Extraction d'information 'a partir de documents Web multilingues : une approche d'analyses structurelles

Tuan Dang Nguyen

UNIVERSITÉ de CAEN/BASSE-NORMANDIE

U.F.R. : Sciences

ÉCOLE DOCTORALE : SIMEM

THÈSE

présentée par

Dang Tuan NGUYEN

et soutenue

le 25 septembre 2006

en vue de l'obtention du

DOCTORAT de l'UNIVERSITÉ de CAEN

spécialité : Informatique

(Arrêté du 25 avril 2002)

Extraction d'information à partir de documents Web multilingues : une approche d'analyses structurelles

MEMBRES du JURY

Directeur :	M. Khaldoun Zreik	Professeur	Université de Paris VIII
Rapporteurs :	M. Imad Saleh	Professeur	Université de Paris VIII
	M. Saïd Tazi	Maître de Conférence	Université de Toulouse
Examinateurs :	Mme Anne Nicole	Professeur	Université de Caen
	M. Jacques Vergne	Professeur	Université de Caen
	M. Jacques Labiche	Professeur	Université de Rouen

Groupe de Recherche en Informatique, Image, Automatique et Instrumentation de Caen — CNRS UMR 6072

Mis en page avec la classe thloria.

Remerciements

L'aboutissement de ma thèse, fruit de quatre ans de dur travail et de grande mélancolie de mon pays natal, est une vraie satisfaction. C'est donc avec un immense plaisir que j'exprime ma profonde gratitude à toutes les personnes qui ont contribué, directement ou indirectement, à la réussite de ma thèse.

Je remercie M. Imad Saleh, Professeur à l'Université de Paris VIII, et M. Saïd Tazi, Maître de Conférence (HDR) à l'Université de Toulouse, pour avoir accepté de juger ce travail, également pour leur lecture minutieuse du manuscrit, et pour leur commentaires qui ont grandement contribué à la qualité finale de mon mémoire.

Je remercie Mme Anne Nicolle, Professeur à l'Université de Caen, M. Jacques Labiche, Professeur à l'Université de Rouen, M. Jacques Vergne, Professeur à l'Université de Caen, pour avoir accepté de participer au jury de cette thèse.

Je remercie de tout mon coeur Mme Nadine Lucas pour m'avoir co-encadré et proposé des idées précieuses sur sa méthode d'étude au début de mes travaux de recherche. Je remercie également M. Bruno Crémilleux pour m'avoir acceuilli à l'équipe DoDoLa.

Je remercie infiniment M. Khaldoun Zreik, Professeur à l'Université de Paris VIII (ex-Professeur à l'Unversité de Caen), pour avoir dirigé ce travail, pour le temps qu'il m'a consacré pendant toutes ces années, et pour sa grande générosité que je me souviens toujours dans ma vie.

Je remercie tous les membres de l'équipe DoDoLa pour leur encouragement et pour leurs amitiés pendant mes années de thèse. Je remercie aussi mes autres copains : Bassam Baki, Sala El Falou, Arnaud Soulet, François Rioul, Hossam Hanna, Jin Yao, Nadia Zérida pour leur gentillesse et pour leur cordialité.

Je remercie en particulier Pierre Renaux pour ses aides très dévouées dans la correction de mon mémoire, et pour sa disponibilité en tout temps qui nous ont montré une amitié sincère et sublime.

Je termine ces remerciements par ma famille et par mon pays qui ont beaucoup attendu mon retour.

Table des matières

Table des figures

Liste des tableaux

Première partie

Introduction générale

Motivations et objectifs

La question de la spécificité culturelle et tout particulièrement celle de la spécificité linguistique influencent encore les travaux de déploiements massifs des procédures de standardisation et de mondialisation. Ceci peut être expliqué par le fait que le nombre de sites multilingues ne cesse de croître malgré leur coût de développement qui est nettement plus élevé que celui du développement des sites monolingues. Cependant, nous constatons que les recherches sur la structure des sites Web multilingues sont très peu explorées. Les travaux dans plusieurs domaines comme la recherche d'information ou l'extraction d'information se concentrent majoritairement sur l'aspect des contenus en les traitant principalement avec les techniques de traitement automatique des langues naturelles. Ce constat est devenu un point de départ pour notre premier intérêt de nous engager à étudier l'aspect structurel des sites Web multilingues.

Depuis les années 1990, les missions du domaine d'extraction d'information ont été précisées grâce aux développements du programme MUC (Message Understanding Conferences). Désormais l'objectif de ce domaine s'est fixé à extraire des motifs et ne traite que certains problèmes linguistiques bien précisés tels que : la reconnaissance des entités nommées, l'analyse des structures lexicales des structures de phrase, l'enlèvement de l'ambiguïté lexicologique, la résolution de coréférence, etc. Les systèmes d'extraction d'information visent maintenant à collecter les données représentant l'information dans les pages Web (documents semi-structurés). La nature de l'information ainsi extraite risque d'être brute et superficielle.

Aussi, nous constatons que les structures de documents (même semi-structurés ou non-structurés) peuvent contenir de nombreuses informations (souvent implicites) que les systèmes actuels ne révèlent pas. En effet, ces informations ne sont ni représentées, ni régies par des moyens d'expressions régulières, c'est pourquoi, elles ne peuvent pas être déterminées par des règles précises, ni apprises à partir des exemples, ni extraites par les extracteurs. Prenant acte de cette limitation, nous nous sommes donnés comme objectif d'extraire, à partir de la structure des informations, de méta-informations englobant (structurant d'une façon ou d'une autre) d'autres informations. Pour cela nous avons opté, dans un premier temps, de tester cette approche, qui est purement structurelle et ne faisant pas appel à des connaissances linguistiques, pour répondre à des questions simples de type : est-ce-que l'information est de nature multilingue ou monolingue ? Quelles sont les frontières entre ces langues ? Quelles sont ces langues ?

Les modèles de représentation « classiques » des structures Web (c'est à dire la structure hiérarchique ou le graphe de documents Web) semblent insuffisants pour extraire des informations « structurelles » à partir des sites multilingues. En effet, ces derniers se démarquent par la présence des points (noeuds, pointeurs, icônes) de changement de langues, ou bien par d'autres phénomènes typiques comme la disposition d'informations complémentaires, des documents partagés par plusieurs langues, etc. Ainsi nous proposons dans cette thèse un modèle de représentation adapté aux hyperdocuments multilingues.

Notre dernière motivation porte sur la question de la reconnaissance des langues dominantes dans un site Web multilingue. A notre connaissance, cette question n'a pas été posée auparavant. Pourtant, la détermination des langues dominantes peut avoir une signification importante à la fois lors de la conception des systèmes de recherche d'information et lors de l'extraction d'information multilingues, etc.

Dans ce mémoire, nous abordons l'aspect multilingue dans un contexte d'extraction d'information à partir des sites Web multilingues. Nous tâcherons d'apporter quelques éléments de réponses à la représentation des hyperdocuments et à l'élaboration d'un système d'extraction d'information basé sur une approche d'analyse structurelle.

Contexte de l'étude

Différentes approches du diagnostic de langues sont proposées pour identifier la « langue la plus employée » dans un document. Ces approches s'intéressent à détecter la langue dans laquelle la majorité (ou la totalité) des textes d'un document sont écrits. Giguet s'est intéressé à étudier les documents plurilingues et a lié la notion de « langues du document » à celle de la « structure multilingue » [Gig98]. Cette contribution a montré l'importance qualitative de la structure dans la description de l'information que le document véhicule.

Les approches uniquement statistiques qui demeurent assez intuitives ne nous semblent pas totalement convaincantes : nous pensons que le caractère multilingue du site Web ne peut pas être reconnu sans étudier sa structure. Dans cette thèse, nous avons observé la faiblesse d'une approche purement statistique en traitant des cas où la majorité des documents, dans un site Web multilingue, sont écrits en une langue qui n'est pourtant pas dominante.

C'est pourquoi l'identification des langues utilisées dans un site Web multilingue n'est qu'une étape pour déterminer ensuite quelles sont les langues dominantes parmi elles.

Pour localiser le contexte de notre étude, nous introduisons brièvement une approche hypbride structurelle (qualitative) et statistique (quantitative) pour déterminer les « langues du site Web ». Partant de ce postulat, nous proposons une solution simple pour identifier les langues dominantes dans un site Web.

Dans le contexte de nos travaux, dont l'objectif est de mettre en exergue l'apport de l'information enfouie dans la structure, nous ne voyons pas l'intérêt de vérifier, de tester ou d'optimiser des méthodes classiques de reconnaissance de langues (écrites). Ces objectifs ont été renforcés par la suivie de plusieurs travaux effectués par Ahmed [ACT04], Padro [PP04], Peng [PS03] avec des problématiques assez proches des notres.

Nous tenons à signaler que nous nous sommes concentré à réaliser un système de type prototype de recherche ayant comme seul objectif de valider nos hypothèses. C'est pourquoi, ce système (Hyperling) ne dispose d'interface homme-machine assez développé.

Problématiques

La prise en compte des structures multilingues devra être un élément déterminant dans les travaux de la recherche d'information, de la fouille du Web, de l'extraction d'information et du Web sémantique. La présence ou non de multiples langues sur un site Web engendre trois types de problèmes dont l'ignorance pourra nuire à la qualité des résultats obtenus :

- la redondance, si le site propose simultanément des traductions en plusieurs langues,
- les parcours bruités lors d'un passage d'une langue à une autre via les vignettes (génération de graphes, conceptuellement, non signifiant),
- la perte de l'information par la négligence de la spécificité structurelle (même implicite) de chaque langue.

Dans ce contexte nous considérons deux problématiques principales : la première observe la faiblesse des travaux et des expériences pouvant détecter le changement de langue au sein d'un site Web multilingue, et la deuxième

problématique observe le manque d'hypothèses « crédibles et validées » pouvant résoudre le problème de déterminer la ou les langues dominantes dans un site Web multilingue.

La détection de changement de langue au sein d'un site Web multilingue a été souvent interprétée, par certains auteurs, par un problème d'alignement de documents multilingues [BS98], [HM01], [RTPG04]. La détection de changement de langue dans notre travail se traduit par l'identification du mécanisme, de la façon et des moyens qu'un site Web multilingue dispose pour passer d'une langue vers une autre.

La notion de « langues dominantes » reste relativement floue et demeure un réel problème lors de l'analyse des sites Web multilingues existants. Nous pensons qu'une simple approche statistique (traitant des sacs des mots) ne permet pas de détecter avec pertinence les langues dominantes. De même, une approche d'analyses structurelles ne permet pas d'aboutir à une conclusion absolue sur les langues dominantes. C'est pourquoi, nous avons opté pour une approche hybride.

Contributions de la thèse

Cette thèse propose un modèle d'hyperdocuments permettant de représenter des sites Web multi ou monolingues en distinguant différents types de structures : la structure hiérarchique interne du document, la structure hypertextuelle du site Web et la structure de relations entre des ancres sources localisées dans les hyperliens.

Dans ce cadre nous avons élaboré et validé plusieurs hypothèses « structurelles » portant sur la densité très puissante des liens (appelés relations par la suite) entre les documents d'une même langue (par rapport aux liens entre les documents de différentes langues) et la convergence des documents d'une même langue. Ces hypothèses ont émergé suite à une série d'observations (manuelles, semi-automatiques ou automatiques) faites sur la structure des sites Web multilingues.

Pour valider ou bien réfuter ces hypothèses, nous avons adopté une démarche expérimentale qui a consisté à développer le système Hyperling dont la fonctionnalité a été testée par une série d'expérimentations sur plusieurs sites Web multilingues professionnels.

Organisation du mémoire

Les contenus essentiels de ce mémoire sont organisés en deux parties principales : les deuxième et troisème parties. Dans la deuxième partie, nous présentons un aperçu sur les Web multilingues. La troisème partie est dédiée à la construction de notre système d'extraction d'information à partir de documents multilingues.

La deuxième partie (cf. partie II) se divise en deux chapitres. Le premier chapitre (cf. chapitre 1) porte sur les caractéristiques des hyperdocuments (sites Web) multilingues. Le deuxième chapitre (cf. chapitre 2) présente l'état de l'art en matière d'extraction d'information à partir des structures d'information multilingues.

La troisème partie (cf. partie III) est composée de quatre chapitres. Le troisième chapitre (cf. chapitre 3) se focalise sur la représentation des hyperdocuments. Le quatrième chapitre (cf. chapitre 4) introduit la modélisation et le fonctionnement du système Hyperling. Les cinquième (cf. chapitre 5) et sixième chapitres (cf. chapitre 6) présentent la construction de deux modules principaux d'Hyperling pour catégoriser les documents Web en fonction de leurs langues et pour reconnaître des langues dominantes.

Dans la conclusion (cf. partie IV), nous rappelons la réalisation de notre étude, les résultats, ainsi que les perspectives de cette recherche.

Deuxième partie

Le multilinguisme sur le Web

Chapitre 1

Caractéristiques des sites Web multilingues

« Réaliser l'accès universel aux « e-contenus » dans toutes les langues, améliorer les capacités linguistiques des utilisateurs et développer des outils pour l'accès multilingue à Internet. »
UNESCO, 2003

1.1 Introduction

Dans ce chapitre préliminaire, nous tenons à évoquer l'importance croissante que les usages et développements de l'univers du Web ont connu cette dernière décennie. Aussi, nous mettons l'accent tout particulièrement sur la diversité et la multitude des langues présentes dans cet univers fortement dynamique (en évolution permanente). L'ouverture de cet univers (qui est international par défaut) nous mène tout naturellement à considérer un phénomène, qui demeure prépondérant : celui du multilinguisme dans la conception et la structuration de l'information.

L'étude des ressources d'information multilingues sur le Web fait l'objet de multiple travaux dans plusieurs disciplines : la recherche d'information (Adriani [Adr00], Bertoldi [BF03], Besançon [BFF04], Dini [DLM+05], Fluhr [Flu05], Sperer [SO00], etc.), l'extraction d'information (Azzam [AHG+99], Declerck [DC03], Kiyoshi [KSG04], Masche [Mas04], Maynard [May03], etc.) et la catégorisation (Cavnar [CT94], Giguet [Gig95], Peng [PS03], etc.). Dans ces disciplines, le multilinguisme est souvent considéré comme un problème de fédération de modèles et de méthodes issu principalement des travaux

de recherche du Traitement Automatique des Langues Naturelles (TALN) comme : les outils d'analyse morphologique et syntaxique, dictionnaires, traducteurs automatiques et les générateurs de résumés automatiques.

Les études bibliographiques que nous avons menées nous ont permis d'observer une absence partielle d'intérêts (manifestés) à l'égard de la structuration de l'information dans les ressources multilingues. Cependant, nous avons constaté que la conception des sites Web multilingues connaît une efflorai-son avérée. De même, les études montrent que les internautes semblent se familiariser et gérer facilement des structures de documents Web de type multilingue (cf. section 1.2).

1.2 La diversité des langues sur le Web

Au début de l'Internet, le monolinguisme (précisément l'usage de la langue anglaise) a dominé presque la totalité de l'information transitant sur ce nouveau média. Ce fait, n'était qu'une conséquence logique de l'histoire de l'Internet qui a été introduit comme un réseau, ARPANET en 1969, du Ministère de la Défense américain[1]. De même, la création du World-Wide Web en 1989-1990, par Tim Berners-Lee, au CERN (European Laboratory for Particle Physics)[2] et la distribution du logiciel de navigation Web Mosaic (l'ancêtre du navigateur Web Netscape) en novembre 1993, ont été fortement répandus d'abord en Amérique du Nord, puis dans le reste du monde.

Répartition en terme d'information (quantité d'information)

L'année 1997 a été marquée par la présence importante de diverses langues sur la place de l'Internet. L'étude sur la répartition des langues, sur l'espace d'Internet, menée par l'équipe Babel[3] (une initiative conjointe d'Alis Technologies et de l'Internet Society) a montré que : l'anglais occupait 82,3% de l'information disponible sur l'Internet suivi par l'allemand 4%, le japonais 1,6%, le français 1,5%, l'espagnol 1,1%, le suédois 1,1%, et l'italien 1,0%.

[1] History of the Internet (http ://www.historyoftheinternet.com/).
[2] About the World Wide Web Consortium (W3C) (http ://www.w3.org/Consortium/).
[3] Palmarès des langues de la Toile, juin 1997 (http ://alis.isoc.org/palmares.html).

Répartition en terme de sites (nombre de sites)

D'autres études statistiques assez approfondies (tenant compte de la structure, la taille, l'utilisation, et le contenu du Web) ont été effectuées dans le cadre du Projet de Caractérisation du Web, réalisées par l'OCLC (Online Computer Library Center) [Sch00]. Ces études ont conclu que 29 langues étaient présentes en 1999, contre 24 langues en 1998. D'après la même source, en 1999, 80% des sites Web étaient en anglais, comparés à 84% en 1998.

Répartition en terme de pages Web (nombre de pages)

O'Neill a observé que de 1999 à 2002, les sites Web en anglais occupaient environ 72% des pages Web. Le nombre de pages en japonais a considérablement augmenté pour atteindre 6% en 2002 (3% en 1999) alors que le nombre de pages en d'autres langues était relativement stable : l'allemand 7%, le français 3%, et l'espagnol 3%. L'italien augmentait de 2% en 1999 à 3% en 2002. Contrairement, le chinois diminuait provisoirement de 3% en 1999 à 2% en 2002, ainsi que le portugais de 2% en 1999 à 1% en 2002 [OLB03].

Toutefois, un rapport de l'UNESCO en 2000 constatait que près de deux tiers des pages Web (68%) étaient rédigés en anglais. Cette proportion atteignait 96% pour les sites de commerce électronique.

En 2003 et selon l'étude effectuée par Global Reach les proportions des langues dans les pages Web étaient : l'anglais 68,4%, le japonais 5,9%, l'allemand 5,8%, le chinois 3,9%, le français 3,0%, l'espagnol 2,4%, la russe 1,9%, l'italien 1,6%, le portugais 1,4%, le coréen 1,3%, et d'autres langues 4,6%.

Répartition en terme d'utilisateur

Cependant, le nombre des utilisateurs non-anglophones a évolué plus rapidement que celui des internautes anglophones. En 1999, 48,7% des utilisateurs de l'Internet n'avaient pas l'anglais comme première langue, ce qui représentait une augmentation de 20% par rapport à 1996. En 2000, selon l'UNESCO, les utilisateurs anglophones ne seraient plus majoritaires sur l'Internet(49%), comparé au 92,2% en 1998.

En septembre 2003, l'étude de Global Reach[4] a montré que les Anglophones ne représentaient que 35,8% des utilisateurs de l'Internet suivis par les Chinois (13,7%), les Espagnols (9%), les Allemands (6,9%), les Français (4,2%),

[4]Global Internet statistics (by language) (http ://www.glreach.com/globstats/).

les Japonais (8,4%), les Coréens (3,9%), les Italiens (3,8%), les Portugais (3,1%), les Arabes (1,7%), les Russes (0,8%).

Le sursaut des utilisateurs asiatiques sur l'Internet a été prévu pour 2005 pour que les Chinois atteignent 20%, les Japonais 9% et les Coréens 4,3%. Les Anglophones n'occuperaient que 29%, les Espagnols 7%, et les Allemands 6%.

Pour réagir à la variété linguistique et à l'inondation des utilisateurs multinationaux sur le Web, des moteurs de recherche multilingue ont vu le jour dès 1995 (dont le plus célèbre était Altavista). Cependant, et jusqu'en 2001, plusieurs moteurs de recherche ont été développés en plusieurs langues, il s'agissait donc des moteurs plurilingues : 25 langues sur Google, 11 langues sur Excite, 19 langues sur Altavista, et 44 langues sur AllTheWeb [Lan01].

Aujourd'hui, il est courant que l'anglais ne soit pas automatiquement la langue officielle pour la promotion de l'information sur l'Internet, chose qui n'était pas aussi imaginable il y a une décennie. Néanmoins, les communautés anglophones (l'anglais américain, l'anglais australien, etc.) demeurent prépondérantes. De plus, la langue anglaise (et ses dérivées) reste la plus répandue en tant que première langue étrangère pour diverses raisons dont principalement le facteur économique. O'Neill a observé que 7% des sites Web en 1998 étaient multilingues, et que le nombre des sites Web multilingues devrait se réduire à 5% en 1999 [OLB03]. Lavoie a effectué une enquête auprès de 156 sites Web multilingues en 1999, ils ont déclaré que la présence des langues sur cet échantillon était : l'anglais 100%, le français 31%, l'allemand 31%, l'italien 21%, l'espagnol 21%, le japonais 10%, les portugais 10%, le suédois 10%, le chinois 7%, etc [LO99]. Cette étude montre également que les langues « puissantes » attirent encore un grand nombre des utilisateurs même si ces derniers préfèrent interroger le Web en leurs propres langues maternelles.

L'importance du multilinguisme peut être expliqué par le nombre d'études effectuées à ce propos en comparaison aux autres aspects relatifs au développement de l'Internet tels que : le multimédia, les bibliothèques électroniques, et les bases de données, etc. qui sont encore peu explorés. Par ailleurs, l'intérêt accru qui est porté au multilinguisme peut être aussi expliqué par le développement grandissant des logiciels de traduction automatique ces dernières

années.

1.3 Typologies des sites Web multilingues

Selon le W3C[5], le multilinguisme caractérise les sites Web qui utilisent plusieurs langues (« A multilingual site is concerned with more than just the language »). Cette définition du W3C couvre également les pages Web plurilingues, c'est-à-dire que plusieurs langues peuvent exister dans une même page (« A multilingual site might also mix multiple languages within the same page »).

La notion de site Web multilingue reste relativement évasée, car elle englobe probablement des phénomènes très variés du multilinguisme. La classification d'un site Web multilingue dépend des caractéristiques et des critères d'évaluation de la propriété et de la spécificité multilingue d'une part et d'autre part les différents types d'ancres sources recensées.

Dans un premier temps et afin de déterminer les propriétés les plus concrètes des sites Web multilingues, nous avons effectué une série d'analyses simples (semi-automatiques) d'un ensemble de sites. Ces analyses nous ont permis de retenir trois caractéristiques primitives relatives aux sites Web multilingues :

- la structure de navigation,
- la structure logique interne des pages Web,
- le contenu.

Ces caractéristiques devront nous permettre d'extraire des suites de corrélations relatives aux différentes langues qui seraient utilisées dans un site Web multilingue. Nous observons, dans la majorité des sites Web multilingues, une forte similarité dans les comportements de la première et troisième caractéristiques qui sont la structure de navigation et le contenu. Tandis que la structure logique interne des documents peut être variable dans les langues pour plusieurs buts qui sont très probablement liées à une question de culture de structuration (interne) logique de l'information.

