Expérimentation d'une classification contextuelle de documents

Wahiba Bahsoun*, Meriem Beylagoun*

Wbahsoun@irit.fr, beylagoun@irit.fr

* Institut de Recherche en Informatique de Toulouse, Equipe SIG/RI

Université Paul Sabatier ,118 Route de Narbonne 31062 Toulouse Cedex 09- France

Mots-clés : Recherche d'information, Classification, Analyse de données.

Keywords : Information Retrival, Classification, data analysis.

Palabras clave : Busqueda de infomacion, Clasificaciôn, Analisis de datos.

Résumé : Cet article présente une étude expérimentale sur la classification thématique de documents textuels, plus précisément la classification des documents. L'objectif est d'offrir une présentation appropriée des documents sélectionnés par un système de recherche de d'information afin que l'utilisateur puisse les parcourir thématiquement en fonction de sa requête. L'outil tétralogie nous a permis de procéder à la validation expérimentale de notre approche. Les résultats obtenus montrent l'intérêt de notre approche.

1 Introduction

Les moteurs de recherche sont aujourd'hui le meilleur moyen pour accéder à cet immense gisement des pages Web. Le problème de ces moteurs ne réside pas tant dans le volume de la connaissance ni dans sa vitesse de croissance mais plutôt dans l'aptitude à retrouver la bonne information. Dans le cas précis du WEB [Lian, 2003], les moteurs de recherche retournent habituellement une liste linéaire contenant des milliers de pages susceptibles de répondre aux besoins en information exprimés par l'utilisateur [Salt, 1971],

Selon les résultats statistiques du groupe Naryanan de l'Université de Stanford(http://www-db.stanford.edu/), le nombre de pages similaires représente environ 22% du nombre de pages Web d'une part, et d'autres part, les informations sont parfois disponibles sur plusieurs types de médias et sous différents formats. Par exemple le même article avec le même contenu peut être présenté sous des formes différentes comme html, pdf, post-script. Les sujets abordés dans les différents documents sont multiples et certains sont éloignés des thématiques recherchées par l'utilisateur soit parce que ces dernières ne sont pas clairement exprimées dans la requête soit parce que le système n'a pas su les prendre en compte correctement. D'ailleurs, même dans le cas des documents répondant à la requête, c'est à dire les documents pertinents, ceux-ci peuvent aborder différents aspects de la pertinence.

Afin de mieux identifier et présenter à l'utilisateur ces différentes thématiques abordées dans les documents de manière générale, et les documents pertinents de manière particulière, nous proposons des solutions permettant de présenter à l'utilisateur une vision synthétique et globale des résultats de sa recherche. A cet effet, on s'oriente vers l'utilisation des méthodes d'analyse de données [Benz, 1993] et de classification.

En recherche d'information, la classification des documents peut être effectuée de deux façons. Elle peut être réalisée a priori sur toute la collection de document, ou bien à posteriori sur seulement la liste des documents restitués en réponse à une requête. C'est ce que nous appelons la classification contextuelle.

Notre travail rentre dans le cadre de la classification contextuelle. Le but est de proposer un moyen permettant de mieux présenter les résultats de recherche à l'utilisateur. Pour mieux présenter nous entendons, arriver à regrouper les documents en tenant compte du thème abordé dans ces documents.

La seconde section présente la problématique abordée dans cet article. Dans la section III nous détaillons l'approche adoptée et dans la section IV nous illustrons les différents résultats obtenus lors nos expérimentations.

2 Problématique

Avec le développement exponentiel d'Internet, les Systèmes de la recherche d'information (SRI) se trouvent face à de nouveaux défis pour l'accès à l'information, plus précisément présenter à l'utilisation une réponse en adéquation avec le problème posé dans la requête. Comme nous l'avons mentionné en introduction, la plupart des moteurs de recherche comme par exemple : Google présentent les résultats de recherche sous forme de liste linéaire. Cette liste peut comporter des documents pertinents et non pertinents traitant différents thèmes. Pour illustrer cette problématique, prenons à titre d'exemple, la requête « Virus » que nous avons lancé sur notre moteur de recherche Mercure sur une collection de documents issus de TREC [Voor, 1999], et que nous avons utilisée dans le cadre de notre expérimentation illustrée dans la section IV de cet article.

