Le Web : une source d'information pour l'intégration de multi-termes dans un processus de Recherche d'Information

Mohamed Hatem Haddad, Mathias Géry, Dominique Vaufreydaz

Le Web : une source d'information

pour l'intégration de multi-termes dans le processus de Recherche d'Information

M. Hatem Haddad — Mathias Géry — Dominique Vaufreydaz

Laboratoire CLIPS-IMAG
Équipe MRIM (Modélisation et Recherche d'Information Multimédia)
B.P. 53, 38041 Grenoble Cedex 9, France
{Hatem.Haddad, Mathias.Gery, Dominique.Vaufreydaz}@imag.fr

RÉSUMÉ. *: le Web représente une source d'information riche et diversifiée. Dans cet article, nous proposons de tirer profit de cette richesse pour la collecte et l'analyse de documents dans la perspective d'une indexation relationnelle basée sur des multi-termes. La chaîne de traitement proposée comprend un robot qui permet la collecte de données pour la construction de corpus textuels, ainsi d'un module linguistique d'analyse du texte pour l'extraction d'information. La comparaison des collections obtenues avec les collections de la compagne d'évaluation Amaryllis montre la richesse linguistique des corpus collectés, et plus particulièrement la richesse des multi-termes extraits.*

ABSTRACT. *Web is a rich and diversified source of information. In this article, we propose to benefit from this richness to collect and analyze documents, with the aim of a relational indexation based on noun phrases. Proposed data processing chain includes a spider collecting data to build textual corpora, and a linguistic module analyzing text to extract information. Comparison of obtained corpus with corpus from Amaryllis conference shows the linguistic diversity of collected corpora, and particularly the richness of extracted noun phrases.*

MOTS-CLÉS : *corpus textuels, analyse du Web, extraction de multi-termes, Recherche d'Information.*

KEYWORDS: *textual corpora, Web analysis, noun phrases extraction, Information Retrieval (IR).*

1. Introduction

L'application de traitements linguistiques sur des corpus textuels a été abondamment étudiée dans le domaine de la Recherche d'Information (RI) comme le mentionne les travaux dans le cadre de la piste *traitement automatique du langage naturel* de TREC. Dans cet article, nous nous intéressons à la représentation du contenu textuel d'un document. Le traitement linguistique visé doit dépasser la simple troncature des mots extraits du corpus ou l'élimination des mots outils. Il est intéressant d'utiliser des multi-termes (un multi-terme est un mot constitué d'au moins deux termes significatifs) pour l'indexation des documents, afin de pallier l'utilisation restrictive des seuls mots isolés, amplement critiquée dans la littérature (Strzalkowski *et al.*, 2000). Par exemple, dans le cas de "pomme de terre", les mots simples "pomme" et "terre" ne conservent pas le sens du multi-terme "pomme de terre" et deviennent donc une source d'ambiguïté s'ils sont utilisés séparément comme termes d'indexation. Les multi-termes sont aussi utiles dans le cas de termes vagues, comme par exemple les deux termes "main" et "sac" qui ne sont pas assez spécifiques pour faire une distinction entre "sac à main" et "main dans le sac".

Avec l'émergence du Web comme source d'information riche et variée, les besoins en RI changent et les méthodes utilisées doivent permettre d'extraire des informations diversifiées répondant à la demande de l'utilisateur. Pour expérimenter de nouvelles méthodes de RI, on utilise des collections de test comme celles de la campagne d'évaluation Amaryllis (INIST, OFIL et LRSA). Ces collections sont composées de dizaines de milliers de documents provenant d'une source commune (articles du journal Le Monde pour OFIL, notices bibliographiques scientifiques pour INIST, monographies sur la culture mélanésienne pour LRSA). Or, nous faisons l'hypothèse que la qualité des informations extraites est directement liée à la qualité des corpus utilisés : l'expérimentation de méthodes de RI, qu'elles soient linguistiques ou statistiques, nécessite une grande richesse de corpus. Cette qualité n'est pas toujours assurée avec des collections de test classiques, en particulier au niveau de la diversité des thèmes abordés. En conséquence, les expérimentations sont souvent restreintes à des traitements spécifiques et ciblés sur le domaine traité, ne permettant qu'une extraction d'informations sur d'un domaine donné.

