Web Content Data Mining : la classification croisée pour l'analyse textuelle d'un site Web

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

Web Content Data Mining : la classification croisée pour l'analyse textuelle d'un site Web

Malika Charrad*, Yves Lechevallier**
Gilbert Saporta***, Mohamed Ben Ahmed*
*Laboratoire RIADI, Ecole Nationale des Sciences de l'Informatique, Tunis
malika.charrad@riadi.rnu.tn
mohamed.benahmed@riadi.rnu.tn
**INRIA-Rocquencourt, 78153 Le Chesnay cedex
yves.lechevallier@inria.fr
***CNAM, 292 rue Saint-Martin, 75141 Paris cedex 03
saporta@cnam.fr

Résumé. Notre objectif dans cet article est l'analyse textuelle d'un site Web indépendamment de son usage. Notre approche se déroule en trois étapes. La première étape consiste au typage des pages afin de distinguer les pages de navigation ou pages « auxiliaires » des pages de contenu. La deuxième étape consiste au prétraitement du contenu des pages de contenu afin de représenter chaque page par un vecteur de descripteurs. La dernière étape consiste au block clustering ou la classification simultanée des lignes et des colonnes de la matrice croisant les pages aux descripteurs de pages afin de découvrir des biclasses de pages et de descripteurs. L'application de cette approche au site de tourisme de Metz prouve son efficacité et son applicabilité. L'ensemble de classes de pages groupés en thèmes facilitera l'analyse ultérieure de l'usage du site.

1 Introduction

Le Web représente aujourd'hui la principale source d'information. Ce gisement contenant une grande quantité de données non-structurées, distribuées et multi-medias a besoin d'être maintenu, filtré et organisé pour permettre un usage efficace. Cette tâche s'avère difficile à réaliser avec la large distribution, l'ouverture et la forte dynamicité du Web. Par conséquent, plusieurs travaux de recherche ont tenté d'analyser le contenu des sites Web et comprendre le comportement des utilisateurs de ces sites. L'approche que nous proposons dans cet article se situe dans ce cadre. Notre objectif est d'analyser un site Web en se basant sur le contenu et indépendamment de l'usage. En d'autres termes, nous cherchons à réduire la quantité d'information contenue dans le site Web en un groupe de thèmes qui pourraient susciter l'intérêt des internautes. Il sera par la suite possible d'analyser le comportement des utilisateurs vis-à-vis de ces thèmes.

2 Approche du Web Content Data Mining

Le Web Content Data Mining (WCDM) est l'application des techniques de Data mining au contenu du Web (textes, images, hyperliens…). Il est défini comme étant une analyse textuelle avancée intégrant l'étude des liens hypertextes et la structure sémantique des pages Web. L'approche que nous proposons pour le WCDM est présentée par le schéma suivant.

