Clusterisation du Web en vue d'extraction de corpus homogènes

Camille Prime-Claverie, Michel Beigbeder, Thierry Lafouge

Clusterisation du Web en vue d'extraction de corpus homogènes

Camille Prime-Claverie[1], Michel Beigbeder[1], Thierry Lafouge[2]

[1] RIM, Ecole Nationale Supérieure des Mines de Saint Etienne, 158 cours Fauriel, 42023 Saint Etienne Cedex 2 (France)
[2] RECODOC, Université Claude Bernard Lyon 1, 43 bd du 11 novembre 1918, 69622 Villeurbanne Cedex (France)
prime@emse.fr, mbeig@emse.fr, lafouge@enssib.fr

RESUME. Les ressources disponibles sur le Web sont de plus en plus diverses aussi bien d'un point de vue thématique, qu'au niveau de leur type, de leur origine géographique, etc. Cependant, les outils de recherche ne prennent pas en compte cette hétérogénéité et ne proposent qu'un accès par mots-clés aux documents du web. Cet article présente une méthode basée sur les hyperliens, permettant d'extraire du graphe Web des sous-corpus de documents homogènes. L'expérience décrite ici utilise la méthode des co-citations et s'intéresse plus spécialement à la notion de genre (type) de document web.

ABSTRACT. Web resources are more and more different, not only regarding thematic content but also related to type of document, geographic origin, level, language, etc. However, web search engines do not take into account this heterogeneity and propose only a thematic access by keywords to the documents. This paper presents a method allowing to extract homogenous corpus of web documents. This method based on link analysis uses co-citation method and focuses more specially on the notion of type of web documents.

MOTS-CLES : méthode des co-citations ; graphe Web ; typologie des pages ; entropie.

KEYWORDS : co-citation method ; link analysis ; genre of web document ; entropy.

1. Introduction

La priorité des moteurs de recherche disponibles sur la toile est de retourner en un minimum de temps, le plus de pages web pertinentes sur un sujet donné. S'appuyant sur les techniques des systèmes de recherche d'information traditionnels (SRI), leur objectif est donc de retrouver et de ranger les pages par ordre de pertinence thématique. Cependant, contrairement aux bases de données traditionnelles, le Web est un magma d'information regroupant des documents hétérogènes à tout point de vue. Ainsi les utilisateurs le consultent avec des attentes et des objectifs bien différents. Prenons l'exemple d'un élève et d'un chercheur recherchant tous les deux de l'information sur la physique nucléaire. Le premier s'orientera avant tout vers des mémoires ou exposés en français d'un niveau vulgarisateur, alors que le second préférera des articles scientifiques probablement écrits en anglais, et pourquoi pas des appels à communication ou d'autres documents en relation avec son activité scientifique. Il paraît donc nécessaire de ne pas se limiter à la description thématique d'un document, mais de considérer aussi ses autres *directions*, comme son niveau, son origine géographique, son type (ou genre), etc. Ce problème, précédemment soulevé par Gravano [GRA 00], ne semble pas pris en compte par les moteurs de recherche généralistes.

Deux orientations sont possibles pour surmonter les difficultés de recherche d'information liées à l'hétérogénéité du Web. La première essaye de constituer des corpus de documents homogènes. Sur ce principe plusieurs outils de recherche spécialisés ont été créés. Ils ne prennent en compte qu'un ou plusieurs types de documents bien déterminés et n'indexent que ceux-ci. L'un des exemples est le moteur CiteSeer (maintenant appelé Research Index) [LAW 99] qui regroupe la plupart des articles scientifiques d'informatique disponibles sur la toile. La seconde orientation, plus ambitieuse, consiste à caractériser (c'est-à-dire indexer) l'ensemble des documents du Web pour une ou plusieurs directions. Gravano et al. [DIN 00] proposent une méthode pour déterminer l'origine géographique des pages web. Crowston et Williams [CRO 00], et Glover et al. [GLO 01] s'intéressent plus spécialement à la notion de genre (type) de document existant sur la toile. Les premiers étudient les genres de communications reproduits ou émergents sur la toile, comme les FAQ ou les homepages. Les seconds présentent une méthode automatique par apprentissage permettant de reconnaître certains types de documents. Plus récemment Kwasnik et al. [KWA 01] étudient comment la prise en compte du genre de document web peut améliorer la recherche d'information.

C'est la deuxième approche qui nous semble la plus intéressante et vers laquelle nous nous orientons. Elle engendre 3 questions : (i) quelles sont les directions des documents à prendre en compte pour améliorer la recherche d'information sur le Web ? (ii) Comment représenter ces directions (vocabulaire libre, langage contrôlé) ? (iii) et comment les renseigner ?