Cette requête sélectionne une liste de documents traitant les termes « virus » aussi bien des documents informatique que ceux qui traitent les termes « virus » liés à la médecine.

Avec cette représentation, l'utilisateur doit visualiser le contenu du document [Tami, 2000] [Bazi, 2002] pour voir le thème traité.

Ce que nous proposons permet de construire clairement ces deux différents thèmes sous formes de classes puis les présenter à l'utilisateur qui fera le choix du thème approprié à son besoin.

3 Classification contextuelle de documents

Le regroupement thématique de documents restitués en réponse à une requête peut être un moyen efficace pour identifier les différents thèmes véhiculés par les documents pertinents. . La problématique est de proposer à l'utilisateur un moyen permettant de mieux présenter les résultats restitués, de regrouper les documents en tenant compte du thème abordé dans le document .de sa recherche. Nos expérimentations ont été réalisées en simulant une recherche Web sur une collection de documents issue du programme TREC [Voor, 1999].

3.1 Approche proposée

L'approche adoptée est composée de trois étapes :

- 1ière étape - On soumet tout d'abord une requête à un moteur de recherche,
- 2ième étape - le moteur nous renvoie une liste ordonnée de documents,

- 3ième étape - nous restructurons cette liste pour mieux identifier et représenter les documents pertinents.

La troisième étape est évidemment la plus fondamentale. Dans cette étape les documents restitués sont d'abord analysés pour extraire les éléments importants puis traités par une fonction de classification. Trois techniques de classification ont été expérimentées : la Classification Ascendante Hiérarchique (CAH), l'Analyse Factorielle de Correspondance (AFC) et l'Analyse en Composantes Principales (ACP).

Nous avons utilisé les méthodes de classification proposées par l'outil tétralogie développé par l'équipe SIG/RI . Cet outil est basé sur l'analyse exploratoire des données et des méthodes de classification automatique. L'approche suivie par tétralogie est essentiellement basée sur la découverte de connaissance à partir de base de données de type bibliographique ou de brevets [Jpbe, 1973], [Karo, 2003], [Komp, 2004].

L'objectif de ce travail est de mesurer l'impact de ces techniques pour l'identification des documents pertinents et de leur similarité ou « dissimilarité » thématique.

Nos présentons dans ce qui suit les expérimentations que nous avons menées et les moyens mis en œuvre.

3.2 Contexte expérimental

Plus précisément nous avons utilisé :

- La collection GOV : nous utilisons une collection de données ciblée provenant d'un site Web américain gouvernemental. Pour bien comprendre le principe de l'extraction des données de la base de données, il est important de préciser comment ce dernier est structuré. Tous les documents et les requêtes de la base de données « GOV » sont présentés de la même façon. Cette structuration homogène des documents et des requêtes permet d'avoir une stratégie de récupération des données. Dans le cadre de nos expérimentations, le format de nos documents/ requêtes manipulés sont des fichiers HTML.

- Le moteur Mercure, Mercure est un SRI basé sur un réseau de neurones multicouches [Boug, 1992]. L'entrée du réseau représente la requête de l'utilisateur, la sortie représente les résultats de recherche obtenus. Les couches du réseau sont constituées par une couche de neurones termes qui représentent la couche d'entrée du réseau, une couche de neurones documents qui représente la couche de sortie et des couches cachées entre la couche des termes et la couche des documents.

- Et l’outil Tétralogie. Les fonctionnalités de l’outil Tétralogie sont utilisées pour la découverte d'informations implicites, élaborées voire cachées. Elles se déclinent en termes de *classification*, d'associations et de séquences, [Dousset, 1996]. [Fayy, 1996].

Nous précisons que les documents restitués sont analysés par Mercure avant de les soumettre à Tétralogie. Cela implique que le processus de recherche d’information : indexation, extraction et sélection des termes pertinents de la base documentaire utilisent le système « Mercure » pour restituer une liste ordonnée de documents dont le degré de pertinence par rapport à la requête a été calculé en fonction des mots-clés de façon classique.