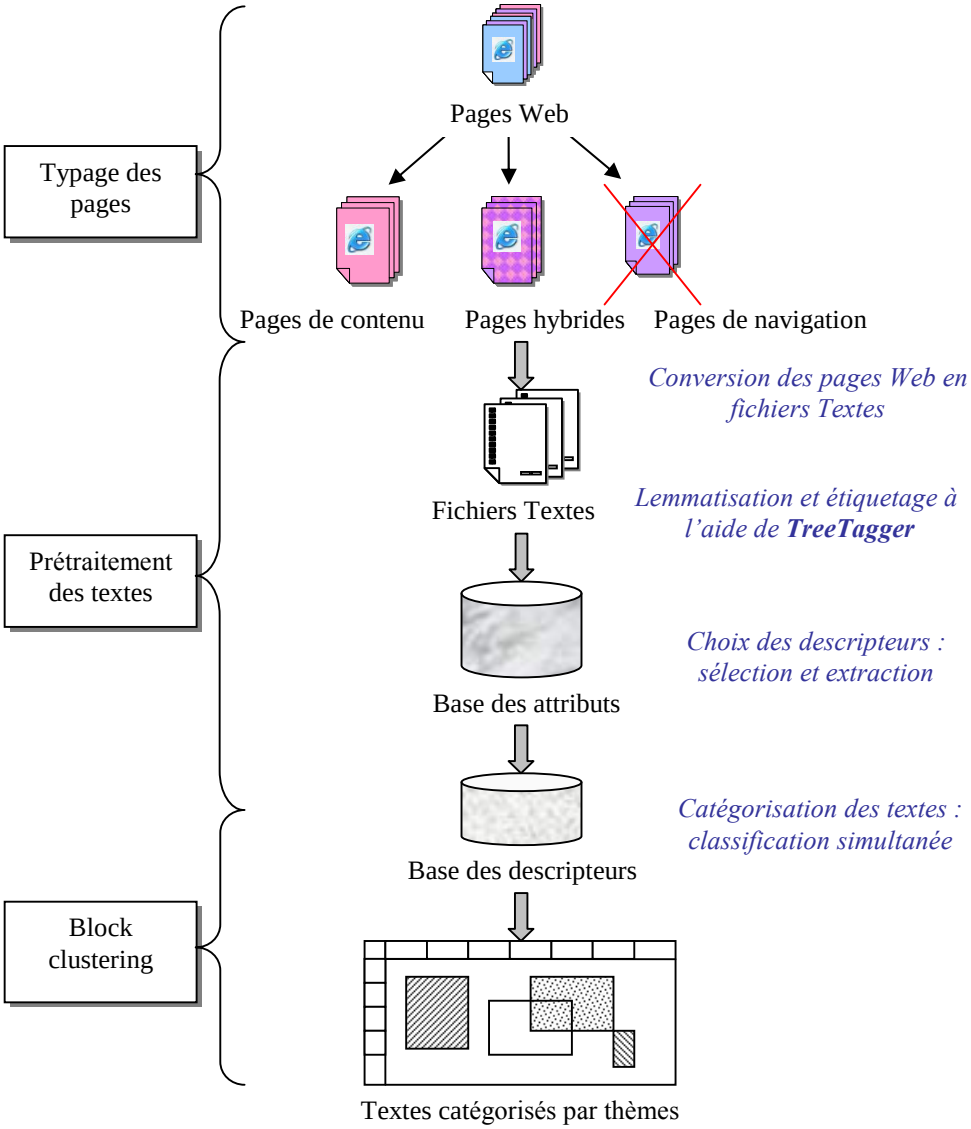

FIG. 1 – *Approche proposée pour le Web Content Data Mining.*

Cette approche se déroule en trois étapes principales. La première est celle de la classification des pages Web en pages de navigation et pages de contenu. L'objectif de cette étape est de limiter le travail postérieur aux pages de contenu. La deuxième est celle du prétraitement des textes et préparation des données et la dernière est celle de la classification simultanée ou block clustering.

3 Typage des pages

L'objectif de cette étape est de distinguer les pages qui servent à faciliter la navigation sur le site, appelées pages de navigation ou pages auxiliaires, des pages contenant de l'information qui pourrait intéresser l'internaute. Ces pages sont appelés « pages de contenu ». Certaines pages Web sont à la fois des pages de contenu et des pages de navigation. Il s'agit de pages hybrides.

3.1 Description des variables

Le typage des pages est effectué à l'aide d'une classification appliquée aux pages du site caractérisées par un ensemble de variables. Dans (Charrad et al. 2006) les variables utilisées sont : le nombre de liens entrants (inlinks), le nombre de liens sortants (outlinks), la durée moyenne de consultation de chaque page et le nombre de visites à chaque page. Ces variables sont déterminées à partir des fichiers Logs résultant de la navigation sur le site. Comme nous nous intéressons à l'analyse textuelle du site indépendamment de l'usage, et que le site étudié est un site de tourisme (site de Metz)[1], nous utilisons les variables suivantes: le nombre de liens entrants (inlinks), le nombre de liens sortants (outlinks), la taille du fichier (size), le nombre d'images par page, la présence d'un contenu textuel (texte).

Comme la variable « texte » est binaire, une première classification est effectuée sur les pages du site pour différencier les pages présentant un contenu textuel des pages présentant seulement des liens ou des images. Ainsi, la première hypothèse utilisée est que les pages ne contenant pas du texte sont des pages de navigation. Par contre, les pages présentant un contenu textuel peuvent servir de pages de navigation ou de pages de contenu. La détermination du type de la page est effectuée par la méthode de K-means.