Pour indexer les documents web, trois types d'information peuvent être utilisées :

- le contenu lui-même des pages web : c'est-à-dire l'ensemble du code source de la page — le texte, les balises, les liens hypertextes, les liens vers les images ou d'autres ressources multimédias — la taille des fichiers, etc.

- le graphe créé par les liens hypertextes reliant les pages les unes aux autres.

- les données provenant de l'usage comme les fichiers de log, les cookies, etc.

Cette classification est proposée par la communauté du web mining [KOS 00]. Remarquons que les données relatives à l'usage sont impossibles à obtenir pour l'ensemble des sites. C'est pourquoi nous orientons notre recherche vers des méthodes utilisant séparément ou combinant les données issues du contenu et du graphe web.

Nous pensons que le graphe formé par les liens hypertextes est porteur d'information et que celui-ci peut être analysé afin de mieux comprendre l'univers du Web et bien sûr d'améliorer l'accès à son contenu. Dans la littérature antérieure l'information portée par le graphe se traduit en terme de :

- référence [BRI 98], [KLE 99], [SAV 96]. Différentes méthodes consistent à calculer le rang des documents réponses en fonction de leurs relations avec les autres. Par exemple, l'algorithme de classement implémenté dans le moteur Google [BRI 98] ordonne les documents en fonction leur visibilité sur le Web. Plus une page est citée par les autres, meilleur est son rang.

- liens sémantiques [KUM 99], [LAR 96] : les techniques mises en place essayent de rapprocher des documents similaires d'un point de vue thématique.

2. Notre approche et notre contribution

Nous voulons montrer qu'une étude approfondie du graphe web peut permettre de rapprocher non seulement des documents proches thématiquement, mais aussi des documents ayant des directions similaires. Dans cet article nous ne proposerons pas de méthode permettant d'indexer automatiquement les pages web pour des directions non thématiques, mais tenterons de montrer qu'il est possible d'extraire du Web des sous-ensembles homogènes pour certaines directions. Notre hypothèse est la suivante. Si un document A contient un lien hypertexte vers un document B, il existe (au moins pour le créateur de la page A) une association entre ces deux documents. Cette association se traduit par des valeurs identiques pour une ou plusieurs directions (c'est-à-dire que deux pages reliées dans le Web partagent au moins un point commun). Cet article présente à la fois une expérience et des premiers résultats permettant de mesurer l'ampleur de cette hypothèse et une réflexion sur une typologie possible des sites web et des pages web. En effet, dans le cadre de cette expérience nous nous sommes plus particulièrement intéressés au genre (type) des pages et des sites web.

3. Etapes du processus expérimental

Différentes études [KUM 99], [LAR 96], [PRI 01] ont déjà montré qu'il était possible de rapprocher des pages ou sites ayant des thématiques proches en utilisant les hyperliens. Qu'en est-il pour les autres directions ? L'expérience décrite ici tente de répondre à cette question. Elle s'est déroulée en quatre étapes. La première, la constitution du corpus, base de test pour notre travail, a pour objectif d'extraire un sous-ensemble du Web, regroupant des pages d'une même thématique mais ayant a priori des propriétés hétérogènes pour les autres directions. L'expérience sera ensuite menée en deux étapes indépendantes :

- la structuration automatique (clusterisation) du corpus utilisant une méthode basée sur les liens (relations entre les pages),

- l'indexation manuelle des pages pour quatre directions relatives au genre (type).

Notre objectif est de montrer que la clusterisation basée sur les hyperliens permet d'obtenir des groupes de pages (clusters) dans lesquels il existe au moins une direction commune, sorte de point commun que partagent toutes les pages d'un groupe. Notre quatrième étape est donc une analyse quantitative de chaque cluster pour mesurer l'homogénéité (la ressemblance) des pages qui le composent.

