WCUM pour l'analyse d'un site web

Malika Charrad, Yves Lechevallier, Mohamed Ben Ahmed, Gilbert Saporta

WCUM pour l'analyse d'un site web

Malika Charrad*,*** Yves Lechevallier**
Mohamed Ben Ahmed*, Gilbert Saporta***

*Ecole Nationale des Sciences de l'Informatique
malika.charrad@riadi.rnu.tn,
**INRIA-Rocquencourt
yves.lechevallier@inria.fr
***Conservatoire National des Arts et Métiers
gilbert.saporta@cnam.fr

Résumé. Dans ce papier, nous proposons une approche WCUM (Web Content and Usage Mining) permettant de relier l'analyse du contenu d'un site Web à l'analyse de l'usage afin de mieux comprendre les comportements de navigation sur le site. L'apport de ce travail réside d'une part dans la proposition d'une approche reliant l'analyse du contenu à l'analyse de l'usage et d'autre part à l'extension de l'application des méthodes de block clustering, appliquées généralement en bioinformatique, au contexte Web mining afin de profiter de leur pouvoir classificatoire dans la découverte de biclasses homogènes à partir d'une partition des instances et une partition des attributs recherchées simultanément.

1 Introduction

La caractérisation des internautes fréquentant un site Web est un problème incontournable pour assister l'internaute et prédire son comportement. Ces considérations ont motivé d'importants efforts dans l'analyse des traces des internautes sur les sites Web. D'autres efforts ont été concentrés sur l'analyse du contenu des pages Web. Sachant que le comportement des utilisateurs sur un site web dépend fortement du contenu des pages du site et inversement le contenu du site devrait répondre aux attentes des usagers du site, nous proposons de faire la liaison entre le contenu et l'usage d'un site web. Notre idée est d'exploiter les différentes informations relatives au contenu d'un site Web et de son usage en vue de l'analyser. Le point de départ de cette approche est le contenu textuel du site et les fichiers logs contenant les traces des utilisateurs.

2 Approche WCUM

L'approche WCUM relie l'analyse du contenu à l'analyse de l'usage d'un site Web (fig. 1). Elle se déroule en deux principales étapes. La première consiste à l'analyse textuelle d'un site Web afin de découvrir les thèmes des pages. La seconde étape consiste à introduire ces thèmes dans l'analyse de l'usage du site. L'application de cette approche nécessite d'une part

l'aspiration du site afin de transformer ses pages en fichiers texte, et d'autre part la collecte des fichiers Logs contenant la trace des utilisateurs sur le site.

