WCUM pour l'analyse d'un site Web

Malika Charrad [1] Yves Lechevallier[2] Gilbert Saporta[3]
Mohamed Ben Ahmed[4]

[1,4]Ecole Nationale des Sciences de l'Informatique, Tunis
[2]INRIA Rocquencourt, Paris
[1,3]Conservatoire National des Arts et Métiers, Paris

Extraction et gestion des connaissances, EGC'2010

Plan

Contexte

- Le Web : source **volumineuse** et **dynamique** des données.
- Augmentation exponentielle du nombre d'utilisateurs : croissance de 380,3 % entre 2000 et 2009 [1] (taux de pénétration = 25,6%).
- Croissance colossale du nombre de documents en ligne : plus de 3 millions de sites web sont crées mensuellement selon l'enquête de Netcraft [2].

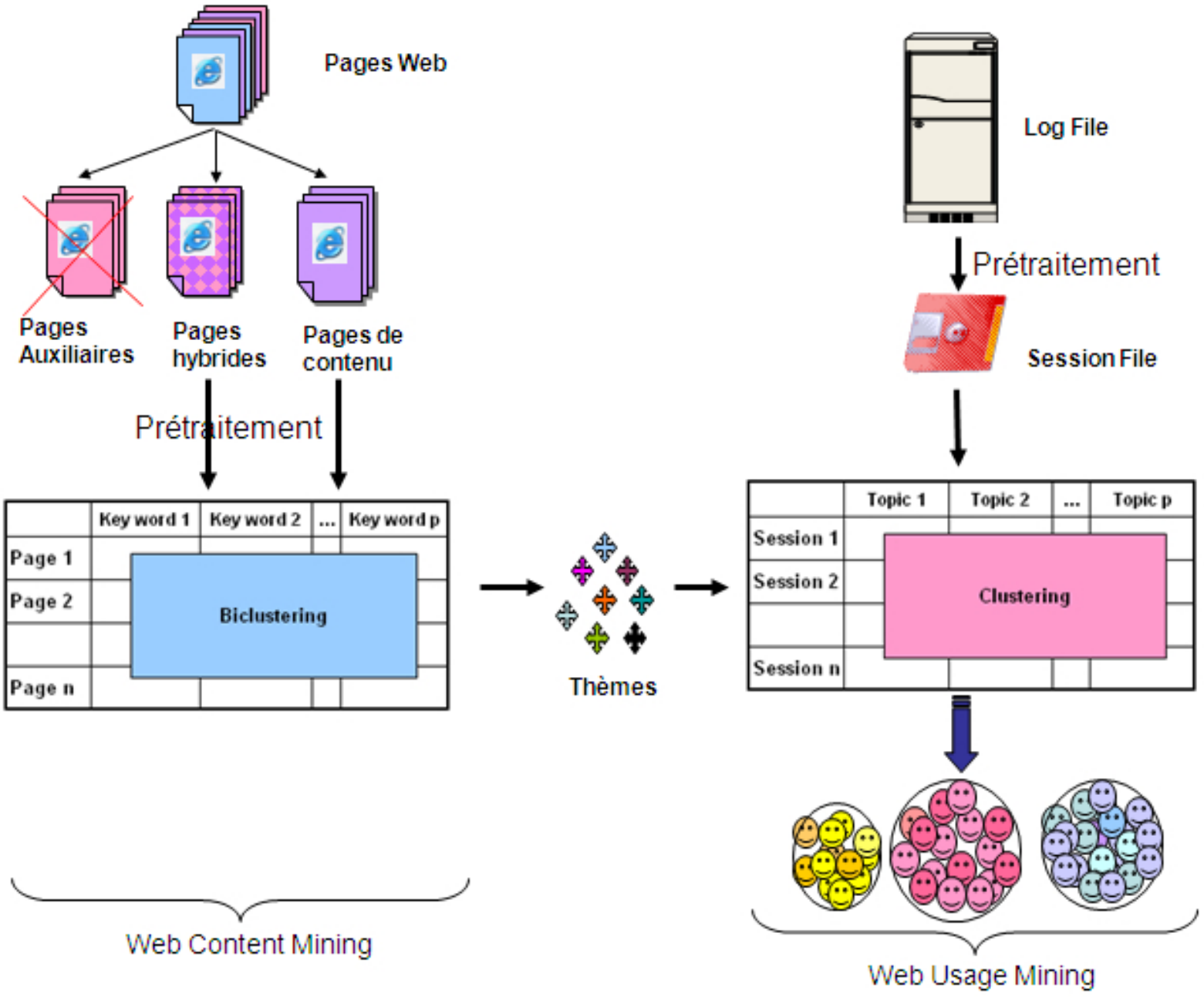
$\Longrightarrow$ L'abondance, la dynamicité et l'aspect polymorphe des ressources en ligne (texte, multimedia, base des données...) ont motivé d'importants efforts dans la recherche en WM.