4 Résultats des Expérimentations

Les expérimentations sont réalisées en deux étapes : l'**indexation** [Li, 1997], [Gery, 1999] et l '**analyse**. Pour effectuer les tests nous avons fixé la valeur 300 pour le nombre de documents à extraire et une dizaine de requêtes formulées par l'utilisateur.

4.1 Première étape : indexation

L'utilisateur lance une requête, Le moteur de recherche Mercure renvoie **300** documents. Chaque document est ensuite analysé pour y extraire une liste de termes importants sans les mots vides.

A l'issue de cette étape, on construit une matrice document - termes.

La figure 1 : présente un extrait d'une matrice construite à partir de la requête : « Children's literature ».

Tétralogie V7.0 Tableur 2D Fichier : Profil

	literatur	children	flp	book	kid	interlibrar	magazin	literacy	philadelphi	adult	child	reader
1	G03-55-3892 3	1	0	2	0	0	0	2	0	1	0	2
2	G41-03-3398 2	1	0	2	1	0	2	0	0	1	0	0
3	G21-31-1587 3	2	0	2	2	0	0	3	0	2	0	0
4	G05-19-0171 3	1	0	0	0	0	0	1	0	1	2	0
5	G24-28-2550 4	2	0	2	2	0	0	0	0	2	0	2
6	G08-44-0570 3	1	0	2	1	0	0	2	0	1	1	0
7	G28-10-3697 7	3	0	4	0	0	0	0	0	5	0	0
8	G02-18-0695 2	1	0	1	0	0	0	0	0	1	1	0
9	G21-47-1313 5	2	0	3	3	0	0	0	0	0	2	0
10	G32-85-3795 3	2	0	2	0	0	0	0	3	2	1	2
11	G16-40-1281 5	2	0	0	0	0	0	0	0	0	0	0
12	G42-10-3845 5	2	0	4	0	0	4	0	0	0	0	0
13	G20-01-1749 8	4	0	0	5	0	0	0	0	0	4	0
14	G21-31-2716 6	3	0	0	0	0	0	0	0	4	0	0
15	G11-68-1283 2	1	0	2	0	0	0	1	0	1	0	1
16	G12-24-3363 3	2	0	0	0	0	0	0	0	1	1	0
17	G07-93-3516 3	2	0	2	0	0	0	2	0	0	1	2
18	G01-76-2000 2	1	0	0	0	0	0	0	0	1	1	0
19	G21-49-2513 2	1	0	1	1	0	2	0	2	0	0	1
20	G42-74-2634 3	1	0	2	0	0	2	0	0	0	0	2
21	G37-31-3675 4	2	0	0	0	0	0	0	0	0	2	2
22	G11-58-0756 2	1	0	2	0	0	0	2	1	0	1	1
23	G12-08-0247 9	4	0	6	0	0	0	0	9	0	0	0

Figure 1 : Résultat de l'indexation pour la requête « Children's literature »

Dans cette Figure : - la Première colonne représente les noms des documents.
- les autres colonnes les termes extraits de ces documents
- la valeur de la matrice indique la fréquence du terme dans le document concerné.

4.2 Seconde étape : Analyse

Une fois que la matrice est construite, on passe à l'étape d'analyse de la matrice qui est le résultat de la première étape. Le but de cette étape est de faire ressortir toutes les connaissances endogènes, informations cachées ou implicites, en se basant sur les méthodes de classification fournies par le logiciel Tetralogie. Chacune des classes représente une carte[1] d'information. Pour visualiser les résultats des analyses, des cartes factorielles en 2D, 3D ou 4D animées de manière interactive ou automatique (zooms, choix des couleurs...) sont utilisées.

Nous présentons dans ce qui suit les résultats de l'analyse obtenus en utilisant la CAH et l'AFC .