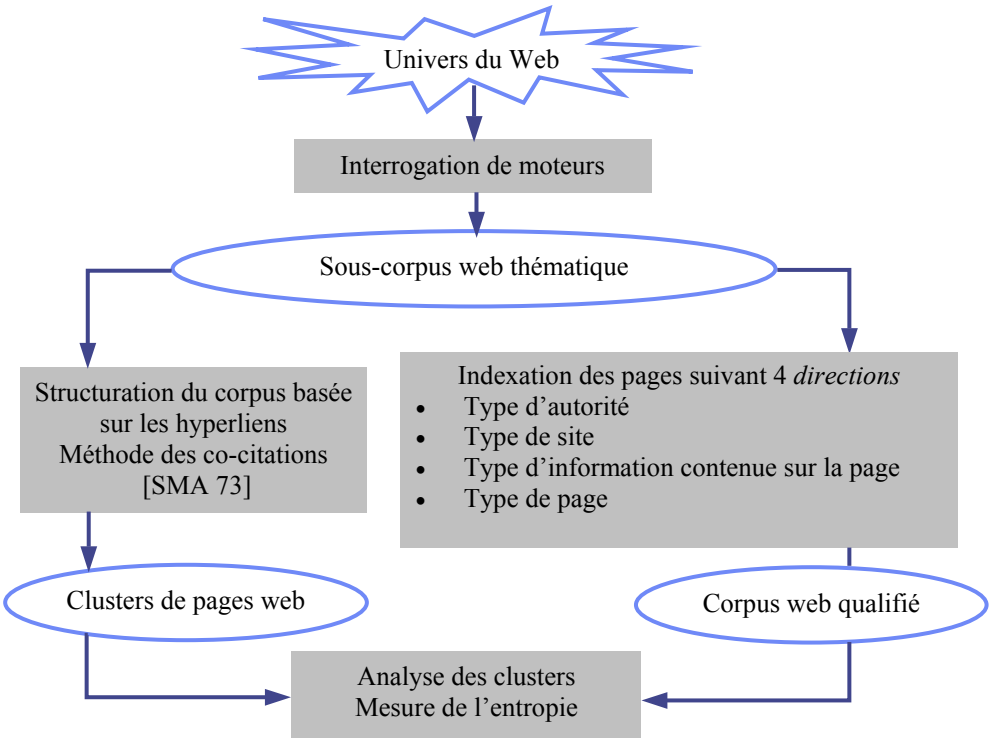

Figure 1. Etapes du processus expérimental

3.1. Etape 1 : Constitution du corpus

Le corpus de pages web, doit être mono-thématique, hétérogène pour les autres directions (type, niveau, etc.) et pour permettre une structuration basée sur les hyperliens doit comporter de nombreuses relations directes et indirectes entre les pages. Pour cette expérience, le thème retenu est l'astronomie, thème fédérateur engendrant de nombreuses publications web de genre différent et provenant de communautés très diverses : scientifiques, amateurs, étudiants, etc. D'autre part, pour faciliter l'indexation, et pour réduire le nombre de réponses, le corpus a été limité aux pages écrites en français. L'interrogation des 2 moteurs Google (http://www.google.com) et Hotbot (http://www.hotbot.lycos.com) en août 2001 avec la requête "astronomie" réduite aux pages françaises, a permis d'obtenir un ensemble de 1541 pages web différentes, noté A dans la suite de l'article. De plus, pour appliquer la méthode de clusterisation choisie (méthode des co-citations voir § suivant), il est nécessaire d'obtenir pour chacune des 1541 pages ses « pères », c'est-à-dire l'ensemble des pages pointant vers elle. 18714 pages « pères » ont été retrouvées grâce à la fonction *link* autorisée par les moteurs Google et Hotbot, cet ensemble sera noté B.

Nous avons été confrontés aux limites dues aux moteurs de recherche. En effet, ceux-ci ne renvoient pas toute l'information disponible. Par exemple, pour la requête "astronomie", Google indique "environ 54000 réponses" mais n'en affiche que 540. Ceci peut s'expliquer par des raisons pratiques (gain de temps), mais il existe aussi une volonté de la part des concepteurs des moteurs de ne pas dévoiler la totalité de leurs informations en particulier pour les requêtes utilisant la fonction *link* [BAR 01]. Il est donc impossible d'obtenir avec ces outils un sous-graphe exhaustif du graphe web.

3.2. Etape 2 : Structuration du corpus par la méthode des co-citations

La méthode des co-citations, utilisée en bibliométrie depuis 1973 [MAR 73] [SMA 73], a pour objectif de créer à partir d'articles scientifiques d'un même domaine de recherche, et plus précisément de leurs références bibliographiques, des cartes relationnelles de documents ou d'auteurs qui reflètent à la fois les liens sociologiques et thématiques de ce domaine. Cette méthode repose sur l'hypothèse que deux références bibliographiques de date quelconque, fréquemment citées ensemble ont une parité thématique. Comme dans le réseau des publications scientifiques [GAR 72] un lien hypertexte peut matérialiser une citation et indiquer une relation intéressante entre la page d'origine et la page vers laquelle il pointe. Plusieurs bibliométres (Ingwersen et al. [ING 98], [BJO 01], Aguillo [AGU 99], Rousseau [ROU 97], Egghe [EGG 00], Boubourides [BOU 99]) proposent des équivalences entre les concepts établis en bibliométrie et le graphe du Web. Plus particulièrement, Larson [LAR 96], Pitkow et Pirolli [PIT 97], Prime et al. [PRI 01] se sont intéressés à la transposition de la méthode des co-citations de documents pour caractériser les univers du Web. Ils mettent en évidence les limites théoriques

et techniques de l'analogie, mais ont montré l'intérêt de la structuration pour rapprocher thématiquement les pages. Une des limites de cette analogie est de considérer tous les liens hypertextes comme des liens de citation ou de référence. En effet, il faut aussi prendre en compte les liens de publicité, mais surtout ceux qui servent à se déplacer dans un même site web : les liens de navigation interne. C'est pourquoi pour cette expérience nous avons supprimé tous les liens intra-serveurs entre les pages pères et les pages de l'ensemble A, espérant ainsi supprimer la majorité des liens de navigation. Après cette élimination notre ensemble B ne contient plus que 11632 pages distinctes.