La structure logique interne des pages Web ne donne pas suffisamment d'information sur la visualisation (affichage) du contenu, pourtant c'est l'élément de base de la gestion d'interaction (d'interface) avec l'utilisateur. La

[5]FAQ : International & multilingual web sites (http ://www.w3.org/International/questions/qa-international-multilingual).

visualisation est la partie perçue par l'utilisateur, qu'elle soit vue ou écoutée (dans notre recherche nous ne nous intéressons qu'à la modalité visuelle). De plus, la structure de navigation ne reflète pas non plus la logique de fonctionnement d'un site Web multilingue.

Une démarche d'analyses approfondies et dédiées au problème de reconnaissance des langues dans un site multilingue, nous a permis de discerner une projection orientée de l'ensemble des caractéristiques recensées d'un site Web, sur trois axes essentiels expliquant le comportement (la classification) d'un site Web[6] (cf. tableau 1.1) :

- la visualisation (partie perçue),
- la logique de fonctionnement,
- le contenu.

Caractéristiques explicatives	**Représentation**
Visualisation	Structure physique des pages Web Interface : couleurs, mise en pages, etc.
Logique	Structure de navigation Interactions utilisateurs - site Web
Contenu	Texte, figure, etc.

TAB. 1.1 – Caractéristiques explicatives d'un site Web multilingue

A partir de de ces constats nous avons introduit la notion du « parallélisme multilingue » comme étant le critère le plus important pouvant expliquer la similarité dans les différentes structures de corrélations explorées.

Notion 1 ***Parallélisme multilingue*** *: Le parallélisme multilingue se confirme par la détection de similarités (ou bien d'équivalences) entre les caractéristiques essentielles des langues pouvant co-exister sur un site Web multilingue.*

En se référant à ce critère de parallélisme, nous distinguons, dans un site Web, trois niveaux pour la propriété « multilingue » (cf. tableau 1.2).

Il est à rappeler que l'évaluation de la propriété multilingue des sites Web s'insère complètement dans l'objectif de cette étude qui consiste à définir une

[6]European Environment Information and Observation Network : Multilingual websites structures and definitions (http ://www.eionet.eu.int/software/design/multilinguality/websitestructures).

Parallélisme dans les langues (visualisation/logique/contenu)	**Propriété multilingue**
Complètement parallèle	Forte
Semi-parallèle	Faible
Peu parallèle	Très faible

Tab. 1.2 – Les trois niveaux de la propriété multilingue du site Web

approche universelle (indépendante des notions linguistiques et par conséquent des langues) de classification (reconnaissances) des sites Web multilingues.

En effet, nous distinguons au moins deux types, assez répandus, des sites Web multilingues :

- le premier type consiste à sélectionner par l'utilisateur la langue de son choix pour présenter le contenu. Sachant que l'utilisateur peut changer de langues quand il veut, à l'aide des hyperliens.
- le deuxième type propose différentes langues (mixées), sur la même page Web, pour représenter simultanément le contenu. Bien évidement, et pour des raisons d'ergonomie et de structuration physique de l'information sur les différents supports, ce type de site Web impose des restrictions sur le nombre de langues utilisées par page et sur le volume du contenu fourni par chaque langue.

Les particularités principales de ces deux types sont présentées dans le tableau 1.3.

Site Web multilingue	**Type 1**	**Type 2**
Visualisation	Compliqué	Simple
Logique		
Contenu		
Changement de langue	Direct ou indirect (cf. tableau 1.5)	Non
Nombre de langues	Nombreux	Limité

Tab. 1.3 – Particularités des types de sites Web multilingues

On pourrait également considérer un type particulier de sites Web qui est l'association (partielle ou totale) des sites Web monolingues où le contenu

de chaque langue serait indépendamment localisé (dans la plupart des cas, chaque site monolingue est accédé par une adresse d'URL distincte). Ce type de sites concerne très souvent des sites d'organismes internationaux ou bien de grandes entreprises comme par exemple Coca-Cola[7], Nike[8], etc. Dans le cadre de notre recherche ce type de sites n'est pas considéré comme véritablement multilingue.

Pour consolider ces hypothèses de classification, nous avons évalué plusieurs sites Web relatifs à des organismes internationaux (cf. annexe A pour savoir leurs noms complets). Les résultats de ces évaluations sont illustrés dans le tableau 1.4. Cependant, il faut noter que nos évaluations consistaient à ne repérer que les langues dites dominantes[9].

1.4 Stratégies de changement de langue dans un site Web multilingue

Le changement de langue dans un site Web multilingue s'opère le plus souvent de manière très simple, c'est-à-dire directement par le visiteur en cliquant sur une ancre, que nous appelons « ancre source ». Généralement, cette ancre source est représentée par une icône du drapeau ou le nom de la langue souhaitée. Lors d'un changement de langue sur un site multilingue, l'utilisateur se retrouve toujours sur la page équivalente dans l'autre langue lorsque celle-ci existe.

1.4.1 Ancre source de changement de langue : primaire, secondaire

Le fonctionnement des ancres sources de changement de langue n'est pas similaire sur tous les sites Web multilingues. La discordance entre les contenus des langues peut conduire à un phénomène d'incohérence. Par exemple, il pourrait y avoir plusieurs pages dans une langue quelconque qui ne soient plus accessibles à partir d'autres langues. C'est le cas typique des pages Web d'une langue qui ne disposent pas de leurs équivalences en d'autres langues (retard du développement ou la traduction a été oubliée), ou bien le cas contraire, c'est-à-dire de pages Web, considérées comme moins importantes,

[7]http ://www.coca-cola.com/

[8]http ://www.nike.com/

[9]Abréviations : an (anglais), ar (arabe), ch (chinois), es (espagnol), fr (français), ru (russe)

Site Web	**Nombre de langues dominantes**	**Propriété multilingue**	**Parallélisme multilingue**
EUROPA	20 langues	Forte	Dans toutes les langues
FAO	5 langues (an, ar, ch, es, fr)	Forte	Dans quatre langues : an, fr, es, ar
ILO	6 langues (an, ar, de, es, fr, ru)	Forte	Dans trois langues : an, fr, es
IMF	3 langues (an, es, fr)	Forte	Entre les langues : fr-an, fr-es, es-an, es-fr
UN	6 langues (an, ar, ch, fr, es, ru)	Faible	Compliqué
UNDP	3 langues (an, es, fr) (non-compris 18 sites régionaux)	Très faible	Compliqué
UNFPA	4 langues (an, ar, es, fr)	Forte	Dans trois langues : an, fr, es
UNICEF	3 langues (an, es, fr) (non-compris 37 sites régionaux)	Faible	Compliqué
WB	19 langues (distribuées en plusieurs sites régionaux)	Non-multilingue	
WTO	3 langues (an, es, fr)	Forte	Dans toutes les langues

Tab. 1.4 – Caractéristiques des sites Web étudiés

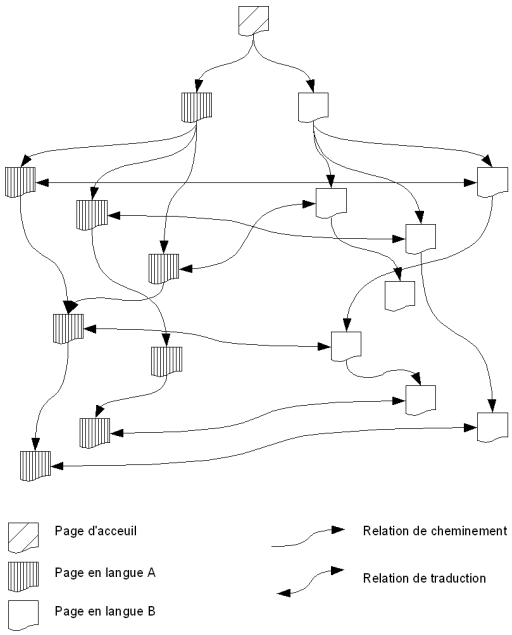


FIG. 1.1 – Changement de langue : méthode directe

n'ayant pas d'ancres sources de changement de langues pointant sur leur traductions, pourtant celles-ci existent.

Pour les sites Web multilingues du premier type, le changement de langue peut se faire de deux manières (cf. tableau 1.5).

Changement de langue	**Méthode de changement de langue**
Directe (cf. figure 1.1)	Ancre de changement de langue redirigeant sur la page Web correspondante de la langue ciblée
Indirecte (cf. figure 1.2)	Ancres de changement de langue redirigeant sur la page Web « accueil » de la langue ciblée

TAB. 1.5 – Méthodes de changement de langue

Dans le cas particulier des très grands sites Web, nous pouvons observer la présence, sur certaines pages, de différentes formes d'ancres source de changement de langue qui assurent les mêmes fonctions. Par exemple dans un site Web bilingue français - anglais, la majorité des pages en anglais utilisent une

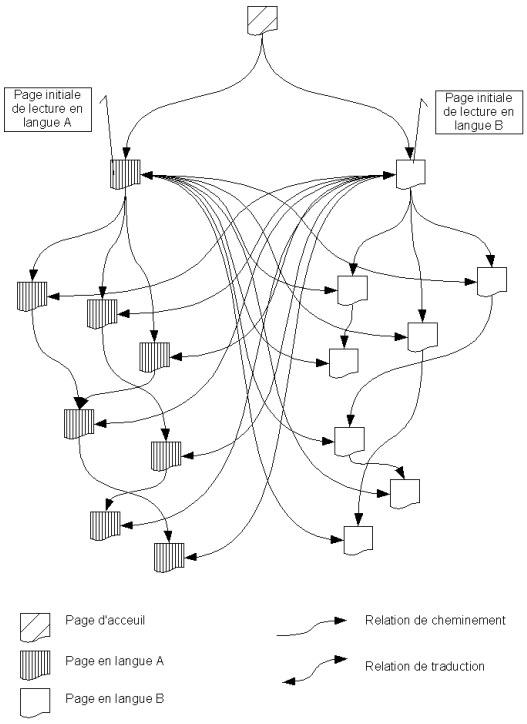


FIG. 1.2 – Changement de langue : méthode indirecte

ancre source nommée « français » pour afficher leurs traductions en français, alors que d'autres pages, toujours en anglais, peuvent utiliser une ancre source nommée « FR » pour diriger vers leurs traductions en français.

Nous distinguons donc les « ancre sources primaires » qui sont les ancres source de changement de langue présentes sur la page d'accueil et des « ancres sources secondaires » qui seraient présentes sur d'autres pages du site sous une forme différente que les ancres sources primaires. Dans la plupart des cas, les ancres sources secondaires sont des abréviations des ancres sources primaires quand celles-ci sont textuelles.

1.4.2 Ancre source partagée par plusieurs langues

Dans plusieurs sites Web multilingues, il existe des ancres sources (pas de changement de langue) qui sont présentes sans modification dans plusieurs langues. Par exemple, une ancre source « Einstein » peut être présente dans toutes les langues d'un site Web multilingue. Nous considérons que cette ancre source est partagée par plusieurs langues.

Notion 2 ***Ancre source partagée par plusieurs langues*** *: Une ancre source partagée est présente dans plusieurs langues du site Web multilingue. Dans chaque langue, elle permet de diriger vers une page de la même langue.*

Dans ce cas de figure, nous distinguons deux types de situations :

- le premier type : une ancre source partagée par plusieurs langues et ne provoquant pas de changement de langue. Donc son effet d'hyperlien, reste monolingue. Par exemple, dans un site Web bilingue anglais - français, l'ancre source « Einstein » est présente sur les pages en français et en anglais. Lorsqu'elle est sur une page en anglais cette ancre pointe vers une page en anglais. De même lorsque cette ancre source « Einstein » est présente sur une page en français, elle pointe vers une page en français (cf. la figure 1.3).
- le deuxième type : il s'agit d'une ancre source partagée par plusieurs langues et pointant toujours sur un document fixe existant en une seule langue. Ce type d'ancre peut provoquer un changement de langue. Par exemple, dans un site Web bilingue anglais - français, une ancre source « Einstein » qui est présente sur une page qu'elle soit en français ou en anglais, pointe vers une page en anglais.

Le deuxième type de situation pose des problèmes sérieux de confusion entre les ancres sources partagées par plusieurs langues et les ancres sources pour la référence interlingue.

1.4.3 Ancre source pour la référence interlingue

Nous rencontrons ce type d'ancre source lorsqu'il s'agit d'un document écrit en une langue qui utilise des documents de référence existant en d'autres langues dans un site Web multilingue.

Notion 3 ***Ancre source pour la référence interlingue*** *: Ce n'est pas une ancre partagée, il s'agit bien d'une ancre se trouvant sur une page en une langue quelconque et qui est utilisée pour pointer vers une autre page (de référence) qui est en langue différente.*

Par exemple c'est le cas d'un site Web bilingue illustré sur la figure 1.4 et qui montre que l'ancre source « Newton » qui est présente sur une page en langue A fait référence à une page en langue B différente de A.

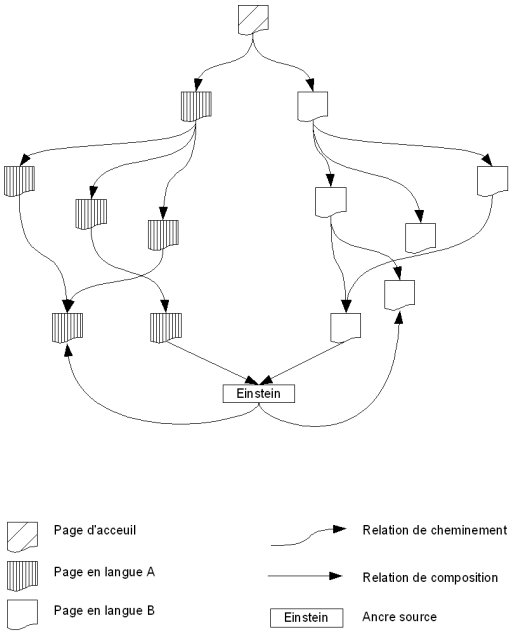


FIG. 1.3 – Exemple d'une ancre source partagée par deux langues

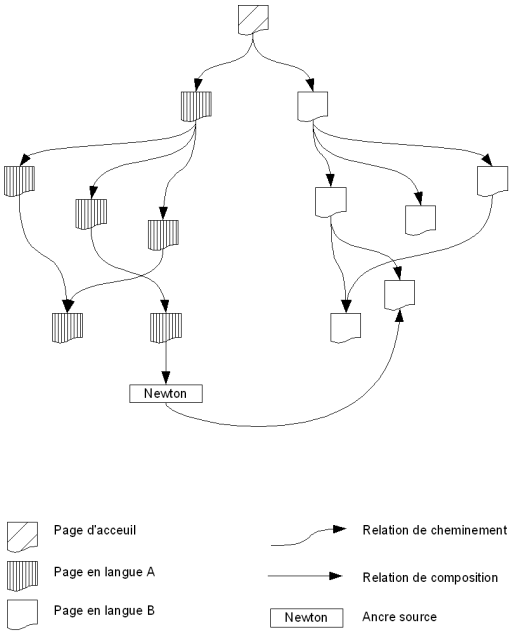


FIG. 1.4 – Exemple d'une ancre source pour la référence interlingue

Ainsi, nous pouvons constater que le deuxième type de situation de présence d'ancre source partagée (cf. section 1.4.2) peut bien être considérée comme un cas particulier d'une ancre source pour la référence interlingue. L'exemple dans la figure 1.5 indique l'ancre source « Einstein » qui est présente en deux langues (A et B) d'un site Web bilingue et ne faisant référence qu'à une seule page en langue B. C'est pourquoi elle peut être considérée comme une ancre source pour la référence interlingue.

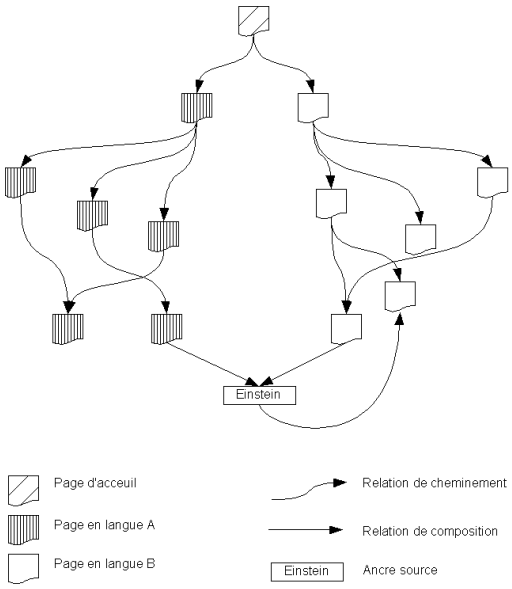


FIG. 1.5 – Exemple de la confusion entre une ancre source pour la référence interlingue avec une ancre source partagée par deux langues

1.5 Document complémentaire

L'une des caractéristiques remarquables qui se présente très souvent dans la structure des sites Web multilingues est l'existence des documents complémentaires.

Notion 4 ***Document complémentaire*** *: Il s'agit d'un document existant en plusieurs langues utilisées sur un site Web multilingues mais il n'existe pas dans toutes les langues.*

Dans ce contexte nous distinguons deux types de situations :
- le premier type : le document n'a pas été traduit en d'autres langues,

– le deuxième type : le document n'est pas important, l'auteur ne souhaite pas (ou oublie) mettre de liens vers ses traductions, pourtant celles-ci existent.

Les documents complémentaires peuvent être également d'une grande importance lorsqu'il s'agit d'un problème de recherche d'information, ou d'extraction d'information à partir des sites Web multilingues.

1.6 Évolution des sites Web multilingues

Il y a plusieurs années, un des plus grands événements du Web était l'arrivé des services de traduction en ligne proposés par AltaVista et par FreeTranslation. Ces services étaient, entre autres, à l'origine de l'accélération de développement des outils de traduction automatique des sites Web. La progression de la qualité des traducteurs automatiques a ouvert une perspective alternative pour la création des sites Web multilingues. Les sites Web monolingues vont pouvoir être traduits, avec un coût d'assistance raisonnable, en plusieurs langues par des logiciels spécialisés.

Cependant, il faut noter que la qualité des traducteurs automatiques reste encore problématique, surtout entre les langues de familles différentes. Ces difficultés sont principalement d'ordre linguistique (c'est-à-dire des problèmes syntaxiques, morphologiques et surtout sémantiques). Autrement dit, les traducteurs automatiques sont certainement incapables de répondre aux demandes de la bonne qualité de traduction. En dehors des textes scientifiques, techniques et commerciaux, la traduction automatique n'aura pas beaucoup de promesses envers les oeuvres culturelles, ainsi que les documents juridiques, diplomatiques et politiques.

1.7 Conclusion

La diversité des langues sur l'Internet, généralement, et sur le Web, particulièrement, est issue de deux dynamiques : la standardisation rapide des technologies d'information dans le monde entier, et la puissance potentielle du Web en tant qu'environnement de distribution de ressources d'information. En même temps, les activités humaines sur le Web suivent le mouvement de l'internationalisation et de la globalisation. Ce fait laisse considérer que les langues européennes (comme l'anglais et le français par leurs forces

culturelles, économiques, politiques, ou par des facteurs historiques) sont encore susceptibles de retenir un nombre considérable d'utilisateurs Internet, au détriment d'autres langues (les langues natives). C'est pour faire face à ce frein culturel, essentiellement, que la naissance et la croissance rapide du multilinguisme sur le Web ont pu voir le jour.

Dans ce chapitre, nous avons abordé différents types de sites Web mutilingues ainsi que les différentes stratégies de navigation entre les langues. Nous avons également mis en exergues que quelques notions, considération et confusions pourront être rencontrées lors du traitement des sites multilingues.

Aussi, nous avons appuyé notre démarche par des expérimentations dont l'objectif était l'évaluation des hypothèses relatives aux notions du parallélisme linguistique et des ancres sources. Nous avons également présentés les différents types d'ancres sources pouvant exister sur un site Web multilingues telles que : les ancres sources de changement de langue, les ancres sources partagées par plusieurs langues et les ancres sources pour la référence interlingue.

Dans le cadre de cette thèse, nous nous sommes intéressé exclusivement aux sites relevant du premier type de site multilingue (langues non mixées), car l'usage et le développement des sites Web multilingues relatifs au deuxième type (langues mixées) sont de plus en plus rares. Tous ces éléments importants permettent de former une connaissance sur la topologie des sites Web multilingues.

Nous consacrons le chapitre suivant, pour exposer l'étude sur l'état de l'art en extraction d'information dans les sites Web multilingues que nous avons mené.

Chapitre 2

Extraction d'information à partir des sites Web multilingues

2.1 Introduction

L'extraction d'information est un terme appliqué à l'extraction automatique des types d'information pré-spécifiés dans les textes en langues naturelles. Elle est ainsi considérée comme un processus permettant de former une structure d'informations (en particulier une base de données) à partir des sources d'informations textuelles et non structurées. Classiquement ce terme est émanant des recherches sur la linguistique computationnelle et le traitement automatique des langages naturelles [GW98], en étant pendant longtemps influencé par les travaux en compréhension du texte [Eik99], jusqu'avant la fin des années de 1980[1].

Dans ce chapitre, nous présentons d'abord une brève histoire de l'extraction, automatique ou semi-automatique, d'information en mettant l'accent sur les tendances d'évolution dans ce domaine. Nous abordons ensuite les stratégies en cours de développement dans cette discipline. Enfin, nous nous intéressons tout particulièrement à l'extraction d'information à partir des sites Web multilingues.

[1]Particulièrement, Popov a considéré l'extraction d'information comme une nouvelle discipline dans le traitement automatique des langues naturelles [PKK+03].

2.2 Approche de représentation diminuée de l'information

Nous rappelons que l'un des objectifs de notre recherche est de rester autant que possible indépendant des spécificités linguistiques des langues lors du traitement des sites Web multilingues. Dans cette partie, la notion « information » est volontairement réduite à deux aspects principaux pour notre problématique de recherche, à savoir : la forme et la structure.

2.2.1 Forme de l'information

L'information est une notion variée selon les disciplines scientifiques, comme la thermodynamique avec le concept d'entropie, la physique avec la théorie du signal, la biologie avec la théorie du génome ou l'économie avec la théorie de décision. Il en est de même en ce qui concerne les secteurs socioprofessionnels comme par exemple le journalisme, l'administration publique, etc. Toutefois, « le principal artisan de la théorie de l'information est Shannon (1948) dont les travaux reposent sur les études de Kupfmuller (1924), Hartley (1928) et Whittaker (1935) » [2]. Dans sa théorie, Shannon a radicalement réduit l'information à un phénomène physique quantifiable : « l'information est la mesure de l'entropie, de la dégradation du signal en présence du bruit ».