Nous avons montré l'utilité du Web pour construire de grands corpus de données pour la reconnaissance de la parole (Vaufreydaz *et al.*, 2001). La grande diversité des thèmes abordés permet d'extraire des modèles de langage riches. Le Web est un gigantesque espace d'information hétérogène : plus de 500 millions de personnes l'utilisent (NUA, 2001). Ces utilisateurs recherchent et consultent, mais aussi produisent des documents. Ainsi, on trouve une quantité de plus en plus importante de données. On estimait la taille du Web à 320 millions de pages en 1998 (Lawrence *et al.*, 1998), et à plus de 2 milliards en 2000 (Murray *et al.*, 2000). De plus, ces documents sont de plus en plus diversifiés : ils traitent de tous les sujets imaginables, dans un grand nombre de langues, et en utilisant différentes formes d'expressions. Une autre caractéristique intéressante du Web est l'aspect dynamique de son

contenu, qui implique une évolution constante de la terminologie utilisée. Ces caractéristiques sont donc très intéressantes pour construire des corpus riches, diversifiés, et de grandes tailles.

Nous présentons dans cet article la construction de corpus textuels à partir du Web et l'extraction de multi-termes à partir de ces corpus. Nous discutons des avantages et des inconvénients du Web par rapport à des collections de type Amaryllis, dans un objectif de représentation plus riche des documents. Pour cela, nous commencerons dans la 2ème section par la présentation de méthodes d'extraction de multi-termes pour la RI, en particulier la méthode utilisée dans le cadre de notre étude. Dans la 3ème section, nous détaillerons la chaîne de traitement complète qui permet, à partir du Web, d'extraire des corpus textuels, de les traiter et d'en extraire des multi-termes. Dans la 4ème section, nous présentons les caractéristiques générales des corpus collectés, comparativement à des corpus classiques. Enfin, nous analyserons dans la 5ème section les résultats obtenus par nos expérimentations d'extraction de connaissances à l'aide de ces corpus.

2. Extraction de multi-termes pour la RI

Les approches d'extraction des multi-termes peuvent être classées en trois catégories : l'approche statistique, l'approche linguistique et l'approche hybride. Fagan, qui utilise une approche statistique se basant sur l'hypothèse que l'emploi de deux mots en cooccurrence est l'expression d'une relation sémantique entre ces mots (Fagan, 1989), montre qu'il y a trop de bruit dans les combinaisons des mots trouvés. Différentes approches, comme la découverte de combinaisons de mots en fonction de leur régularité d'apparition (Church *et al.*, 1990), prennent en considération la distance entre les mots ou l'ordre de leur apparition. Ces approches statistiques permettent de couvrir de manière exhaustive toutes les combinaisons possibles des termes dans une fenêtre allant d'un bigramme de termes au document tout entier. Certaines combinaisons, valables d'un point de vue statistique, peuvent ne pas se justifier sémantiquement. Ceci est du principalement à l'ignorance du contexte linguistique de ces termes. Un autre inconvénient est le grand nombre de combinaisons de termes qu'une approche statistique peut extraire d'un corpus par rapport à une approche linguistique ou hybride. Parmi les travaux se basant sur une approche linguistique, on peut citer les systèmes *Lexter* (Bourigault, 1994) et *Fastr* (Jacquemin, 1997). *Lexter* repose sur des marques de frontières externes de syntagmes nominaux maximaux. Il commence par assigner automatiquement une étiquette grammaticale aux mots de la phrase. Ensuite, il utilise les bornes définies de syntagmes nominaux pour repérer les groupes nominaux. *Fastr* utilise des méta-règles pour repérer dans un corpus les variantes de termes à partir d'une liste de termes initiaux. Ces méta-règles se rapportent à la syntaxe, à la coordination de termes, à la modification et la substitution, à la permutation et à des règles morpho-syntaxiques et sémantico-syntaxiques. Les approches hybrides sont organisées en deux classes différentes selon l'ordre des traitements effectués : celles qui font un

filtrage statistique suivi d'un filtrage syntaxique tel que le système *Xtrac* (Smadja 1993), et celles qui font un filtrage syntaxique suivi d'un filtrage statistique tel que *Prométhée* (Morin, 1999).