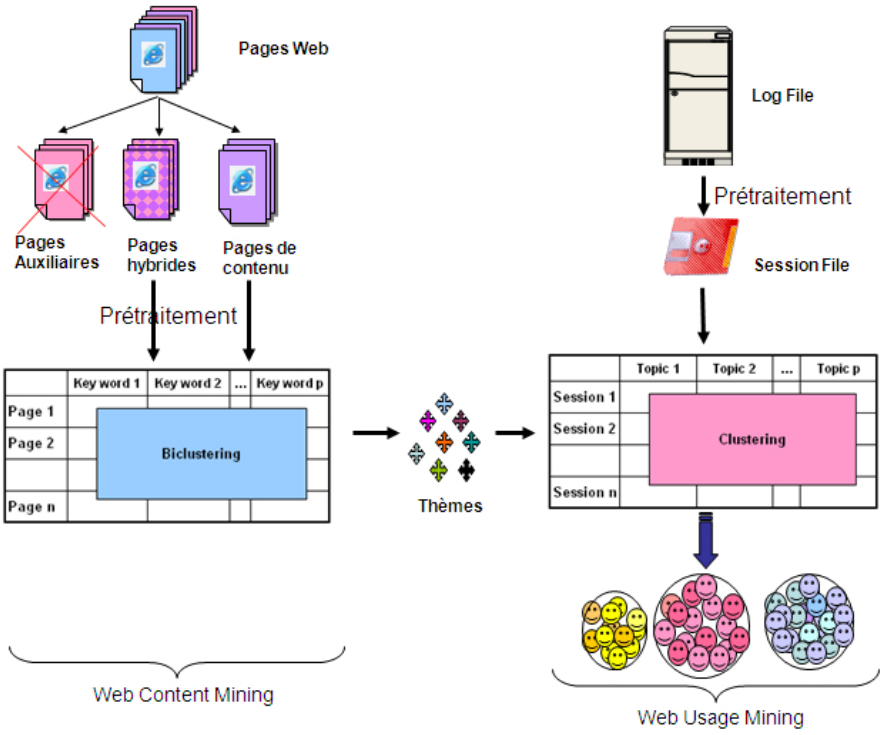

FIG. 1 – *Approche WCUM*

2.1 WCM : Analyse textuelle

2.1.1 Prétraitement des pages Web

L'analyse textuelle d'un site Web consiste en premier lieu à distinguer les pages de navigation des pages de contenu. Les pages de navigation, ou pages auxiliaires, servent à faciliter la navigation sur le site alors que les pages de contenu présentent une information éventuellement utile aux utilisateurs. Certaines pages de contenu contiennent plusieurs hyperliens permettant de naviguer sur le site. Ces pages présentant les caractéristiques communes de pages de contenu et de pages auxiliaires sont appelées "pages hybrides" . Cette étape a pour objectif d'exclure les pages auxiliaires de l'analyse et de limiter le travail de prétraitement aux pages de contenu. La classification des pages est basée sur un ensemble de variables tels que le nombre de liens entrants et sortants et la taille des documents...etc (Charrad et al., 2008). La deuxième étape consiste au prétraitement linguistique et à la sélection de descripteurs afin d'aboutir à une représentation matricielle du site. Le prétraitement nécessite tout d'abord la conversion

des pages Web en fichiers Textes, et le remplacement des images qu'ils contiennent par leurs légendes. L'étiquetage et la lemmatisation à l'aide de TreeTagger permettent de remplacer les verbes par leur forme infinitive, les noms par leur forme au singulier et certaines formes des verbes tels que les participes présents et les participes passés par leurs racines. Afin de réduire la dimension de l'espace vectoriel des vecteurs représentant les textes, il s'avère nécessaire de supprimer :

- Les formes de ponctuation,
- Les mots vides tels que les prépositions, les déterminants, les numéros, les conjonctions, les pronoms et les abréviations,
- Les mots inutiles à la classification tels que les adverbes et les adjectifs. Ainsi, seuls les noms et les verbes sont conservés dans la base des descripteurs.
- Les mots très fréquents : Nous avons adopté la méthode proposée par Stricker (2000). En effet, le rapport $R(m,t) = TF(m,t)/CF(m)$ tel que $TF(m,t)$ est l'occurrence du mot m dans un texte t et $CF(m)$ est l'occurrence du mot m dans l'ensemble des documents, permet de classer les mots par ordre décroissant. Plus le mot m est fréquent, plus le ratio est faible et, inversement, plus un mot est rare, plus le ratio est élevé. Dans le cas limite où un mot n'apparait qu'une seule fois dans l'ensemble de documents, ce ratio vaut 1 et le mot est classé en tête de liste.
- Les mots très peu frequents : ce sont les mots dont le nombre de documents dans lesquels ils apparaissent est inférieur à un certain seuil. Dans notre cas, nous supprimons les mots qui apparaissent dans une seule page du site Web.

Le prétraitement des textes aboutit à la construction d'une matrice croisant les descripteurs aux pages avec le nombre d'occurrences du descripteur dans une page du site comme poids. Un algorithme de classification croisée, CROKI2 (classification CROisée optimisant le Khi2 du tableau de contingence) proposé par Govaert (1983), est ensuite appliqué à la matrice pour découvrir des biclasses de pages et de descripteurs permettant d'attribuer un thème à chaque groupe de pages.

2.1.2 CROKI2 pour la classification croisée

Dans la littérature, la majorité des travaux sur la classification des documents appliquent des méthodes de classification simple sur l'une des deux dimensions (documents ou termes). Dans ce cas, un document appartenant à une classe L est décrit par tous les termes et chaque terme appartenant à une classe K caractérise tous les documents. Ainsi, en faisant porter la structure sur un seul ensemble, la détermination des liens entre les deux partitions est difficile. Dans notre cas, nous cherchons à identifier des classes de documents qui sont mieux décrits par un sous-ensemble de descripteurs, ce qui nécessite de découvrir dans les données des blocs de pages et de termes qui sont fortement correlés. Par conséquent, les algorithmes de classification simultanée sur les lignes et les colonnes sont plus adaptés à ce type de problème. L'algorithme CROKI2 proposé pour la classification croisée d'un tableau de contingence permet la découverte de blocs homogènes à partir d'une partition des descripteurs et une partition des pages recherchées simultanément. Il repose sur l'optimisation du critère du χ^2 de contingence. Disposant d'un tableau de contingence défini sur deux ensembles I et J, il s'agit de trouver une partition P de I en K classes et une partition Q de J en L classes telles que le χ^2 de contingence du nouveau tableau de contingence construit en regroupant les lignes et les colonnes suivant les partitions P et Q soit maximum. L'algorithme proposé construit une suite de couples de