La première phase de la méthode consiste à calculer pour chaque couple de pages de l'ensemble A leur fréquence de co-citation notée Cij, c'est-à-dire, le nombre de fois où ces deux pages sont citées ensemble par des pages de l'ensemble B. Les résultats sont inscrits dans une matrice de co-citation, matrice carrée symétrique. La diagonale contient les occurrences de citation des pages de A notées Ci, c'est à dire le nombre de pages de B citant une page de A. Parmi les 1541 pages de A seules 198 sont cocitées, c'est-à-dire qu'elles sont citées avec une autre page de A.

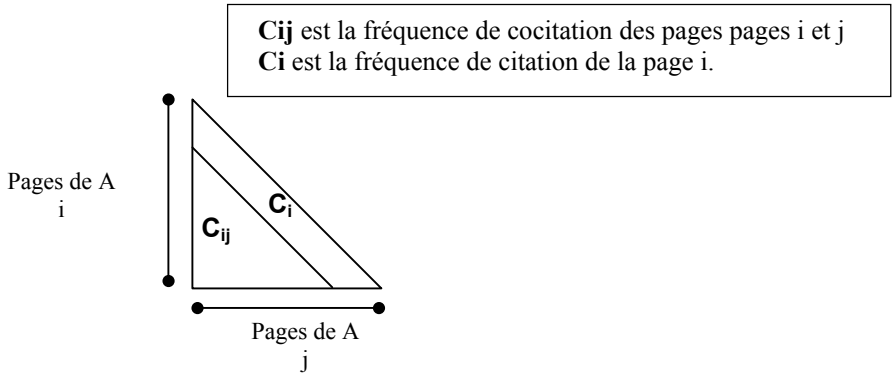

Figure 2. Matrice de co-citation

Nous déterminons ensuite la proximité entre deux pages, avec un indice similarité qui traduit mathématiquement l'idée suivante : deux éléments sont proches, si par rapport à leurs fréquences respectives, leur fréquence de co-occurrence est importante. Il existe plusieurs indices possibles. Par convention, ils varient de 0 à 1, sont égaux à 1 lorsque les deux éléments apparaissent toujours ensemble, et à 0, lorsque ceux-ci n'apparaissent jamais ensemble. Comme la distance, les indices de similarité sont symétriques, I(i,j) = I(j,i). Par contre, ils ne vérifient pas l'inégalité triangulaire. L'indice de similarité choisi est l'indice d'équivalence, sans aucune pondération,

$$E_{ij}=\frac{C_{ij}^2}{C_i \times C_j}\ .$$

Le regroupement des références les plus proches utilise des méthodes de classification automatique issues de l'analyse de données. Il s'agit d'une classification hiérarchique ascendante [HAR 75]. Parmi les différentes méthodes possibles (le simple lien, le lien complet, le chaînage moyen) nous avons choisi la méthode du lien complet. Au seuil le plus bas, avant que toutes les pages ne soient regroupées ensemble, nous avons 54 clusters de 2 à 8 pages et 38 singletons.

3.3. Etape 3 : Indexation manuelle du corpus suivant 4 directions liées au genre

Dans la plupart des systèmes de recherche d'information disponibles sur le Web, les unités informationnelles retournées aux utilisateurs sont les pages web. Celles-ci, nœuds du réseau hypertexte Web, représentent généralement des documents élémentaires dont le contenu exprime un nombre limité d'idées [BAL 96]. Ainsi, elles sont souvent autonomes d'un point de vue sémantique. Cependant, elles ne suffisent pas toujours pour comprendre et appréhender le document dont elles font partie, et ne permettent pas de l'indexer correctement, c'est-à-dire de répondre aux questions : de quoi parle ce document, pour qui, dans quel but ? etc. De plus, la notion même de document sur le Web, est une question en soi. Selon le principe de multiplicité et d'emboîtement énoncé par P. Levy [LEV 90], un document web peut contenir d'autres documents web [DOA 99]. D'autre part, les nœuds du réseau représentent des unités documentaires de niveaux différents (principe d'hétérogénéité), et ne sont pas comparables. Bien que nous ne soyons pas capables de définir clairement ce qu'est un document web, et encore moins les repérer, il existe dans le Web des ensembles de pages, homogènes d'un point de vue documentaire et repérables facilement. Il s'agit des sites web, ensembles de pages cohérents (objectifs et thèmes communs), créés et maintenus par une même autorité. Au niveau de la forme, les pages d'un site partagent la même charte graphique et ceux-ci possèdent toujours une page d'accueil, point d'entrée permettant d'atteindre les ressources du site.