Notre approche consiste à extraire des unités lexicales complexes de type nominal appelés multi-termes. Au contraire de la plupart des travaux cités, notre approche doit être le plus généraliste possible afin de traiter n'importe quel domaine d'application, et plus particulièrement le Web. Pour cela, elle se base sur les patrons morpho-syntaxiques les plus utilisés dans une langue, la langue française dans le contexte de notre étude. En définissant les patrons spécifiques à d'autres langues, l'approche peut être facilement exploitée sur d'autres langues ainsi que sur des corpus multilingues. Etant donnée la masse d'information, un traitement linguistique approprié s'impose. La mise en œuvre d'un traitement linguistique en profondeur repose sur des analyseurs robustes et exhaustifs de la langue, trop complexes pour l'objectif visé et contraignants dans le cadre d'un SRI. C'est pourquoi nous adoptons une analyse plus superficielle qui élimine la détermination de la structure linguistique profonde et ne tient compte que de l'extraction des multi-termes.

3. Chaîne de traitement

Nous présentons la chaîne de traitement qui permet de collecter des données brutes sur le Web, de les analyser pour en extraire des corpus normalisés, et d'utiliser ensuite ces corpus pour l'extraction de multi-termes. Le schéma général de la chaîne de traitement est représenté dans la Figure 1.

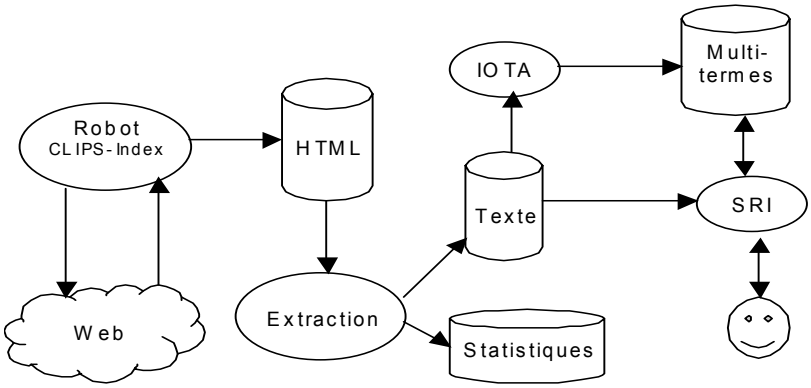

Figure 1. *Chaîne de traitement*

Nous présentons les 3 principales composantes de la chaîne de traitement : le robot (pour la collecte de données sur le Web), l'analyseur/constructeur de corpus et le module d'analyse linguistique du SRI IOTA.

3.1. *Le robot CLIPS-Index*

Le robot CLIPS-Index permet de parcourir le Web afin de collecter et de stocker des pages. Son objectif est de collecter un maximum d'information sur un domaine Internet donné, parmi une grande quantité de données hétérogènes. Les principaux standards du Web (HTML, HTTP, URL) ne sont pas toujours respectés, ce qui impose de trouver un bon compromis entre l'efficacité et la gestion des erreurs. Le robot est basé sur une architecture multi-thread qui permet de lancer des centaines de requêtes HTTP simultanément, en assurant la synchronisation entre les processus afin d'éviter la collecte d'URLs en double. Cela nécessite la gestion d'un stock de dizaines de millions d'URLs. Malgré sa puissance de collecte, le robot doit causer le moins de désagréments possibles sur les serveurs Web interrogés. Pour cela, il respecte le protocole d'exclusion des robots (Koster 1996), ainsi qu'un délai minimum d'attente entre chaque requête sur un même serveur afin d'éviter sa surcharge. CLIPS-Index enchaîne les étapes suivantes :

– Choix d'une URL parmi un stock d'URLs disponibles.

