Classification automatique : Applications au Web Mining

Yves Lechevallier[1], Doru Tanasa[2], Brigitte Trousse[2], Rosanna Verde[3]

[1]INRIA-Institut National de Recherche en Informatique et en Automatique, Domaine de Voluceau-Rocquencourt B.P.105, 78153 Le Chesnay Cedex, France, Yves.Lechevallier@inria.fr
[2]INRIA-Institut National de Recherche en Informatique et en Automatique, Sophia Antipolis- B.P.93, 2004 Route des Lucioles, 06902 Sophia Antipolis Cedex, France, {Doru.Tanasa, Brigitte.Trousse}@inria.fr
[3]Dip. Strategie Aziendali e Metodologie Quantitative - SUN – Seconda Università di Napoli, Piazza Umberto I, 81043 Capua, Italie rosanna.verde@unina2.it

RÉSUMÉ *Dans ce travail nous présentons une approche classificatoire appliquée aux données du Web Usage Mining..*

MOTS-CLÉS :*Classification automatique, classe, Web Usage Mining.*

1. Introduction

Le développement du Web a entraîné au cours de ces dernières années une explosion des données liées à son activité. Pour analyser ce nouveau type de données, sont apparues de nouvelles méthodes d'analyse regroupées sous le terme Web Mining dont les trois axes de développement actuels sont les suivants :

- *Web Content Mining* : analyse textuelle avancée, intégrant les particularités du Web telles que les liens hypertextes et la structure sémantique des pages,
- *Web Structure Mining* : analyse de la structure de liens hypertextes de pages Web en vue d'une catégorisation des pages et sites Web et/ou une classification de sites Web,
- *Web Usage Mining* : analyse des comportements de navigation

La création de sites de grande taille nécessite de prendre en compte la navigabilité du site du point de vue de l'utilisateur. L'étude de parcours à partir des logs issus de fichiers serveurs ou de traces propriétaires peut aider le responsable du site Web à repenser la structure et l'ergonomie du site pour repérer les problèmes des utilisateurs et améliorer la navigabilité.

2. À nouveau champ, de nouvelles structures de données

Les données analysées par le Web Usage Mining proviennent aujourd'hui principalement des fichiers « log http ». La structure du site (graphe des liens hypertexte) et l'information sur les utilisateurs du site (leurs profils) peuvent constituer d'autres sources supplémentaires d'information.

2.1 Présentation des fichiers log HTTP

Suivant le protocole client-seveur http, le poste client qui souhaite accéder à une ressource va émettre une requête adressée au serveur et contenant l'adresse d'allocation de la ressource :

```
GET http://www.inria.fr/accueil.html
```

A l'autre bout, le serveur de l'INRIA interprète la requête http, accède à la ressource demandée et la retourne au client. Comme dans la plupart des programmes informatiques, l'ensemble des opérations effectuées par le serveur sont enregistrées dans des fichiers « log » qui permettent de disposer d'une trace détaillée de l'activité du serveur. Nous utilisons le format de logs HTTP le plus répandu, l'ECLF (*Extended Common Log Format*) [LUO 95].

2.2 Pré-traitement des fichiers log HTTP : Nettoyage des données

Le nettoyage des données pour les fichiers log Web consiste à supprimer les requêtes inutiles de fichiers «log ». Ces requêtes concernent souvent les images et les fichiers multimédia. L'identification de robots Web et la suppression des requêtes provenant de ces robots sont les autres tâches de cette étape. Pour plus de détails concernant l'étape de prétraitement de fichier log HTTP le lecteur intéressé peut se rapporter à [TAN 03].

2.3. Difficultés de construction des sessions

L'unité d'analyse étant la séquence de pages visualisées et non la simple requête, il est préalablement nécessaire de regrouper les requêtes contenues dans les fichiers log pour reconstituer les sessions de visites. Bien que cette tâche paraisse à première vue assez aisée, l'analyste est confrontée à un certain nombre de problèmes techniques.

