comparaison de différentes techniques d'optimisation pour l'apprentissage non-supervisé. Technical report, Université Pierre et Marie Curie.

2. Laurent Candillier. La classification non supervisée. Technical report, Equipe GRAppA, Lille 3, Sepembre 2004.

3. B. Cheikh. Etat de la connaissance sur le web usage mining. Technical report, Novembre 2002.

4. R. Rakotomalala D.A. Zighed. Extraction de connaissances à partir de données (ECD). Techniques de l'Ingénieur, H 3 744, 2003.

5. Francois-Xavier Jollois. Contribution de la classification automatique à la fouille de données. PhD thesis, Université de Metz, 12 Décembre 2003.

6. J. MacQueen. Some methods for classification and analysis of multivariate observations. In 5th Berkeley symposium on Mathematical statistics and probability, pages 281–297, Berkeley, 1967.

7. Christophe Saint-Jean. Classification paramétrique robuste partiellement supervisée en reconnaissance des formes. PhD thesis, Université de la Rochelle, 17 Décembre 2001.

8. Doru Tanasa, Brigitte Trousse, and Florent Masseglia. Mesures de l'internet, chapter Fouille de données appliquée aux logs web : état de l'art sur le Web Usage Mining, pages 126–143, Les Canadiens en Europe, 2004. Sous la direction d'Eric Guichard.

9. B. Ziani Y. Ouinten. Application des techniques de fouilles de données à l'analyse des fichiers log. In COSI'07, Oran, 11-13 Juin 2007.