– Collecte de la page HTML correspondante.

– Analyse de la page : extraction de la liste des URLs.

– Stockage de la page.

– Ajout des nouvelles URLs au stock d'URLs disponibles.

CLIPS-Index est un robot puissant et efficace : avec un PC de bureau (500MHz, 128Mo RAM), il peut trouver, charger, analyser et stocker jusqu'à 3 millions de pages par jour. L'extracteur d'URLs prend en compte les erreurs commises par les auteurs de sites Web par rapports aux standards. Ainsi, nous avons collecté 38'994 pages sur le domaine ".imag.fr" le 5 octobre 2000, alors que le même jour Altavista affirmait indexer 24'859 pages sur ce même domaine.

3.2. *Traitement et normalisation des corpus Web*

La phase suivante consiste à normaliser les données brutes collectées sur le Web, les étapes les plus importantes étant les suivantes :

– L'extraction du texte à partir du HTML, qui doit être robuste pour tenir compte du peu de respect des normes, et doit donner un texte correctement ponctué en vue de traitements linguistiques à l'échelle de la phrase.

– L'élimination des doublons (alias de noms de serveurs, sites miroirs, etc).

– L'extraction du lexique et le calcul de la couverture lexicale du corpus.

– L'extraction de statistiques diverses, par exemple la langue des documents ou des informations ayant trait à la structure des pages Web.

En sortie de l'analyseur, on obtient des fichiers au format désiré (par exemple, au format TEI pour IOTA ou au format spécifique du SRI SMART).

3.3. *Le système IOTA*

Nous avons utilisé le système IOTA (Chiaramella *et al.*, 1986) pour l'analyse morpho-syntaxique et l'extraction des multi-termes. L'analyseur morpho-syntaxique est un analyseur de surface qui utilise un dictionnaire associé à un modèle morphologique. Un traitement particulier est appliqué aux formes non reconnues. Il permet en utilisant des schémas de résolution répertoriés manuellement, correspondant à des cas d'ambiguïté typiques et dont la résolution est connue, de leur attribuer une interprétation potentielle (Chevallet *et al.*, 1997). La sortie de ce module est une collection de textes étiquetés, c'est-à-dire dont tous les mots ont été catégorisés. La fréquence globale d'un terme dans une collection ainsi que sa fréquence selon une fenêtre sont calculée.

La deuxième étape utilise cette collection étiquetée et en extrait un ensemble de multi-termes. Les multi-termes candidats sont extraits par repérage de frontières de catégories syntaxiques (on considère par exemple qu'un multi-terme commence et par un substantif ou un adjectif). Un filtre syntaxique permet de ne garder que les multi-termes valides par rapport à un ensemble de patrons morpho-syntaxiques. Ces derniers correspondent à des patrons génériques de la langue française ("substantif substantif", "substantif préposition substantif", etc.).

4. Collecte des corpus

4.1. Choix des corpus

Nous avons choisi de comparer les corpus classiques de la campagne Amaryllis (INIST et OFIL) avec deux corpus extraits du Web :

– "Tunisie" : un corpus de pages collectées sur le domaine ".tn", pour obtenir un corpus qui ne soit pas trop volumineux pour être traité rapidement, qui contienne une majorité de documents francophones, et qui soit représentatif d'un pays.

– "Journaux" : un corpus de pages provenant de sites Web de journaux, afin de construire un corpus textuel de grande taille, francophone, et de bonne qualité dans l'utilisation de la langue française.

Un filtre, paramètre de CLIPS-Index, permet de sélectionner les sites Web à collecter. Il est exprimé à l'aide d'une expression régulière sur le nom du site. Celui utilisé pour "Tunisie" permet de se restreindre aux noms de sites se terminant par ".tn". Le filtre utilisé pour "Journaux" a été construit automatiquement par un "*extracteur de noms de sites thématisés*", dont le rôle est de parcourir des parties de la hiérarchie de l'annuaire Yahoo! et d'en extraire un filtre (par exemple "/Actualites_et_medias/Journaux_et_magazines/").