3.2 Analyse des résultats

La méthode K-means (l'algorithme de Forgy (1965)) est appliquée aux pages présentant un contenu textuel c.à.d appartenant à la première classe. Les variables utilisées pour caractériser ces pages sont le « Nombre Outlinks », le « Nombre Inlinks », le « Size » et le « Nombre Images ». Les caractéristiques des sous-classes obtenues permettent d'attribuer une étiquette à chacune d'entre elles. En effet, la sous-classe C1 comporte environ 9% de pages caractérisées par la présence d'un nombre élevé de liens entrants et sortants utilisés pour passer d'une page à une autre, une taille importante et un nombre faible d'images. Ces pages sont destinés principalement pour passer d'une page à une autre. Elles correspondent alors à des pages de navigation bien qu'elles présentent un contenu textuel. Par contre, la sous-classe C2 comporte environ 27% de pages caractérisées par un nombre élevé de liens

[1] www.tourisme.mairie-metz.fr

entrants et d'images, en plus du contenu textuel, et un nombre faible de liens sortants. Ainsi, plusieurs pages pointent vers les pages de la sous-classe C2 mais elles pointent vers un nombre faible de pages. Ces pages correspondent à des pages de contenu. La sous-classe C3 présente les caractéristiques de pages de contenu (C2) et de pages de navigation (C3) à la fois, on les considère comme des pages hybrides. En résumé, trois classes sont obtenues à la fin de la classification : les pages auxiliaires résultant de la première et la seconde phase de classification, les pages de contenu et les pages hybrides.

Variables de la classification	Texte		Nombre Outlinks	Size	Nombre Inlinks	Nombre Images
Classe 1 (92%)	+	C1 (9%)	++	++	++	-
		C2 (27%)	-	+	++	+
		C3 (64%)	+	+	+	+
Classe 2 (8%)	-					

TAB. 1 – *Caractérisation des classes.*

L'application de l'analyse en composantes principales du logiciel Tanagra 1.4.10 (Rakotomalala, 2005) au même tableau de données permet de réduire l'espace vectoriel initial à un nouvel espace vectoriel composé de trois axes factoriels qui expliquent environ 84% de l'inertie totale. La projection des classes obtenues par le K-means sur le premier plan factoriel permet de vérifier la séparabilité entre ces classes (figure 2). Les Ronds correspondent aux pages auxiliaires, les triangles correspondent aux pages de contenu et les croix correspondent aux pages hybrides.

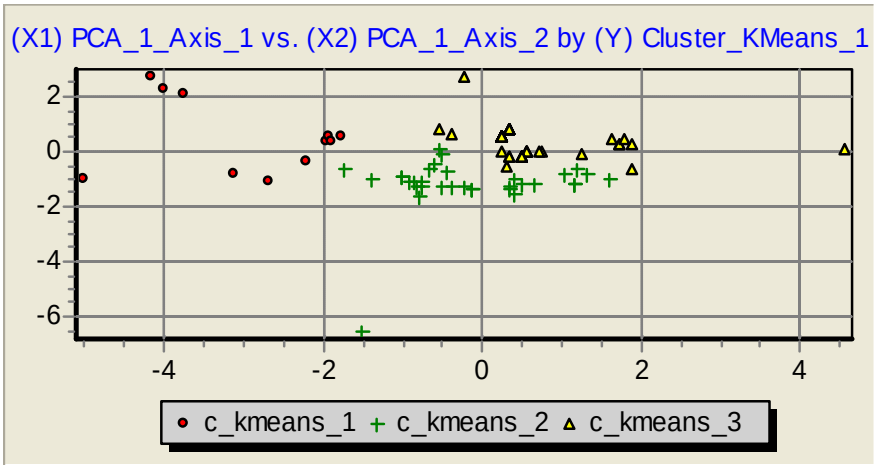

FIG. 2 – *Projection des pages sur les axes factoriels.*

3.3 Comparaison entre typage manuel et typage par Kmeans

Le typage manuel consiste à analyser manuellement le site Web et attribuer une étiquette à chaque page selon qu'elle soit une page de contenu, une page hybride ou une page de navigation. L'attribution de l'étiquette dépend seulement du motif de la visite que pourrait faire l'internaute au site. En effet, une page de contenu pour un visiteur peut être considérée comme étant une page de navigation pour un autre.
Le typage manuel du site de Metz a montré que 12% des pages du site sont considérées comme des pages de navigation, 11% comme des pages de contenu et 77% des pages sont des pages hybrides. Nous remarquons que les résultats de la classification sont proches des résultats du typage manuel. Ces résultats montrent que le site de Metz présente peu de pages auxiliaires et de pages de contenu. La majorité de pages sont de type hybride.