La théorie de Shannon, a été l'objet de nombreuses critiques portant sur les applications de la théorie statistique de la communication. En opposition à la position de Shannon, certains auteurs ont poursuivi dans des voies différentes. Bar-Hillel a développé une « théorie de l'information sémantique » basée sur la logique des propositions, indépendamment de toute transmission [BH52]. Les travaux de Barwise sur la théorie des situations se sont situés au niveau de la pragmatique, et considèrent que l'information est fortement liée au contexte [Bar89]. Jakobson a proposé un « schéma de la communication humaine » comprenant un émetteur, un récepteur, un contexte, un contact entre eux, un code commun et enfin un message [Jak63]. Kerbrat-Orecchioni a reformulé ce modèle en y ajoutant la notion d'univers du discours comprenant : conditions concrètes de la communication, contraintes sur le thème du discours, la nature particulière de l'émetteur et du destinataire [KO80].

Toujours dans l'esprit de la théorie de Shannon, l'information, dans son usage technique, incarne deux sens distincts :

[2] Encyclopédie Hachette Multimédia 2005.

– l'information est une quantité, au sens strict de la théorie de l'information, mesurée à l'aide d'une formule qui est sensiblement la même, (mais avec un signe inversé) que celle utilisée par le physicien Ludwig Boltzmann à la fin du XIXe siècle pour mesurer l'entropie des gaz.
– le terme « information » est également utilisé pour désigner un symbole numérique (0 ou 1) qui est codé de façon binaire [Bre93].

Le dernier usage du mot « information » découle d'une distinction essentielle entre la forme et le sens du message[3]. Dans cette thèse, nous avons bien des raisons pour adopter a priori une interprétation diminuée de l'information basée plutôt sur sa forme plutôt que sur ses sens. Aussi nous nous intéressons, dans le contexte de notre recherche, à la structure (de cette forme) de l'information comme élément de base pour les traitements que nous proposons (c'est-à-dire les analyses structurelles pour la reconnaissance des caractéristiques multilingues d'un site Web).

2.2.2 Structure de l'information

A propos de la structure de l'information, Géry s'y appuie pour distinguer quatre types d'information : l'information atomique, l'information structurée, l'hyper-information et l'hyper-information contextuelle. Pour lui, l'information ne peut exister indépendamment d'un contexte. L'information est fortement liée aux types de documents qui l'incarnent (la représentent) : le document atomique, le document structuré, l'hyperdocument et l'hyperdocument contextuel (dynamique tel qu'un site Web) [Gér02].

Dans ce contexte, le concept « document » est lié à la façon dont sont organisés les composants véhiculant l'information. Le document comprend alors une structure décrite par un « code » approprié comme par exemple des notations comme : SGML (Standard Generalized Markup Language), HTML (Hypertext Markup Language) ou bien XML (Extensible Markup Language). Le document exige aussi des représentations (audio, visuelle, etc.) appropriées sur différents supports matériels (informatique, imprimé, etc.). Selon Estival un document serait : « Toute connaissance mémorisée, stockée sur un support, fixée par l'écriture ou inscrite par un moyen mécanique, physique, chimique, électronique, constitue un document » [EM81].

Toute évolution du concept document est concrétisée par la mise en place de nouveaux modèles du document comme par exemple le texte, l'hypertexte

[3]Encyclopédie de l'Agora (http ://agora.qc.ca/).

ou l'hypermédia, etc. Un modèle du document est une norme définie proposant un système formel (ou des notations) permettant de décrire et construire un document. Un langage est susceptible de définir plusieurs modèles de documents.

La description de la structure d'un document consiste à identifier et à décrire chacun des éléments textuels - ou non textuels - qui le constituent. En effet, on distingue, en général, deux types de structure : la structure physique et la structure logique. En matière de structure physique du document, on décrira sa mise en page, on définira les différentes zones de texte, leur agencement les unes par rapport aux autres ainsi que l'ensemble de leurs caractéristiques typographiques : police, couleur, gras, italique, etc. Pour la structure logique, on décrira plutôt le rôle, le comportement et la nature de chaque élément d'un document ainsi que l'ensemble des liens hiérarchiques et/ou logiques qui les lient les uns aux autres.

2.3 Extraction d'information

Historiquement, la première idée de l'extraction d'information avait été inaugurée par le linguiste américain Z. Harris pendant les années 1950 et a été officiellement introduite dans le cadre du programme MUC (Message Understanding Conferences) vers la fin des années 1980 [GW98], [GS96]. En effet, ces repères historiques étaient particulièrement distinctes car à l'époque, ces idées d'analyses structurelles et quantitatives se sont faites remarquées par leur originalité et leur audace. L'histoire de l'extraction d'information, d'après notre recherche bibliographique, peut être répartie en trois périodes ou intervalles temporeles : avant le programme MUC, pendant le programme MUC et après le programme MUC.

2.3.1 Première période : avant le programme MUC

Initialement, l'extraction d'information ne concernait qu'un certain nombre de projets que Gaizauskas a classé en travaux avant le programme de MUC[4] dans lesquels cet auteur a cité deux projets de recherche à long terme de type de traitement des langues naturelles [GW98] :

- LSP (Linguistic String Project), a été démarré au milieu des années 1960 et a durée jusqu'au début des années 1980 à l'université de New

[4]Gaizauskas a distingué les développements de l'extraction d'information en trois grandes catégories : les premiers travaux avant le programme de MUC, les travaux menés dans le cadre du programme de MUC et les travaux hors de programme de MUC [GW98].

York. Ce projet consistait à développer une grammaire computationnelle de l'anglais pour créer des formes d'information régularisées (c'est-à-dire des motifs) dans le domaine médical.
- FRUMP, est basé sur le modèle de R. Schank et a été réalisé à l'université de Yale, pour la compréhension de la langue, en particulier des textes d'histoire [DeJ82].

Suite à ces projets, les années 1980 ont vu les premiers développements des systèmes commerciaux [GW98] :
- ATRANS, a été conçu pour le traitement automatique des messages de transfert d'argent entre les banques. Il adopte l'approche effectuée à l'université de Yale pour remplir un motif afin de lancer des transferts d'argent automatiques après une intervention de vérification manuelle (humaine) [LG86].
- JASPER, a été développé par Carnegie Group pour Reuteurs. Il analyse des communiqués de presse sur PR Newswire pour alimenter « un module » proposant des informations sur les revenus et les dividendes des entreprises.
- SCISOR, a été développé par General Electric pour analyser les fusions et acquisitions des corporations [JR90].

Il y a eu également deux autres projets de recherche académique [GW98] :
- le projet développé par J. Cowie pour extraire des descriptions régulières (c'est à dire les motifs) des plantes dans les guides de fleurs. L'approche de J. Cowie a été expérimentée sur un domaine très spécifique et s'est appuyé sur un ensemble de mots-clés manuellement choisis.
- le projet développé par G. P. Zarri pour traduire automatiquement les textes historiques en français en un métalangage capturant certaines relations sémantiques sur les détails biographiques.

La caractéristique commune de ces premiers projets était l'application du remplissage des motifs avec l'information extraite à partir des textes en langues naturelles dont le traitement restait encore manuel, et propre aux domaines spécifiques.

2.3.2 Deuxième période : durant le programme MUC

En première période (avant MUC), l'extraction d'information était influencée par des travaux de recherches en compréhension du texte, basés essentiellement sur des approches et des techniques linguistiques.

Par la suite, les recherches se sont concentrées sur l'extraction d'information à partir des données textuelles dans le but de résoudre des problèmes linguistiques, à l'exception de quelques projets transitionnels ayant vu le jour à la fin de la première étape de MUC et grâce à l'accélération de productions de documents Web semi-structurés sur l'Internet.

Trois tendances dominantes ont marqué cette deuxième période [GW98] : l'adaptation de traitements linguistiques aux spécificités des systèmes (des automates), l'acquisition automatique des règles d'extraction et l'intégration de modules relativement indépendants.

Les systèmes les plus connus, issues de ce mouvement, d'après MUC-3 (1991), sont : TACITUS [Hob91], Proteus [GS93] et PIE [Lin95]. Ainsi, le domaine de l'extraction d'information s'est attribué une mission officielle et s'est donné un ensemble coordonné de tâches focalisant essentiellement sur l'extraction des motifs. Cette évolution a été bien démontrée par les projets qui ont été présentés dans les différentes rencontres allant de MUC-4 (1992) à MUC-6 (1995). A titre d'illustration nous citons le système SRI [AHB+95], FASTUS [HAT+92], SRA [Kru95] et TIPSTER [Gri95]. La majorité de ces derniers n'intégrait pas de traitements sophistiqués en linguistique computationnelle, ce qui n'était pas tout a fait le cas des systèmes comme LASIE [GWH+95], PLUM [Wei95], et MITRE [ABD+95].

Les cinq tâches fondamentales de l'extraction d'information, qui ont été définies par le programme MUC, sont :

- NE (Named Entity) : reconnaissance des entités nommées,
- CO (Coreference) : résolution des coréférences,
- TE (Template Element) : construction des éléments de motif,
- TR (Template Relation) : construction des relations de motif,
- ST (Senario Template) : production des motifs de scénario.

Pour établir un modèle standard d'extraction d'information, Hobbs a proposé un système comprenant dix modules [HAT+92] :

- segmentation du texte,
- pré-traitement d'un segment de texte en phrases,
- filtrage des phrases,
- pré-analyse des structures lexicales comme par exemple les groupes nominaux, groupes verbaux et appositions,
- analyse des éléments lexicaux et des structures de phrase,
- combinaison de fragments,

- interprétation sémantique,
- enlèvement de l'ambiguïté lexicologique,
- résolution de coréférence,
- création des motifs.

Cette période a été marquée par la réalisation de plusieurs projets Européen tels que POETIQUE (Portable Extendable Traffic Information Collator), SINTESI (Sistems INtegrato per TESti in Italiano), TREE (TRans European Employment), et FACILE (Fast Accurate Categorisation of Information using Language Engineering). Ainsi que certains projets du programme LC CEC (Language Engineering, Commission of the European Communities) à savoir : AVENTINUS et ECRAN.

2.3.3 Troisième période : après le programme MUC

Dans cette période de nombreux systèmes ont été présentés : WHISK [Sod99], RAPIER [Cal98], SRV [Fre98], WIEN [Kus97], SoftMealy [HD98] et STALKER [MMK99]. Nous y observons trois nouvelles tendances [Cun05] : la portabilité des systèmes d'extraction d'information, l'extraction automatique des contenus et l'annotation pour le Web sémantique.

La portabilité des systèmes d'extraction d'information

L'adaptation des systèmes existants aux nouveaux domaines d'application était une tâche difficile dans laquelle trois grands courants de travail se sont distingués [Cun05] :

- l'apprentissage des règles d'extraction à partir des exemples annotés [Car97],
- le développement des algorithmes d'apprentissage automatique [Gri01],
- le développement des règles et des modèles par l'observation et le traitement des manipulations effectuées par du personnel qualifié [CS04].

L'extraction automatique des contenus

Le programme ACE (Automatic Content Extraction), commencé en septembre 1999, a permis le lancement d'une nouvelle génération d'applications robustes en traitement des langues naturelles en favorisant un développement plus rapide, assuré par des systèmes autonomes de traitements de corpus annotés [May03]. Les résultats potentiels de ce programme sont relatifs à la recherche documentaire, l'exploitation de données, le développement de

grandes bases de connaissances et l'annotation automatique pour le Web Sémantique.

L'annotation pour le Web sémantique

L'annotation des pages Web ainsi que la création des ontologies sont devenues des tâches automatiques ou semi-automatiques. Cela a donné naissance à tout un nouveau domaine de recherche intitulé OBIE (Ontology-Based Information Extraction) [BW04]. OBIE s'est donné deux défis principaux [Bri98] :

- l'identification de nouveaux concepts et des exemples dans le texte pour enrichir l'ontologie du Web,
- la classification des plateformes d'annotation sémantique en plusieurs catégories primaires, basées sur le motif ou l'apprentissage automatique[5] ou une combinaison de deux approches.

De nombreux systèmes ont été développés pour surmonter ces défis [RH05] : AeroDAML [KH01], Armadillo [DCW03], KIM [PKK+04], Magpie [DD04], MnM [VVMD+02], MUSE [MTB+03], Ont-O-Mat [HSC02] et SemTag [DEG+03].

AeroDAML [KH01] utilise une approche basée sur le motif pour assigner des noms propres et des relations communes aux classes correspondantes et attributs désignés par l'ontologie de DAML[6]. Ce système comprend le composant AeroText qui est un API (Application Programming Interface) Java pour l'extraction d'information. AeroText organise l'usage de l'ontologie en deux niveaux : le niveau supérieur qui consiste en une hiérarchie des noms de WordNet [Fel98], et le niveau inférieur qui est une base de connaissances. AeroText se compose de quatre composants principaux : un compilateur pour transformer des données linguistiques en une base de connaissances, un moteur pour traiter les documents source, un IDE (Integrated Development Environnement) pour construire et tester la base de connaissances et une base de connaissances commune contenant des règles indépendantes du domaine pour extraire des noms propres et des relations.

Armadillo [DCW03] est une évolution du système Amilcare intégrant un module d'induction d'adapteur (induction wrapper) aux sites Web ayant une structure très régulière. Armadillo a une approche basée sur le motif pour

[5]Les plateformes d'annotation sémantique basées sur l'apprentissage automatique utilisent deux approches : probabiliste et inductive.

[6]DAML - DARPA Agent Markup Language (http ://www.daml.org/).

chercher des entités nommées. Aucune annotation manuelle n'est exigée. Ce système fait appel aux services Web de Google et CiteSeer pour vérifier et confirmer ou refuser les entités trouvées.

KIM [PKK+04] est composé d'une ontologie, une base de connaissances, une annotation sémantique, une indexation et un serveur. KIM utilise le répertoire de SESAME RDF [BKvH02] pour stocker l'ontologie et la base de connaissances. Le processus d'annotation sémantique se fonde sur une ontologie pré-construite KIMO et une base de connaissances d'inter-domaines. Le composant d'extraction d'information pour l'annotation sémantique réutilise des composants de l'outil GATE [CMBT02].

Magpie [DD04] associe automatiquement une couche sémantique à une ressource Web, plutôt que de faire l'annotation manuelle, en s'appuyant sur une ontologie proposée par [Gru93]. Magpie est considéré comme une avancée vers le navigateur Web sémantique.

MnM [VVMD+02] fournit un environnement pour annoter manuellement un corpus d'apprentissage. Il intègre également des mécanismes d'induction basés sur l'algorithme Lazy-NLP. Il livre les résultats sous la forme d'une bibliothèque de règles d'induction permettant d'extraire l'information à partir des textes (de corpus).

MUSE [MTB+03] a été conçu pour la reconnaissance des entités nominatives et des coréférences. Pour sa mis en application, il utilise le framework GATE [May03]. Les modules d'extraction d'information (Processing Resources) forment un canal de traitement pour découvrir les entités. L'étiquetage sémantique est accompli à l'aide du JAPE [CMBT02].

Ont-O-Mat [HSC02] est une implémentation du framework d'annotation sémantique S-CREAM (Semi-automatic CREAtion of Metadata). Ont-O-Mat adopte les outils d'extraction d'information de type Amilcare. Ce dernier utilise le module ANNIE (A Nearly-New IE system) proposé dans le cadre de GATE pour extraire l'information. ANNIE transmet les résultats à Amilcare, qui en induit des règles d'extraction d'information. Ultérieurement, le module d'annotation de l'Ont-O-Mat est remplacé par l'algorithme PANKOW (Pattern-based Annotation through Knowledge On the Web) [CHS04], assez proche de celle utilisée par Armadillo [DCW03].

SemTag [DEG+03] est le module d'annotation sémantique d'une plateforme appelée Seeker. Il annote des pages Web en trois phases : repérage, apprentissage, et étiquetage. SemTag/Seeker est un système extensible dont les nouvelles implémentations peuvent remplacer l'algorithme TBD (Taxonomy-based Disambiguation). SemTag utilise la taxonomie TAP, qui couvre une gamme d'informations lexicologiques et taxonomiques issues des articles non spécialisés et assez variés (musique, cinéma, sport, santé, etc.).

2.4 Stratégies d'extraction d'information

La conception des systèmes d'extraction d'information est basée sur deux approches fondamentales : l'ingénierie de connaissances et l'apprentissage automatique [App99].

Dans l'ingénierie de connaissances, les règles grammaticales sont manuellement construites en utilisant des connaissances linguistiques. Le développement de tel systèmes est véritablement laborieux d'autant plus que leur performance dépend très souvent de l'expérience humaine.

Bien que le développement de l'approche d'apprentissage automatique reste plus rapide que celui de l'ingénierie de connaissances, l'apprentissage automatique exige néanmoins un volume de données assez conséquent (quantitativement mais aussi qualitativement) [Eik99]. Très souvent on utilise des méthodes d'apprentissage supervisé. Roth a proposé une méthode probabiliste pour reconnaître des entités et des relations [RtY02]. Suzuki a adopté les graphes pour la représentation des données [SHSM03]. Toutefois, Yangaber a proposé une méthode d'apprentissage non-supervisé [YG98].

L'adoption de telle ou telle approche d'extraction d'information (ingénierie de connaissance ou apprentissage automatique) dépend de la structure des documents à traiter. L'ingénierie de connaissances est historiquement appliquée aux textes non-structurés [AMM97], [CL96]. Au contraire, l'apprentissage automatique est adéquat pour les textes semi-structurés ou structurés [Fre98], [HD98], [Kus97], [MMK99], [Sod99].

L'extraction d'information se divise actuellement en deux branches principales dépendant de la nature « structurelle » des ressources : textes (données non-structurées) et documents Web (données semi-structurées/structurées) [MMK99].

L'extraction d'information à partir des textes a fait l'objet de nombreux développements [MMK99] : AutoSlog [Ril93], LIEP [Huf95], PALKA [KM95], CRISTAL [SFAL95], CRISTAL-Webfoot [Sod97] et HASTEN [Kru95].

L'extraction d'information à partir du Web peut être vue comme une extension ou une généralisation « complexe » de l'extraction d'information à partir des ressources textuelles. Les documents Web sont en majorité semi-structurés bien qu'il est possible de trouver des documents structurés ou non-structurés[7]. Les approches d'extraction traditionnelle c'est-à-dire à partir des textes non-structurés ou bien à partir des bases de données (fortement structurées) ne semblent pas être appropriées aux documents Web.

Les méthodes applicables dans ce champs d'étude visent alors à analyser des méta-données disponibles dans les documents semi-structurés, comme des balises HTML [Sod99], des délimiteurs [MMK99] ou de simples unités syntaxiques [Kus97]. La nécessité des applications pouvant extraire et intégrer l'information à partir des multiples sources Web a conduit à développer un nouveau champ d'étude où l'extraction d'information est réalisée par des outils spéciaux s'appelant « adaptateur » ou « extracteur » qui n'utilisent pas de contraintes linguistiques. L'adaptateur analyse la source d'information dans les pages Web, et transforme le contenu extrait sous une forme prédéfinie.

Laender a distingué plusieurs courants de développement d'adaptateur selon qu'ils soient basés sur : les langages de description, l'analyse de la structure du document Web, la modélisation de la structure du document Web, le traitement automatique des langues naturelles, le mécanisme d'induction ou la base d'ontologie [LRNdST02].

2.4.1 Adaptateurs à base de langues descriptives

Une des premières approches primitives est de définir des langues descriptives pour assister l'utilisateur à construire des adapteurs. Nous pouvons parler des systèmes les plus connus : Minerva [CM98], TSIMMIS [HGMN+97], Web-OQL [AM99] et LIXTO [BFG01].

[7] La propriété structurelle du document Web est une notion relative et dépend de critères qui la caractérisent. Hsu a proposé ses critères pour distinguer les documents Web en trois types : non-structurés, semi-structurés et structurés [HD98].

Minerva [CM98] est un module important du système Araneus. C'est un outil pour construire des adaptateurs possédant une grammaire descriptive décrite en EBNF (Extended Backus Naur Form) : pour chaque document, un ensemble de « productions » est défini : chaque « production » définit la structure d'un symbole non-terminal dans la grammaire. Minerva est doté d'un langage pour la recherche et la restructuration des documents, appelé Editor, qui assure également les fonctions de base d'un éditeur de texte.

TSIMMIS [HGMN+97] est un système qui permet à l'utilisateur de spécifier des règles d'extraction de données semi-structurées à partir d'une page Web. Il comporte des adaptateurs pouvent être configurés par des fichiers de spécifications, écrits par l'utilisateur. Chaque fichier de spécification est décrit par une séquence des commandes qui définissent les étapes d'extraction. Chaque commande est écrite sous la forme de [variables, source, motif] où « variables » indique des variables contenant les résultats d'extraction, source désigne le document Web et motif décrit des données à reconnaître.

Web-OQL [AM99] est un langage de requête de type déclaratif qui est capable d'extraire des motifs choisis dans les pages HTML. Pour ce faire, un adaptateur générique analyse la page d'entrée et présente le résultat sous la forme d'un arbre abstrait de syntaxe HTML, appelé hypertree, représentant le document. Avec cette syntaxe, il est possible d'écrire des requêtes qui localisent les données désirées dans l'arbre et transforment ces données en une structure comme le tableau.

LIXTO [BFG01] est un système dont l'objectif est d'aider l'utilisateur à créer de manière semi-automatique des adaptateurs via une interface visuelle et interactive. Il propose un langage de description des règles d'extraction (Elog) qui est basé sur la logique du premier ordre. LIXTO peut aussi transformer des données extraites à partir d'une page HTML en XML.

2.4.2 Génération (par induction) d'adaptateurs à partir des pages étiquetées

Dans cette approche, les méthodes de construction d'adaptateurs utilisent l'apprentissage à partir des pages exemples étiquetées. Dans l'ordre nous présentons les systèmes WIEN [Kus97], STALKER [MMK98] et SOFTMEALY [HD98].

WIEN [Kus97] propose la première formalisation de la génération par induction d'un adaptateur. Il s'agit d'un ensemble d'outils pouvant étiqueter automatiquement des documents. L'apprentissage inductif est sollicité ici pour générer un adaptateur (ensemble de règles) à partir d'un ensemble de pages étiquetées.

STALKER [MMK98] est un système d'apprentissage non-supervisé des règles d'extraction. STALKER introduit le concept des arbres-EC (Embedded Catalog Tree) permettant de décrire la structure logique du document. Il propose également la transformation des documents en séquence de symboles, la représentation de règles d'extraction sous forme d'automates, l'adaptation d'une méthode d'induction par raffinements successifs et l'élargissement de la notion de délimiteur.

SOFTMEALY [HD98] cherche principalement à résoudre des problèmes liés à l'ordre dans lequel les attributs sont représentés (apparition non séquentielle) et aux attributs manquants. Les règles d'extractions doivent tenir compte des différentes permutations entre les attributs apparaissant dans les occurrences d'une relation à extraire. SOFTMEALY propose une approche qui n'est plus basée sur des délimiteurs mais sur des séparateurs. Un séparateur permet de caractériser une position à la fois à partir du texte se trouvant juste avant cette position et du texte se trouvant juste après. Un séparateur prend alors en compte à la fois ce qui se trouve à gauche et à droit de la position qu'il détermine. Cette position correspond soit au début ou à la fin d'une valeur. Ainsi, même le format du contenu de cette valeur est pris en compte par le séparateur.