4.2. Collecte des corpus

Nous avons collecté les corpus “Tunisie” et “Journaux” à différentes dates. Le Tableau 1 montre les statistiques des collectes, qui se sont déroulées jusqu’à épuisement des URLs sur les domaines recherchés. Cela explique les performances anormalement basses (entre 2,87 et 9,44 documents par seconde, contre plus de 30 normalement), car les dernières URLs d’un domaine sont plus délicates à dénicher.

Corpus	Date de la collecte	Durée de la collecte	Nombre de documents	Docs par sec.
Tunisie	16 mars 2001	1 h 08	38’523	9,44
Tunisie 2	22 août 2001	1 h 50	60’787	9,21
Tunisie 3	24 janvier 2002	7 h 49	109’162	3,88
Journaux	7 novembre 2001	17 h 43	244’364	3,83
Journaux 2	11 janvier 2002	38 h 29	397’854	2,87

Tableau 1. *Caractéristiques des collectes*

4.3. Caractéristiques générales des corpus textuels

Nous disposons de 2 corpus classiques que nous avons comparé à 5 corpus extraits du Web. Chaque corpus HTML est analysé pour en extraire un corpus de textes distincts, les documents doublons étant détectés et éliminés. Le Tableau 2 montre les caractéristiques générales des corpus textuels.

Corpus	Nombre de docs	Taille HTML/TEI		Taille Texte	
		Collection	Ko/page	Collection	Ko/page
INIST	163’308	100 Mo	0,63	79 Mo	0,50
OFIL	11’016	33 Mo	3,06	32 Mo	2,93
Tunisie	27’959	161 Mo	5,90	27 Mo	1,00
Tunisie 2	43’651	397 Mo	9,31	55 Mo	1,30
Tunisie 3	79’361	863 Mo	11,13	165 Mo	2,13
Journaux	198’158	4’391 Mo	22,69	896 Mo	4,63
Journaux 2	345’860	7’728 Mo	22,88	1'491 Mo	4,41

Tableau 2. *Caractéristiques générales des corpus textuels*

Le rapport entre la taille du corpus HTML et celle du corpus textuel va de 4,9 à 7,2. Ce phénomène vient de l’utilisation de plus en plus importante de balises HTML et autres pour la présentation des pages, avec des outils de production de HTML qui insèrent de plus en plus de données. Par exemple, une page HTML minimum occupe 74 octets avec la norme HTML, 304 octets avec Netscape Composer et plus de 2’000 octets avec Word !

5. Analyse des corpus et résultats

5.1. Répartition des langues

L'extraction de la langue d'un document, est basée sur les fréquences des mots les plus courants de chaque langue (anglais, français, italien, allemand, espagnol, etc.), selon si le document comporte une plus grande proportion de « le, la, les, un, une, des, dans, etc. » ou de « and, any, for, not, of, the, etc. ». Pour chaque langue, ces mots fréquents sont extraits d'un corpus de textes de référence.

On observe une très grande majorité de pages francophones dans la Figure 2, en particulier dans les corpus "Journaux" et "Journaux 2". La grande proportion de pages dont la langue n'a pas été extraite ("inconnue") s'explique, pour les corpus tunisiens, par un nombre important de pages vides d'éléments textuels, souvent remplacés par des images, alors que les pages de "Journaux" et "Journaux 2" contiennent presque toujours du texte.