partitions (P^n,Q^n) optimisant le χ^2 du tableau de contingence en appliquant alternativement sur I et sur J une variante de la méthode des Nuées Dynamiques de Diday (1971).

2.2 WUM : Analyse de l'usage

La première étape dans un processus de Web Usage Mining, une fois les données collectées, est le prétraitement des fichiers Logs qui consiste à nettoyer et transformer les données. La deuxième étape est la fouille des données permettant de découvrir des règles d'association, un enchainement de pages Web apparaissant souvent dans les visites et des "clusters" d'utilisateurs ayant des comportements similaires en terme de contenu visité. La dernière étape dans le processus est celle d'analyse et d'interprétation. Nous proposons dans ce papier d'exploiter les résultats de l'analyse textuelle pour l'analyse de l'usage.

2.2.1 Prétraitement des fichiers Logs

Le prétraitement des fichiers logs a comme objectif la structuration et l'amélioration de la qualité des données provenant de ces fichiers pour les préparer à une analyse des usages. Les objets à reconstruire ou à identifier dans un processus de prétraitement de fichiers logs web sont les clics, les utilisateurs, les robots web, les sessions, les navigations et parfois les épisodes. Le prétraitement des données se décompose en deux phases principales : une phase de nettoyage des données et une phase de transformation. Le nettoyage consiste à filtrer les données inutiles à travers la suppression des requêtes ne faisant pas l'objet de l'analyse, telles que les requêtes aux images ou aux fichiers multimédia, et celles provenant des robots Web. La transformation des données regroupe plusieurs tâches telles que l'identification des utilisateurs et la construction des sessions et des visites. L'identification des internautes est effectuée à l'aide des adresses IP, des Cookies ou du couple (pseudonyme, mot de passe). La reconstitution des sessions se fait en regroupant les requêtes émises par cet utilisateur. Chaque session est décomposée en visites en se basant sur le critère empirique de Kimball et Merz (2000).

2.2.2 Fouille de données

Cette étape consiste à appliquer des techniques de fouille des données sur le fichier de sessions afin d'extraire des connaissances sur les comportements des utilisateurs du site. Ces techniques varient entre les méthodes factorielles, les méthodes de classification automatique telles que les règles d'association pour la découverte de motifs fréquents de navigation (Marascu et Masseglia, 2006), les cartes de Kohonen pour la classification des utilisateurs (Charrad, 2005), (Lechevallier et al., 2003), (Fu et al., 2000), (Benedek et Trousse, 2003), (Srivastava et al., 2000) et les méthodes de classification supervisée telles que les arbres de décision, les réseaux de neurones et le raisonnement à base de mémoire.

3 Expérimentations et résultats

Nous proposons d'appliquer l'approche WCUM à un site Web de tourisme. Le prétraitement des textes aboutit à la construction d'une matrice croisant 418 descripteurs à 125 pages. Chaque cellule dans la matrice correspond au nombre d'occurrences du descripteur dans la page.

3.1 Résultats de l'analyse textuelle

L'application de l'algorithme CROKI2 à cette matrice aboutit à un ensemble de biclasses. La sélection des meilleures nécessite le recours aux critères suivants :

– **Pertinence de la biclasse :** la pertinence P de la biclasse est mesurée à travers la part de l'inertie conservée par la biclasse, notée B_{kl}, dans l'inertie totale B.

$$P = B_{kl}/B$$

avec

$$B_{kl} = f_{k.}f_{.l}(\frac{f_{kl}}{f_{k.}f_{.l}-1})^2$$

et

$$B = \sum_{k,l} B_{kl}$$

– **Homogénéité de la biclasse :** l'homogénéité, H, de la biclasse est mesurée par la part d'inertie B_{kl}, conservée par la classe par rapport à l'inertie initiale T_{kl} des points de la classe. La valeur obtenue comprise entre 0 et 1 est d'autant plus grande que la biclasse est homogène.

$$H = (B_{kl}/T_{kl})$$

avec

$$T_{kl} = \sum_{i \in P_k} \sum_{j \in Q_l} f_{i.}f_{.j}(\frac{f_{ij}}{f_{i.}f_{.j}-1})^2$$

A chaque classe de descripteurs, un thème est attribué en fonction des termes qui le composent (tab. 1).