- *Identification des utilisateurs :*

Pour regrouper les requêtes, il est nécessaire de savoir quels utilisateurs les ont émises. Si l'utilisateur a accepté de s'enregistrer et s'identifie avec un login, alors le repérage est immédiat, mais cela ne concerne qu'une très faible minorité des visites. Une autre méthode répandue mais nécessitant l'acceptation de l'utilisateur, consiste à écrire dans la mémoire du navigateur, c'est-à-dire sur le poste client, un fichier d'identification nommée *cookie* qui sera réutilisé dans chacune des requêtes et permettra au serveur d'en identifier la provenance. En fait on ne dispose que de l'adresse IP qui est identique pour tous les utilisateurs partageant un même router ou pour ceux accédant à l'Internet via le même serveur proxy. Dans ce cas il est difficile de parler d'identification d'utilisateur.

- *Identification de sessions :*

Dans le cas où l'utilisateur aurait été identifié par une des deux premières méthodes décrites plus haut toutes les requêtes qui proviennent de cet utilisateur constitueront sa session. Dans les autres cas nous considérons le couple ip/agent pour construire les différentes sessions.
Le début de session est défini par le fait que la provenance de l'utilisateur (URL enregistrée dans le referrer) est extérieure au site. Par contre, aucun signal n'indique la déconnexion du site, ce qui pose un problème pour déterminer la fin des sessions. Les critères proposés sont en fait des seuils temporels d'inactivité allant de 25-30 minutes à 24 heures.

- *Reconstitution des parcours :*

Après avoir déterminé le début et la fin de la session et avoir filtré les requêtes auxiliaires, reste à reconstituer l'ordre chronologique de visualisation des pages sur le site, c'est-à-dire le *parcours* du visiteur. En effet, si on veut étudier des séquences de pages vues, et non simplement leurs associations il faut tenir compte du caractère séquentiel des sessions étudiées. Tant que la page de provenance (referrer) correspond à la page précédemment visualisée, le tracé du parcours sur le site est aisé. Une confusion peut cependant intervenir du fait que les dernières pages visualisées sont stockées dans la mémoire du navigateur et que par conséquent, lorsque le visiteur repasse par des pages précédemment visualisées, le poste client n'adresse aucune requête au serveur. Dans ce cas fréquent, il est nécessaire de *lire entre les lignes* du fichier « log » pour reconstituer le segment de parcours non enregistré.

3. Classification

Après les phases de nettoyage et de transformation de données qui permettent de construire un tableau de description des sessions nous abordons la phase d'analyse. L'objectif de cette phase est de découvrir différents comportements d'utilisateurs ou des catégories de comportement de navigation par diverses approches [SÄU 01] : Analyse des séquences fréquences, segmentation, modèle prédictifs, réseau neuronal et classification automatique.

3.1. Travaux existants

Il existe des nombreuses méthodes de classification utilisées dans la fouille de données. Cependant, peu de méthodes ont été appliquées aux données du Web : BIRCH dans [FU 99], CLIQUE dans [PER 98], EM dans [CAD 00] car il est difficile, voire impossible, d'adapter certaines méthodes aux particularités des données Web compte tenu de la taille de ces tableaux tant pour les sessions que pour les pages différentes.
Dans [MOB 02] les auteurs considèrent deux méthodes de classification, mais qui ne prennent pas en compte l'ordre des *pages* dans les sessions. Dans [FU 99] les sessions des utilisateurs sont généralisées en utilisant une induction, basée sur les attributs, qui réduit la dimension des données. Les pages sont organisées par une structure hiérarchique liée à l'adresse physique de la page Web. Les données ainsi généralisées sont classées en utilisant un algorithme efficace BIRCH de classification hiérarchique, introduit par [ZHA 96]. Une classification non-supervisé basée sur un réseau de neurones est utilisée dans [BEN 03] pour grouper les sessions similaires (issues d'un site annuaire thématique) en classes.