2.4.3 Génération d'adaptateurs par l'extraction de motifs : analyses de la structure du document

Au constat de la régularité des séquences de balises suivant les mêmes formats dans la structure des documents, l'extraction de motifs via l'analyse de la structure des documents permet de générer des expressions (des motifs) décrivant le format général dans lequel se trouvent des données à extraire. Quelques systèmes peuvent être présentés dans cette approche comme par exemple W4F [SA99], XWRAP [LPH00], ROADRUNNER [CMM01] et IEPAD [CL01].

W4F [SA99] est un outil pour construire des adaptateurs en divisant le processus de développement en trois phases : l'utilisateur décrit d'abord

la façon d'accéder au document, puis il décrit les données recherchées pour déclarer enfin la structure pour stocker les résultats. Quand un document est trouvé sur le Web, d'après des règles de recherche, W4F le transmet à un analyseur qui en construit un arbre DOM (Document Object Model). L'utilisateur peut écrire des règles d'extraction grâce au langage HEL (HTML Extraction Language) pour extraire des données de l'arbre. Les données extraites sont stockées sous la forme NSL (Nested String List), un format de W4F, avant d'être transformées en d'autres formats pour différentes applications.

XWRAP [LPH00] représente les documents sous forme d'arbres en construisant une bibliothèque de composantes et une interface interactive pour guider l'utilisateur à créer des adaptateurs en Java pour chaque source spécifique.

ROADRUNNER [CMM01] explore les attributs dans les pages HTML pour construire automatiquement des adaptateurs. La méthode préconisée consiste à comparer la structure HTML des deux ou plusieurs pages appartenant à une même classe pour créer un schéma des données contenues dans ces pages. A partir de ce schéma, une grammaire est induite pour reconnaître des instances d'attributs identifiées pour ce schéma dans les pages Web.

IEPAD [CL01] consiste à découvrir automatiquement des règles d'extraction dans les pages Web. Le système peut automatiquement identifier la frontière entre les champs en se basant sur les motifs fréquents et sur l'alignement de séquence multiple. La découverte des motifs fréquents est réalisée grâce à l'arbre PAT (une structure de données). Les motifs fréquents sont extensibles grâce à l'alignement pouvant ainsi couvrir un nombre accru d'exemples.

2.4.4 Génération d'adaptateurs par des techniques de traitements automatiques des langues naturelles

Certains systèmes sont construits par l'utilisation des techniques de traitements automatiques des langues naturelles. Ces techniques sont appliquées dans les systèmes comme WHISK [Sod99], RAPIER [Cal98] et SRV [Fre98], pour apprendre des règles d'extraction des données existantes dans les textes. Ces règles sont basées sur des contraintes syntaxiques et sémantiques.

WHISK [Sod99] est un système d'apprentissage automatique qui produit des règles d'extraction pour une grande variété de documents non-structurés ou structurés. Les motifs d'extraction sont des expressions (spéciales) régulières ayant deux composantes : l'une décrit le contexte du motif, et l'autre indique les délimiteurs du motif à extraire.

RAPIER [Cal98] apprend les motifs d'extraction qui utilisent l'information syntaxique et l'information des classes sémantiques. Son modèle comprend un motif de pré-remplissage, un motif de post-remplissage (qui jouent le rôle des délimiteurs gauches et droits) et un motif de remplissage décrivant la structure d'information à extraire.

SRV [Fre98] est un outil pour apprendre des règles d'extraction à partir de données textuelles (textes). Il s'agit d'un procédé de traitement basé sur un ensemble d'attributs thématisés (token-oriented features). Un attribut peut être simple ou relationnel. Un attribut simple est une fonction assignant une valeur discrète à un terme. Un attribut relationnel assigne un terme à un autre terme. L'apprentissage des règles consiste à identifier et à engendrer des attributs trouvés dans les exemples. SRV est aussi capable d'extraire des données dans les pages HTML à l'aide des attributs spécifiques.

2.4.5 Génération d'adaptateurs à partir des motifs

Cette approche consiste à chercher dans les documents des portions de données qui peuvent être utilisées pour remplir les motifs pré-construits. Deux illustrations de cette approche sont les systèmes NoDoSE [Ade98] et DEByE [RNLdS99].

NoDoSE [Ade98] est un outil interactif pour déterminer semi-automatiquement des structures de documents pour extraire des données semi-structurées. L'utilisateur décompose de façon hiérarchique la structure du document via l'interface en choisissant des groupes de données et en les décrivant. A chaque niveau de la décomposition, l'utilisateur crée un objet avec une structure complexe, puis le décompose en d'autres objets avec une structure plus simple. NoDoSE apprend la façon dont l'utilisateur identifie des objets en induisant une grammaire des documents à partir des objets construits.

DEByE [RNLdS99] est un outil interactif qui reçoit un ensemble d'objets retenus à partir d'une page Web et génère des motifs d'extraction permettant d'extraire de nouveaux objets dans des pages similaires.

2.4.6 Génération d'adaptateurs à partir d'ontologie

Cette approche ne se base pas sur la structure des données dans un document pour générer des règles ou des motifs pour l'extraction d'information. Dans cette approche, l'extraction est faite directement sur les données. Pour un domaine spécifique, une ontologie est utilisée pour déterminer des morceaux de données dans un document, et objets sont construits à partir de ceux-ci. Un système connu dans cette approche est BYU [ECJ+99].

BYU [ECJ+99] est un outil qui a été développé par le Data Extraction Group de l'université de Brigham Young. BYU analyse une ontologie préconstruite manuellement par des experts pour produire automatiquement une base de données à partir des documents associés.

De plus de ces approches, Habegger cite les adaptateurs générés à partir de relations extraites ou à partir des bases de connaissances [Hab04].

2.4.7 Adaptateurs générés à partir de relations extraites

Dans cette approche, l'objectif est de construire un ensemble de motifs permettant d'extraire un sous-ensemble des instances d'une relation donnée. Les motifs sont alors applicables sur l'ensemble des pages du Web. Un exemple dans cette approche est le système DIPRE [Bri98].

DIPRE [Bri98] est basé sur l'hypothèse de l'existence d'une dualité entre les motifs et les relations : pour une relation donnée, il existe un ensemble de motifs permettant de retrouver une partie des occurrences de la relation. A partir des relations initiales, l'algorithme identifie automatiquement de nouveaux motifs. Les motifs sont alors utilisés pour extraire de nouvelles relations. Le processus est itéré à plusieurs reprises pour aboutir à un point fixe, s'il existe, pour lequel aucun nouvel exemple de la relation n'est généré ou jusqu'à ce que l'utilisateur soit satisfait du nombre d'exemples extraits.

2.4.8 Adaptateurs générés à partir des bases de connaissances

L'objectif de cette approche n'est pas d'obtenir un adapteur spécifique à une source donnée. Cette approche utilise néanmoins des connaissances du domaine en visant à construire un adaptateur générique pouvant s'appliquer

à des pages appartenant à un domaine donné. Dans cette approche, deux systèmes sont proposés : l'un par Gao [GS99] et l'autre par Seo [SYC01].

Gao [GS99] a présenté une méthode de représentation hybride pour les schémas des données semi-structurées, dans laquelle un schéma est représenté comme une hiérarchie de concept et un ensemble des unités de la connaissance. Un algorithme a été développé pour construire un adaptateur générique, qui utilise les schémas créés et exploite les structures de page.

XTROS [SYC01] représente les connaissances du domaine appliqué, et construit automatiquement un adapteur pour chaque source d'information. L'algorithme de génération d'adapteurs consiste à identifier des lignes logiques d'un document en utilisant les connaissances du domaine, puis de chercher le motif le plus fréquent dans la séquence des lignes logiques. Par la suite, l'adapteur est construit en se basant sur la position et la structure de ce motif.

En dehors de ces classifications par l'approche, l'automation est considéré comme un critère pour comparer les systèmes. Originellement, l'approche naturelle pour développer une procédure d'extraction d'information pour une source Web est de construire manuellement l'ensemble des motifs ou règles d'extraction. Dans cette approche manuelle, un système peut établir des règles d'extraction prédéfinies sans préciser comment elles sont obtenues, ou il met en place des outils d'aide pour assister l'utilisateur dans la création de règles. Cependant, la construction manuelle d'adaptateurs est une tâche fastidieuse surtout face au nombre de sources pour lesquelles une telle tâche est nécessaire. C'est pour cette raison que l'idée d'automatiser cette tâche est apparue. Ce fut Kushmerick qui a proposé une première méthode d'automatisation [Kus97].

Selon Laender [LRNdST02], trois groupes de systèmes se distinguent par leurs niveaux d'automatisation :

- des systèmes manuels : Minerva [CM98], TSIMMIS [HGMN+97], WebOQL [AM99] et BYU [ECJ+99].
- des systèmes automatiques : XWRAP [LPH00] et RoadRunner [CMM01].
- des systèmes semi-automatiques : W4F [SA99], WHISK [Sod99], RAPIER [Cal98], SRV [Fre98], WIEN [Kus97], SoftMealy [HD98], STALKER [MMK98], NoDoSE [Ade98] et DEByE [RNLdS99], [LRNdS02].

2.5 Extraction d'information multilingue

Depuis MUC-6, DARPA a rejoint MET (Multi-lingual Entity Task) pour la première tâche de reconnaissance des entités nommées multilingues. Cependant, l'extraction d'information multilingue reste encore une notion très proche de celle de l'extraction d'information inter-lingue [RSY02]. Parfois, elles sont mutuellement utilisées dans une approche fondamentale qui est « la projection inter-lingue ».

De nombreux projets de recherches en extraction d'information multilingue ont été présentés comme ECRAN [Poi99] et MIETTA [XNS00]. Les aspects communs de ces modèles s'appuient sur des outils de traduction automatique pour traiter les langues prévues et des modules indépendants de langues (multilingues) pour procéder à certaines tâches similaires dans ces langues.

Masche a proposé un modèle assez complet pour l'extraction d'information multilingue en se basant sur les idées suivantes [Mas04] :

1. Les documents, rédigés en différentes langues, sont reconnus par un module d'identification de langue naturelle et traduits en une langue préférée où il existe un système d'extraction d'information monolingue correspondant.
2. Les modules indépendants de la langue (c'est à dire les modules multilingues) sont construits pour réaliser des tâches communes et prédéfinies (la segmentation de mots, la reconnaissance des entités nommées), tandis que d'autres tâches sont restées monolingues (l'analyse morphologique, le traitement syntaxique, l'analyse de la coréférence).
3. Les motifs extraits sont traduits en plusieurs langues selon le besoin de l'utilisateur.

A travers ces modèles, nous constatons très clairement que l'idée essentielle de l'extraction d'information multilingue risque d'évoquer la réalisation répétitive de certaines tâches en plusieurs langues. Cette approche risque de souffrir de quelques points faibles :

- la négligence de la structure du corpus multilingue si elle existe comme dans le cas du site Web multilingue,
- la redondance de l'information répétée dans plusieurs langues,
- le manque d'information n'existant pas dans une ou quelques langues du corpus multilingue.

Ceci nous mène à considérer quelques nouveaux problèmes fondamentaux à résoudre pour améliorer la performance de systèmes d'extraction d'information multilingue :

- déterminer les informations complémentaires existant dans une ou plusieurs langues et qui n'existent pas dans d'autres langues du corpus,
- identifier la correspondance traduite entre les documents dans les différentes langues du corpus.

2.6 Conclusion

Le domaine d'extraction d'information a bien déterminé ses stratégies de développement, depuis le début des années 1990. De nombreux domaines y font appel et inversement[8].

La conséquence la plus significative du programme MUC, par rapport à notre projet, est de donner la priorité à l'extraction des motifs informationnels dans une démarche d'extraction d'information à partir de documents. Cette direction permet entre autre de séparer la démarche d'extraction d'information de celle de la compréhension de textes. De la sorte, nous favorisons les analyses structurelles aux traitements linguistiques de documents et surtout des hyperdocuments.

Actuellement nous distinguons trois tendances dans le développement du domaine de l'extraction d'information : la portabilité du système d'extraction d'information, l'extraction automatique du contenu et l'annotation automatique basée sur l'ontologie. Au coeur de ces courants, les adaptateurs représentent un phénomène technique passionnant incarnant un niveau de développement assez élevé dans l'extraction d'information ces dernières années.

Aussi, nous avons observé l'émergence d'un nouveau mouvement qui dominerait probablement cette discipline de recherche dans les années à venir : l'extraction d'information multilingue. En s'appuyant essentiellement

[8]Plusieurs auteurs considèrent la relation bilatérale entre l'extraction d'information avec la fouille de texte, la fouille du Web et la recherche d'information [KB00] : d'une côté, l'extraction d'information est une tâche dans la phase de pré-traitement pour d'autres disciplines mentionnées ci-dessus, et de l'autre côté, celles-ci sont une part du processus d'extraction d'information.

sur des outils de traduction automatique, dont les qualités demeurent problématiques, la définition et la validation des bases théoriques et méthodologiques de ce nouveau champs d'étude restent à l'ordre du jour, d'où l'intérêt des travaux de recherche menés dans ce domaine.

Troisième partie

Reconnaissance des langues dominantes dans un site Web multilingue

Chapitre 3

Représentation des hyperdocuments

3.1 Introduction

Dans ce chapitre nous analysons la structure du Web selon trois niveaux : la structure interne d'une page Web, la structure externe de la page Web et la structure macroscopique du Web. Dans cette approche d'analyse nous distinguons la structure hiérarchique de la structure hypertextuelle du web.

Nous présentons ensuite les principales propriétés du Web mises en évidences jusqu'à aujourd'hui. Ces propriétés sont abordées sous deux points de vue : macroscopique (le graphe Web peut-il être décomposé en grandes parties ?) et microscopique (le graphe Web contient-il des petites structures locales particulières ?). Les propriétés statistiques ainsi que les modèles les plus importants dans l'étude du graphe Web sont aussi discutés.

Enfin, nous nous intéressons à la structure « superficielle » du Web, c'est-à-dire un ensemble de pages statiques réalisées en HTML, pour introduire notre modèle de représentation des sites Web. Nous mettons également l'accent sur quelques notions intéressantes telles que *l'ancre source* et *le graphe d'ancres sources*.

3.2 Structure du Web

Le World Wide Web, conçu par Tim Berners-Lee en 1989, communément appelé le Web, est un système de type hypertexte (plutôt hypermédia) fonctionnant sur l'Internet et permettant de consulter, à l'aide d'un navigateur

(logiciel client), des pages Web mises en ligne dans des sites Web. Le Web représente des milliards des pages interconnectées, écrites le plus souvent en HTML (HyperText Markup Language). Les normes, basées sur HTML, permettent de décrire la structure du Web, avec d'une part la description de la structure hiérarchique des pages Web et d'autre part la description élaborée des hyperliens (la structure hypertextuelle du Web).

Comme nous l'avons évoqué ci avant, nous considérons que la structure du document Web ou bien du site Web peut être répartie en plusieurs niveaux [Gér02] :

1. La structure interne aux pages, qui est décrite par les balises HTML.
2. La structure externe aux pages, qui est décrite par le réseau d'hyperliens. Cette structure couvre à la fois :
   - une structure hiérarchique, c'est-à-dire la structure arborescente interne à un site. Les liens hypertextes peuvent être utilisés pour décrire la structure interne d'un document, auquel ses différentes parties sont fragmentées en plusieurs pages HTML. Les liens sont alors utilisés seulement pour faciliter la lecture et la maintenance.
   - une structure hypertextuelle, c'est-à-dire la structure de graphe interne à un site. Cette structure organise les documents (pages HTML) au sein d'un même site Web, permettant une consultation hypertexte des sites.
3. La structure macroscopique du Web, c'est-à-dire la structure de graphe externe aux sites. En effet, les lecteurs peuvent aussi naviguer de site Web en site Web en suivant des liens de référence, qui à première vue ne semblent pas décrire de structure particulière.

Géry constate qu'un site Web intègre des objets de type « document structuré » (structure arborescente, sens de lecture linéaire) et des objets de type « hypertexte » (structure de graphe, sens de lecture non-linéaire) [Gér02].

De la sorte, et dans le cadre de notre projet de thèse, nous admettons qu'un site Web représente un ensemble de documents structurés mais aussi une importante application d'hypertexte.

3.2.1 Structure hiérarchique du Web

Par définition, un document structuré est composé d'un ensemble d'éléments (ou objets) organisé dans une logique la plus souvent hiérarchique (la structure logique).

La notion de document structuré comprend, dans le cadre de cette thèse, trois composants principaux : le contenu, les structures et les stratégies de lecture [Gér02].

1. Le contenu d'un document structuré désigne les informations textuelles ou multimédia, représentées sous la forme d'un ensemble de composants (des figures, des images, des tableaux, des paragraphes, etc.).
2. Les normes de représentation de documents structurés, telles que l'ODA (Office Document Architecture) et le SGML (Standard Generalized Markup Language), distinguent deux types de structures : la structure physique et la structure logique, qui sont définies de la manière suivante :
   – La structure physique correspond à l'organisation d'affichage des données qui composent le document. Elle dépend de l'environnement de présentation du document, comme le format du papier ou l'écran d'un ordinateur de type : ordinateur individuel, ordinateur de poche, téléphone cellulaire, ...
   – La structure logique correspond à l'organisation hiérarchique des données du document. Elle propose, implicitement, une stratégie de lecture. Elle est la plus souvent indépendante de l'environnement de présentation (affichage).
3. La stratégie de lecture d'un document structuré consiste à enchaîner la lecture des parties successives, dans un sens connu implicitement, jusqu'à la conclusion ou la prise d'une décision d'arrêt de lecture.

Cette notion de *structure hiérarchique* est restée très présente dans la conception des pages HTML mais aussi des sites web (page d'accueil, annuaires, etc.). Nous faisons la distinction entre structure hiérarchique des pages et structure hiérarchique des sites en raison de la possibilité de décrire la structure logique au sein d'une page HTML et entre des pages HTML.

Structure hiérarchique intra-page

Les pages HTML (ou équivalent) possèdent une structure interne, appelée *structure hiérarchique intra-page*, qui permet de définir des éléments de différentes granularités.

Plusieurs approches ont été développées pour extraire ou identifier la structure hiérarchique intra-page d'un hyperdocument (site web) telle que l'utilisation de la structure logique, décrite à l'aide des balises HTML (ou tout autre type de langage de description structuré, comme SGML) :

- Fuller propose de fragmenter un document textuel, exprimé à l'aide de SGML, en un ensemble de noeuds et de relations de composition pour transformer cette structure en un hypertexte [FMSDW93].
- Riahi suggère l'usage d'une base de données orientée objets, basée sur des unités informationnelles, qui sont extraites et structurées en fonction des balises HTML [Ria98].
- Carchiolo modélise la structure logique interne des sites Web en combinant la structure décrite à l'aide des balises HTML et la similarité structurelle des parties de documents [CLM00].
- Géry analyse la structuration interne des pages HTML selon trois niveaux de granularité HTML : la phrase, le paragraphe et la section [Gér02].

D'autres approches font appel à des motifs pour l'intégration des données semi-structurées provenant de bases hétérogènes au sein d'un même modèle de documents [GY96], [jHtY97], [AMM97]. Nous nous sommes intéressés également aux travaux de Salton basés sur la recherche de similarité entre les parties de textes (données textuelles) pour détecter des hyperliens sémantiques à l'intérieur même d'un document [SAS96].

Structure hiérarchique intra-site

Dans la structure d'un site Web, il y a au moins deux types d'hyperliens : les référentiels et les organisationnels (structurels). Les hyperliens référentiels établissent des relations de cheminement entre les « documents sources » et les « documents destinations » en faisant des chemins de lecture. Par contre, les hyperliens organisationnels construisent la structure hiérarchique d'un site Web sous forme d'arbre : le document parent est relié par hyperlien organisationnel à un document enfant et vice-versa.

Grâce à des notations standardisées telle que le URL (Uniform Resource Locators) on peut établir des hyperliens, entre diverses ressources, décrivant une structure hiérarchique interne d'un site Web (appelée *structure hiérarchique intra-site*) où les différentes parties sont fragmentées en plusieurs documents HTML (au lieu de se localiser dans un même document). Or, ces normes ne permettent pas de prédire si le site Web représente un seul document structuré (lecture linéaire), ou si il représente un ensemble de documents organisés sous forme hypertextuelle (lecture par navigation) [Gér02].

Botafogo a montré qu'il est possible de différencier automatiquement les hyperliens hiérarchiques (organisationnels) des hyperliens de référence, en

extrayant une racine et la hiérarchie qui en découle [BRS92]. Il considère qu'une racine permet d'accéder à tous les noeuds sauf ceux qui sont isolés, qu'elle est à une distance faible des autres noeuds, et qu'elle possède un nombre considérable de fils. Les deux premières considérations sont vérifiables si le noeud possède un fils. La dernière considération permet d'éliminer les noeuds qui ont uniquement un rôle d'index (sans être réellement racine du site) [Gér02].

Aguiar insiste sur la difficulté de la tâche d'identification des hyperliens structurels dans un site Web, pour proposer deux hypothèses : 1) les hyperliens structurels existent mais sont mélangés avec d'autres types d'hyperliens, il faut envisager une méthode pour trier les hyperliens ; 2) les hyperliens structurels n'existent pas nécessairement, il faut les extraire [AB00]. Cet auteur, en optant pour la seconde hypothèse, propose une méthode basée sur l'analyse statistique de la distribution des termes dans les pages et entre les pages, ainsi que la distribution des hyperliens entre les pages pour extraire ces hyperliens structurels [Gér02].

La possibilité d'extraire une structure hiérarchique interne à un site Web a été renforcée par les travaux de Géry, qui proposent un algorithme utilisant des heuristiques simples sur la syntaxe des URLs, en accordant de l'importance à la structure hiérarchique des répertoires du serveur Web [Gér02].

3.2.2 Structure hypertextuelle du Web

Un site Web est un hypertexte du fait qu'il possède des noeuds (les pages HTML) qui sont connectés par des hyperliens (définis à l'aide d'URLs). Le Web est donc considéré comme un ensemble d'hypertextes dans lequel chaque site Web est un hypertexte distinct, qui est structuré indépendamment des autres et qui présente une organisation autonome de ses informations.