4 Prétraitement des textes

L'objectif du prétraitement est de représenter chaque page du site par un vecteur de descripteurs qui donne une idée sur son contenu. Le prétraitement est réalisé en deux étapes. La première étape est celle de représentation des documents. La seconde est celle du choix des descripteurs.

4.1 Représentation des textes

Plusieurs méthodes ont été proposées dans la littérature pour la représentation des textes. La méthode la plus utilisée est la représentation vectorielle, appelée «bag-of-words » ou « sac de mots » dans laquelle chaque texte est représenté par un vecteur de n termes pondérés (Salton et Buckley, 1998). À la base, les n termes sont les n différents mots apparaissant dans les textes de l'ensemble d'entraînement. Scott et Matwin (1999) ont fait des nombreux essais pour représenter les textes pour des fins de classification. Ils ont utilisé les groupes nominaux (des suites de noms et d'adjectifs) pour construire les termes de l'espace vectoriel et une application pour l'analyse de la nature grammaticale des mots du texte. Ils ont aussi évalué l'impact de regrouper les mots synonymes en un même méta-descripteur. La notion d'hyperonymes a été aussi mise à l'épreuve pour regrouper des mots. Aucune de ces méthodes n'a produit de résultats équivalents ou supérieurs à l'approche «bag-of-words». Lewis (1992) a également représenté les textes à l'aide de groupes nominaux mais les résultats n'étaient pas satisfaisants.

En adoptant l'approche « bag of words », le poids associé à chaque terme peut être une valeur binaire, indiquant la présence ou l'absence du terme dans le document (1 si le mot est présent dans le texte, 0 sinon), ou un entier positif représentant le nombre d'occurrences du terme dans le document. L'inconvénient d'introduire ce comptage est qu'il accorde un poids important aux termes qui apparaissent très souvent à travers toutes les classes des documents et qui sont peu représentatifs d'une classe en particulier. Une autre méthode largement utilisée pour calculer le poids d'un terme est la fonction TFIDF (acronyme de Term Frequency Inverse Document Frequency») (Salton et Buckley, 1998).

4.2 Choix des descripteurs

L'inconvénient de l'utilisation directe du vocabulaire contenu dans les textes d'entraînement est la dimension très élevée de l'espace vectoriel. L'utilisation de tous les mots contenus dans les textes peut influencer négativement la précision de la classification. D'autre part, un mot présent dans un nombre élevé de textes ne permet pas de décider de l'appartenance d'un texte qui le contient à l'une ou l'autre des catégories car son pouvoir de discrimination est faible.

Pour pallier ces problèmes, certaines techniques de réduction de la dimension du vocabulaire ont été mises en place. Ces techniques se divisent en deux grandes familles.

- Les techniques basées sur **la sélection de descripteurs** («feature selection») : Ces techniques conservent seulement les descripteurs jugés utiles à la classification, selon une certaine fonction d'évaluation. L'avantage de la sélection de descripteurs consiste à éliminer les descripteurs réellement inutiles ou les descripteurs erronés («noisy») (mots mal orthographiés par exemple). Plusieurs techniques de sélection de descripteurs ont été développées en vue de réduire la dimension de l'espace vectoriel. Chacune de ces techniques utilise des critères lui permettant de rejeter les descripteurs jugés inutiles à la tâche de classification. Le processus de sélection de descripteurs débute généralement par la suppression de mots très fréquents tels que les mots grammaticaux ou les mots de liaisons. La recherche de radical («stemming») et la lemmatisation sont d'autres méthodes utilisées pour créer un vocabulaire réduit. Leur but est de regrouper en un seul descripteur les multiples formes morphologiques des mots qui ont une sémantique commune. Les mots très peu fréquents, qui n'apparaissent qu'une ou deux fois dans l'ensemble de documents, sont également supprimés, car il n'est pas possible de construire des statistiques fiables à partir d'une ou deux occurrences (Stricker, 2000).
- Les techniques basées sur **l'extraction de descripteurs** («feature extraction») : Ces techniques créent des nouveaux descripteurs à partir des descripteurs de départ, en faisant des regroupements ou des transformations afin de réduire le nombre de descripteurs redondants. Le processus d'extraction de descripteurs consiste à créer à partir des descripteurs originaux un ensemble de descripteurs synthétiques en effectuant des regroupements de termes « term clustering » ayant une sémantique commune (Stricker, 2000). Les classes obtenues deviennent les descripteurs d'un nouvel espace vectoriel. Blum et Mitchell (1998) rapportent des résultats intéressants à propos de ce regroupement.