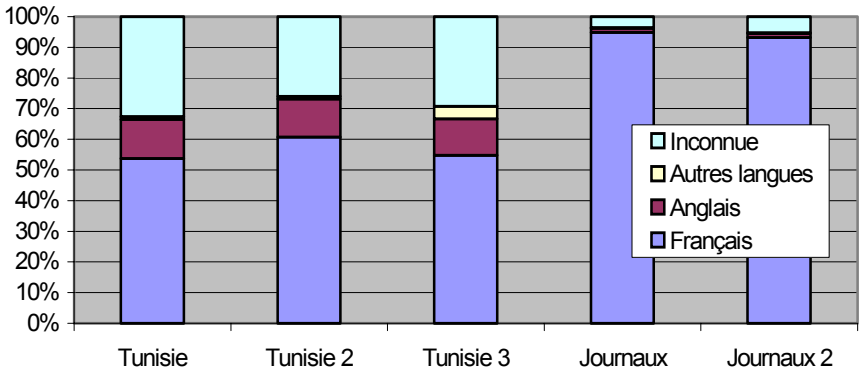

Figure 2. *Répartition des langues*

5.2. Lexique et couverture du français

Le Tableau 3 récapitule le nombre de termes distincts apparaissant dans chaque collection, ainsi que le nombre total de termes par collection et par document. Nous obtenons effectivement des corpus volumineux, jusqu'à 30 fois plus que ceux d'Amaryllis pour "Journaux 2". Les documents des collections de journaux ("OFIL", "Journaux", "Journaux 2") sont en moyenne de plus grande taille.

Afin d'évaluer la diversité des corpus, nous avons estimé leur couverture lexicale de la langue française, en calculant le pourcentage de formes lexicales d'un lexique témoin qui apparaissent dans chaque corpus. Ce lexique de 400'000 formes lexicales a été construit à partir du lexique de l'Association des Bibliophiles Universels

(ABU) (comportant plus de 270.000 formes lexicales), et du lexique BDLex (De Calmès *et al.*, 2000) duquel nous avons dérivé plus de 300'000 formes lexicales.

Corpus	Nombre de Termes	Occurrences	
		(par collection)	(par document)
INIST	174'659	8,31 millions	50,89
OFIL	119'434	5,15 millions	467,55
Tunisie	113'418	4,21 millions	150,61
Tunisie 2	164'569	8,46 millions	193,70
Tunisie 3	393'919	25,04 millions	315,57
Journaux	536'361	133,97 millions	676,07
Journaux 2	850'659	257,04 millions	743,19

Tableau 3. *Nombre de termes*

La Figure 3 montre la couverture lexicale de chacun des corpus étudiés, qui est beaucoup plus importante pour les corpus "Journaux" et "Journaux 2" que pour les corpus classiques.

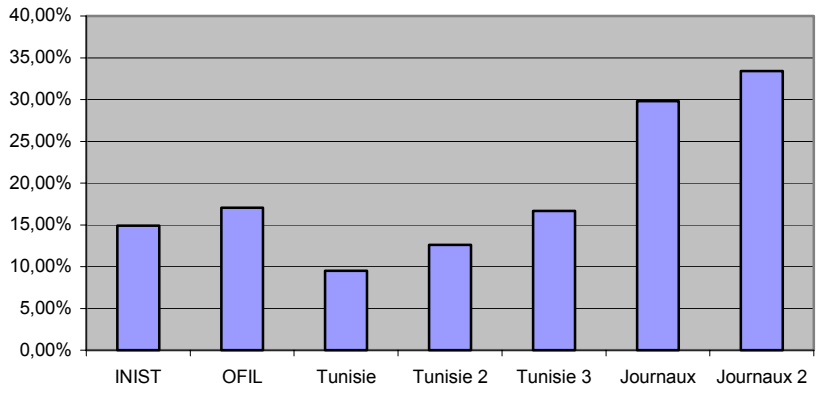

Figure 3. *Couverture du français*

5.3. Distribution des catégories grammaticales

La distribution des catégories grammaticales des termes est presque identique pour les différents corpus. Comme le montre le Tableau 4, les catégories dominantes sont les substantifs, les adjectifs et les noms propres. Cette distribution nous a guidé

dans le choix de patrons morpho-syntaxiques qui prennent en compte principalement ces trois catégories.

	OFIL	INIST	Tunisie	Tunisie 2	Tunisie 3	Journaux	Journaux 2
Substantif	30,60%	33,62%	29,21%	29,30%	28,14%	28,88%	28,52%
Adjectif	27,25%	31,55%	26,48%	27,00%	26,15%	27,84%	27,71%
Nom propre	18,38%	11,75%	12,96%	12,89%	13,79%	16,37%	16,42%
Reste	23,76%	23,09%	31,35%	30,80%	31,91%	26,91%	27,35%

Tableau 4. *Distribution des catégories grammaticales des termes*

5.4. Extraction des multi-termes

L'extraction des multi-termes a donné un nombre plus important de ces derniers dans les collections Web, par rapport aux collections Amaryllis, comme illustré dans la Figure 4. Cette constatation s'explique par le fait qu'il y a un certain nombre de documents dans les collections Web, principalement dans les collections tunisiennes, qui ne contiennent pas ou très peu de texte.