Nous choisissons donc d'indexer chaque page en utilisant à la fois l'information contenue sur celle-ci et celle se trouvant sur le site dont elle est fait partie. Compte tenu des contraintes lors de la constitution de corpus, certaines de nos directions sont déjà fixées ou varient faiblement (la langue, l'origine géographique par exemple). Nous avons donc limité notre étude aux directions relatives au genre (type). Nous en avons définit quatre, qui sont : le type d'autorité responsable du site, le type de site, le type d'information contenue sur la page et le type de page.

- Type d'autorité : pour mieux comprendre l'apport informationnel d'un site, savoir qui est à l'initiative de sa création peut être un indice important. Nous avons distingué : *l'institution, l'entreprise, l'association et la personne individuelle*.
- Type de site : Il dépend du rôle informationnel que veut jouer le site. Nous avons recensé 4 types distincts.

- Le plus courant, le site vitrine (*homeserveur*) favorise l'information autodescriptive, celle décrivant l'autorité responsable du site. Sorte de « plaquette », l'objectif premier est de se présenter. Les thèmes abordés en priorité sont : Qui sommes-nous ? Nos activités, nos produits, nos partenaires, comment nous joindre etc. Cependant, ces sites peuvent aussi héberger dans des niveaux inférieurs (à plusieurs « clics » de la page d'accueil) des documents non autodescriptifs.
- Le *site de recherche* propose un accès aux ressources du Web. Les exemples les plus évidents sont les moteurs et les annuaires généraux. Les moteurs spécialisés n'indexant qu'un seul type de documents comme Citeseer ou un seul type de médium (moteurs de recherche d'images) sont aussi des sites de recherche.
- Se comportant comme un éditeur, le *site de ressources* organise et propose ses ressources propres (contrairement aux sites de recherche). Ils se présentent souvent comme des bibliothèques ou des bases de données.
- Les *services web* proposent des services liés à la vie sur le Web et l'Internet, comme des messageries, forums de news, etc.

- Type d'information contenue sur la page : nous avons l'information *autodescriptive*, relative à l'initiateur du site, et l'information *non autodescriptive*.
- Type de page : il dépend des caractéristiques physiques de la page. Nous en avons retenu 5 : la *page d'accueil*, les *portails* (pages comportant de nombreux liens externes), les *index* organisant les ressources internes d'un site (nombreux liens internes), les *pages de contenu* (plus de texte que de liens), les pages de *formulaires*.

Deux difficultés ont été rencontrées lors de l'indexation. La première provient du manque d'information autodescriptive, en particulier sur certains sites de ressources et de recherche où il s'est avéré impossible de savoir qui se trouvait à l'origine du site. La deuxième est apparue lorsqu'une page semblait appartenir à plusieurs catégories, en particulier pour le type de site. Certains sites proposent d'importantes bases de données concernant les produits ou les services de l'autorité qui les a créés (comme le site de la SNCF pour les horaires de trains). Nous les considérons comme des homeserveurs et non comme des sites de ressources, car l'information proposée est relative à leur activité, sauf pour les sites dont l'activité de l'initiateur concerne l'information et la communication. Ainsi, tous les sites d'éditeurs (*Encyclopedia Universalis* par exemple), de journaux (*Le Monde*) proposant en priorité et en grande majorité de l'information documentaire (accès aux articles) sont des sites de ressources. Voici les résultats de l'indexation manuelle des quatre directions pour notre ensemble A.

Type d'autorité	nb pages	type site	nb pages
Association	57	Homeserveur	125
Entreprise	42	Site de recherche	22
Institution	39	Site de ressources	39
Personne	37	Service Web	5
Indéterminé	23	Indéterminé	7
Type page	**nb pages**	**Type information**	**nb pages**
Page d'accueil	131	Autodescriptif	104
Page de contenu	28	Non autodescriptif	16
Index	13	Indéterminé	78
Indéterminé	8		
Portail	17		
Formulaire	1		