4.3 Résultats du prétraitement

Dans notre cas, 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. Ces textes sont par la suite traités par l'algorithme TreeTagger[2] qui a été développé à l'Institut de Linguistique Computationelle de l'Université de Stuttgart (Schmid, 1994).

[2] Les publications relatives à cet algorithme ainsi que les codes source sont disponibles sur le site : http://www.ims.uni-stuttgart.de/projekte/corplex/DecisionTreeTagger.html

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,
- Les termes de type non reconnu par TreeTagger sont examinés manuellement afin de ne garder que les noms et les verbes. D'autre part, les termes auxquels TreeTagger attribue l'étiquette « Nom » sont examinés afin de supprimer les noms propres que TreeTagger n'arrive pas à identifier. 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'apparaît 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 fréquents : 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. Ceci réduit la base des descripteurs à 652 descripteurs au lieu de 1500.
  - dont le nombre d'occurrences dans la base est égal à 1 *i.e* les mots qui apparaissent une seule fois dans toute la base.

Le prétraitement des textes aboutit à la construction d'une matrice croisant 418 descripteurs à 125 pages avec le nombre d'occurrences du descripteur dans une page du site comme poids.

	P1	P2	…	P_q
Descripteur 1	0	2		4
Descripteur 2	1	5		0
…				
Descripteur $_n$	0	0		3

TAB. 2 – *Tableau des données.*

5 Block clustering

5.1 Pourquoi le block clustering ?

Les méthodes de classification automatique appliquées à des tableaux mettant en jeu deux ensembles de données agissent de façon dissymétrique et privilégient un des deux ensembles en ne faisant porter la structure recherchée que sur un seul ensemble. Ainsi, la détermination des liens entre les deux partitions est difficile. La recherche simultanée de partitions sur les deux ensembles a donné naissance à des méthodes de classification simultanée (block clustering) telles que les méthodes de classification directe de Hartigan (1972, 1975), les méthodes de classification croisée de Govaert (1983) et les méthodes de biclustering utilisée généralement en bioinformatique. Ces méthodes de classification simultanée fournissent des blocs homogènes à partir d'une partition des instances et une partition des attributs recherchées simultanément.

5.2 Application de Croki2

Comme notre tableau des données est un tableau de contingence, nous avons appliqué l'algorithme CROKI2 (classification CROisée optimisant le Khi2 du tableau de contingence) proposé par Govaert (1983) pour les tableaux de contingence. L'objectif de cet algorithme est de trouver une partition P de I en K classes et une partition Q de J en L classes telle que le χ^2 de contingence du nouveau tableau construit en regroupant les lignes et les colonnes suivant les partitions P et Q soit maximum. Si $P = (P_1,...,P_K)$ est la partition de I en K classes, $Q = (Q_1,...,Q_L)$ est la partition de J en L classes, le nouveau tableau de contingence T_1(P,Q) s'écrit : $T_1(k,l) = \sum_{i \in P_k} \sum_{j \in Q_l} s_{ij}$ avec $k \in [1,...,K]$ et $l \in [1,...,L]$. Les fréquences marginales définies sur le tableau T1 s'écrivent : $f_{kl} = \sum_{i \in P_k} \sum_{j \in Q_l} f_{ij}$, $f_{k.} = \sum_{i \in P_k} f_{i.}$, $f_{.L} = \sum_{j \in Q_L} f_{.j}$

L'algorithme CROKI2 consiste à déterminer une série 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.

Les entrées de l'algorithme sont : le tableau de contingence, le nombre de classes en ligne et en colonne et le nombre de tirages de départ. Les sorties de l'algorithme sont : la valeur du Khi2 du tableau initial, la valeur du Khi2 du tableau (P,Q), le pourcentage d'inertie (ou d'information) conservée et le tableau de contingence initial réordonné en ligne et en colonne suivant les classes de deux partitions.

