Une application des cartes topologiques auto-organisatrices à l'analyse des fichiers Logs[0]

Aïcha El Golli[*], Brieuc Conan-Guez[*], Fabrice Rossi[1*], Doru Tanasa[**], Brigitte Trousse[**], Yves Lechevallier[*]

Projet AxIS

[*]INRIA-Rocquencourt
Domaine De Voluceau, BP 105
78153 Le Chesnay Cedex, France

[**]INRIA-Sophia
2004 route des Lucioles, BP 93
06902 Sophia Antipolis,France

{*Aicha.Elgolli, Brieuc.Conan-guez, Fabrice.Rossi, Doru.Tanasa, Brigitte.Trousse, Yves.Lechevallier*}*@inria.fr*

RÉSUMÉ. *Dans ce travail nous pr´esentons une classification des rubriques visit´ees par des internautes grâce à une approche classificatoire utilisant une adaptation des cartes topologiques auto-organisatrices aux tableaux de dissimilarit´es*

MOTS-CLÉS : *Classification, cartes topologiques auto-organisatrices, Web Usage Mining*

1. Introduction

Les cartes topologiques auto-organisatrices de Kohonen [KOH 97] sont parmi les méthodes de classification non supervisées les plus utilisées. En effet, outre leur faculté à regrouper les données similaires au moyen de prototypes comme en quantification vectorielle et/ou en classification, elles autorisent la conservation de la topologie, d'où leur capacité à produire des représentations ordonnées, qu'on appelle prototypes ou vecteurs référents, sur une carte. Le calcul de ces vecteurs référents se base sur la notion de centre de gravité et malheureusement ce concept n'est pas applicable aux données complexes [GAN 04] [2]. Une adaptation de la version batch des cartes topologiques aux tableaux de dissimilarités a été proposée dans [ELG 03], [ELG 04] afin de permettre son application à différents types de données. Dans cette adaptation seule la définition d'une mesure de dissimilarité est nécessaire au déroulement de la méthode.

Soit d la mesure de dissimilarité choisie, rappelons les principales étapes de l'algorithme itératif des cartes topologiques de Kohonen sur tableaux de dissimilarités, à savoir l'étape d'affectation et l'étape de représentation. Chaque neurone c appartenant à la carte C est représenté par un ensemble $a_c = \{z_1 ... z_q\}$ d'élements de Ω de cardinal fixe q et appelé individu référent. Durant la phase d'affectation chaque élément $z_i \in \Omega$ est associé à un neurone gagnant c. Ce neurone est défini comme le neurone qui minimise la fonction d'adéquation d^T entre son individu référent a_c et l'élément z_i.

$$f(z_i) = \min_{c \in C} d^T(z_i, a_c) = \min_{c \in C} \Big(\sum_{r \in C} K^T(\delta(r,c)) \sum_{z_j \in a_r} d^2(z_i, z_j) \Big)$$

0. Publi´ee dans les actes des Onzi`emes journ´ees de la Soci´et´e Francophone de Classifi cation.
Disponible `a http://apiacoa.org/publications/2004/sfc04.pdf
1. Les coordonn´ees actuelles de Fabrice Rossi sont disponibles `a l'URL http://apiacoa.org/
2. Un compte rendu r´edig´e par Brigitte Trousse est disponible `a l'URL : http ://www-sop.inria.fr/axis/fdc-egc04/fdc-cr.html

$K^T(\delta(r,c))$ étant une fonction de voisinage qui dépend de la distance $\delta(r,c)$ entre le neurone gagnant c et le neurone r sur la carte. Après affectation de tous les éléments $z_i \in \Omega$, la phase de représentation permet de chercher le système des individus référents minimisant une fonction coût E. Pour cela, à chaque neurone r de la carte C on associe l'individu référent a_r minimisant la fonction E_r suivante :

$$E_r = \sum_{z_i \in \Omega} K^T(\delta(f(z_i), r)) \sum_{z_j \in a_r} d^2(z_i, z_j)$$