Table 1. Résultats de l'indexation

3.4. Etape 4 : Analyse des clusters

Le résultat de l'étape 2 est une partition de l'ensemble de départ A. Nous allons comparer la distribution des différentes valeurs pour les 4 variables dans le corpus complet et dans chaque cluster. Nous pouvons mesurer la diversité de chaque ensemble en utilisant l'entropie de l'information de Shannon [SHA 48]. Rappelons que l'entropie d'un système se calcule par la formule suivante, où N est le nombre d'éléments du système, S le nombre de valeurs que peuvent prendre les éléments, et Ni l'effectif de chaque valeur,

$$H = -\sum_{i=1}^{S} \frac{N_i}{N} \ln \frac{N_i}{N} .$$

Elle permet de mesurer l'apport informationnel d'un système. « Plus un système est composé d'un grand nombre d'éléments différents, plus sa quantité d'information est grande, car plus grande est son improbabilité de le constituer tel qu'il est en assemblant au hasard ses constituants [ATL 79] ». Notons que l'entropie est nulle lorsque tous les éléments ont la même valeur, et maximale lorsqu'ils sont tous différents. La redondance [MAR 58] normalise la fonction d'entropie, et mesure l'ordre d'un système plutôt que son désordre : $R = (H_{max} - H) / (H_{max} - H_{min})$., elle varie de 0 à 1, et est égale à 1 lorsque l'entropie du système est minimum (le système est le plus ordonné possible) et nulle lorsque que H est maximum[1]. Le

[1] Pour des distributions discrètes Hmax tend vers ln(S) lorsque N est grand. Pour N petit il est préférable de calculer Hmax comme suit, avec q et r résultat de la division euclidienne de N par S,

$$H = -r\frac{q+1}{N}\ln\left(\frac{q+1}{N}\right) - (S-r)\frac{q}{N}\ln\left(\frac{q}{N}\right) .$$

tableau ci-dessous nous donne pour chacune des directions l'entropie et la redondance de l'ensemble A.

Direction	nb valeurs	H	H max	Redondance (R_A)
Type d'autorité	5	1,57	1,61	0,024
Type de site	5	1,06	1,61	0,34
Type de page	6	0,88	1,79	0,51
Type d'information	3	0,91	1,09	0,17

Table 2 : Valeurs d'entropie et de redondance de l'ensemble A pour les 4 directions.

Les valeurs de redondance des quatre directions ont aussi été calculées pour chaque cluster. Nous considérons que la classification découpe de manière ordonnée l'ensemble A pour une direction donnée si une majorité de clusters ont des valeurs de redondance significativement supérieure à celle de A , en pratique supérieure à R_A+0,2, et noté $R_c >> R_A$. Toutefois, rappelons que cet ensemble n'est pas distribué de la même manière pour les 4 directions, et donc que la probabilité d'uniformité dans certains clusters est relativement forte. Ceux-ci auraient pu être formé de manière aléatoire, sans intervention particulière de la classification basée sur les liens. C'est pourquoi nous avons mesuré pour les 4 directions en utilisant la distance du khi2, l'écart entre la distribution de chaque cluster avec la distribution du corpus ramenée à la taille du cluster (qui serait la distribution des modalités tirées au hasard), appelée *distribution théorique*. Ne pouvant pas pratiquer de test statistique sur des effectifs aussi petits, nous avons déterminé graphiquement le seuil à partir duquel nous pensons que les clusters sont assez distants de leurs distributions théoriques pour ne pas avoir été générés aléatoirement. Nous obtenons pour les différentes directions les résultats mentionnés dans la table 3.

		Type d'autorité		Type de serveur		Type de page		Type d'information	
		nb C*	nb P*	nb C	nb P	nb C	nb P	nb C	nb P
Important écart à leur distribution théorique	$R_{cluster} >> R_A$	25	**86**	14	**53**	7	20	32	**113**
	($R_{cluster}$ =1)	*(20)*	*(63)*	*(9)*	*(26)*	*(6)*	*(12)*	*(24)*	*(70)*
	$R_{cluster}$ < ou proche de R_A	0	0	8	17	17	**39**	6	12
Faible écart à la distribution théorique. Cluster alétoire ?	$R_{cluster} >> R_A$	10	34	21	65	20	68	9	21
	$R_{cluster}$ < ou proche de R_A	19	40	11	25	10	33	7	13
Total		54	160**	54	160	54	160	54	160

* nb C : nombre de clusters ; nb P : nombre de pages

** 160 pages regroupées en 54 clusters et 38 singletons.

Table 3. Tableau des résultats

Nous pouvons lire dans ce tableau que :

- pour les 3 directions, type d'autorité, type de serveur et type d'information plus de 70% des pages appartiennent à des clusters où la redondance est nettement supérieure à celle de A (lignes 1 et 4 du tableau). De plus, parmi les clusters ayant un écart important à leurs distributions théoriques (lignes 1 à 3), on constate que le nombre de pages appartenant à des clusters où la redondance est inférieure à celle de A reste faible (ligne 3).
- pour la direction type de page, seules 55% des pages appartiennent à des clusters où la redondance est nettement supérieure à R_A (lignes 1 et 4), et la majorité d'entres elles appartiennent à des clusters ayant un faible écart à leur distribution théorique (ces clusters auraient donc pu être formés de manière aléatoire) (ligne 4). Parmi les clusters ayant un écart important à leurs distributions théoriques (lignes 1 à 3), une très grande majorité ont une redondance inférieure ou proche de R_A.