1. www.internetworldstats.com
2. www.news.netcraft.com

Motivation

- La majorité des travaux en WM s'intéressent soit à l'analyse du contenu (WCM), soit à l'analyse de l'usage (WUM).
- Dans les rares travaux qui ont associé le WCM au WUM, seules les données issues des fichiers logs sont utilisés (Srivastava et al.,2000), (Zeng et al., 2002), (Liu et al., 2005) et (Koutsonikola et Vakali, 2009).

$\Rightarrow$ Nous proposons dans ce papier d'associer l'analyse du contenu à l'analyse de l'usage en utilisant le contenu textuel des pages et les données extraites des fichiers logs.


Pages Web
Pages Auxiliaires
Pages hybrides
Pages de contenu
Prétraitement
Key word 1
Key word 2
...
Key word p
Page 1
Page 2
Page n
Biclustering
Thèmes
Web Content Mining
Log File
Prétraitement
Session File
Topic 1
Topic 2
...
Topic p
Session 1
Session 2
Session n
Clustering
Web Usage Mining

Analyse textuelle d'un site Web

Typage des pages

- Pages auxiliaires : facilitent la navigation sur le site.
- Pages de contenu : pages contenant de l'information qui pourrait intéresser les visiteurs.
- Pages hybrides : pages de contenu utilisées aussi pour la navigation sur le site.

Méthodes de typage

- Typage par classification
- Typage en utilisant l'algorithme HITS (Kleinberg, 1999)

Analyse textuelle d'un site Web

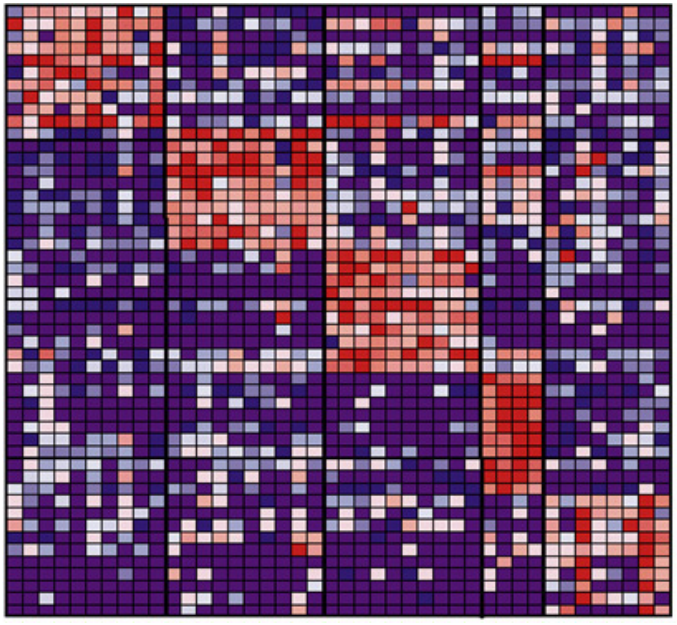
Prétraitement des textes

- Conversion des pages Web en fichiers textes
- Remplacement des images par leurs légendes
- Etiquetage et Lemmatisation à l'aide de TreeTagger
- Réduction de la dimension de l'espace de descripteurs par la suppression des formes de ponctuation, des mots vides (pronoms, prépositions, déterminants,...), les mots très fréquents, les mots très peu fréquents.

	Page 1	Page 2	...	Page m
Terme 1	5	0	...	8
Terme 2	1	1	...	4
...	...	...	...	...
Terme n	0	3	...	0

Block clustering

Les méthodes de Block clustering ou bipartitionnement s'appliquent **simultanément** sur les deux dimensions et produisent des blocs homogènes dans lesquels chaque individu est caractérisé par un sous-ensemble d'attributs et chaque attribut caractérise un sous-ensemble d'individus.



Pourquoi le Block clustering ?