2. Données et problème

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, de nouvelles méthodes d'analyse sont apparues sous le terme du Web Mining. Dans cet article nous présentons une partie d'une étude de l'activité du site Web de l'Institut National de Recherche en Informatique et Automatique (INRIA). Le premier objectif de cette étude est l'analyse de la perception de l'activité de l'INRIA par les internautes via celle de son site. Le deuxième objectif est d'apporter des éléments significatifs en vue de l'amélioration de la qualité du site, et de la réponse qu'il apporte aux besoins des utilisateurs. La principale source d'information des visiteurs d'un site Web provient des fichiers Logs listant toutes les requêtes HTTP des clients dans l'ordre de leurs visites. La grande quantité de données et la faible qualité de l'information se trouvant dans ces fichiers nécessitent leurs prétraitements.

Pour notre application, nous avons pris les fichiers Logs HTTP du serveur Web national de l'INRIA et ceux du serveur de Sophia Antipolis issus des 15 premiers jours de l'année 2003. Un utilisateur qui recherche de l'information, navigue parmi tous les serveurs de l'INRIA d'une façon relativement transparente car les pages des différents serveurs Web sont fortement liées entre elles. Il y a de grandes chances que le visiteur ne remarque même pas que le serveur Web a changé. Pour l'analyste du Web Usage Mining, ce changement est très important car il permet d'analyser le comportement de l'utilisateur dans sa recherche de l'information. Ayant un fichier Log Web par serveur, l'analyste doit donc reconstituer le chemin suivi par l'utilisateur sur les différents serveurs sur lesquels ce dernier a navigué. Notre solution est de fusionner tous ces fichiers Logs Web, puis de reconstituer les visites des internautes [TAN 03], [TAN 04].

Deux grandes étapes constituent le prétraitement des fichiers Logs, à savoir la transformation des données et le nettoyage des données. Le résultat du prétraitement est une base de données relationnelle [ARN 03].

L'étape de transformation des données consiste à fusionner les fichiers Logs, rendre anonymes les Ip (ou les noms des domaines) dans le fichier Log obtenu et à grouper les requêtes par session (même Ip, même Agent). Ensuite, les sessions sont divisées en navigations en choisissant un seuil $\Delta t = 30min$.

Le nettoyage des données pour les fichiers Logs consiste à supprimer les requêtes pour les ressources Web qui ne font pas l'objet de l'analyse (les fichiers images par exemple) et les requêtes ou visites provenant des robots Web.

La structure des sites (graphe des liens hypertextes) et l'information sur les utilisateurs des sites (leurs profils) constituent des sources d'information supplémentaires à la base de données relationnelle obtenue suite au prétraitement. A partir de la base de données obtenue grâce au prétraitement, nous avons décidé de sélectionner les navigations d'une durée supérieure à soixante secondes. Nous avons aussi éliminé les pages dont le code statut représente une erreur. Dans nos traitements, nous avons choisi d'analyser les navigations des sites du siège (www) et de Sophia (SOP), l'équivalent de 300 000 pages visitées. A ces 300 000 pages visitées correspondent 3969 navigations visitant donc les pages du siège et aussi celles de Sophia [LEC 03]. A chaque page visitée correspond une rubrique 1 et une rubrique 2.

$$http://\underbrace{www-sop}.inria.fr/\underbrace{axis}/\underbrace{personnel}/Brigitte.Trousse/bri-eng.html$$

Site rubrique 1 rubrique 2

Nous avons créé une taxonomie sur les ”rubriques 1”. En effet, chaque rubrique 1 appartient à une rubrique sémantique. Par exemple : les rubriques ”axis”, ”sinus”, ”sloop”, sont des projets de l'INRIA Sophia et donc appartiennent à la rubrique sémantique ”projet”. Nous avons donc créé une table relative aux rubriques sémantiques qui à chaque rubrique fait correspondre sa rubrique sémantique.