5.3 Analyse des résultats

L'application de l'algorithme CROKI2 du logiciel SICLA permet d'obtenir la nouvelle matrice suivante (Fig.5) dont les couleurs sombres représentent les meilleures biclasses. Les couleurs plus claires représentent les biclasses moins bonnes.
Le choix de ces meilleures biclasses est basé sur les trois critères suivants :

- La part d'inertie conservée par la classe par rapport à l'inertie initiale des points de la classe (B_{kl}/T_{kl}). Cette valeur indique la qualité de représentation d'une classe. La valeur obtenue comprise entre 0 et 1 sera d'autant plus grande que la classe sera homogène, avec $T_{kl} = \sum_{i \in P_k} \sum_{j \in Q_l} f_{i.} f_{.j} (f_{ij} / f_{i.} f_{.j} - 1)^2$, $B_{kl} = g_{k.} g_{.l} (g_{kl} / g_{k.} g_{.l} - 1)^2$ où $g_{kl} = \sum_{i \in P_k} \sum_{j \in Q_l} f_{i.} f_{.j}$ et $g_{.l} = \sum_{j \in Q_l} f_{.j}, g_{k.} = \sum_{i \in P_k} f_{i.}$
- La part d'inertie conservée par la biclasse sur l'inertie totale (B_{kl}/B). Cette valeur indique l'importance de la biclasse, avec $B = \sum_{k,l} B_{kl}$.
- Les nombres des éléments de la classe en ligne et la classe en colonne constituant la biclasse. Dans notre cas, les classes vides en ligne ou en colonne et celles composées de 2 éléments ou moins ne sont pas retenues.

Comme aucun critère n'est proposé pour choisir le nombre de classes en lignes et en colonnes dans l'algorithme CROKI2, nous utilisons le coefficient T de Tschuprow $T = \sqrt{\frac{\chi^2}{n\sqrt{(K-1)(L-1)}}}$, avec $0 \le T \le 1$. En effet, Le meilleur découpage produit deux partitions avec un Tschuprow qui tend vers 1. Si on suppose T = 1, et que nous cherchons le même nombre de classes en ligne et en colonne, i.e. K=L, on aura $K = \frac{\chi^2}{n} + 1$

En utilisant les critères cités ci-dessus, nous obtenons les meilleurs biclasses présentées dans le tableau suivant (Tab.3).

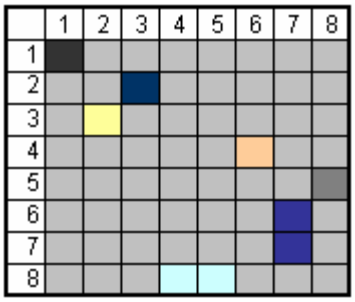

FIG. 3 – *Matrice des biclasses.*

Biclasse	Bkl/B	Bkl/Tkl
(1,1)	9	97%
(2,3)	10	26%
(3,2)	7	22%
(4,6)	6	18%
(5,8)	6	30%
(6,7)	3	65%
(7,7)	2	87%
(8,4)	2	26%
(8,5)	4	41%

TAB. 3 – *Informations sur les biclasses.*

L'examen de ces biclasses a pour objectif d'attribuer un thème à chaque groupe de pages. Par exemple, la biclasse (2, 3) composée de la classe 2 de descripteurs et la classe 3 des pages a pour thème « spécialités de cuisine » sachant que presque tous les descripteurs sont en relation avec l'alimentation (amande, crème, eau de vie, flamber, fruit, glacer, mirabelle, recette, purée…etc). Les biclasses (8,4) et (8,5) présentent le même thème « hébergement » puisque il s'agit de la même classe en ligne. Par conséquent, il est possible de regrouper les pages de la classe 4 et de la classe 5 dans une même classe (4+5). La nouvelle biclasse

obtenue après fusion des biclasses (8,4) et (8,5) a pour thème « informations sur les hôtels ou hébergement». Par contre, les biclasses (6,7) et (7,7) présentent en commun la même classe en colonne. Ainsi, les pages de la classe 7 auront deux thèmes différents « horaires et tarifs des lieux à visiter » et « réservation ».