- La majorité des travaux en Text mining sont basés sur la classification simple des documents ou des descripteurs.
- Dans le cas de la classification simple, chaque document appartenant à une classes de documents est décrits par tous les descripteurs et chaque descripteur appartenant à une classe de descripteurs caractérise tous les documents.
- Notre objectif est d'identifier des groupes de documents qui sont bien décrits par un sous-ensemble de descripteurs ce qui nécessite la construction des biclasses composés de pages et de descripteurs qui sont fortement correlés.

Algorithme CROKI2

- L'algorithme CROKI2 (classification **CRO**isée optimisant le **Khi2** du tableau de contingence) proposé par Govaert (1983) a pour objectif de trouver une partition P sur les lignes et une partition Q sur les colonnes telle que le Khi2 de contingence du nouveau tableau construit en regroupant les lignes et les colonnes suivant les partitions P et Q soit maximum.
- Le critère général à optimiser est : Le critère général à optimiser est :

$$\chi^2(P,Q) = \sum_{k=1}^{K}\sum_{l=1}^{L} \frac{(f_{kl} - f_{k.}f_{.l})^2}{f_{k.}f_{.l}}$$

- L'algorithme CROKI2 est basé sur l'algorithme Kmeans. Il nécessite donc de fixer le nombre de classes à priori.

Choix des biclasses

- Homogénéité de la biclasse : mesurée par la part d'inertie conservée par la biclasse par rapport à l'inertie initiale.

$$HB = (W_{kl}/T_{kl})$$

avec

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

et

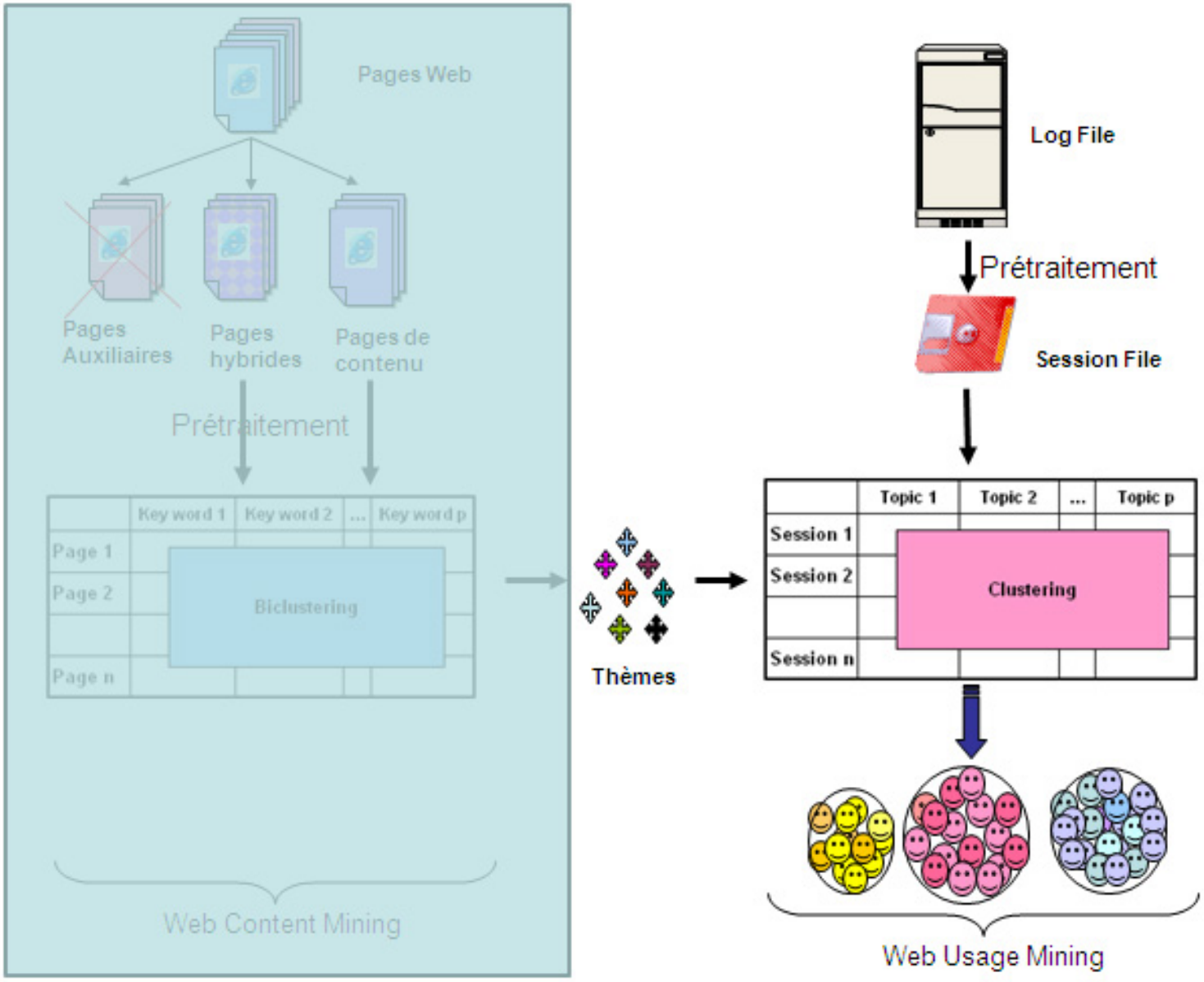
$$T_{kl} = \sum_{i\in P_k}\sum_{j\in Q_l} f_{i.}f_{.j}(\frac{f_{ij}}{f_{i.}f_{.j}-1})^2$$