3.2. Notre approche

Nous aborderons uniquement l'aspect classification automatique et conceptuel, dans ce cas il s'agit de structurer l'ensemble des sessions ou l'ensemble des pages en typologies afin de dégager des comportements similaires et de les identifier. Cependant dans le cas où les objectifs sont définis par des groupes de pages (*rubriques*) nous proposons de modéliser chaque objectif par un objet symbolique. Les algorithmes utilisées sont de type Nuées Dynamiques [DID 71]. Ils recherchent simultanément une partition P de E en k classes et un vecteur L de k prototypes minimisant :

$$\Delta(P^*, L^*) = Min\left\{\Delta(P, L) \,\middle|\, P \in P_k, L \in D^k\right\}$$

avec P_k l'ensemble des partitions de E en k classes non vides. Ce critère Δ exprime l'adéquation entre la partition P et le vecteur des k prototypes. Il est souvent défini comme la somme des distances entre tous les objets s de E et le prototype g_i de la classe C_i la plus proche. L'algorithme procède, alternativement par une étape de représentation suivie d'une étape d'allocation.

Classification des sessions

La phase de classification de l'ensemble des sessions sera suivie par la description des classes et leur positionnement sur un plan factoriel. A partir des ces descriptions nous montrerons comment les modéliser sous la forme de concepts ou bien mettre en œuvre des arbres de décision afin d'obtenir pour chaque classe des règles.

Classification symbolique

A partir de la variable «referer » on peut définir par des requêtes des objets symboliques qui représentent la description de cette classe qui est identifiée à partir d'une connaissance experte. Dans ce cas nous devons utiliser des méthodes de classification, développées dans le cadre de l'analyse symbolique qui s'applique sur des variables multivaluées. Le concept de *prototype* est ici un modèle de représentation d'une classe et il servira à la construction d'indicateurs d'interprétation de ces classes.

4. Perspectives

Une des limites de toutes ces techniques de Web Usage Mining est qu'elles sont difficilement interprétables du fait qu'elles ne décrivent les parcours qu'en termes de noms de documents HTML principalement connus par les concepteurs du site. Un marquage sémantique des pages faciliterait donc la lecture des résultats obtenus, et pourrait intervenir dans la conception même de ces outils de Data Mining.

5. Références

[BEN 03] BENEDEK A., TROUSSE B., « Visualization Adaptation of Self-Organizing Maps for Case Indexing », In *27th Annual Conference of the Gesellschaft fur Klassifikation*, Cottbus, Germany, 12-14 mars 2003.

[CAD 00] CADEZ I. V., HECKERMAN D., MEEK C., SMYTH P., AND WHITE S., « Visualization of navigation patterns on a web site using model-based clustering », In *Proceedings of the sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 280 – 284, Boston, Massachusetts, 2000.

[DID 71] DIDAY, E « La méthode des Nuées dynamiques ».Rev. *Stat. Appliquée*, Vol XIX, p19-34, 1971.

[FU 00] FU Y., SANDHU K., SHIH M., « A generalization-based approach to clustering of web usage sessions », In *Proceedings of the 1999 KDD Workshop on Web Mining*, San Diego, CA, vol. 1836 of LNAI, pag 21 – 38, Springer, 2000.

[LOU 95] LUOTONEN A., « The Common Logfile Format », http://www.w3.org/Daemon/User/Config/Logging.html, 1995.

[MOB 02] MOBASHER B., DAI H., LUO T., AND NAKAGAWA M., « Discovery and evaluation of aggregate usage profiles for web personalization », *Data Mining and Knowledge Discovery*, 6(1):61 – 82, janvier 2002.

[PER 98] PERKOWITZ M., ETZIONI O., « Adaptive web sites: Automatically synthesizing web pages », In *AAAI/IAAI*, pages 727 – 732, 1998.

[SÄU 01] SÄUBERLICH F., HUBER K.-P., « A Framework for Web Usage Mining on Anonymous Logfile Data », SAS Institue GmbH, 2001.

[TAN 03] TANASA D., TROUSSE B., « Le prétraitement des fichiers log Web dans le Web Usage Mining Multi-sites », In *Journées Francophones de la Toile*, juin – juillet 2003.

[ZHA 96] ZHANG T., RAMAKRISHNAN R., LIVNY M., « Birch: An efficient data clustering method for very large databases », In H. V. Jagadish and Inderpal Singh Mumick, editors, *Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data*, Montreal, Quebec, Canada, juin 4-6, 1996, pages 103 – 114, ACM Press, 1996.