(2,3)		(8,4)		(8,5)		(6,7)		(7,7)	
amande	specialites/caramel_htm	centre-ville	hebergement/aut2et_html	centre-ville	hebergement/aut0et_html	accompagnateur	visite/resguid1_html	âge	visite/resguid1_html
chantilly	specialites/mira_htm	classe	hebergement/mtz2et_html	classe	hebergement/aut3et_html	autocar	visite/reshand1_html	circuit	visite/reshand1_html
crème	specialites/mousse_htm	classement	hebergement/mtz3et_html	classement	hebergement/mtz_html	entrée	visite/resscol1_html	cour	visite/resscol1_html
eau-de-vie	specialites/origines_htm	distance	resto/resto_html	distance	hebergement/mtz0et_html	fermer		dimanche	
flamber	specialites/spec_htm	étoile	specialites/luth_html	étoile		guide		jour	
fruit	specialites/tartes_htm	formule		formule		heure		musée	
glacer		hôtel		hôtel		langue		nombre	
lait		nord		nord		mardi		office	
mirabelle		sud		sud		matin		or	
pâte		zone		zone		mobilité		personne	
purée						pays		reservation	
recette						réduction		tarif	
sucre						semaine		visiter	

FIG. 4 – *Exemple de biclasses.*

Le tableau suivant présente les meilleures biclasses et les thèmes qui leur sont associés.

Biclasses	Thèmes
(1,1)	Adresses utiles
(2,3)	Spécialités cuisine
(3,2)	Informations sur les services
(4,6)	Edifices et monuments
(5,8)	Calendrier
(6,7)	Horaires et tarifs des lieux à visiter
(7,7)	Réservation
(8,4)	Informations sur les hôtels (Hébergement)
(8,5)	

TAB. 4 – *Thèmes associés aux biclasses.*

6 Analyse factorielle des correspondances

Dans cette section, nous appliquons au tableau de contingence l'analyse factorielle des correspondances. La projection des pages et des descripteurs sur le même plan factoriel (figure 5) permet d'associer des groupes de pages à des groupes de descripteurs. A titre d'exemple, Le nuage de points situé à gauche est composé des descripteurs (points bleus) appartenant à la classe 8 de descripteurs et des pages (triangles rouges) appartenant aux classes 4 et 5 de pages. Ce nuage de points correspond aux deux biclasses (8,4) et (8,5) déterminées à l'aide de CROKI2. Le nuage de points en bas comporte les descripteurs appartenant à la classe 5 de descripteurs et les pages appartenant à la classe 8 de pages. Ce nuage correspond à la biclasse (5,8). L'ensemble des points au centre est composé des descripteurs appartenant à la classe 1 et la classe 3 et des pages appartenant à la classe 1 et la

classe 2 de pages. Ce nuage correspond aux deux biclasses (1,1) et (3,2). Le nuage situé en haut est un mélange de descripteurs et de pages appartenant aux biclasses (6,7), (7,7), (4,6) et (2,3). Ainsi, on trouve dans les résultats de l'analyse factorielle des correspondances certains résultats obtenus dans la section précédente.