A chaque classe des pages, une classe de descripteurs est associée pour former la biclasse. Chaque classe de pages appartient à au moins une biclasse. A titre d'exemple, la classe 6 de pages appartient à la fois aux biclasses (3,6) et (6,6). Deux thèmes sont donc associés aux pages composant cette classe. Le thème principal est identifié en se basant sur l'homogénéité et la pertinence des biclasses. Le tableau ci-dssous (tab. 2) croise quelques classes de pages aux thèmes qui leur sont associés.

D'autre part, l'URL de chaque page est organisé de façon hiérarchique sous forme de rubriques et de sous-rubriques représentant, selon le point de vue du concepteur, le contenu de la page. En examinant la structure arborescente des pages, la décomposition de l'URL de chaque page en rubriques et la comparaison de ces rubriques avec les résultats de la classification des pages basée sur le contenu permet de vérifier si les rubriques reflètent le contenu des pages.

3.2 Résultats de l'analyse de l'usage

En considérant les fichiers Logs du site dont on a analysé le contenu, nous procédons au nettoyage des requêtes non valides (dont le statut est inférieur à 200 ou supérieur à 399), les requêtes provenant des robots Web, les requêtes dont la méthode est différente de "GET" et les scripts. L'identification des sessions est effectuée en utilisant le couple (IP, User-Agent). Par suite, deux requêtes provenant de la même adresse IP mais de deux user-agents différents

Classes de descripteurs	Mots-clé	Thèmes
Classe 6	Bergerie, Brasserie, Centre, Distance, Fax, Hôtel, Magasin, Nord, Port, Zone, Restaurant, Sud, Technopôle, Tél, Village	Hôtels et Restaurants
Classe 5	Activité, Fête, Bal, Football, lieu, Manifestation, Occasion, Réunion	Activités et Manifestations
Classe 1	Amande, Crème, Eau, Flamber, Fruit, Gastronomie, Glacer, Lait, Mirabelle, Oeuf, Purée, Recette, Sucre,Hôspitalité	Recettes de cuisine
Classe 2	Arme, Art, Artiste, balade, Château, Découverte, Eglise, Exposition, Galerie, Guerre, Habit, Histoire, Illustrer, Maréchal, Monument, Moyen-Age, Palais, Pasteur, Peintre, Peinture, Promeneur, Renaissance, République, Saint, Siècle, Spectacle, Trésor	Histoire et Monuments
Classe 3	Cathédrale, Bibliothèques-médiathèques Boulevard, Capitale, Direction, Edifices, Gare, Guide, Hôpital, Information	Autres Adresses

TAB. 1 – *Exemples de thèmes*

	Th 1	Th 2	Th 3	Th 5	Th 6	Th. principal
Classe 2	X	X	X			Thème 2 : Histoire et Monuments
Classe 4	X		X			Thème 1 : Spécialités de cuisine
Classe 6			X		X	Thème 6 : Hôtels et Restaurants

TAB. 2 – *Exemples de Biclasses*

appartiennent à deux sessions différentes. Chaque session est décomposée en visites. Une visite est composée d'une suite de requêtes séquentiellement ordonnées, effectuées pendant la même session et ne présentant pas de rupture de séquence de plus de 30 minutes. Suite au nettoyage et transformations des requêtes provenant des fichiers logs, une matrice croisant les sessions aux pages est construite. Or d'après l'analyse du contenu, chaque page est affectée à une biclasse donc à un thème. Par suite, il est possible de croiser les visites (ou navigations) aux thèmes.

	Thème 1	Thème 2	...	Thème m
Navigation 1	20	0	...	2
Navigation 2	0	11	...	0
...	...	...	...	...
Navigation n	0	43	...	10

TAB. 3 – *Matrice des pages et des descripteurs*

Chaque cellule de la matrice correspond au nombre total de visites effectuées aux pages appartenant au thème j au cours de la navigation i. Un algorithme de classification simple est appliqué à la matrice $Navigations \times Themes$ pour découvrir des classes d'utilisateurs ayant un comportement similaire sur le site.