La notion d'hypertexte[1] demeure un modèle de représentation de l'information dans lequel cette dernière est organisée sous la forme des unités indépendantes, autonomes et interconnectées, appelés *noeuds*. Chaque noeud correspond à une page Web et peut être en principe relié à une multitude

[1]Le principe des hypertextes a été inventé par Bush [Bus45]. L'idée principale d'un système hypertexte est donc de donner la possibilité à l'utilisateur de gérer (consulter et modifier) un document ou un ensemble de documents de manière non linéaire, en organisant les informations de manière associative. Ultérieurement, Nelson a popularisé le concept et forgé le terme « hypertexte » en imaginant un réseau de machines coopérantes donnant accès à ensemble de connaissances réparties [Nel72].

d'autres noeuds par des hyperliens. Le noeud est donc l'unité minimale d'information dans un hypertexte. Le support d'un noeud d'information peut être une page si l'information est textuelle. Quand l'information n'est pas uniquement textuelle, le support d'un noeud peut être un graphique, une animation, une image, une séquence de vidéo ou de son, etc. Les hyperliens peuvent exister aussi entre des hypertextes, autrement dit, des noeuds externes aux sites.

Structure hypertexte intra-site

La structure hypertexte interne d'un site Web, appelée *structure hypertexte intra-site*, organise les documents (pages HTML) au sein d'un même site Web. Cette structure offre la possibilité de parcourir un site Web en choisissant au fur et à mesure les chemins de lecture, contrairement aux documents structurés qui comportent un chemin de lecture imposé [Gér02].

Bray a analysé une collection de 11 millions de pages HTML et a montré que, si la densité des hyperliens est importante (en moyenne une page comporte 14 hyperliens sortants, et seulement 25% des pages présentes sur le web sont des « feuilles »), les pages forment des « grappes », qu'il formalise par le concept de site Web. Un site est alors un groupe de pages très reliées entre elles (elles se connectent fortement), mais peu reliées au reste du Web [Bra96] (elles connectent très peu aux pages à l'extérieur du site Web). En effet, quatre pages sur cinq pointent uniquement sur des pages appartenant à un même site. De plus, ces sites sont souvent isolés : 80% d'entre eux sont référencés par moins d'une dizaine d'autres sites, et 80% d'entre eux n'en référencent aucun.

La structure hypertexte intra-site exige de déterminer le rôle plutôt que la position d'une page dans une structure hiérarchique. Ainsi, Pirolli propose une classification des pages Web d'un site selon leur rôle dans l'hypertexte [PPR96], en montrant qu'il est possible de déterminer le type d'une page par une combinaison entre l'analyse de la topologie du réseau d'hyperliens, la similarité entre les documents, les statistiques d'utilisation du site (nombre d'accès, navigation, etc.), ainsi que divers autres critères statistiques : titre, auteur, taille de la page, etc. Chaque page est représentée par l'ensemble des caractéristiques qui correspondent à ces éléments, et qui sont stockées dans un vecteur. Les vecteurs sont ensuite comparés à une liste de vecteurs prédéfinis représentant les caractéristiques des différents types de la classification.

Spertus considère d'une manière générale, qu'il existe une structure du Web, en particulier dans les sites Web, et que cette structure pourrait être extraite à partir des URLs [Spe97]. Spertus se base sur une classification semblable des pages et établit un certain nombre de règles permettant d'obtenir des informations sur les pages d'un site. Ces règles se basent sur une information contenue dans les hyperliens, qui permet une classification de ceux-ci par une analyse syntaxique de l'URL.

Structure macroscopique du Web

En prenant en compte uniquement les hyperliens sortant des sites Web, c'est à dire en considérant les pages dans le contexte global du Web et non plus localement à un site, la structure macroscopique est celle qui organise les sites Web entre eux (les liens entre eux).

La plupart des méthodes d'extraction de structure au niveau macroscopique du Web s'intéressent à des groupes de pages plutôt qu'à des pages prises individuellement, comme par exemple des grappes de sites, qui ont parfois une structure typique, comme par exemple les *anneaux* du Web [Bra96], [CK97].

Géry distingue deux types d'approche pour extraire une structure macroscopique du Web [Gér02] :

1. Traitement d'une page ou d'un site par rapport au Web global : L'origine de cette approche a commencé dans l'analyse de citations ou de co-citations dans la littérature scientifique : la bibliométrie adaptée au Web [Kes63], [Sma74], [WM89]. Il existe plusieurs méthodes qui cherchent à extraire les pages Web jouant un rôle particulier dans le réseau d'hyperliens, en se basant sur un « score » pour extraire des pages qui font autorité (référencées par beaucoup de pages) ou des pages rayonnantes (qui référencent beaucoup de pages) [Bri98]. Ce score est éventuellement amélioré en intégrant une notion de qualité [The01] ou de réputation [RM00]. Ces notions demeurent toutefois subjectives en ne se basant que sur le réseau d'hyperliens pour les évaluer. Enfin, on peut aussi se baser sur des scores combinant autorité et rayonnement [Kle99a], [Kle99b].
2. Traitement d'un groupe de pages ou d'un groupe de sites : La structure macroscopique du Web est extraite en analysant la connectivité du réseau d'hyperliens inter-sites. Selon Kleinberg, ce sont des structures

de communauté qui identifient une communauté d'intérêts [GKR98], [KKR+99].

Les premiers résultats d'une analyse de la topologie du Web à grande échelle ont montré une connectivité forte du réseau d'hyperliens. Selon une étude effectuée par Albert portant sur 325.000 pages et 1,5 millions d'hyperliens, la moyenne de la plus courte distance entre deux noeuds de la collection - vue comme un graphe orienté - serait de $d = 0,35 + 2,06 * log(N)$, avec N le nombre de noeuds [AJB99]. Albert extrapole cette estimation au Web entier, dont la taille était évaluée à l'époque à 800 millions de documents, pour estimer le diamètre du Web à 18,59 hyperliens [Gér02].

3.3 Graphe Web

La structure du Web est représentée souvent (« naturellement ») à l'aide d'un graphe dont les noeuds sont des pages Web et les arcs (orientés) sont des hyperliens [BK00], [KRR+00b], [KRR+00a].

Le Web est un exemple de réseau social dont la théorie concerne les propriétés de la connexité et de la distance dans le graphe [WF94]. Depuis 1996, on a adopté les principes du réseau social pour analyser des graphes du Web comme par exemple la mesure de la popularité du PageRank [Bri98] et HITS [Kle99a]. L'étude du graphe Web joue ainsi un rôle central pour la compréhension des phénomènes sous-jacents et pour de nombreuses applications comme la robustesse du réseau [AJB99], [MdMLD02], l'optimisation des moteurs de recherche [ERC+00], [GKR98].

3.3.1 Visions du graphe Web

L'étude du graphe Web passe tout d'abord par des méthodes d'exploration du Web fournissant différentes vues et « structures sous-jacentes ». L'exploration du Web rencontre néanmoins des difficultés qui amènent à obtenir des vues partielles et biaisées du Web [GL03b]. En l'absence d'informations sur le biais introduit par la méthode d'exploration, plusieurs façons de décrire la structure du graphe Web ont été proposées : approches issues d'une vision macroscopique ou d'une vision microscopique [LCD+05]. Ces deux catégories d'approche ont été appliquées pour optimiser des moteurs de recherche et pour faire émerger des communautés d'intérêts sur le Web [GL02], [Kle99a], [Kle99b], [KKR+99].

Vision macroscopique

Une vision macroscopique permet de décrire le graphe Web en le décomposant en plusieurs sous ensembles (regroupements).

Broder et Kumar ont effectuée une importante expérimentation sur une collection des données fournies par Altavista (203 millions de pages HTML et 1,5 milliards d'hyperliens) incluant des liens permettant d'accéder à des sites isolés du reste du Web dont les URLs ont été fournies à Altavista directement par les auteurs de sites [BK00], [KRR+00b]. Le diamètre de la collection choisie est de 28 hyperliens, mais cette valeur extrapolée à l'ensemble du Web serait de plus de 500 hyperliens selon Broder. De plus, la probabilité qu'il existe un chemin entre deux pages du réseau prises au hasard est seulement de 25%. Si ce chemin existe, alors sa longueur moyenne est de 16 hyperliens. Les résultats de cette expérimentation ont été de grande utilité : elle a permis de mettre en avant la macrostructure dite du noeud papillon ; elle a montré que la connectivité du Web est beaucoup moins forte que ce que l'on pensait [Gér02]. En conclusion de cette étude, Broder a discerné quatre grands ensembles de pages HTML (cf. figure 3.1 [GL03b]) :

1. Noyau fortement connexe (« SCC »- Strongly Connected Component) : il existe une zone du Web composée de pages fortement liées entre elle. Dans cette zone, on peut naviguer de n'importe quelle page à n'importe quelle autre en suivant des hyperliens. Cette zone comporte 56 millions de pages.
2. Origine (« IN ») : (44 millions de noeuds) cette zone est composée de pages qui permettent d'accéder aux pages de la zone SCC, mais qui ne sont pas accessibles depuis les pages de la zone SCC.
3. Extrémité (« OUT ») : de taille équivalente à IN (44 millions de noeuds) les pages de OUT ne renvoient pas vers les pages SCC.

L'étude du web a permis également de révéler trois autres structures :
- des composants isolés : certaines zones du Web sont isolées des zones principales « SCC », « IN » et « OUT ». Aucun hyperlien n'y mène, et aucun hyperlien ne relie ces pages aux zones principales.
- les tubes : certaines zones du Web, de tailles plus réduites, relient les pages de la zone « IN » directement aux pages de la zone « OUT », sans passer par la zone « SCC ».
- les branches : il s'agit de zones atypiques qui relient des sites isolés de l'ensemble, soit à la zone « OUT », soit à la zone « IN ». Ces zones parfois éloignées s'étendent comme des « vrilles de vigne » et contiennent

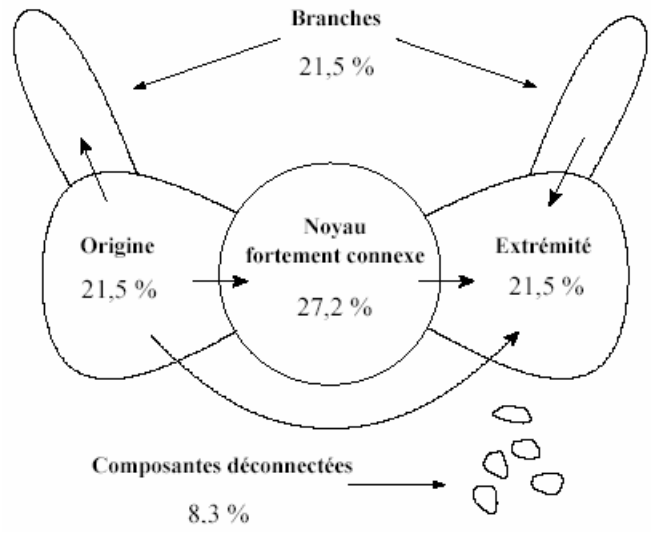


FIG. 3.1 – Structure macroscopique du graphe Web

des pages à partir desquelles on ne peut pas atteindre le noyau et qui ne sont pas accessibles à partir de ce noyau.

Vision microscopique

L'étude microscopique du graphe Web se concentre sur les particularités locales. En ce sens, il faut dans un premier temps, et partant de la structure locale du graphe Web, définir une communauté [GL03b], [KL01]. Par la suite il faut proposer des méthodes pour détecter automatiquement des communautés.

Une communauté est décrite comme un couple d'ensembles de pages Web tels que toutes les pages du premier ensemble pointent vers toutes les pages du second. De plus, les pages du second ensemble ne pointent pas les unes vers les autres. Cette structure dense correspond à une communauté centrée autour d'un sujet de prédilection [Kle99a] : le premier ensemble contient les pages de « fans » qui mettent des hyperliens vers leurs « stars » (cf. figure 3.2 [GL03b]). Cette définition est souvent assimilée à la théorie des pages qui font autorités [KL01].

D'autres définitions interprètent la communauté comme une collection de pages Web qui possèdent plus d'hyperliens entre les pages de la collection qu'avec les pages externes [FLGC02] (cf. figure 3.3 [GL03b]).

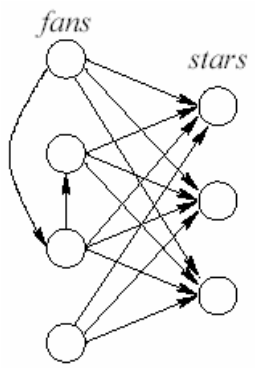


FIG. 3.2 – Structure microscopique du graphe Web I

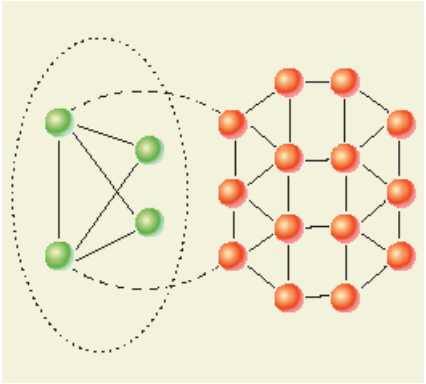


FIG. 3.3 – Structure microscopique du graphe Web II

Cette définition sollicite des techniques dédiées aux problèmes de partitionnement de graphes. D'ailleurs, à l'issue d'une analyse textuelle des hyperliens de ce type de communauté, nous remarquons que ces pages sont concentrées de la même façon que celles de la communauté définie en premier.

D'autres types de structures locales existent, comme par exemple les clans qui correspondent aux ensembles de pages Web pour lesquelles il suffit d'un très petit nombre de clics pour aller d'une page à l'autre [GL03b].

3.3.2 Propriétés statistiques du graphe Web

Les premières expérimentations de Barabasi et de Kumar, réalisées à l'aide d'un échantillon d'environ 40 millions de pages Web, ont montré que la distribution des degrés dans un graphe Web est en loi de puissance [BA99], [KRR+00b]. D'une part, ceci signifie qu'il y a peu de sommet de très hauts degrés, tandis qu'il y en a beaucoup avec de petits degrés. D'autre part, on constate que la plupart des pages sont relativement peu référencées alors que certaines le sont énormément. L'étude de Broder sur un échantillon de 200 millions de pages Web a confirmé ces conclusions [BK00].

D'autres expérimentations de Broder sur deux échantillons de 203 millions et de 271 millions de pages Web, ont montré également que les distributions des degrés en loi de puissance sont très similaires [BK00].

Depuis quelques années, de nombreuses propriétés et mesures statistiques du graphe Web ont été proposées, que nous pouvons en citer en particulier [GL03a] :

- la distance moyenne, c'est-à-dire la moyenne des distances de tous les couples de sommets (en ne considérant que les plus courts chemins entre ces sommets),
- le coefficient de « clustering », c'est-à-dire la probabilité pour que les voisins d'un sommet soient voisins entre eux,
- la distribution des degrés, c'est-à-dire, pour chaque valeur, le nombre de sommets ayant ce nombre d'hyperliens.

La combinaison de la distance moyenne et du coefficient de « clustering » est souvent appelée la propriété « small-word ». Les mesures ont permis de constater que, pour le graphe Web [GL03b] :

- la distance moyenne est courte,
- le coefficient de « clustering » est élevé,

– les distributions des degrés suivent des lois de puissance,

Ces propriétés statistiques ne sont pas spécifiques au graphe Web mais elles relèvent également de nombreux autres contextes : le graphe des acteurs, le graphe des appels téléphoniques, le graphe de dépendance des espèces, le graphe des connexions des neurones dans un cerveau, etc. Tous ces graphes, et bien d'autres encore [AJB99], ont aussi une distance moyenne courte, un fort coefficient de « clustering » et une distribution des degrés en loi de puissance [GL03a].

3.3.3 Modèles réalistes du graphe Web

Les propriétés du graphe Web sont prouvées par diverses expérimentations ayant pour objectifs de comprendre et de reproduire des phénomènes émergeants sur le Web. A ce propos différents modèles ont été proposés pour différents objectifs comme par exemple, la construction des topologies réalistes pour la simulation, la compréhension des phénomènes sous-jacents, etc. [GL04a], [GL04a] [Str01].

Le hasard Le modèle le plus simple était fondé sur les graphes aléatoires définis comme suivant : étant donné deux entiers n et m, le graphe $G_{n,m}$ est un graphe à n sommets obtenu en tirant aléatoirement m paires de sommets qui formeront les arêtes [ER59]. Ce modèle est très limité dans la mesure où il produisait des graphes qui ne sont pas adaptés à la « réalité », à l'exception au calcul de la distance moyenne qui est souvent faible cependant conforme aux observations faites sur les graphes Web.

Clustering Watts propose un modèle de regroupement « clustering » qui est défini comme suivant : considérant un anneau de sommets, chacun étant relié à ses k plus proches voisins où k est un entier donné [WS98]. Par la suite, chaque arête sera liée avec une probabilité p donnée, c'est-à-dire qu'une extrémité de chaque arête est remplacée avec la probabilité p, par une nouvelle extrémité choisie aléatoirement. Ce modèle possède un fort coefficient de « clustering » ainsi qu'une distance moyenne faible. Cependant, il ne rend pas compte du phénomène de la distribution des degrés en loi de puissance et ne restitue donc pas parfaitement les propriétés des graphes Web rencontrés en pratique.

Attachement préférentiel Albert et Dorogovtsev introduisent un modèle qui est capable de reproduire les propriétés du graphe Web à l'aide d'un

processus de construction appelé attachement préférentiel : les sommets sont rajoutés un à un et reliés aléatoirement aux sommets préexistants [AJB99], [DMS00]. Toutefois, les sommets auxquels le nouveau sommet est relié sont choisis avec une probabilité qui croît en fonction de leur degré. Nous constatons que, d'une part, cette procédure permet d'obtenir des graphes dont la distribution des degrés suit une loi de puissance, et que la distance moyenne entre les sommets est faible. D'autre part, cette procédure ne respecte pas le coefficient de « clustering » des graphes Web « réels ». Ce modèle, aussi bien que les précédents ne correspond pas complètement aux propriétés des graphes Web rencontrés en pratique.

Structure bipartie Nous pouvons parler de deux autres modèles fondés sur la structure bipartie d'un graphe, proposée par Guillaume et composée de trois propriétés principales du graphe Web « réel » : un graphe aléatoire bipartie avec la distribution des degrés prédéfinis et un graphe dynamique bipartie associé à la procédure de l'attachement préférentiel [GL04b]. La présence de trois propriétés est vue comme une conséquence de la structure bipartie. L'objectif consiste à mettre en place une procédure de décomposition du graphe en une structure bipartie et ensuite à l'utiliser pour produire les trois principales propriétés du graphe Web réel.

3.4 Représentation des hyperdocuments

Nous présentons dans cette section les principales approches de modélisation de la structure d'un document ou d'un hyperdocument.

Historiquement, la représentation des documents est influencée par les méthodes d'indexation, de classification/catégorisation et de recherche/restitution d'information qui adopte l'approche basée sur les mots (l'ensemble de mots ou le sacs de mots). Cette approche s'est très tôt montré peu convenable (perte d'information) au type de documents structurés, définis à l'aide des langages de balisage. Ce fait a amené, dans un premier temps, de nombreux travaux à proposer une modélisation des documents structurés avec un SGBD (Système de Gestion de Base de Données) relationnel ou objets permettant de représenter la structure des documents et de les interroger à l'aide de requêtes SQL/OQL.

A titre d'exemple, il est très utile dans ce contexte de citer le langage WebSQL, qui stocke les documents et leurs attributs externes dans une table

Document et le réseau de liens dans une table Anchor [MMM96]. Nous pouvons citer également le langage POQL permettant de stocker des documents SGML dans une base de données orientée objets [CR94]. Dans une représentation fortement typée telle que POQL, la correspondance stricte est établie entre la DTD et le schéma de la base à chaque type de noeud SGML correspond une classe d'objets. Cependant, l'utilisation des relations (au sens de base de données) pour décrire des documents structurés n'est pas en fait très souple. La difficulté principale réside dans le passage d'une structure hiérarchique à un ensemble de relations la représentant. Les documents se conforment à une structure rigide encapsulée dans le schéma de la base de données.

La prise en compte des hyperliens entre les hyperdocuments est inspirée des travaux sur les réseaux sociaux [WF94], [KSS97]. Les premières expérimentations de cette approche ont été effectuées dans le domaine de recherche d'information sur l'Internet [BH98]. Aussi Chakrabarti et Fürnkranz ont proposé d'établir des modèles pour classifier des pages Web en exploitant des textes associés aux hyperliens (dans les ancres sources ou le voisinage des hyperliens) pour améliorer la performance des méthodes de classification [CDK+98], [Für99].

La structure interne des pages HTML est utilisée dans certains moteurs de recherche sur le Web pour affiner l'indexation. Typiquement, Boyan propose de considérer plus attentivement les termes présents, par exemple, dans le titre, les en-têtes, les méta-données, les mots écrits en italique ou en gras, les ancres, etc. [BFJ96].

Un hyperdocument comporte une multitude d'hyperliens entre les documents. Chaque hyperlien matérialise une *relation* entre deux documents. D'après Géry, nous distinguons trois types de relations dans un site Web [Gér02] :

- les relations de composition décrivant l'organisation structurelle des documents (la structure hiérarchique ou logique).
- les relations de cheminement décrivant une structure délinéarisée et définissant une lecture non linéaire[2],
- les relations de référence externes au site.

Pour extraire des entités, des composants et des relations dans la structure du site Web, nous analysons cette dernière selon 4 niveaux de granularité :

[2]Une suite de relations de cheminement est appelée *chemin de lecture*

– le site,
– les documents,
– les sections,
– les paragraphes.

La représentation d'hyperdocument exige de modéliser à la fois les structures hiérarchiques internes aux documents et la structure hypertexte intra-site.

3.4.1 Documents structurés

La représentation des documents structurés que nous présentons dans cette section consiste à modéliser la structure hiérarchique des composants, qui sont des objets définis dans le contexte d'une page Web à l'aide des balises HTML. En plus, notre modèle de représentation permet aussi de préciser un phénomène assez universel dans lequel des composants sont partagés (réutilisés) par plusieurs documents.

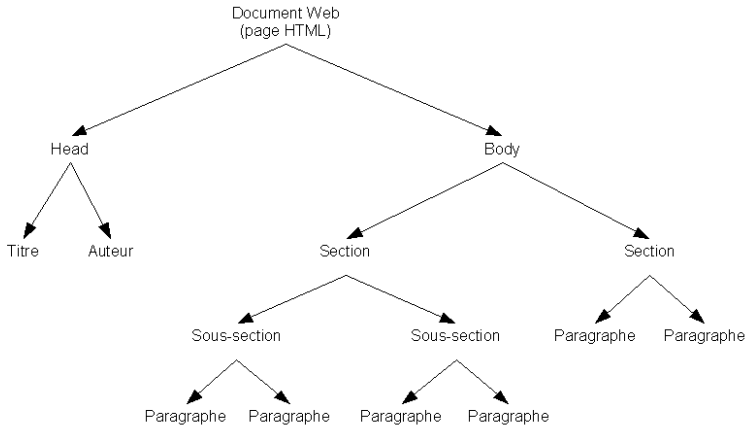


FIG. 3.4 – Représentation d'un document structuré

Au sein des paragraphes se trouvent des composants élémentaires comme des images, des sons, des vidéos, etc. Les composants élémentaires sont considérés comme des objets nominatifs et insécables (ne pouvant pas contenir aucun autre composant).