4.2.1 Utilisation de la CAH

On commence nos expérimentations par l'application d'une classification ascendante hiérarchique (CAH), avec la distance euclidienne entre éléments sur la matrice précédemment créée. Les classes sont obtenues en agrégeant les deux éléments les plus proches. Ces classes sont à leur tour agrégées en fonction de leur distance deux à deux. Le processus est itératif, il est réitéré jusqu'à l'obtention d'une classe unique. Cette classification génère un arbre que l'on peut couper à différents niveaux pour obtenir des classes plus ou moins grandes ou plus en plus nombreuses comme illustrées dans la figure **2**

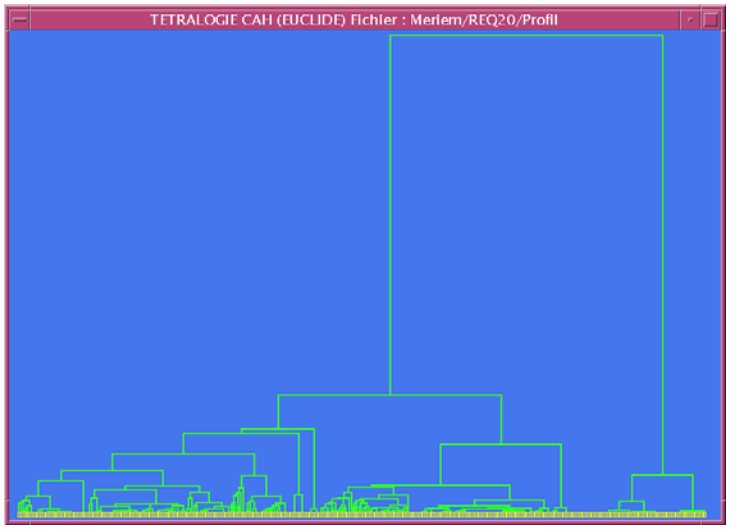

Figure **2** : Application de la CAH.

La figure **3** représente une partie de la CAH appliquée aux résultats d'évaluation de l'indexation à la requête « Children's literature ». Une fois la CAH construite une étape primordiale de l'approche est la détermination des classes de documents qui seront présentés à l'utilisateur. On choisit de couper l'arbre de la CAH à différents niveaux pour avoir une meilleure classification. Pour construire ces classes il suffit de couper la CAH à un niveau donné. Ceci n'est pas sans conséquence. En effet, si la CAH est coupée au niveau de la racine, on se retrouve avec une classe unique contenant tous les documents. Si au contraire on choisit de couper au niveau le plus bas, on se retrouve avec autant de classes que de documents. Le choix du niveau où l'on coupe la hiérarchie est donc important. Après différentes expérimentations notre choix s'est fixé sur un niveau correspondant à **5** classes.
La figure **3** montre que chaque classe lui est associée une liste de documents.