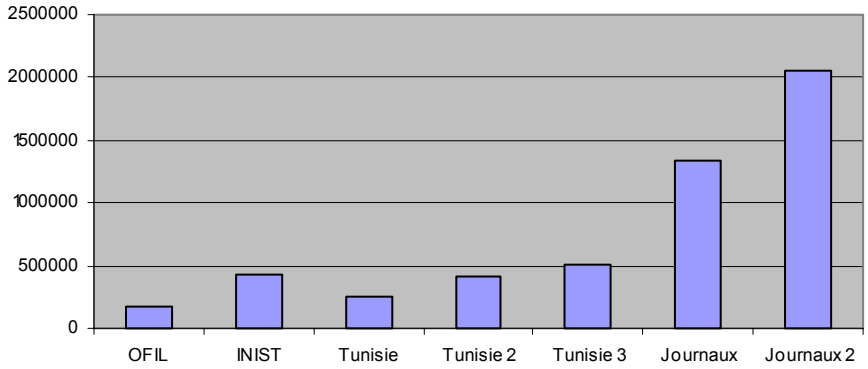

Figure 4. *Nombre de multi-termes extraits*

Par contre, le nombre moyen de multi-termes dans un document est plus important dans le cas de la collection OFIL, et très faible dans le cas de la collection INIST comme illustré dans la Figure 5. Cela s'explique par le fait que INIST est une collection scientifique où les phrases sont très courtes avec un style descriptif simple. Il est intéressant de constater que pour une même collection, les fréquences de certains multi-termes ont largement augmentées. Par exemple, le multi-terme "enseignement supérieur" était pratiquement inexistant dans la collection "Tunisie" et passe à une fréquence de 1773 dans la collection "Tunisie 3" contre seulement

992 dans “Tunisie 2”. Des nouveaux multi-termes apparaissent d’une collecte à une autre, et reflètent par exemple un événement médiatique, tel le multi-terme “jeux méditerranéens” dont la fréquence est de 972 dans “Tunisie 2” et 178 dans “Tunisie”.

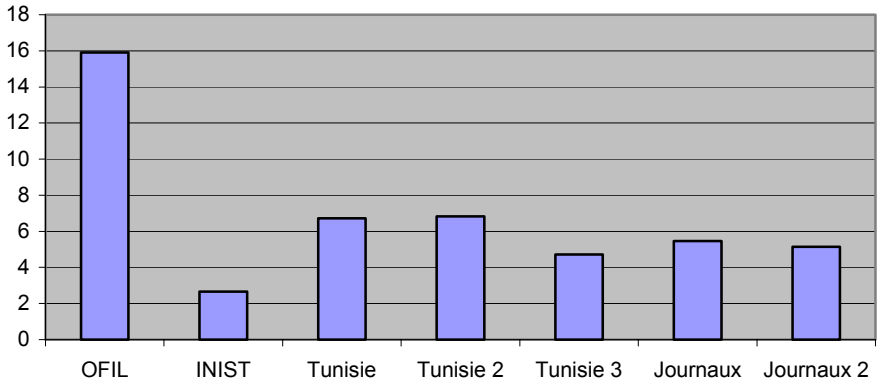

Figure 5. *Nombre de multi-termes/document*

6. Conclusion

Dans cet article, nous avons étudié l’intérêt du Web comme source d’information pour la construction de corpus textuels riches et diversifiés, avec comme objectif final l’extraction d’information pour la RI. Nous avons comparé qualitativement et quantitativement les corpus obtenus avec les corpus classiques OFIL et INIST d’Amaryllis. La quantité phénoménale d’information disponible sur le Web nous permet incontestablement de construire des corpus de taille très importante. Nous les avons utilisé pour l’application de méthodes linguistiques d’extraction d’information, mais cette grande quantité de données est aussi particulièrement appréciable pour l’application de méthodes statistiques d’extraction d’information (Géry *et al.*, 1999). Toutefois, l’avantage principal de ces corpus provient de l’aspect dynamique du Web et de la grande diversité des domaines traités. En effet, cela nous permet d’extraire des informations couvrant de nombreux domaines de la connaissance et suivant l’évolution du vocabulaire.