Cette méthode de représentation nous permet de décrire les composants communs qui peuvent être partagés par plusieurs documents (cf. figure 3.5).

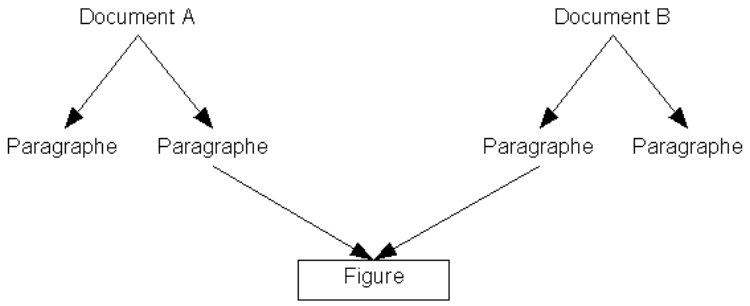


FIG. 3.5 – Exemple d'un composant partagé par deux documents Web

3.4.2 Graphe d'ancres sources

Les documents Web (pages HTML) sont connectés par des hyperliens qui sont ancrés dans la page source. Les ancres « source » se trouvent sous la forme d'un titre textuelle (un mot ou une phrase) ou d'une image cliquable (un icône).

Nous constatons que dans un graphe de documents Web [BK00], [KRR+00b], [KRR+00a], les noeuds et les arcs ne sont pas étiquetés. Cette notion du graphe Web ne peut pas exprimer le sens d'utilisation des arcs (les relations matérialisées par des hyperliens) connectant les noeuds (les documents), c'est pourquoi nous ne pouvons pas définir avec précision la similarité entre les sous graphes générés par des hyperliens utilisant les mêmes ancres sources. En plus, pour représenter des hyperdocuments multilingues, ce graphe ne permet pas de montrer les frontières entre les groupes de documents selon leurs langues. La méthode d'exploration du graphe Web ne peut pas trouver de « repères » qui marquent le changement de langues dans le parcours des chemins de lecture.

Pour résoudre ces deux problèmes, nous introduisons la définition des « relations entre ancres sources » :

Définition 1 ***Relation entre ancres sources*** *: Un hyperlien, connectant un document source et un document destination, définit des relations entre l'ancre source de cet hyperlien localisé dans le document source vers toutes les ancres sources localisées dans le document destination.*

En nous basant sur les relations entre ancres sources, nous introduisons la définition du « graphe d'ancres sources » comme suit :

Définition 2 ***Graphe d'ancres sources*** *: Un graphe d'ancres sources est un couple* (V, E)*, où* V *est un ensemble de noeuds qui sont les ancres sources, et* $E \subseteq V \times V$ *est un ensemble d'arcs qui sont les relations entre ancres sources.*

Dans un graphe d'ancres sources, nous considérons que les hyperliens, qui sont utilisés de façon similaire pour diriger vers un document, partagent les mêmes ancres sources (cf. figure 3.6). Cependant, il existe des hyperliens qui partagent une ancre source pour connecter à différents documents (cf. figure 3.7).

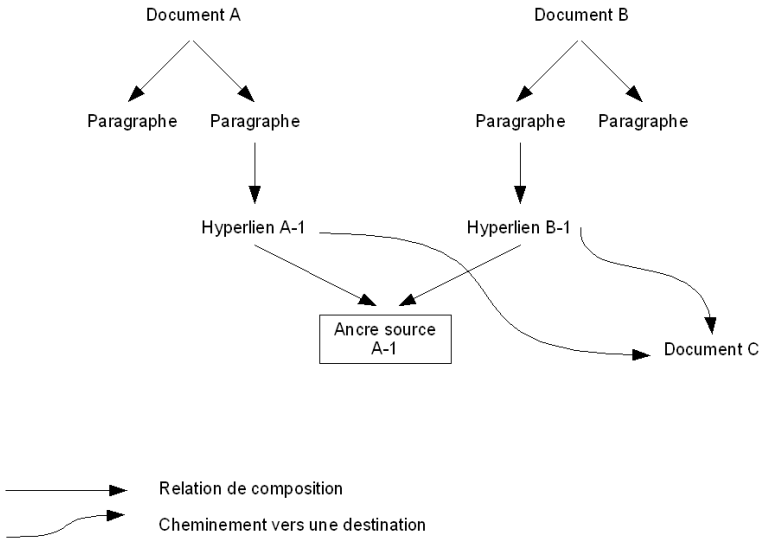


FIG. 3.6 – Exemple de deux hyperliens partageant une ancre source pour se diriger vers un document

Dans un tel graphe, il existe des ancres sources très spécifiques, qui deviennent des « repères » de frontière entre les différentes langues pouvant exister ou co-exister sur un site Web. Ces ancres sources spécifiques sont des ancres de changement de langue dont chacune est utilisée par presque la totalité des documents d'une langue présente sur un site Web multilingue pour permettre de passer d'une langue vers d'autres langues. C'est pourquoi, statistiquement ces ancres de changement de langue ont leurs degrés (entrant et sortant) très élevés dans un graphe d'ancres sources par rapport à tous les autres types d'ancres sources normales.

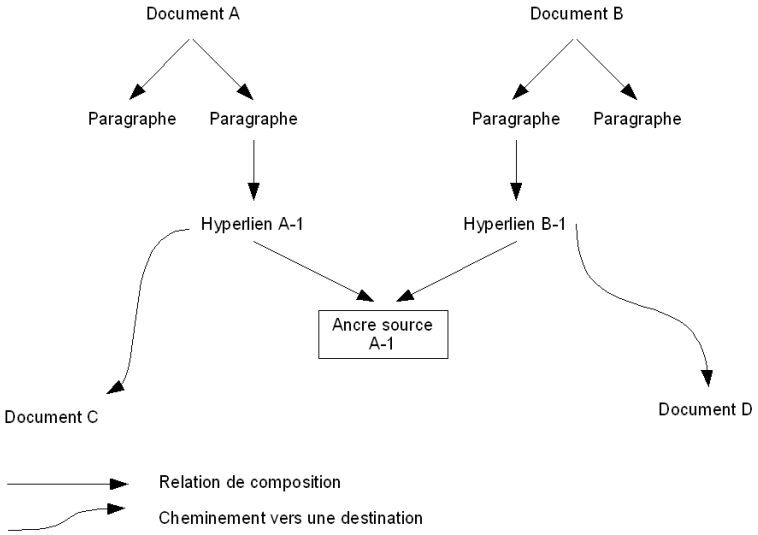


FIG. 3.7 – Exemple de deux hyperliens partageant une ancre source pour se diriger vers différents documents

En fixant un seuil statistique pour le degré des ancres sources (des noeuds dans un graphe d'ancres sources), nous pouvons éliminer les ancres sources spécifiques. Ceci divise le graphe d'ancres sources en plusieurs régions de langue que l'exploration peut distinguer grâce aux parcours des chemins discontinus entre les frontières des langues.

Principe de transformation d'un hyperlien en relations entre ancres sources

Un hyperlien connectant deux documents est transformé sous la forme de relations entre les ancres sources localisées dans ces deux documents (cf. figure 3.9).

Le processus de transformation d'un hyperlien en relations entre ancres sources consiste à :

- extraire l'ancre source A de l'hyperlien dans le document source,
- déterminer tous les ancres sources B_i des hyperliens dans le document destination,
- créer des relations entre l'ancre source A avec les ancres sources B_i telles que : $A \rightarrow B_i$.

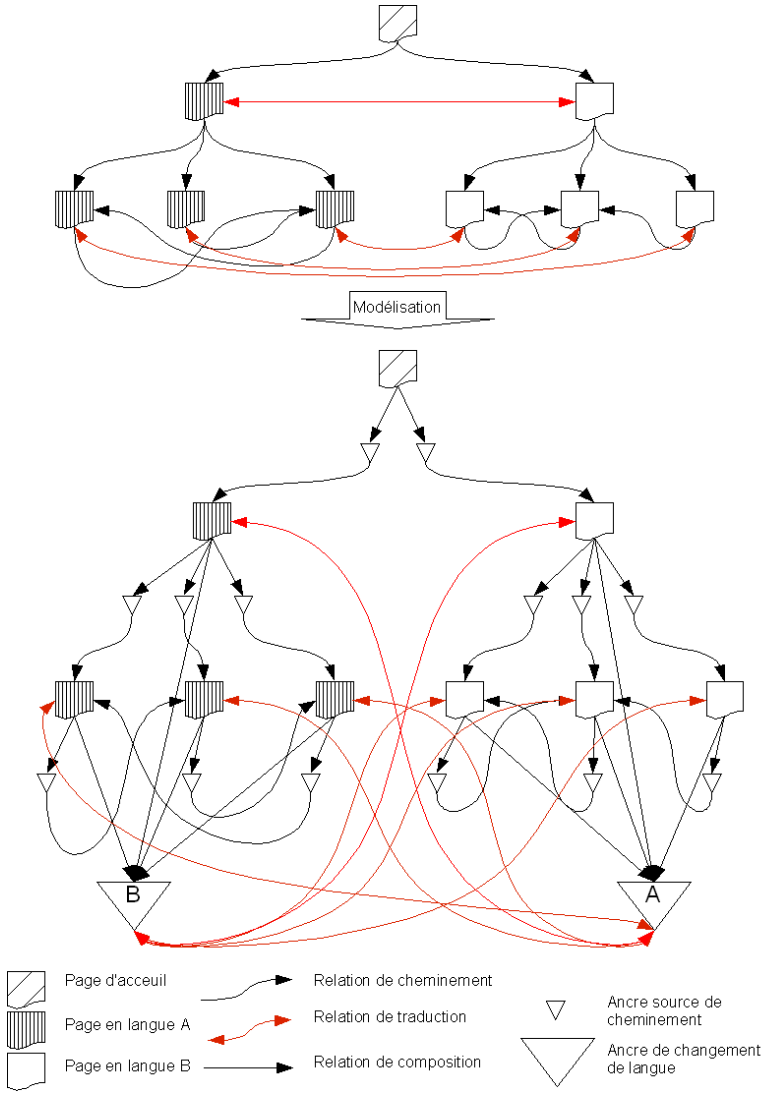


FIG. 3.8 – Ancres de changement de langue dans un hyperdocument bilingue

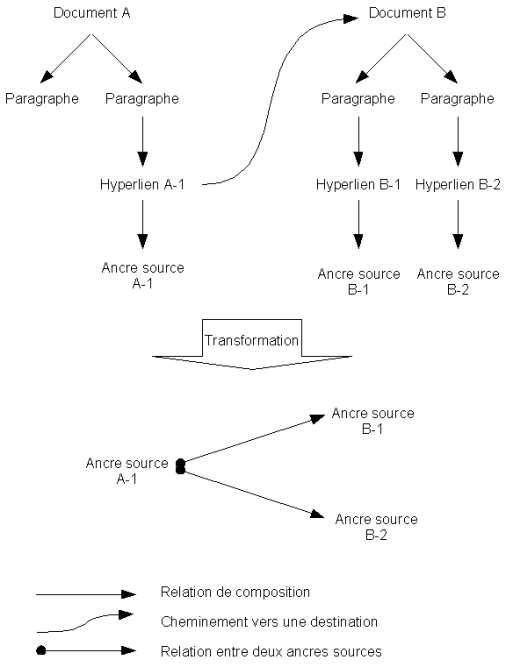


FIG. 3.9 – Transformation d'un hyperlien en relations entre ancres sources

Processus de création d'un graphe d'ancres sources

Le processus de création d'un graphe d'ancres sources consiste à faire les opérations suivantes :

- déterminer tous les hyperliens,
- transformer les hyperliens en relations d'ancres sources.

3.4.3 Modèle d'hyperdocuments

Après ces analyses et revues des travaux dans ce domaine, nous proposons un modèle d'hyperdocument qui consiste à représenter :

- les structures hiérarchiques internes aux documents en représentant les relations de composition,
- la structure hypertexte du site Web en représentant les relations de cheminement,
- la structure de relations entre ancres sources.

Notre modèle d'hyperdocument, que nous appelons le GDS (Graphe de Documents Structurés) se compose donc en deux niveau de représentation (cf. figure 3.10).

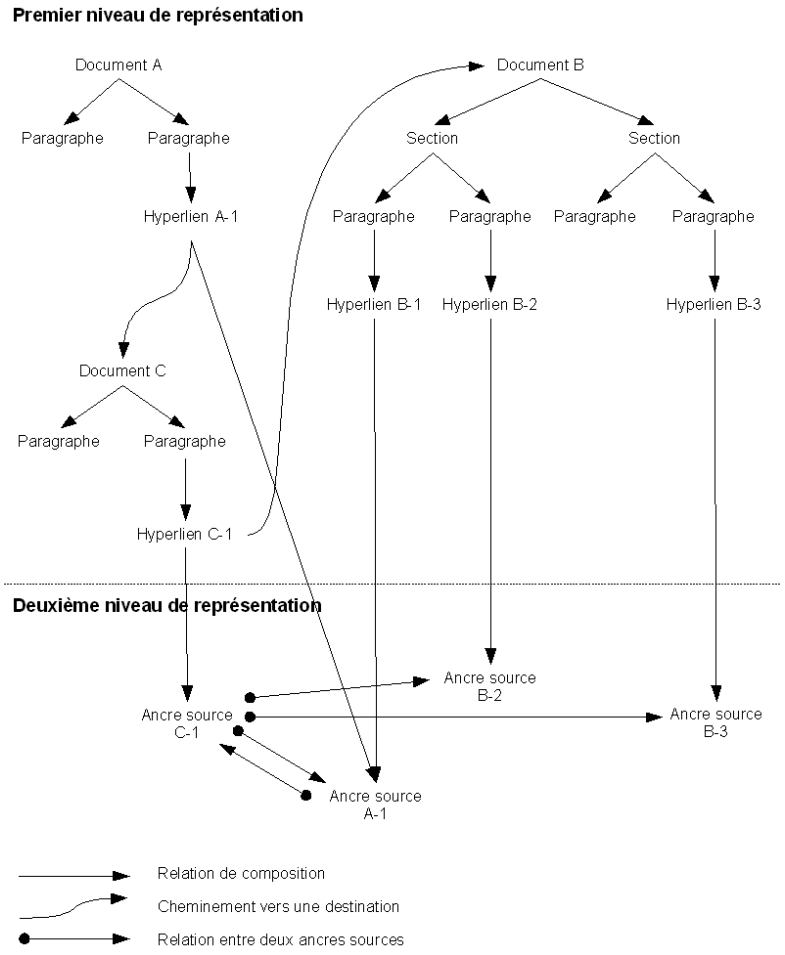


FIG. 3.10 – Représentation d'un hyperdocument

Le premier niveau du modèle représente la structure hypertexte du site Web ainsi que les structures hiérarchiques des documents. Ce premier niveau de représentation est intégré dans un graphe d'ancres sources qui représente le deuxième niveau de notre modèle.

3.5 Conclusion

De nombreuses approches ont été proposées pour l'étude de la structure et de la modélisation du Web. La dualité des documents structurés/hypertextes implique non seulement l'existence d'une structure du Web, mais l'existence de plusieurs structures : structure hiérarchique, structure hypertexte et structure macroscopique [Gér02].

Nous avons proposé un modèle de représentation structurelle des hyperdocuments (sites Web), appelé GDS. Ce modèle comporte deux niveaux de représentation : le premier niveau représente la structure hypertextuelle de l'hyperdocument ainsi que les structures hiérarchiques internes aux documents (pages Web), et le deuxième niveau est un graphe d'ancres sources qui tente de préciser les relations entre les documents.

Notre modèle d'hyperdocument GDS a été adopté dans notre système Hyperling qui sera présenté dans les trois chapitres 4, 5, 6 suivants.

Chapitre 4

Système Hyperling : modélisation et fonctionnement

4.1 Introduction

Dans ce chapitre nous introduisons les objectifs et les caractéristiques du système Hyperling. La finalité d'Hyperling est l'établissement d'une méthode d'exploitation de l'information structurelle (tel que l'hyperdocument) pour reconnaître des langues dominantes dans un site Web multilingue.

Dans un premier temps, nous évoquons la problématique de la conception du système Hyperling. Ensuite nous abordons le processus de reconnaissance des langues dominantes dans un site Web multilingue avant de présenter l'approche de modélisation retenue. Enfin, nous décrivons le fonctionnement du système.

4.2 Objectifs et caractéristiques

Dans notre projet nous observons que la structure des documents et des sites Web incorpore des informations qui sont indispensables pour toute démarche d'optimisation de la recherche d'information, de l'extraction d'information ou de la fouille des sites Web. La crédibilité de cette hypothèse peut être renforcée par le développement accéléré des documents structurés ainsi que des hyperdocuments.

Pour illustrer notre propos nous avons développé un système, intitulé Hyperling, capable de déterminer, sans aucune connaissance linguistique préalable et explicite, des langues dominantes sur un site Web multilingues. Hy-

perling est développé dans le cadre d'un projet de recherche global sur la recherche et l'extraction d'information à partir de documents électroniques dynamiques monolingues ou multilingues.

Les caractéristiques d'Hyperling sont définies selon les trois critères :

- adopter des traitements indépendants des langues (de point de vue d'ingénierie linguistique),
- concevoir des méthodes distributionnelles statistiques basées sur l'apprentissage automatique,
- développer des approches d'analyses structurelles des documents et des sites Web multilingues.

En respectant ces critères, la réalisation du système Hyperling n'a aucune nécessité de l'usage des techniques de traitement automatique des langues naturelles. Au travers de ce projet de recherche, nous avons voulu exploiter au maximum des méthodes distributionnelles statistiques, d'apprentissage automatique et d'analyses structurelles pour évaluer la qualité d'information « implicite » pouvant être découvert à partir de la structure des sites Web.

Hyperling est un système d'extraction d'information adoptant une approche d'analyse structurelle. Il ne fait pas partie des systèmes classiques d'extraction de motifs à partir de données fortement structurées. Hyperling est un système de fouille structurelle du Web (Web Structure Mining)[1] ayant pour but de découvrir de nouvelles informations à partir de la structure du Web. L'extraction d'information dans ce contexte est définie comme un processus de repérage, formalisation et de traitements des structures de données pouvant comporter de l'information pertinente. En ce sens, on peut considérer que la fouille structurelle du Web comme faisant partie des processus d'extraction d'information [KB00].

[1]Etzioni a proposé la première fois la Fouille du Web (Web Mining) en la définissant comme l'utilisation des techniques de Fouille de Données pour découvrir et extraire automatiquement l'information à partir des documents et des services du Web [Etz96]. La structuration des hyperliens a fait l'objet de nombreux travaux de recherche : Bharat [BH98], Brin [Bri98], Chakrabarti [CDK+98], Fürnkranz [Für99]. Ces travaux ont donné lieu à un axe de recherche et développement en plein essor depuis 2000, à savoir la Fouile des Structures Web.

4.3 Problématiques du développement

La compréhension de la propriété multilingue a besoin d'une analyse profonde dans la composition structurelle du site Web, plutôt que des statistiques sur l'utilisation de chaque langue dans l'ensemble des pages. En dehors de cette difficulté, la notion du site Web « multilingue » n'est pas très claire par rapport à la notion « monolingue ». De même, la notion des « langues dominantes », dans un site Web multilingue, reste encore à être précisée.

4.3.1 Confusion entre « monolingue » et « multilingue »

Un document comprenant plusieurs langues dans son contenu (parties textuelles, tableaux, etc.) est soit un document multilingue soit un document plurilingue (si ce sont des traductions fidèles en plusieurs langues). Par contre, un document écrit en une seule langue est monolingue. En dehors de ces deux notions, la notion multilingue peut être attribuée à un corpus (une collection de documents) ou à un site Web (un hyperdocument composé de plusieurs documents interconnectés), où chaque document peut être monolingue.

Dans ce travail de recherche, quand nous évaluons la propriété monolingue ou multilingue d'un site Web, nous n'abordons pas la nuance entre la notion de document multilingues et la notion plurilingues.

Néanmoins, la différenciation entre les sites monolingues et les sites multilingues n'est pas toujours claire. Pour faciliter cette tâche, nous proposons de classer les documents Web en trois catégories selon leurs fonctions dans un site Web :

1. Document de contenu : ce sont de documents dont le but n'est pas d'aider la navigation, mais de délivrer de l'information en contenant des parties textuelles. Les documents de contenus d'une même langue sont fortement interconnectés.
2. Document d'index : ce sont de documents d'aide à la navigation, comme les tables des matières ou les listes des hyperliens ou équivalents.
3. Document de référence : Les documents de référence sont des documents de contenu mais ils ne pointent pas vers d'autres documents.

En observant les deux types de documents, ceux de contenus et ceux de références, nous pouvons expliquer en grande partie la confusion entre le site « monolingue » et le site « multilingue » :

1. Quand les documents de contenu sont écrits en une seule langue, le site Web est monolingue même si les documents de références étaient écrits en d'autres langues.
2. Quand la majorité des documents de contenus sont écrits en une seule langue, même si une petite partie de ces documents est écrite en d'autres langues, le site Web est monolingue.
3. Quand les documents de contenu sont écrits en plusieurs langues, il faut examiner le critère du parallélisme multilingue (cf. section 1.3) pour qualifier le niveau de la propriété multilingue du site Web. Il est à noter que même si le critère du parallélisme multilingue est proposé, l'évaluation du niveau de la propriété multilingue d'un site Web restera encore non-quantitative.

Cette analyse permet d'affirmer que la propriété monolingue ou multilingue dépend totalement des documents de contenus dans un site Web. En plus, l'existence de plusieurs langues dans un site Web ne reflète absolument pas sa propriété multilingue.

4.3.2 Vague des langues « dominantes »

Un site Web multilingue est une structure typique dans lequel les documents écrits en une langue dominante doivent être très nombreux et aussi fortement interconnectés. La reconnaissance des langues dominantes ne consiste pas à faire une statistique des documents écrits en chaque langue du site Web multilingue sans vérifier l'importance de la densité des relations entres les documents dans chaque langue.

Un autre facteur qui affecte aussi la reconnaissance des langues dominantes est le phénomène des documents complémentaires. En principe, un site Web multilingue exige la correspondance de la majorité des contenus en toutes les langues (c'est-à-dire il y a des traductions d'un document dans toutes les langues). Cependant, plusieurs sites Web multilingues ne respectent pas ce critère générant ainsi une sorte d'ambiguïté causée par des documents dits *complémentaires*, qui existent dans quelques langues sans être présents en d'autres langues du site Web (cf. section 1.5). Ces sites Web multilingues ont appliqué cette stratégie pour publier des informations importantes en quelques langues très utilisées par de grandes communautés d'utilisateurs, et ne fournissant qu'un minimum d'informations en d'autres langues jugées peu universelles.