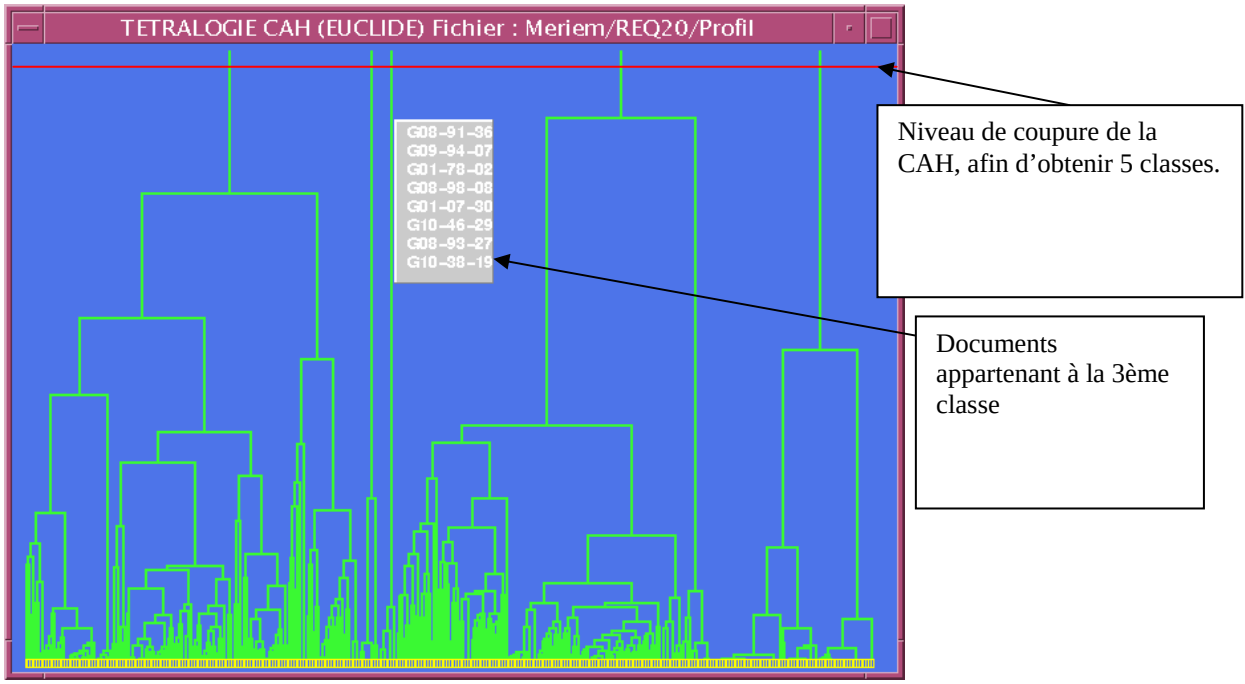

Figure 3: Représentation des documents sur une classe

La figure **4** présente pour chaque classe les termes importants qui lui sont affectés.

		literatur	children	fip	book	kid	interlibrar	magazin	literacy	philadelphi	adult	child
1	Classe1	692	354	0	48	80	9	7	12	8	175	205
2	Classe2	695	344	2	253	162	1	129	95	74	158	52
3	Classe3	208	112	479	168	195	345	245	228	240	2	107
4	Classe4	55	20	0	0	0	0	0	0	0	0	0
5	Classe5	58	24	0	0	42	0	0	0	0	0	0
6												

Figure 4 : Représentation des termes dans les classes

Une fois les classes identifiées, le but de notre démarche est d'arriver à les représenter graphiquement et rechercher les liens pouvant exister entre les classes. Pour ce faire on utilise une AFC. Plus précisément l'objectif est de présenter dans un même espace l'ensemble des termes de la requête et les classes identifiées.

4.2.2 Utilisation de l'AFC

L'outil Tétralogie offre l'avantage de permettre une rotation des axes, afin de découvrir des corrélations cachées entre les variables analysées. Le choix de la disposition des axes est réalisé de sorte que les partitions du nuage de points soient visualisées le plus clairement possible. Le but de cette manipulation est d'avoir une sorte de masque permettant d'identifier rapidement un groupe d'éléments situé prés d'un ensemble de termes représentant en l'occurrence la requête, en se référant à la position du terme sur la vue. Une fois le positionnement obtenu, on visualise ensuite le résultat complet de l'AFC, c'est-à-dire en considérant à la fois les termes et les documents. Ensuite on revient sur la vue des positionnements, et on exporte les positions de la première vue sur la deuxième vue, ce qui a pour effet de réorganiser la nouvelle vue en fonction des positions qui ont été transmises par la première vue.

4.3 Illustration

Pour montrer l'intérêt de notre approche, on considère la requête suivante « Virus », on effectue une recherche via le Moteur de Recherche Mercure, puis on effectue l'analyse des **300** documents sélectionnés. En appliquant la CAH et l'AFC, nous obtenons deux classes. Nous avons extrait **40** termes récapitulatifs de chaque classe. Ces termes sont représentés dans le tableau de la figure **5** ci-dessous.

Figure 5 : Les termes représentatifs pour les 2 classes.