D'autre part, les résultats ont montré également que 40 clusters (115 pages) ont au moins une direction totalement ordonnée (redondance égale à 1).

4. Discussion

L'information portée par le graphe web est à la fois riche et complexe. Nous avons vu que celui-ci supportait des informations sémantiques et de références, nous pouvons dire qu'il véhicule aussi de l'information liée à la typologie des documents web. Cependant, cette typologie ne se traduit pas par des caractéristiques physiques des pages qui sont spécifiées dans l'expérience par la direction type de page [PIR 96]. En effet, l'hypertextualisation produit des documents découpés en nœuds élémentaires hétérogènes, et les règles de découpage particulièrement sur le Web, ne sont pas normalisés. Si bien que la co-citation de documents homogènes n'engendre pas forcément la co-citation de pages se ressemblant physiquement.

D'autre part, nous remarquons qu'à l'intérieur des clusters le type d'information reste stable. Il existe peu de clusters mélangeant à la fois de l'information autodescriptive et non-autodescriptive. En analysant plus loin le contenu de nos clusters, nous constatons aussi que lorsque l'information est de type autodescriptive, et que les pages du clusters sont hébergés sur des homeserveurs, l'autorité du site est du même type (institutions, entreprises, etc.). Ainsi, nous avons plusieurs clusters regroupant des centres de recherche, d'autres regroupant des club d'amateurs, etc. Par contre, lorsque le cluster regroupe des pages pour lesquelles l'information est non-autodescriptive, le type d'autorité du site varie davantage. Ceci met en évidence que le Web est un lieu d'expression qui mêle à la fois des acteurs et des documents. En citant une page web qui décrit l'autorité qui l'a créé, ce n'est pas l'apport documentaire de la page qui est pointée, mais plutôt son initiateur en temps qu'acteur, non pas du cybermonde, mais du monde réel [ROS 99]. L'étude du graphe par la méthode des co-citations permet donc d'extraire des sous-corpus de documents relativement homogènes, mais surtout d'identifier et de distinguer des réseaux d'acteurs et des réseaux de documents.

5. Conclusion

Cette étude exploratoire visait l'extraction de sous-ensembles web homogènes pour certaines directions avec la méthode des co-citations. Dans le cadre de cette expérience, nous nous sommes plus particulièrement intéressés aux directions liées à typologie des documents web. Les résultats de l'étude sont assez encourageants et montrent que cette structuration permet de détecter à la fois des réseaux d'acteurs et des réseaux de documents. Les limites rencontrées sont dues au manque d'information retournée par les moteurs de recherche, et nous pensons que l'obtention d'un sous-graphe exhaustif du Web permettrait de clusteriser davantage de pages. Cette méthode permet de mieux comprendre et appréhender l'univers du Web et ouvre des perspectives pour améliorer la recherche d'information et d'envisager une indexation liée à la typologie des pages des clusters par des méthodes automatique ou semi-automatique.

6. Bibliographie

[AGU 99] Aguillo I., "Statistical Indicators on the Internet : The European Science Technology Industry System in the World-Wide Web" at <http://diotima.math.upatras.gr/weborg/aguillo2>, 1999.

[ATL 79] Atlan H., *Entre le cristal et la fumée*, Essai sur l'organisation du vivant, Seuil, 1979, p.45.

[BAL 96] Balpe J.P., Lelu A., Papy F., Saleh I., *Techniques avancées pour l'hypertexte*, éditions Hermès, 1996.

[BAR 01] Bar-Ilan J., "How Much Information the Search Engines Disclose on the Links to a Web Page ? A Case Study of the "Cybermetrics" Home Page", *Proceedings of the 8th International Conference on Scientometrics and Infometrics, ISSI 2001, Sydney, Australia, July 16-20, 2001,* p. 63-73.

[BJO 01] Björneborn L., Ingwersen P., "Perspectives of Webometrics", *Scientometrics*, vol. 50 (1) 2001, p. 65-82.

[BOU 99] Boudourides M., Sigrist B., Alevizos P., "Webometrics and self-organisation of the Europeans information society", *at <http://hyperion.math.upatras.gr/webometrics>* 1999.

[BRI 98] Brin S., Page L., "The Anatomy of a Large-Scale Hypertextual Web Search Engine", Proc. 7th International Worl Wide Web Conference, 1998.

[CRO 00] Crowston K., Williams M., "Reproduced and Emergent Genres of Communication on the World Wide Web", *The Information Society*, vol. 16, (3) 2000, p. 201-215.