Les grands sites Web multilingues internationaux présentent très souvent des contenus en quelques langues internationales (c'est-à-dire des langues universellement utilisées sur l'Internet comme l'anglais, le français ou l'espagnol), tandis que les sites Web multilingues nationaux utilisent normalement leurs langues natives et le plus souvent ils ne fournissent que partiellement l'information sur le contenu en d'autres langues internationales. Dans ces cas, nous pouvons dire que certaines langues semblent plus importantes que d'autres dans un site Web multilingue. Cependant, il n'est pas évident de détecter ce type de comportement dans les sites Web.

Nous pouvons résumer sur les difficultés principales rencontrées lors de la reconnaissance des langues dominantes dans un site Web multilingue par les points suivants :

1. Quand il existe plusieurs langues dans un site Web, la découverte de toutes les langues utilisées ne peut pas renseigner sur les langues dominantes.
2. Quand il existe plusieurs langues dans un site Web, le nombre de documents écrits en une langue ne permet pas d'affirmer l'importance de cette langue dans le site Web.
3. L'influence des documents complémentaires.

4.4 Modélisation du système Hyperling

Dans cette partie, nous exposons les hypothèses initiales de la conception du système Hyperling. Par la suite, nous analysons les processus de reconnaissance des langues dominantes qui à leurs tours ont permis d'élaborer d'autres hypothèses. Nous concluons par la présentation de l'architecture générale du système Hyperling.

4.4.1 Hypothèses fondamentales

Le système Hyperling s'appuie sur une hypothèse qui consiste à observer la densité des relations « monolingues » entre les documents écrits en une même langue et la densité des relations « interlingues » entre les documents écrits en différentes langues dans un site Web multilingue. Cette hypothèse est présentée comme suit :

Hypothèse 1 *Dans un site Web multilingue, il existe beaucoup de relations monolingues, et il y a très peu de relations interlingues par rapport à celles-là.*

Pour mieux illustrer cette hypothèse, nous proposons de définir les notions suivantes : la relation entre deux documents Web, la relation monolingue et la relation interlingue.

Notion 5 ***Relation entre deux documents Web*** *: Un hyperlien, qui connecte un document source à un document destination, matérialise une relation entre eux.*

Notion 6 ***Relation monolingue*** *: Une relation entre deux documents d'une même langue est une relation monolingue.*

Notion 7 ***Relation interlingue*** *: Une relation entre deux documents de langues différentes est une relation interlingue.*

D'après nos premières expérimentations, l'hypothèse 1 observe une forte densité des relations monolingues par rapport à celle des relations interlingues dans les grands sites Web multilingues. Autrement dit, et selon une approche statistique expérimentale, les documents d'une même langue sont fortement connexes.

A partir de l'hypothèse 1, nous proposons ensuite l'hypothèse 2 :

Hypothèse 2 *Dans un site Web multilingue, les documents convergent selon leurs langues.*

Ces hypothèses ont été concrétisées par le développement du système Hyperling qui catégorise les documents d'un site Web multilingue en plusieurs catégories de langue.

4.4.2 Processus de reconnaissance des langues dominantes

En considérant l'hypothèse de la convergence des documents dans chaque langue (cf. l'hypothèse 2), nous proposons une méthode de reconnaissance des langues dominantes dans un site Web multilingue. Cette méthode ne permet pas seulement de classer des documents en catégories de langue, mais elle consiste à prouver la densité des relations monolingues (entre les documents d'une même langue) comme élément clé pour trouver des catégories dites dominantes.

Nous utilisons le modèle d'hyperdocuments GDS (Graphe de Documents Structurés) pour représenter la structure du site Web, au lieu de travailler directement sur un graphe de documents (premier niveau de représentation du GDS). Nous considérons que le graphe d'ancres sources (deuxième niveau de représentation du GDS) représente mieux la structure du site Web que le graphe de documents.

En travaillant sur le graphe d'ancres sources, nous nous appuyons sur une hypothèse supplémentaire qui est déduite de l'hypothèse 2 :

Hypothèse 3 *Dans un graphe d'ancres sources représentant un hyperdocument multilingue, les ancres sources convergent selon leurs langues.*

Pour identifier les langues dominantes, nous proposons l'hypothèse 4.

Hypothèse 4 *Les langues dominantes dans un site Web multilingue sont déterminées par les langues principales des catégories d'ancres sources qui convergent selon leurs langues.*

Le processus de reconnaissance des langues dominantes suit les phases suivantes :

1. Catégoriser des documents Web en fonction de leurs langues.
2. Evaluer si le site Web est monolingue ou peut être multilingue.
3. Quand le site Web est multilingue :
   - représenter la structure du site Web sous la forme d'un GDS,
   - définir un contexte distributionnel pour chaque ancre source dans un graphe d'ancres sources (deuxième niveau de représentation du GDS),
   - vectoriser les contextes distributionnels des ancres sources,
   - réduire la dimension des vecteurs,
   - appliquer un algorithme de catégorisation aux vecteurs,
   - préciser les catégories de vecteurs dominantes,
   - identifier les langues dominantes à l'aide des hypothèses.

Ce processus est la base de la conception de l'architecture générale du système Hyperling.

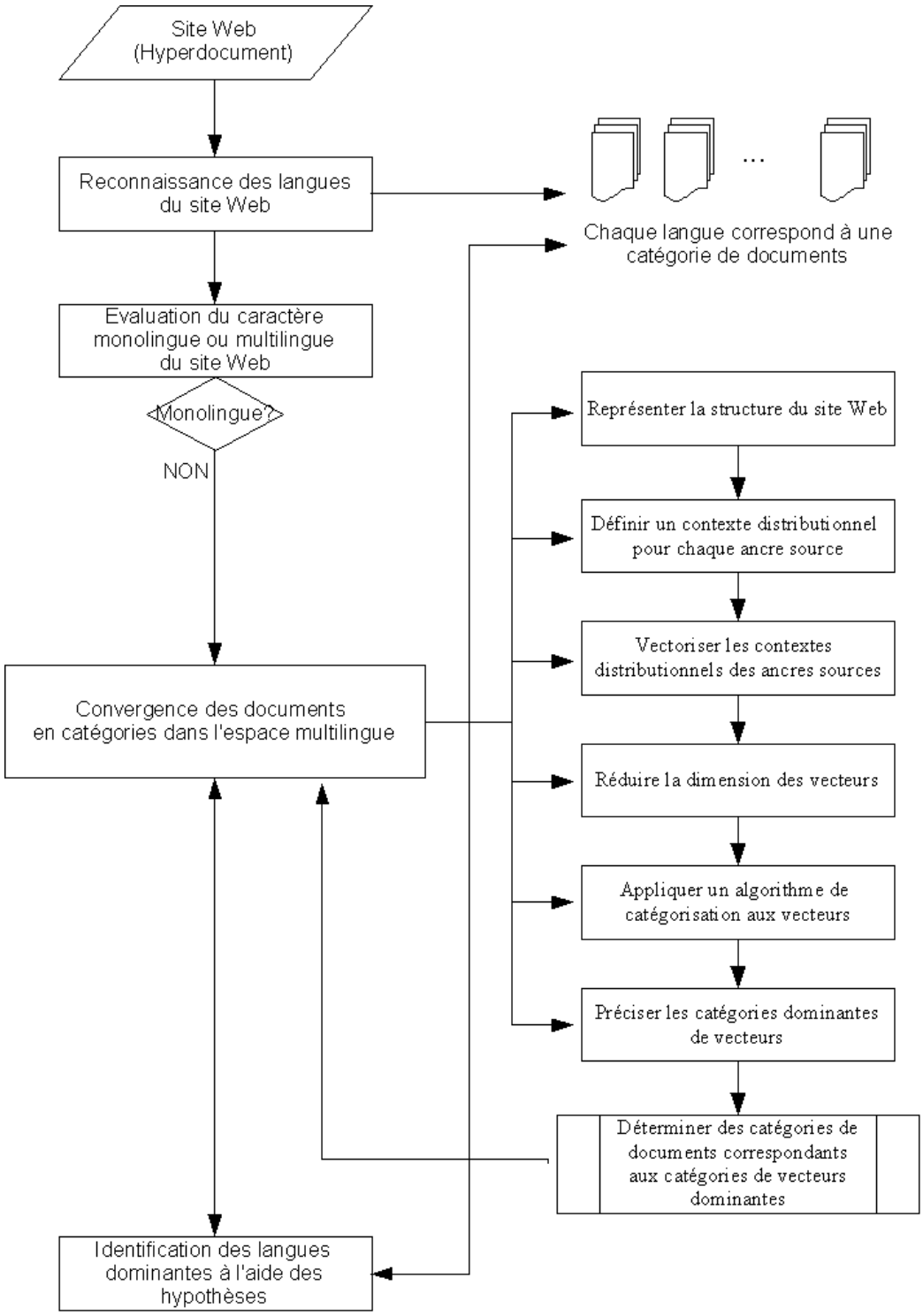


FIG. 4.1 – Processus de reconnaissance des langues dominantes

4.4.3 Architecture générale

L'architecture du système Hyperling a pour but de modéliser le processus de reconnaissance des langues dominantes (cf. section 4.4.2). Cette architecture est composée des composants principaux suivants :

Composant 1 ***Module de catégorisation des documents Web selon leurs langues*** **:** Ce composant consiste à identifier la langue de chaque document à partir de ses textes. Chaque langue correspond à une catégorie de documents. L'efficacité de ce module est optimal s'il s'agit d'un site monolingue.

Composant 2 ***Module d'évaluation du caractère monolingue ou multilingue*** **:** Ce module consiste à évaluer si un site Web est monolingue ou peut être multilingue.

Composant 3 ***Analyseur de structure*** **:** Ce composant correspond à l'analyse de la structure hypertexte intra-site et des structures hiérarchiques intra-pages.

Composant 4 ***Modèle d'hyperdocuments*** **:** Le système a besoin de définir un modèle de documents pour représenter les hyperdocuments. Ce modèle consiste à représenter la structure hypertexte intra-site et les structures hiérarchiques intra-pages. Afin de montrer la pertinence et la faisabilité de notre approche sur les sites Web, nous avons mis en oeuvre le modèle d'hyperdocuments au sein d'un système d'extraction d'information structurelle complet.

Composant 5 ***Modèle de fouille des structures Web*** **:** Ce modèle correspond à la modélisation des mécanismes d'extraction d'information structurelle. Ce module utilise des techniques d'apprentissage, de catégorisation ainsi que des algorithmes de fouille structurelle (c'est-à-dire des techniques du domain « Web Structure Mining »). La construction de ce module dépend de la spécificité du besoin, en exigeant des hypothèses fondamentales.

Composant 6 ***Module d'identification des langues dominantes*** **:** Ce module consiste à identifier les langues dominantes à l'aide des hypothèses.

L'architecture générale du système Hyperling est présentée dans la figure 4.2.

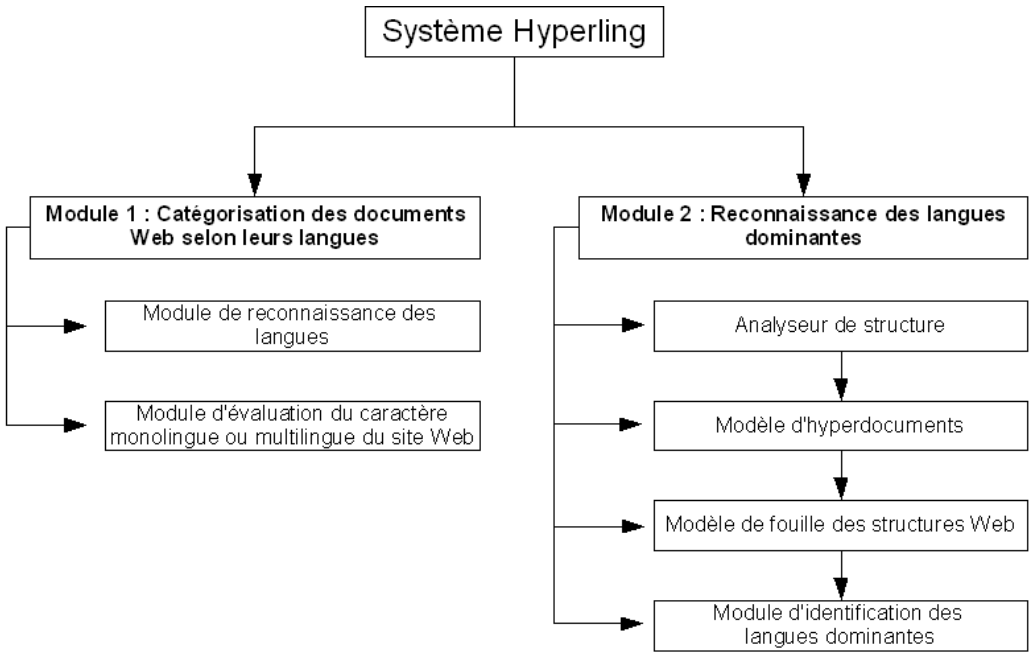


FIG. 4.2 – L'architecture générale du système Hyperling

L'implémentation de l'architecture générale d'Hyperling comporte deux modules principaux (cf. tableau 4.1).

	Composants
Module 1	Composant 1 : Module de catégorisation des documents Web selon leurs langues Composant 2 : Module d'évaluation du caractère monolingue ou multilingue du site Web
Module 2	Composant 3 : Analyseur de structure Composant 4 : Modèle d'hyperdocuments Composant 5 : Modèle de fouille des structures Web Composant 6 : Module d'identification des langues dominantes

TAB. 4.1 – Construction des modules du système Hyperling

4.5 Fonctionnement du système Hyperling

Les tâches principales du système Hyperling sont résumées dans le tableau 4.2.

Tâche	**Fonction**
1	Construire un texte de reconnaissance pour chaque document Web
2	Regrouper les textes de reconnaissance en fonction de leurs langues
3	Evaluer le caractère monolingue ou multilingue du site Web
4	Parcourir et analyser la structure hypertexte du site Web
5	Analyser la structure hiérarchique interne des documents Web
6	Représenter la structure du site Web
7	Définir un contexte distributionnel pour chaque ancre source
8	Vectoriser les contextes distributionnels des ancres sources
9	Réduire la dimension des vecteurs
10	Appliquer un algorithme de catégorisation aux vecteurs
11	Préciser les catégories dominantes de vecteurs
12	Identifier les langues dominantes

TAB. 4.2 – Les tâches principales du système Hyperling

Le schéma de fonctionnement du système Hyperling est présenté dans la figure 4.3.

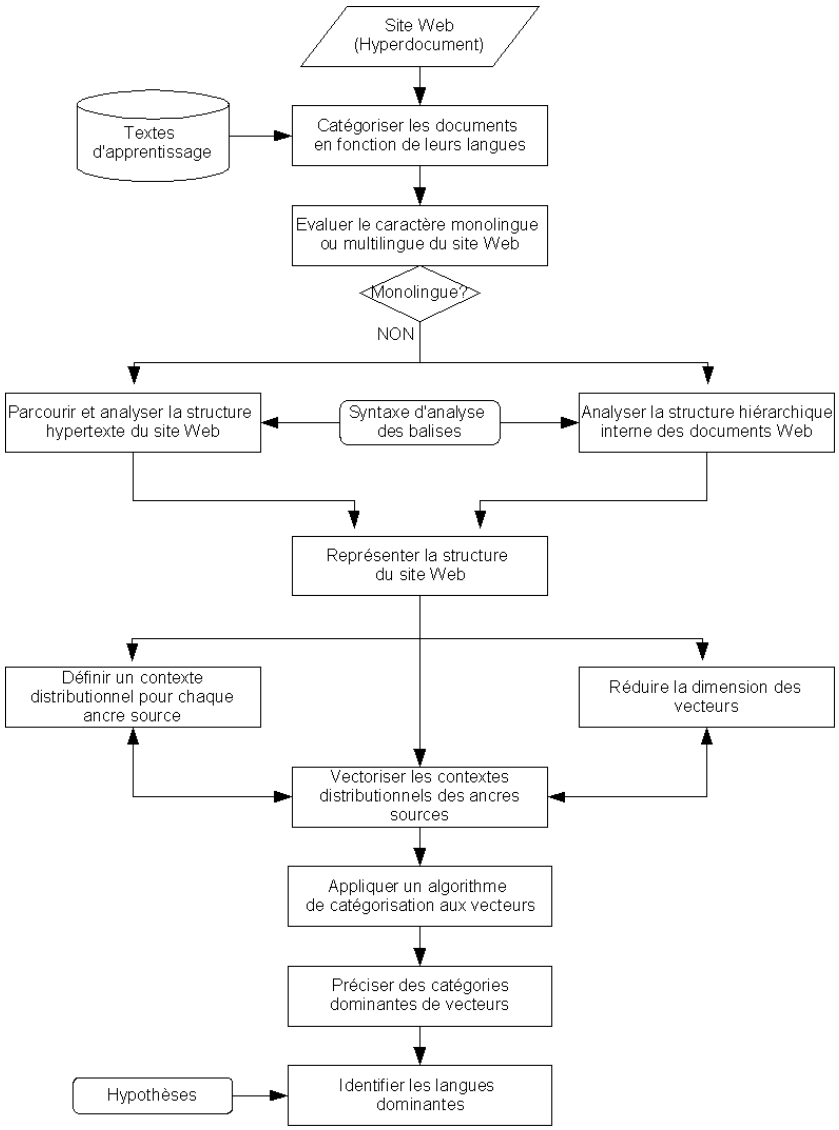


FIG. 4.3 – Fonctionnement du système Hyperling

4.6 Conclusion

Dans ce chapitre nous avons présenté des aspects fondamentaux du système Hyperling. La problématique du développement de ce système a été abordée en insistant sur la confusion entre les sites Web monolingues/multilingues et sur le vague dans la définition des langues dominantes.

En ce qui concerne la modélisation du système, nous avons introduit deux hypothèses fondamentales (l'importance de la densité des relations monolingues et la convergence des documents selon leurs langues dans un site Web multilingue). Nous avons également proposé le processus de reconnaissance des langues dominantes.

L'architecture générale du système Hyperling comportant deux modules principaux ainsi que les tâches principales du fonctionnement de ce système ont été également présentées.

Nos premières expérimentations ont porté sur des grands sites Web multilingues des organisations internationales, comme IMF (International Monetary Fund), UNDP (United Nations Development Program), UNFPA (United Nations Population Fund), UNICEF (United Nations Children's Fund) et WTO (World Bank). Elles ont montré une tendance de convergence très forte des documents dans chaque langue dominante [NZ04], [NZ05] [ZN05], [NZ06]. Ces résultats ont été interprétés pour donner une conclusion sur le nombre des langues dominantes dans un site Web multilingue. De plus, nous avons doté le système d'outils complémentaires lui permettant de reconnaître quelles sont ces langues dominantes.

Dans les deux chapitres suivants, nous présentons en détails la conception et le développement de ces deux modules du système Hyperling.

Chapitre 5

Catégorisation des documents Web selon leurs langues

5.1 Introduction

Dans le chapitre précédent (cf. chapitre 4), nous avons introduit le processus de reconnaissance des langues dominantes dans un site Web multilingue (cf. section 4.4.2) dans lequel la première tâche est de regrouper les documents d'un site Web selon leurs langues. Cette tâche permet entre autre d'identifier toutes les langues utilisées (avant de reconnaître des langues dominantes parmi elles). À partir de l'identification des langues présentes dans un site Web multilingue, le système Hyperling peut évaluer le caractère « monolingue » ou « multilingue » du site Web en proposant un algorithme pour considérer des cas confus entre ces deux notions que nous avons discutées dans la section 4.3.1.

Dans ce chapitre nous présentons le fonctionnement du module de catégorisation de langues qui se divise en trois étapes : pré-traitement (préparation des textes de reconnaissance relatifs aux documents Web en question), traitement (regroupement des textes de reconnaissance en fonction de leurs langues) et post-traitement (évaluation du caractère monolingue ou multilingue du site Web). Nous développons ensuite l'architecture du module en décrivant ses deux composants principaux (reconnaissance des langues et évaluation du caractère monolingue ou multilingue du site Web). En fin de ce chapitre nous délivrons le contenu de différentes expérimentations que nous avons menées.

5.2 Fonctionnement du module

Ce module réalise les tâches 1 à 3 dans l'architecture générale du système Hyperling (cf. tableau 5.1).

Tâche	Etape	Fonction
1	Pré-traitement	Construire un texte de reconnaissance pour chaque document Web
2	Traitement	Regrouper les textes de reconnaissance en fonction de leurs langues
3	Post-traitement	Evaluer le caractère monolingue ou multilingue du site Web

TAB. 5.1 – Synthèse des tâches dans le fonctionnement du module de catégorisation des documents selon leurs langues

Le processus de fonctionnement du module se décompose en trois étapes principales : le pré-traitement (préparation des textes de reconnaissance), le traitement (regroupement des textes de reconnaissance en fonction de leurs langues) et le post-traitements (évaluation du caractère monolingue ou multilingue du site Web) (cf. tableau 5.1).

5.2.1 Pré-traitement

Cette étape de pré-traitement prépare les données pour l'étape de traitement. En effet, les méthodes traditionnelles de reconnaissance des langues travaillent sur des textes bruts (et non sur la structures du document). Cette étape vise ainsi à tenir compte de cette différence dans la préparation des textes de reconnaissance, c'est à dire intégrer des critères structurels dans la préparation des données à regrouper.

Chaque document est alors représenté par un texte de reconnaissance formé des fragments textuelles issues de ses composants. Les documents sont analysés structurellement (selon des critères structurels) pour extraire ces parties textuelles à partir des titres, ancres sources, paragraphes, tableaux et des listes. Les documents ne contenant aucune partie textuelle ou ne comportant qu'un petit volume de texte qui est inférieur à un seuil pré-fixé seront éliminés.

5.2.2 Traitement

Cette étape de traitement consiste à regrouper les textes de reconnaissance en fonction de leurs langues. Dans cette étape, nous avons adapté le modèle de catégorisation de langues basé sur n-grams, proposé par Cavnar et Trenkle [CT94].

L'étape de traitement comporte les opérations suivantes :
- pour chaque langue, générer un profile de fréquences des n-grams à partir des textes d'apprentissage fournis pour cette langue,
- générer un profile de fréquences des n-grams pour chaque texte de reconnaissance,
- calculer la distance entre un profile des n-grams d'un texte de reconnaissance avec tous les profiles des n-grams de langue afin de classer ce texte de reconnaissance à une langue ayant la distance la plus proche.

La méthode de catégorisation des textes de reconnaissance est présentée dans la section 5.3.1 de ce chapitre.