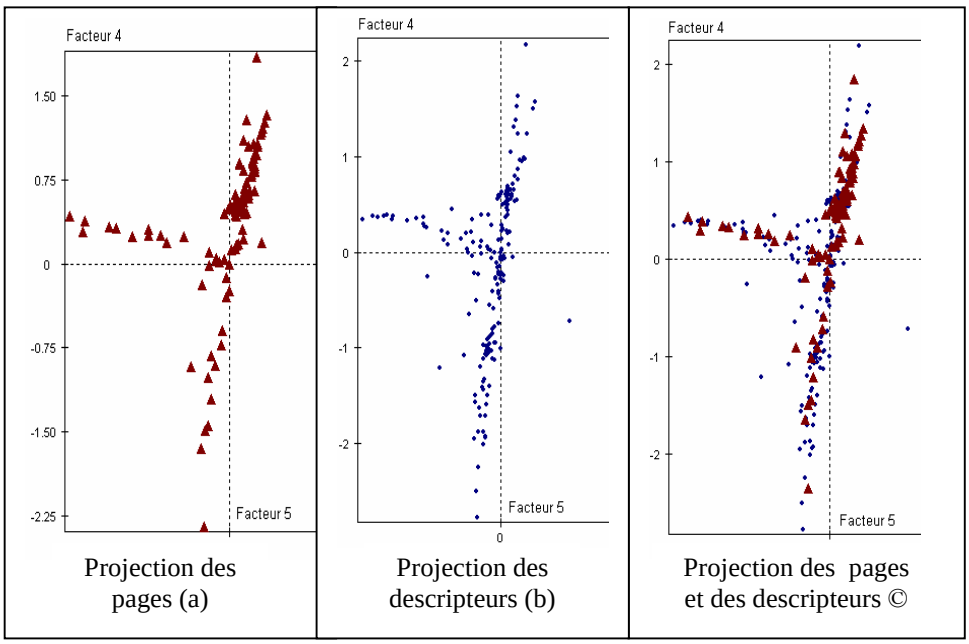

FIG. 5– *Projection des descripteurs et des pages sur les axes factoriels.*

7 Conclusion

Dans cet article nous avons proposé une approche d'analyse du contenu textuel d'un site Web basée sur la catégorisation simultanée des pages et des descripteurs de pages et indépendamment de l'usage du site. Les thèmes découverts vont servir par la suite à l'analyse du site du point de vue de ses usagers. Les résultats obtenus par cette approche prouvent son applicabilité sur des sites Web non volumineux.

Références

Blum, A. et T. Mitchell (1998). *Combining Labeled and Unlabeled Data with Cotraining*. Proceedings of the 11th Annual Conference on Computational Learning Theory.

Charrad, M., M. Ben Ahmed et Y. Lechevallier (2006). *Extraction des connaissances à partir des fichiers Logs*. Lille : Atelier fouille du Web EGC2006.

Da Silva, A., F. De Carvalho, Y. Lechevallier, B. Trousse et R. Verde (2007). *Classification simultanée des lignes et des colonnes d'un tableau de contingence : Application aux données d'usage du Web*. Colloque international de Statistique appliquée pour le développement en Afrique SADA'07, 1-6.

Forgy, E.W (1965). *Cluster analysis of multivariate data: efficiency versus interpretability of classifications*. Biometrics, 21, 768-769.

Govaert, G. (1983). Classification croisée. Thèse de doctorat d'état, Paris.

Hartigan, J. (1972). *Direct Clustering of a Data Matrix*. JASA, 67:337, 123-129.

Hartigan, J. (1975). *Clustering Algorithms*. John Wiley & Sons.

Lewis D. (1992). *Representation and Learning in Information Retrieval*. Thèse de doctorat, Université de Massachusetts.

Rakotomalala R ;(2005) TANAGRA*, une plate-forme d'expérimentation pour la fouille de données*. Revue MODULAD, .70-85.

Salton, G. et C. Buckley (1998). *Term-weighting Approaches in Automatic Text Retrieval*. Information Processing and Management, 24(5): 513-523.

Schmid H. (1994). *Probabilistic PARTOFSpeech Tagging Using Decision Tree.* In Proc. Of the International Conference on New Methods in Language Processing, 44-49

Scott, S. et S. Matwin (1999). *Feature Engineering for Text Classification*. San Francisco: Proceedings of ICML-99, 16th International Conference on Machine Learning, Morgan Kaufmann.

Stricker, M. (2000). *Réseaux de neurones pour le traitement automatique du langage : conception et réalisation de filtres d'informations*. Thèse de Doctorat de l'Université Pierre et Marie Curie, Paris VI.

Trousse, B., M.A. Aufaure, B. Le Grand, Y. Lechevallier and F. Masseglia (2007). *Web Usage Mining for Ontology Management*. In Nigro Hèctor Oscar, Gonzalez Cisaro Sandra Elisabeth G and Xodo Daniel Hugo editors, Data Mining with Ontologies: Implementations, Findings and Frameworks, Information Science Reference, 37-64.

Summary

In this paper, our aim is to analyse textual data of a web site. Our approach consists of three steps: Web pages classification, preprocessing of web pages content and block clustering. The first step consists in classifying web site pages into to major categories: auxiliary pages and content pages. In the second step, web pages content is preprocessed in order to select descriptors to represent each page in the web site. As a result, a matrix of web site pages and vectors of descriptors is constructed. In the last step, a simultaneous clustering is applied to rows and columns of this matrix to discover biclusters of pages and descriptors. An experiment on Metz tourism Web site shows that the approach is practical and efficient. The result of this experiment is a group of pages clusters where each cluster is represented by a theme. These clusters will help to analyse web site usage.