La quantité et la qualité des informations extraites nous offrent de nombreuses perspectives. Nous développons un modèle de RI basé sur l’indexation relationnelle intégrant les multi-termes dans les index de documents (Chevallet *et al.*, 2001). Sa mise en œuvre dans le cadre du Web nécessite l’utilisation de corpus appropriés. La richesse de ces corpus nous offre la possibilité d’extraire des connaissances qui reflètent pour une période donnée le vocabulaire employé et son utilisation par les producteurs de documents.

7. Bibliographie

ABU, Association des Bibliophiles Universels, http://abu.cnam.fr.

Bourigault D., LEXTER, Un Logiciel d'EXtraction de TERminologie. Application à l'acquisition de connaissances à partir de textes, Thèse de doctorat, EHESS Paris, 1994.

Chevallet J.-P., Nie J.-Y., Intégration des analyses du français dans la Recherche d'Information, *Actes de la conférence RIAO*, Montréal, 25-27 juin 1997, p. 761-772.

Chevallet J.-P., Haddad H., Proposition d'un modèle relationnel d'indexation syntagmatique : mise en œuvre dans le système IOTA, *Actes de la conférence INFORSID*, Martigny, Suisse, 30 mai-1er juin 2001, p. 465-483.

Chiaramella Y., Defude B., Bruandet M.-F., Kerkouba D., IOTA: a full test Information Retrieval System, *Actes de la conférence ACM-SIGIR*, Pise, Italie, 8-10 sep. 1986.

Church K.W., Hanks P., Word association norms mutual information and lexicography, *Computational Linguistics*, vol. 16, n°1, 1990, p. 22-29.

De Calmès M., Pérennou G., BDLEX : a lexicon for spoken and written french, *Actes de la conférence LREC*, Grenade, Espagne, 28-30 mai 1998, p. 129-136.

Fagan J.-L., The effectiveness of a nonsyntactic approach to automatic phrase indexing for document retrieval, *JASIS*, vol. 40, n°2, 1989, p. 115-132.

Géry M., Haddad H., Knowledge discovery for automatic query expansion on the World Wide Web, *Actes de la conférence WWWCM*, Paris, France, 15-18 nov. 1999, p. 334-347.

Jacquemin C., Variation terminologique : reconnaissance et acquisition automatiques de termes et de leurs variantes en corpus, HDR, Université de Nantes, 1997.

Koster M., A method for Web robots control, rapport technique, 1996, IETF.

Lawrence S., Giles C.L., Searching the World Wide Web, *Science*, vol. 280, n°5360, 1998, p. 98-100.

Morin E., Extraction de liens sémantiques entre termes à partir de corpus de textes techniques, Thèse de doctorat, Université de Nantes, 1999.

Murray B. H., Moore A., Sizing the Internet, rapport technique, 2000, Cyveillance Inc..

NUA, Nua Internet Surveys, rapport technique, http://www.nua.ie/surveys, 2001, NUA.

Smadja F., Retrieving collocations from text: Xtract, *Computational Linguistics*, vol. 19, n°1, 1993, p. 143-177.

Strzalkowski T., Stein G. C., Bowden Wise G., Bagga A., Towards the Next Generation Information Retrieval, *Actes de la conférence RIAO*, Paris, 12-14 avril 2000, p. 1196-1207.

Vaufreydaz D., Géry M., Internet evolution and progress in full automatic french language modelling, *Actes de la conférence ASRU*, Madonna di Campiglio, Italie, 9-13 déc. 2001.