3. Traitements et analyses

Nous avons choisi de nous intéresser à la classification des rubriques afin de trouver des corrélations. L'approche que nous adoptons s'appuie sur les navigations des internautes. Pour cela, ayant les 3969 navigations grâce à la base de données relationnelle, nous avons construit un tableau décrivant chaque navigation par la liste des ”rubriques 1” consultées. A partir de ce tableau on construit un tableau binaire dont les individus sont les 196 ”rubriques 1” et les variables sont les navigations : une navigation N_i visitant la rubrique R_j et pas la rubrique R_k sera codée respectivement par 1 pour R_j et 0 pour R_k dans le tableau (voir tableau 1). Ayant deux vecteurs binaires R_1

Navigations / Rubriques	N_1	N_2	...	N_{3969}
R_1	0	1	...	0
R_2	1	0	...	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
R_{196}	0	0	...	0

TAB. 1 – Tableau binaire décrivant les 196 rubriques visitées (1) ou pas (0) par une navigation

et R_2, pour définir une similarité ou une dissimilarité spécifique, il est nécessaire d'introduire les quatres quantités suivantes :

- soit a le nombre de fois où $R_1^j = R_2^j = 1$;
- soit b le nombre de fois où $R_1^j = 0$ et $R_2^j = 1$;
- soit c le nombre de fois où $R_1^j = 1$ et $R_2^j = 0$;
- soit d le nombre de fois où $R_1^j = R_2^j = 0$;

La similarité choisie dans notre cas entre les rubriques est la suivante : $S(R_1, R_2) = \frac{a}{a+b+c}$
Ceci correspond à l'indice de similarité de Jaccard. Cet indice indique la probabilité de visite de la rubrique R_1 et de la rubrique R_2 sachant qu'on a visité au moins l'une des deux.

Ayant donc, le tableau de dissimilarités entre les 196 rubriques, nous avons appliqué la méthode des cartes topologiques sur tableau de dissimilarités [ELG 03]. Les paramètres utilisés pour le déroulement de notre algorithme sont les suivants :

Paramètres	valeurs
Dissimilarité	$1 - S(R_1, R_2) = 1 - \frac{a}{a+b+c}$
Ensemble d'apprentissage	196
Nombre de neurones	$12 : 4 \times 3$
cardinal individus référents : q	1

Les résultats obtenus sont assez intéressants. Dans la classification obtenue on s'est intéressé à la rubrique sémantique ”projet” et les classes obtenues sont relativement fidèles à l'organisation des sites des projets de l'INRIA. En effet, avant le 1^{er} Avril 2004 les projets de l'INRIA sont groupés par ”Thème”, il existait 4 thèmes à savoir :

- Thème 1 : Réseaux et systèmes
- Thème 2 : Génie logiciel et calcul symbolique
- Thème 3 : Interaction homme-machine, images, données, connaissances
- Thème 4 : Simulation et optimisation de systèmes complexes

manifestation	Projet(Thème 1)	Projet(Thème 3)	inria
manifestation	Projet(Thème 1)	Projet(Thème 4)	Projet(Thème 2)
Projet(Thème 2)	Projet(Thème 4)	Projet(Thème 4)	Projet(Thème 4)

FIG. 1 – La carte (4× 3) obtenue : représentation de la correspondance sémantique des individus référents (pour les projets on représente le thème auquel ils sont attachés)

En prenant les individus référents des classes et en se référant aux rubriques sémantiques correspondantes, on obtient la carte de la figure 1.

Et pour mieux voir les associations et les corrélations entre les projets, voici le détail des classes obtenues pour la rubrique sémantique ”projet”. On représente les individus référents en gras :