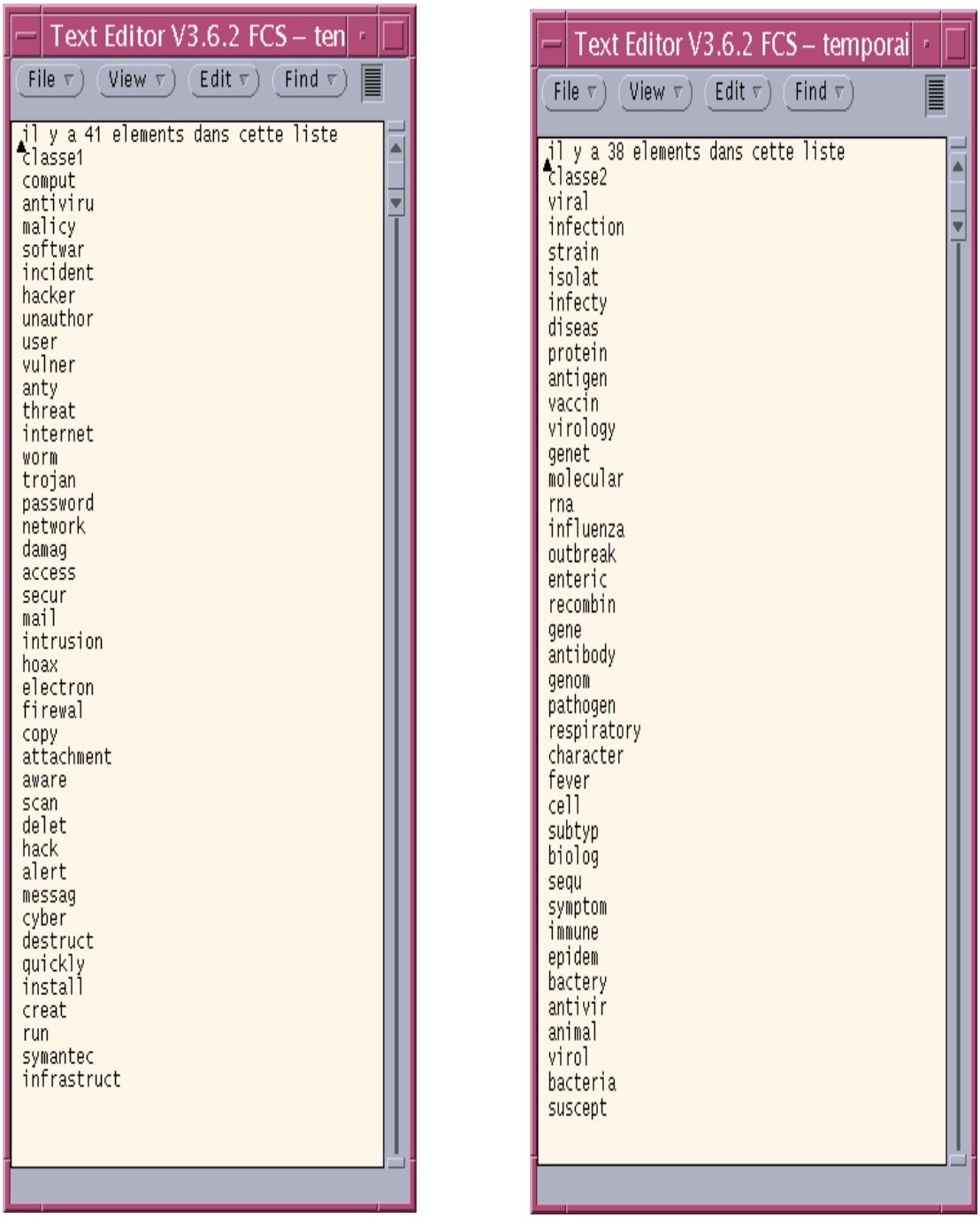

Figure 6 : Termes communs aux classes

En analysant le contenu des tableaux ci-dessus on constate que :

- la classe **1** est composée des termes qui sont tous liés à l'informatique comme « **comput**, **antivirus, user, password,…** ». On peut désigner cette classe par la thématique « Informatique »,.

- la classe **2** est composée par des termes liés à la médecine et au domaine biologique par exemple « **viral, infection, vaccin, genom**, … », alors on peut lui attribuer le thème de « Médecine ».

5 Conclusion et perspectives

L'article présente une technique de classification contextuelle de documents. Son objectif est d'offrir des solutions permettant une structuration appropriée des documents trouvés afin que l'utilisateur puisse les parcourir thématiquement en fonction de sa requête.

Nous avons montré que l'utilisation des méthodes de classification en l'occurrence la CAH et l'AFC permettent effectivement de regrouper les documents pertinents dans les mêmes classes. Ceci est un premier résultat intéressant car au delà de l'aspect présentation des résultats, le fait de regrouper les documents pertinents permet d'améliorer les performances du Système de Recherche d'Information.

Les perspectives envisageables à ce travail portent essentiellement sur les points suivants :

Un certain nombre d'opérations sont effectués manuellement, en l'occurrence le choix de la coupure de l'arbre de la CAH pour déterminer le nombre de classes, ceci n'est évident pas envisageable pour un utilisateur occasionnel, il serait judicieux de pouvoir le faire automatiquement.