- Pertinence de la biclasse : la part de l'inertie conservée par la biclasse dans l'inertie totale.

$$PB = W_{kl}/W$$

avec $W = \sum_{k,l} W_{kl}$

Analyse de l'usage



Prétraitement des fichiers Logs

- Nettoyage des données : consiste à filtrer les données inutiles à travers la suppression des requêtes ne faisant pas l'objet de l'analyse (requêtes invalides, requêtes aux images...) et celle provenant des robots Web.
- Transformation des données : Cette phase regroupe plusieurs tâches telles que l'identification des utilisateurs et des sessions et l'identification des visites.
  - L'identification des utilisateurs est effectuée par le couple (IP, User-Agent)
  - Une session est composée de l'ensemble de pages visitées par le même utilisateur durant la période d'analyse.
  - Une navigation (ou visite) est composée d'une série de requêtes séquentiellement ordonnées, effectuées pendant la même session et ne présentant pas de rupture de séquence de plus de 30 minutes (Kimball, 2000).

Classification floue

Ce choix de procéder à la classification floue au lieu de la classification simple est justifié par le fait qu'un utilisateur peut être intéressé par plusieurs thèmes pendant la même navigation. Par conséquent, une navigation peut être affectée à plusieurs classes à la fois mais avec des degrées différents.

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

Classification floue

Algorithme Fuzzy c-means

- Comme les autres algorithmes de classification non supervisée, il utilise un critère de minimisation des distances intra-classe et de maximisation des distances inter-classe, mais en donnant un certain degré d'appartenance à chaque classe pour chaque individu.
- Il génère les classes par un processus itératif en minimisant le critère suivant :

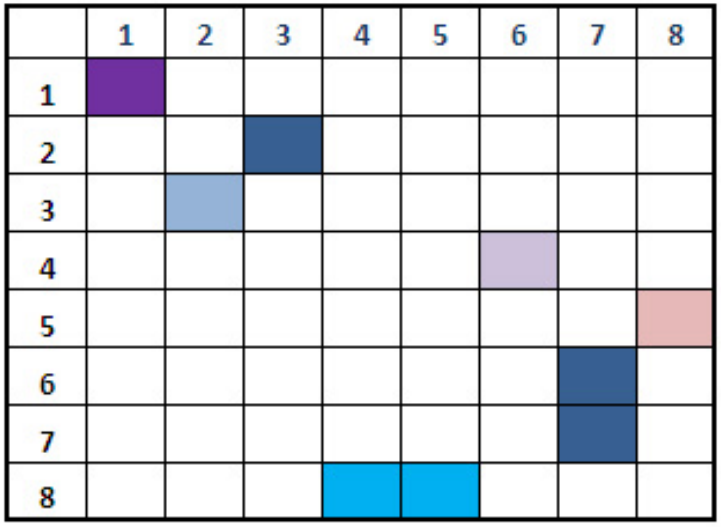
$$J_m = \sum_{j=1}^{c} \sum_{i=1}^{n} (\mu_{ij})^m d_j^2(x_i, \vartheta_j)$$

- Cet algorithme nécessite la connaissance préalable du nombre de clusters.

Application à un site Web de tourisme

Résultats de l'analyse textuelle du site

Le prétraitement et la sélection des descripteurs de pages aboutissent 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.