Th`eme 1 meije Th`eme 2 Koala, croap Th`eme 3 odyssee Th`eme 4 Opale **Classe 9**	Th`eme 1 SOP-mistral [3], SOP-Mimosa, SOP-sloop, SOP-rodeo, rodeo, mascotte, **SOP-mascotte**, sloop, SOP-planete, SOP-oasis **Classe 10**	Th`eme 3 robovis, epidaure, ariana, acacia, orion, aid, SOP-robovis , SOP-epidaure, SOP-odyssee, SOP-acacia, **SOP-orion**, SOP-ariana, SOP-aid, SOP-axis, SOP-visa **Classe 11**	**Classe 12**
Th`eme 1 tropics Th`eme 3 reves Th`eme 4 Omega **Classe 5**	Th`eme 1 Mimosa, tick, **SOP-tick** **Classe 6**	Th`eme 4 **comore**, mefi sto miaou, SOP-mefi sto, SOP-smash **Classe 7**	Th`eme 2 Prisme, SOP-Prisme, SOP-lemme, **SOP-galaad**, SOP-cafe, SOP-saga, SOP-safi r **Classe 8**
Th`eme 2 cafe, lemme, **certilab** Th`eme 4 Chir, Fractales, opale **Classe 1**	Th`eme 1 Mistral, planete, SOP-meije Th`eme 2 oasis, saga, safi r, SOP-Koala Th`eme 4 **caiman**, sinus **Classe 2**	Th`eme 4 icare, SOP-sinus, **SOP-icare**, SOP-miaou, SOP-caiman **Classe 3**	Th`eme 1 SOP-tropics Th`eme 2 SOP-certilab Th`eme 3 SOP-reves Th`eme 4 **SOP-Omega**, SOP-sysdys **Classe 4**

Les classes 6 et 10 par exemple sont composées exclusivement de projets appartenant aux mêmes thèmes. Dans la classe 11, on constate la présence simultanée du projet *Aid* et du projet *Axis*. En effet, le projet *Axis* a remplacé le projet *Aid* et la visite de l'un entraine souvent la visite de l'autre, car il y a un lien mutuel sur les deux pages. On retrouve le même comportement pour le projet *Odyssee* et le projet *Robovis*. La classe 12 ne contient aucun projet.
***Remerciement** : Nous tenons à remercier Mihai Jurca (Axis Sophia) pour son aide dans le développement de l'outil de prétraitement Axis LogMiner*[4].

4. Bibliographie

[ARN 03] ARNOUX M., LECHEVALLIER Y., TANASA D.AND TROUSSE B., VERDE R., Automatic Clustering for the Web Usage Mining, PETCU D., ZAHARIE D., NEGRU V., JEBELEANU T., Eds., *Proceedings of the 5th Intl. Workshop on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC03)*, Editura Mirton, Timisoara, 1-4 October 2003, p. 54 – 66.

[ELG 03] EL GOLLI A., CONAN-GUEZ B., Adaptation des cartes topologiques auto-organisatrices aux tableaux de dissimilarit´es, *Xèmes Rencontres de la Soci´et´e Francophone de Classification*, , 2003, p. 99-102.

[ELG 04] EL GOLLI A., CONAN-GUEZ B., ROSSI F., a self organizing map for dissimilarity data, accept´e `a IFCS (International Federation of Classifi cation Societies), 2004.

[GAN 04] GANÇARSKI P., TROUSSE B., Actes de l'atelier : Fouille de donn´ees complexes dans un processus d'extraction de connaissances, EGC, 2004.

[KOH 97] KOHONEN T., *Self-Organizing Maps*, Springer Verlag, New York, 1997.

[LEC 03] LECHEVALLIER Y., TANASA D., TROUSSE B., VERDE R., Classifi cation automatique : Applications au Web Mining, *M´ethodes et Perspectives en Classification, Xèmes Rencontres de la Soci´et´e Francophone de Classification*, , 2003, p. 157-160.

[TAN 03] TANASSA D., TROUSSE B., Le pr´etraitement des fi chiers log Web dans le Web Usage Mining Multi-sites, *journ´ee Francophones de la toile*, , 2003.

[TAN 04] TANASA D., TROUSSE B., Advanced Data Preprocessing for Intersites Web Usage Mining, *IEEE Intelligent Systems*, vol. 19, n^o 2, 2004, p. 59–65.

3. Le prefi xe SOP- signifi e que le projet a ´et´e consult´e `a partir du site de Sophia
4. Description de l'outil disponible `a l'URL : http ://www-sop.inria.fr/axis/axislogminer/