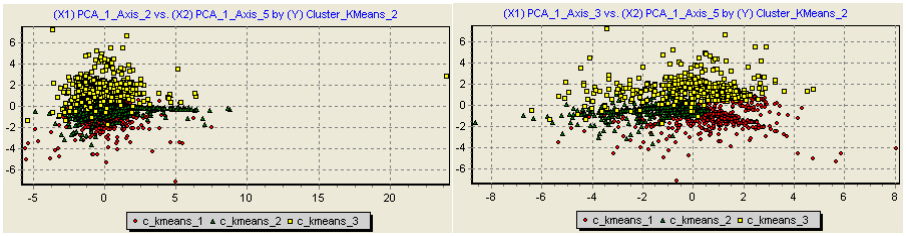

FIG. 2 – *Projection des classes d'utilisateurs*

L'application du Kmeans à la matrice croisant les navigations aux thèmes permet d'identifier trois classes d'utilisateurs. La classe C3 (fig. 2) est composée d'utilisateurs intéressés par les pages traitant du thème 4 (Informations utiles). Ils sont alors à la recherche des informations sur les horaires d'ouverture et de fermetures de certains établissements, des tarifs et des calendriers. Les internautes de la classe C2 sont par contre intéressés par les pages ayant pour thème "Recettes de cuisine", "histoire et monuments" et "Hôtels et Restaurants" (càd Thème1, Thème2 et Thème6). La classe C1 regroupe les visiteurs dont le motif est la recherche des adresses utiles (Thème 3), des manifestations et des activités culturelles (thème 5) et des informations utiles (thème4). Comme la classe majoritaire est C2 (70% des visiteurs), on déduit que ce sont les thèmes 1,2 et 6 qui intéressent le plus les visiteurs. Il s'en suit que les pages traitant de ces thèmes devraient être accessibles facilement et reliés par des hyperliens pour faciliter la navigation sur le site.

4 Conclusion

Dans ce papier, nous avons proposé une approche reliant l'analyse du contenu à l'analyse de l'usage. L'apport de cette approche est qu'elle permet d'identifier les thèmes qui intéressent les visiteurs et de tester si le contenu du site répond à leurs attentes. D'autre part, elle permet de réorganiser les pages de manière à faciliter le parcours du site.

Références

Benedek, A. et B. Trousse (2003). Adaptation of self-organizing maps for case indexing. *In 27th Annual Conference of the Gesellschaft fur Klassifikation, Germany*, 31–45.

Charrad, M. (2005). *Techniques d'extraction des connaissances appliquées aux données du Web*. Mémoire de mastère, Ecole Nationale des Sciences de l'Informatique de Tunis.

Charrad, M., Y. Lechevallier, G. Saporta, et M. B. Ahmed (2008). Web content data mining : la classification croisée pour l'analyse textuelle d'un site web. *Revue des Nouvelles Technologies de l'Information (Cépaduès) 1*, 43–54.

Diday, E. (1971). Une nouvelle méthode de classification automatique et reconnaissance des formes : la méthode des nuées dynamiques. *Revue de Statistique Appliquée 19 2*, 19–33.

Fu, Y., K. Sandhu, et M. Shih (2000). A generalization-based approach to clustering of web usage sessions. *In Proceedings of the 1999 KDD Workshop on Web Mining, San Diego, Springer*, 21–38.

Govaert, G. (1983). *Classification croisée*. Thèse de doctorat, Université Paris 6.

Kimball, R. et R. Merz (2000). Le data webhouse : Analyser des comportements clients sur le web. *Editions Eyrolles, Paris*.

Lechevallier, Y., D. Tonasa, B. Trousse, et R. Verde (2003). Classification automatique : Applications au web mining. *In Yadolah Dodge and Giuseppe Melfi, editor, Méthodes et Perspectives en Classification, Presse Académiques Neuchâtel*, 157–160.

Marascu, A. et F. Masseglia (2006). Extraction de motifs séquentiels dans les flots de données d'usage du web. *Extraction et Gestion des Connaissances (EGC'06), Lille*, 627–638.

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

Summary

The Web Content and Usage based (WCUM) Approach proposed in this paper deals with tha analysis of both content and usage of the web site to better understand the behaviour of web site users. Our main contribution consists in associating the content analysis with usage analysis and adaptating block clustering algorithms, traditionally used in bioinformatics, to web mining problems in order to discover homogeneous blocs of instances and attributes.