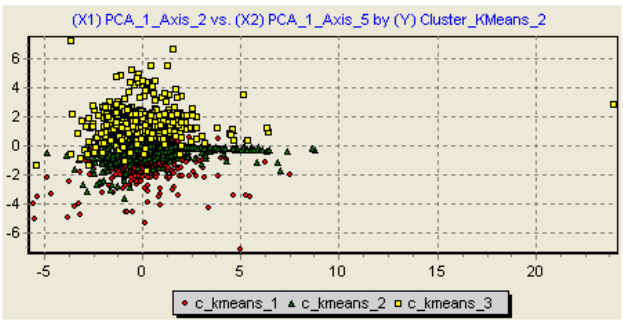
Biclasse	Wkl/W	Wkl/Tkl
(1,1)	9	97%
(2,3)	10	26%
(3,2)	7	22%
(4,6)	6	18%
(5,8)	6	30%
(6,7)	3	65%
(7,7)	2	87%
(8,4)	2	26%
(8,5)	4	41%

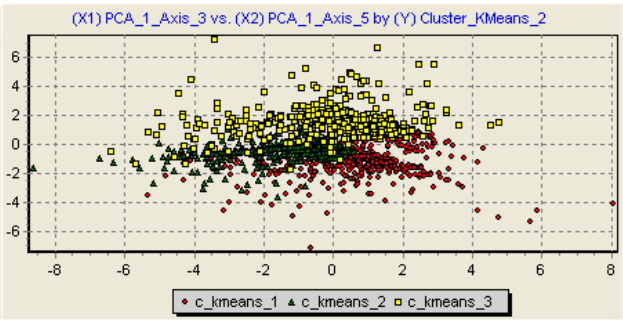
Résultats de l'analyse textuelle du site

Classes de descripteurs	Descripteurs	Thèmes
Classe 1	Arme, Art, Artiste, Château, Eglise, Exposition, Galerie, Guerre, Habit, Histoire, Maréchal, Monument, Moyen-Age, Palais, Pasteur, Peintre, Peinture, Renaissance, République, Siècle, Spectacle, Trésor	Histoire et Monuments
Classe 2	Amande, Crème, Eau, Flamber, Fruit, Gastronomie, Glacer, Lait, Mirabelle, Oeuf, Purée, Recette, Sucre,Hôspitalité	Recettes de cuisine

Résultats de l'analyse de l'usage

- La classe C1 est composée de navigations intéressés par les pages traitant du thème 4 (Informations utiles).
- Les navigations de la classe C2 sont par contre effectuées à des pages ayant pour thème "Recettes de cuisine", "histoire et monuments" et "Hôtels et Restaurants" (càd thème1, thème2 et thème6).
- La classe C3 regroupe les visites dont le motif est la recherche des adresses utiles (Thème 3), des manifestations et des activités culturelles (thème 5) et des informations utiles (thème4).





Conclusion

- Résumé des données textuelles contenues dans les pages web par le block clustering,
- Pallier la difficulté d'appliquer des méthodes de classification à des matrices creuses (*problem of matrices sparsity*),
- Suivi des changements dans le contenu et l'influence de ce changement sur le comportement des utilisateurs,
- Evaluation de la structure du site à partir de l'analyse du contenu et de l'analyse de l'usage,
- Indépendance totale de la méthode de classification aussi bien au niveau de l'analyse textuelle qu'au niveau de l'analyse de l'usage,
- Indépendance par rapport à la langue du site Web.

Bibliographie


Anli, A., Da Silva, A. and Lechevallier Y. : Spécification et développement du module d'analyse de l'usage. Livrable sp8.2 projet eiffel, INRIA Rocquencourt, 2008.


Charrad, M., Ben Ahmed,M., Lechevallier, Y. : Web Usage Mining : WWW Page classification from Log files. International Conference on Machine Intelligence, Tunisia, (2005).


Charrad, M., Lechevallier, Y., Saporta, G. and Ben Ahmed, M. : Web Content Data Mining : la classification croisée pour l'analyse textuelle d'un site Web. Revue des Nouvelles Technologies Informatiques (RNTI), Cépaduès-éditions, Vol.I, pp. 43–54, (2008).


Charrad, M., Lechevallier, Y., Saporta, G. and Ben Ahmed, M. : Le bi-partitionnement : Etat de l'art sur les approches et les algorithmes. Ecol'IA'08, Hammamet, Tunisie (2008).


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