[DIN 00] Ding J., Gravano L., Shivakumar N., "Computing geographical scopes of web resources", *In Proceedings of the 26th International Conference on Very Large Databases*, *VLDB 2000*, September 10-14, Cairo, Egypt, 2000.

[DOA 99] Doan B.L., Beigbeder M., "Virtual WWW Documents : a concept to explicit the structure of WWW sites", *In Proceedings of the 21st Annual Colloquium on IR Research, 19-20 April,* 1999.

[EGG 00] Egghe L., "New informetric aspects of the Internet : some reflections, many problems", *Journal of information science*, vol. 26 (5) 2000, p. 329-335.

[GAR 72] Garfield E., "Citation analysis as a tool in journal evaluation", *Science*, 178, 1972, p. 471-479.

[GLO 01] Glover E., Flake G., Lawrence S., Birmingham W., Kruger A., Giles L., Pennock D., "Improving category specific web search by learning query modifications", *Symposium on Applications and Internet, SAINT 2001, San Diego, California*, 2001,

[GRA 00] Gravano L., "Characterizing Web Resources to Improved Search", *Position paper for the First NSF-DELOS Workshop on Information Seeking, Searching, and Querying in Digital Libraries*, 2000.

[HAR 75] Hartigan J., *A. Clustering Algorithms*, John Willey, New York, 1975.

[ING 98] Ingwersen P., "The calculation of web impact factors", *Journal of Documentation* 54 (2) 1998, p. 236-243.

[KLE 99] Kleinberg J., "Authoritative sources in a hyperlinked environment", *Journal of the ACM*, vol. 46 (5) September 1999, p. 604-632.

[KOS 00] Kosala R., Blockeel H., "Web Mining Research : A Survey", *SIGKDD Explorations*, vol. 2 (1) 2000, p. 1-15.

[KUM 99] Kumar R., Raghavan P., Rajagopalan S., Tomkins A., "Trawling the Web for emerging cyber-communities", *In the Proceedings of the Eighth World Wide Web Conference,* 1999.

[KWA 01] Kwasnik B.H., Crowston K., Nilan M., Roussinov D., "Identifying Document Genre to Improve Web Search Effectiveness", *The Bulletin of the American Society for Information Science and Technology*, vol. 27, (2), 2001.

[LAR 96] Larson R., Bibliometrics of the world wide web : An Exploratory analysis of the intellectual structure of the cyberspace. *In Proceedings of the Annual Meeting of the American Society of Information Science, Baltimore, Md., Oct. 19-24*, 1996.

[LAW 99] Lawrence S., Bollacker K., Giles C.L., "Indexing and Retrieval of Scientific Literature", *In Proceedings of the Eighth International Conference on Information and Knowledge Management, CIKM 99,* 1999, p. 139-146.

[LEV 90] Lévy, P., *Les technologies de l'intelligence*, Editions de la Découverte, 1990.

[MAR 73] Marshakova I.V., "Document coupling system based on references taken from Science Citation Index", in Russia, Nauchno - Teknicheskaya Informatsiya, Ser.2 No.6,3, 1973.

[MAR 58] Margalef R., *Information theory in ecology*, General Systems, 3, 1958, p. 36-71.

[PIR 96] Pirolli P., Pitkow J., Rao R., "Silk from a Sow's Ear : Extracting Usable Structures from the Web", *Proceedings of the Conference on Human Factors in Computing Systems, CHI 96,* 1996.

[PIT 97] Pitkow J., Pirolli P., "Life, death and lawfulness on the electronic frontier", *In Proccedings of the ACM SIGCHI Conference on Human Factors in Computing System, CHI'97*, 1997, p. 118-125.

[PRI 01] Prime C., Bassecoulard E., Zitt M., "Co-citations and co-sitations: a cautionary view on an analogy", *Proceedings of the 8th International Conference on Scientometrics and Infometrics, ISSI 2001, Sydney, Australia, July 16-20*, 2001, p. 529-540.

[ROS 99] Rostaing H., Boutin E., Mannina B., "Evaluation of internet Resources : Bibliometric techniques Applications", *Cybermetrics'99, Colima, July 9*, 1999.

[ROU 97] Rousseau R. "Sitations : an exploratory study", *Cybermetrics*, 1, (1) 1997.

[SAV 96] Savoy J. "Citation schemes in Hypertext information retrieval" In Agosti M. and Smeaton A. editors, *Information Retrieval and Hypertext*. 1996 Kluwer.

[SHA 48] Shannon C.E., "A mathematical theory of communications", *Bell System technical Journal*, 27, 1948, p. 379-423, 623-656.

[SMA 73] Small H.G., "Co-citation in the scientific literature", *Journal of the American Society for Information Science*, vol. 24, 1973, p. 265-269.