De plus tous les documents restitués sont classés automatiquement, il serait également judicieux de trouver un moyen qui permet *de stocker*, *de mémoriser* les classes qui comportent des documents pertinents. Ceci par exemple en calculant une similitude entre la requête et le descripteur de la classe, la manière de construire ce descripteur reste à étudier.

6 Références bibliographiques

[Bazi, 2002] Mustapha Baziz, "Application des ontologies pour l'expansion des requêtes dans un Systeme de Recherche d'information. Rapport de stage DEA; Juin 2002.

[Bene, 2003] 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.

[Benz, 1973] J.P.Benzecri. l'analyse de données. Tome 1 La taxinomie, Tome 2 l'analyse des correspondances, Tome 3 linguistique et lexicologie, Dunod Edition, 1973

[Benz, 1992] J.P.Benzecri ; Correspondance analysis handbook, Marcel Dekker Ed ; NewYork, 1992

[Boug, 1992] M.Boughanem, Les systèmes de recherche d'information d'un modèle classique à un modèle connexionniste, thèse de doctorat de l'Université P.Sabatier, Toulouse, 1992.

[Dkak, 1998] T.Dkaki, B.Dousset, J.Mothe. Analyse d'informations issues du Xeb avec tétralogie. Veille stratégique, scientifique et technologique : VSST'98, p159-170, (Toulouse, France), octobre 1998.

[Fayy, 1996] U.M.Fayyad, G.Piatetsky-shapiro, P.Uthurusamy. advances in knowledge discovry and datamining. AAAI Press, ISBN ; ISBN 0-262-56097-6, 1996

[GSap, 1988] G.Saporta, *"Probabilités, analyse de données et statistique"*, ed. Technip, 1988

[Gery, 1999] Gery M., *"Smartweb : recherche de wones de pertinence sur le world wide web"*, Actes du 17ème Congrès INFORSID, LaGarde, 2-4 juin, pp 133-147, 1999

[Jpbe, 1973]. J.P.Benzécri, *La taxinomie* (T1) *"L'analyse des correspondances* (T2) *"*. Dunod Paris 1973.

[Kata, 2004] Katarzyna Wegrzyn-Wolska, « Le document numérique dynamique : une « étoile filante » dans l'espace documentaire ». Colloque EBSI-ENSSIB 2004.

[Karo, 2003] Said Karouach, « Visualisation interactives pour le découverte de connaissances : concepts, méthodes et outils », Thèse de doctorat, Université Paul Sabatier, Juillet 2003.

[Komp, 2004]. Nongdo Désiré Kompaore : Rapport de D.E.A « Utilisation de la découverte de connaissances dans un contexte de veille scientifique et technique », IRIT, SIG, 2004

[Lian, 2004] Liang DONG et Wahiba Bahsoun, "Analyse statistique pour la classification des pages WEB ». VSST (Veille stratégique scientifique et technologique), 2004.

[Lian, 2003] Dong Liang, « Analyse statistique pour la classification des pages Web », rapport de DEA, Paul Sabatier Toulouse, 2003.

[Li, 1997] Li Y., Rafski L., *"Beyond relevance ranking: hyperlink vector voting"*, 5th International Conference on Computer Assisted Information Retrieval, RIAO'97, Montréal (Canada), pp 648-651, 25-27 juin 1997.

[Moth, 1994] J. Mothe. Modèle Connexionniste pour la Recherche d'Information, Expansion Dirigée de Requêtes et Apprentissage, Thèse de l'Université Paul Sabatier Toulouse. 1994.

[Salt, 1971], G.Salton, 1971 « teh smart terieval system : experiments in automatic document processing » prentice-hall Inc, NJ.

[Salt, 1983]. G.Salton, Mj Macgill « extended booleen information retrieval, communication of the ACM », vol26, N12, pp 1022-1036, 1983.

[Tami, 2000] Linda Tamine. Déc. 2000 "*Optimisation de requêtes dans un Système de Recherche d'Information »*. Thèse de doctorat de l'université Paul Sabatier de Toulouse.