5.2.3 Post-traitement

Cette étape de post-traitement est très importante, elle estime le caractère monolingue ou multilingue du site Web à l'aide d'un algorithme d'évaluation (cf. section 5.3.2).

5.3 Architecture du module

Ce module de catégorisation de langues comprend deux composants : le composant de reconnaissance des langues et le composant d'évaluation du caractère monolingue ou multilingue du site Web. Ses composants assument les tâches 1 à 3 dans l'architecture générale du système Hyperling (cf. tableau 5.2).

L'architecture du module est présenté dans la figure 5.1.

5.3.1 Module de reconnaissance des langues

L'identification automatique des langues naturelles est une tâche qui a été abordée par plusieurs approches statistico-linguistiques basées sur des caractères diacritiques spéciaux [New87], des caratéristiques syllabiques [Mus65],

Composant	Tâche	Etape
Module de reconnaissance des langues	1, 2	Pré-traitement Traitement
Module d'évaluation du caractère monolingue ou multilingue du site Web	3	Post-traitement

TAB. 5.2 – Synthèse des tâches distribuées aux composants du module de catégorisation des documents selon leurs langues

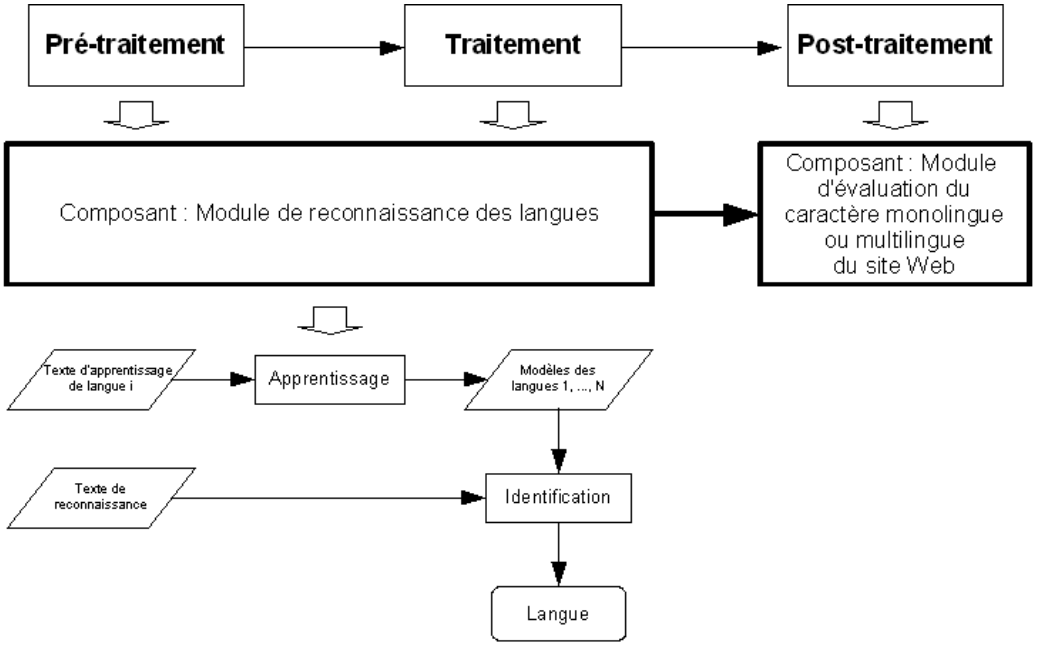


FIG. 5.1 – Architecture du module de catégorisation des documents selon leurs langues

morphologiques et syntaxiques [Zie91], des séquences de lettres caractéristiques [Dun94], des n-grams de caractères [Bee88] ou des séquences de mots [Bat92], etc.

Dans notre démarche, comme nous l'avons signalé, nous nous intéressons aux approches statistiques et tout particulièrement à la méthode de n-gram [Hay93], [CHJS94]. Nous résumons dans la suite quelques approches statistiques qui nous semblent pertinentes par rapport à notre contexte.

Modèle de Markov Les modèles de type Markov Caché (Hidden Markov Model) sont souvent utilisés pour la reconnaissance des langues parlées [ZS94], [LG94] ou écrites [UN90]. Dans ces modèles, chaque état s_i représente un tri-gram de caractères $c_{1i}c_{2i}c_{3i}$. Les paramètres principaux de ce modèle sont la probabilité de transition et la probabilité initiale :

$a_{ij} = P(q_{t+1} = s_j | q_t = s_i)$: la probabilité de transition de l'état i à l'état j.

$\pi_i = P(q_1 = s_i)$: la probabilité de commencer une chaîne dans l'état i.

Ces probabilités sont estimées par le calcul de la fréquence relative de chaque transition et de chaque état initial dans les données d'apprentissage :

$$a_{ij} = \frac{\#(s_i \rightarrow s_j)}{\#s_i}$$

$$\pi_i = \frac{\#s_i(t=0)}{\#chaines}$$

Pour classifier un texte, le système calcule ses probabilités dans tous les modèles de langue :

$$P(q_1, ..., q_T) = \pi_{q_1} \prod_{t=1}^{T-1} a_{q_t q_{t+1}}$$

Puis, le système choisit la langue ayant la plus grande probabilité.

Classificateurs de Naïve Bayes Dans ce modèle, un document d est normalement représenté par un vecteur de K attributs $d = (v_1, ..., v_K)$. Le modèle Naïve Bayes suppose que toutes les valeurs d'attribut v_j sont indépendantes d'une catégorie donnée c.

Un classificateur MAP (Maximum a posteriori) est construit comme suivant :

$$c^* = arg^{max}_{c \in C} \prod_{K}^{j=1} P(v_j|c)$$

L'estimation de $P(v_j|c)$ est souvent ajusté par la technique de lissage Laplace :

$$P(v_j|c) = \frac{N_j^C + a_j}{N^c + a}$$

où N_j^C est la fréquence de l'attribut j dans D^c, $N^c = \sum_j N_j^c$ et $a = \sum_j a_j$. Un cas particulier de lissage Laplace est le lissage « add one », obtenu par $a_j = 1$.

Classificateurs de SVM (Support Vector Machine) Etant donné un ensemble de N exemples linéairement séparés $S = x_i \in \Re^n | i = 1, 2, ..., n$, dont chacun appartient à une de deux classes $y_i \in +1, -1$, l'approche SVM cherche l'espace optimal $w.x + b = 0$ qui sépare avec la marge la plus large des exemples négatifs et positifs. Le problème est formulé comme :

$$minimiser \frac{1}{2} \|w\|^2$$

pour viser à $y_i(w.x_i + b \geq 1)$.

La méthode de n-grams La méthode de catégorisation des textes basée sur le modèle n-gram est appliquée à l'identification des langues naturelles où chaque catégorie de textes correspond à une langue [CT94].

Un n-gram est une chaîne de n caractères. En général, une chaîne de longueur K peut produire, en ajoutant des espaces au début et à la fin de la chaîne, différents autres chaînes : $K + 1$ bi-grams, $K + 1$ tri-grams, $K + 1$ quad-grams, etc. Chaque profile de n-grams est une liste des n-grams créés à partir d'un texte, triée par ordre décroissant de leurs fréquences d'occurences.

Pour catégoriser un texte, le système crée un profile de n-grams pour ce texte, puis compare ce profile avec chaque profile de n-grams de langue que le système a créé à partir des textes d'apprentissage. Cette comparaison est mesurée par la distance entre deux profiles, qui consiste à compter les différences entre la position $rank(t_i, texte)$ du n-gram t_i dans le profile du

texte de catégorisation par rapport avec la position $ranl(t_i, l_j)$ de ce n-gram dans le profile de langue j. La distance entre deux profiles est calculée par la somme de toutes les distances des n-grams :

$$D_j = \sum_{i=1}^{N} |rank(t_i, texte) - rank(t_i, l_j)|$$

où N est le nombre des tri-grams.

Le système calcule la distance entre le profile du texte de catégorisation avec tous les profiles des langues et choisit la langue ayant la distance la plus proche.

Vecteurs de fréquences des tri-grams Cette méthode consiste à comparer le vecteur de fréquences des tri-grams du texte de catégorisation avec les vecteurs de fréquences des n-grams des langues [Dam95].

Un tri-gram t_i est formé par trois caractères consécutifs, donc $t_i = c_{1i}c_{2i}c_{3i}$. Un vecteur de fréquences des tri-grams est un vecteur dans l'espace $N - dimension$, où N est le nombre des n-grams générés,

$$\vec{v} = (v_1, ..., v_N)$$

et v_i est la fréquence du t_i.

Dans l'apprentissage, le système calcule la fréquence relative de chaque tri-gram dans l'ensemble des textes d'apprentissage disponibles pour chaque langue. Un vecteur $\vec{l^j}$ est construit pour chaque langue.

Pour catégoriser un texte, le système construit un vecteur de fréquences des tri-grams $\vec{w}$ pour ce texte, puis compare ce vecteur avec chaque vecteur de fréquences des tri-grams de langue $\vec{l^j}$. Cette comparaison est calculée comme suit :

$$\vec{w}.\vec{l^j} = \frac{\sum_{i=1}^{N} w_i l_i^j}{|\vec{w}| \left|\vec{l^j}\right|}$$

Le système choisit la langue ayant la valeur maximale pour ce calcul.

L'approche retenue pour Hyperling L'efficacité de ces méthodes dépend de plusieurs paramètres comme par exemple la taille des textes d'apprentissage, la taille des textes de reconnaissance, les mots OOP (out-of-place), l'utilisation des grams de caractères ou de mots, etc. Parmi ces paramètres, la taille des textes d'apprentissage et la taille des textes de reconnaissance sont deux facteurs indépendants de l'implémentation du système qui peuvent augmenter l'efficacité du système [ACT04], [PP04], [PS03]. Les techniques de lissage (smoothing techniques) sont aussi appliquées pour la construction des modèles de langue basés sur n-grams [CT98].

Pour construire le composant de reconnaissance des langues dans Hyperling nous avons opté pour une méthode proposée par Cavnar et Trenkle [CT94] qui est basée sur le modèle de n-grams. Cette approche a été approuvée par un nombre important de différentes implémentations et dérivés.

S'inspirant de cette approche pour regrouper les textes de même langue, Hyperling réalise les opérations suivantes :

1. Pour chaque langue, générer un profile de n-grams à partir des textes d'apprentissage disponibles pour cette langue,
2. Pour chaque texte de catégorisation, générer son profile de n-grams,
3. Pour chaque profile de n-grams :
   - calculer les distances entre son profile de n-grams avec tous les profiles de n-grams des langues,
   - déterminer la distance minimale,
   - catégoriser le texte correspondant à ce profile de n-grams à la langue ayant la distance minimale.

La création des profiles de n-grams consiste à « scanner » les textes de catégorisation et à compter toutes les occurrences des n-grams existant dans ces textes. Chaque texte correspond donc à un profile de n-grams. Cette opération suit les étapes suivantes :

1. Segmenter chaque texte en termes (tokens) basés seulement sur les lettres et l'apostrophe (les chiffres et les ponctuations sont éliminés),
2. Générer tous les n-grams possibles de taille de 1 à 7 caractères.
3. Trier les n-grams en ordre décroissant de leurs fréquences d'occurences.

La comparaison des profiles de n-grams consiste en deux opérations :

1. Calculer la distance entre deux profiles de n-grams.
2. Déterminer la distance minimale entre deux profiles de n-grams.

Le calcul de la distance entre deux profiles de n-grams est basé sur la mesure de « out-of-place » [CT94]. La mesure « out-of-place » est une notion très simple qui propose de déterminer la différence entre deux positions d'un n-gram dans deux profiles de n-grams triés par ordre décroissant.

La détermination de la distance minimale entre deux profiles de n-grams consiste à :

1. Calculer les distances entre un profile de n-grams du texte de catégorisation avec tous les profiles de n-grams des langues.
2. Choisir la langue ayant la distance minimale pour ce texte de catégorisation.

Le processus de catégorisation des documents selon leurs langues est présenté dans la figure 5.2 [CT94].

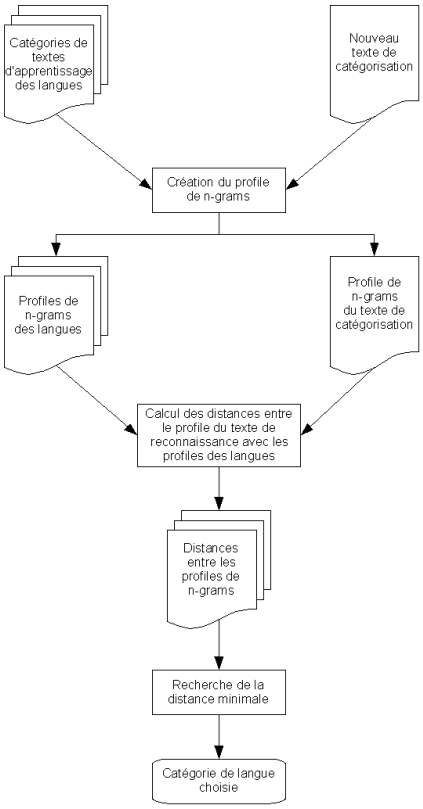


FIG. 5.2 – Processus de catégorisation des documents selon leurs langues

5.3.2 Module d'évaluation du caractère monolingue ou multilingue du site Web

La catégorisation des documents en fonction de leurs langues permet par conséquent d'identifier le nombre de langues utilisées dans un site Web. À partir de ce résultat, nous proposons donc un algorithme pour évaluer le caractère monolingue ou multilingue du site Web :

Algorithme d'évaluation du caractère monolingue ou multilingue du site Web :

1. S'il existe une seule langue dans le site Web, le site Web est absolument monolingue,
2. Sinon, (il y a plusieurs langues dans le site Web)
   - S'il y a une langue qui occupe la majorité de documents (à l'aide d'un seuil pré-défini), ce site Web est considéré comme monolingue,
   - Sinon, le site Web est multilingue et il faut reconnaître des langues dominantes.

Quand le site Web est multilingue, nous appliquons ultérieurement un traitement de reconnaissance des langues dominantes. En fait, il est éventuellement possible que ce site Web a une seule langue dominante.

5.4 Expérimentations

Pour créer des profiles de langues, le système Hyperling utilise des textes d'apprentissage collectés à partir des titres résumés dans les actualités de Google[1]. Les textes d'apprentissage en codage Unicode correspondent aux langues : l'allemand, l'anglais, l'espagnol, le français, l'italien et le vietnamien. Ces textes couvrent différents thèmes : Top Stories, International, France (ou d'autres pays), Economie, Science/Technologie, Sports, Culture, à l'exception du vietnamien dont les textes ont été extraits à partir des divers articles dans le journal en ligne « La Jeunesse » [2].

L'efficacité de cette méthode, proposée par Cavnar et Trenkle [CT94], a été testée et comparée avec d'autres méthodes par Ahmed [ACT04], Padro [PP04] et Peng [PS03].

[1] http ://news.google.com/
[2] http ://www.tuoitre.com.vn/

Langue	**Taille des textes d'apprentissage**	**Nombre des n-grams**	**Taille des n-grams**
allemand	92 KB	370.848	2-7
anglais	85 KB	320.406	
espagnol	104 KB	388.128	
français	80 KB	297.372	
italien	94 KB	355.500	
vietnamien	50 KB	119.220	

Tab. 5.3 – Caractéristiques des textes d'apprentissage

Pour tester et valider nos hypothèses ainsi que les modules implantés dans Hyperling, nous avons examiné des sites Web multilingues (cf. tableau 5.4) :

Sites Web (Langues dominantes)	**Nombre de documents**	**Taille HTML**
IMF (anglais, français, espagnol)	8.912	1,29GB
UNDP (anglais, français, espagnol)	9.869	1,49GB
UNICEF (anglais, français, espagnol)	5.063	160MB
WTO (anglais, français, espagnol)	7.246	344MB
Ambassade de France en Espagne (français, espagnol)	3.228	31,8 MB
Ambassade de France au Vietnam (français, vietnamien)	1.159	13,6 MB

Tab. 5.4 – Caractéristiques des sites Web multilingues

Le système Hyperling a été expérimenté également sur différents sites Web monolingues, comme par exemple des sites Web : TF1[3], France 2[4], M6[5], VNExpress[6], etc.

[3]http ://www.tf1.fr/
[4]http ://www.france2.fr/
[5]http ://www.m6.fr/
[6]http ://www.vnexpress.net/

Les résultats des expérimentations sur ces sites Web sont très favorables et permettent d'identifier leurs langues utilisées ainsi que d'évaluer leur caractère monolingue ou multilingue.

5.5 Conclusion

La catégorisation des documents selon leurs langues est une première tâche qui permet d'identifier le nombre de langues utilisées dans un site Web. À partir de cette connaissance, le système Hyperling peut évaluer la propriété monolingue ou multilingue du site Web.

En dehors de ces informations acquises, nous pouvons découvrir la structure essentielle du site Web multilingue en observant les frontières entre les catégories de documents correspondant aux langues, et l'utilisation des ancres de changement de langue ou des ancres de référence interlingue. Ces ancres sources spécifiques sont très typiques pour les sites Web multilingues et elles caractérisent la structure générale de ces derniers.

Dans le chapitre suivant, nous présentons le second module du système Hyperling qui consiste à identifier les langues dominantes dans un site Web multilingue.

Chapitre 6

Reconnaissance des langues dominantes

6.1 Introduction

Dans cette thèse, nous observons que les sites Web multilingues représentent, entre autres, des entités structurelles. En ce sens, la tâche d'identification des langues dominantes dans un site Web multilingue, ne se limite pas à des critères statistiques simples du type du nombre de documents dans chaque langue pour déterminer les langues dominantes.

Dans ce chapitre nous développons l'implémentation du second module principal d'Hyperling qui permet d'identifier les langues dominantes dans un site Web multilingue à partir des critères structurels.

Nous présentons d'abord le fonctionnement du module. Ensuite nous nous concentrons sur le schéma du module en décrivant l'architecture et le fonctionnement des composants principaux, c'est à dire : analyseur de structure, modèle d'hyperdocuments, module de fouille des structures Web et module d'identification des langues dominantes. Enfin, nous discutons les résultats obtenus par les expériences que nous avons mené sur des sites Web multilingues pour valider le système Hyperling.

6.2 Fonctionnement du module

Ce second module du système Hyperling effectue les tâches 4 à 12 dans l'architecture générale du système Hyperling (cf. tableau 6.1).

Le processus de fonctionnement du module se décompose en trois étapes principales : le pré-traitement, le traitement et le post-traitement. La distribution des tâches dans les étapes du processus de fonctionnement du module est présentée dans le tableau 6.1.

Tâche	**Etape**	**Fonction**
4	Pré-traitement	Parcourir et analyser la structure hypertexte du site Web
5		Analyser la structure hiérarchique interne des documents Web
6	Traitement	Représenter la structure du site Web
7		Définir le contexte distributionnel pour chaque ancre source
8		Vectoriser les contextes distributionnels des ancres sources
9		Réduire la dimension des vecteurs
10		Appliquer un algorithme de catégorisation aux vecteurs
11	Post-traitement	Préciser les catégories dominantes de vecteurs
12		Indentifier les langues dominantes

TAB. 6.1 – Synthèse des tâches dans le fonctionnement du module de reconnaissance des langues dominantes

6.2.1 Pré-traitement

L'étape de pré-traitement prépare des données pour l'étape de traitement. Cette étape assume les deux tâches (4 et 5) dans l'architecture générale du système Hyperling (cf. tableau 6.1). Cette étape consiste à :

- parcourir et analyser la structure hypertexte du site Web,
- analyser la structure hiérarchique interne des documents Web.

En phase de pré-traitement des ressources d'information (des sites Web), Hyperling s'appuie sur des méthodes d'analyses structurelles et statistiques (approuvées par diverses communautés d'apprentissage et de fouille des données).

6.2.2 Traitement

L'étape de traitement couvre les tâches 6 à 10 dans l'architecture générale du système Hyperling (cf. tableau 6.1). L'objectif de cette étape consiste à réaliser les opérations suivantes :

- représenter la structure du site Web sous la forme d'un GDS (Graphe de Documents Structurés),
- déterminer un contexte distributionnel pour chaque ancre source (un noeud) dans le graphe d'ancres sources, qui est le deuxième niveau de représentation du GDS,
- vectoriser chaque ancre source à base de son contexte distributionnel,
- réduire la dimension des vecteurs,
- appliquer un algorithme de catégorisation aux vecteurs.

A l'utilisation d'un graphe d'ancres sources, nous pouvons proposer une méthode pour déterminer un contexte distributionnel pour chaque ancre source. Cette méthode consiste à utiliser des chemins partant de l'ancre source « racine » (le noeud initial) du graphe et provenant à une ancre source observée. Les ancres sources apparaissent dans les chemins arrivant à une ancre source observée forment le contexte de celle-ci.

A partir des contextes distributionnels, nous proposons de vectoriser chaque ancre source pour refléter sa distribution à l'utilisation de fréquences de ses coexistences avec d'autres ancres sources sur différents chemins. L'ensemble des vecteurs devient l'entrée d'un algorithme de catégorisation.

6.2.3 Post-traitement

L'étape de post-traitement correspond aux tâches 10 et 11 dans l'architecture générale du système Hyperling (cf. tableau 6.1). Cette étape consiste à :

- préciser les catégories de vecteurs dominantes,
- identifier les langues dominantes.

Dans la section suivante, nous présentons l'architecture de ce module de reconnaissance des langues dominantes dans un site Web multilingue.

6.3 Architecture du module

Le module de reconnaissance des langues dominantes dans un site Web multilingue se compose de quatre composants principaux :

- un analyseur de structure,
- un modèle d'hyperdocuments,
- un modèle de fouille des structures Web,
- un module d'identification des langues dominantes.

Chaque composant réalise une ou des tâches dans l'architecture générale du système Hyperling (cf. tableau 6.2).

Composant	**Tâche**
Analyseur de structure	4, 5
Modèle d'hyperdocuments	6
Modèle de fouille des structures Web	7, 8, 9, 10
Module d'identification des langues dominantes	11, 12

TAB. 6.2 – Synthèse des tâches distribuées aux composants du module de reconnaissance des langues dominantes

L'architecture du module est présenté dans la figure 6.1. Ce schéma explique la construction générale du module ainsi que la fonction de chaque composant dans le processus de fonctionnement du module.

La construction des composants du module est introduite dans les parties suivantes.

6.3.1 Analyseur de structure

Nous avons développé un outil d'analyse et d'extraction de structures à partir des sites Web. À partir de données brutes HTML, l'analyseur extrait des corpus normalisés (texte, liens, images, etc.), la structure des sites et des pages.

L'analyseur fait l'extraction selon 4 niveaux de granularités différentes (cf. tableau 6.3). Le processus d'extraction est basé essentiellement sur des éléments syntaxiques des pages HTML, comme par exemple l'utilisation de balises HTML précises pour délimiter des paragraphes, et des hyperliens.

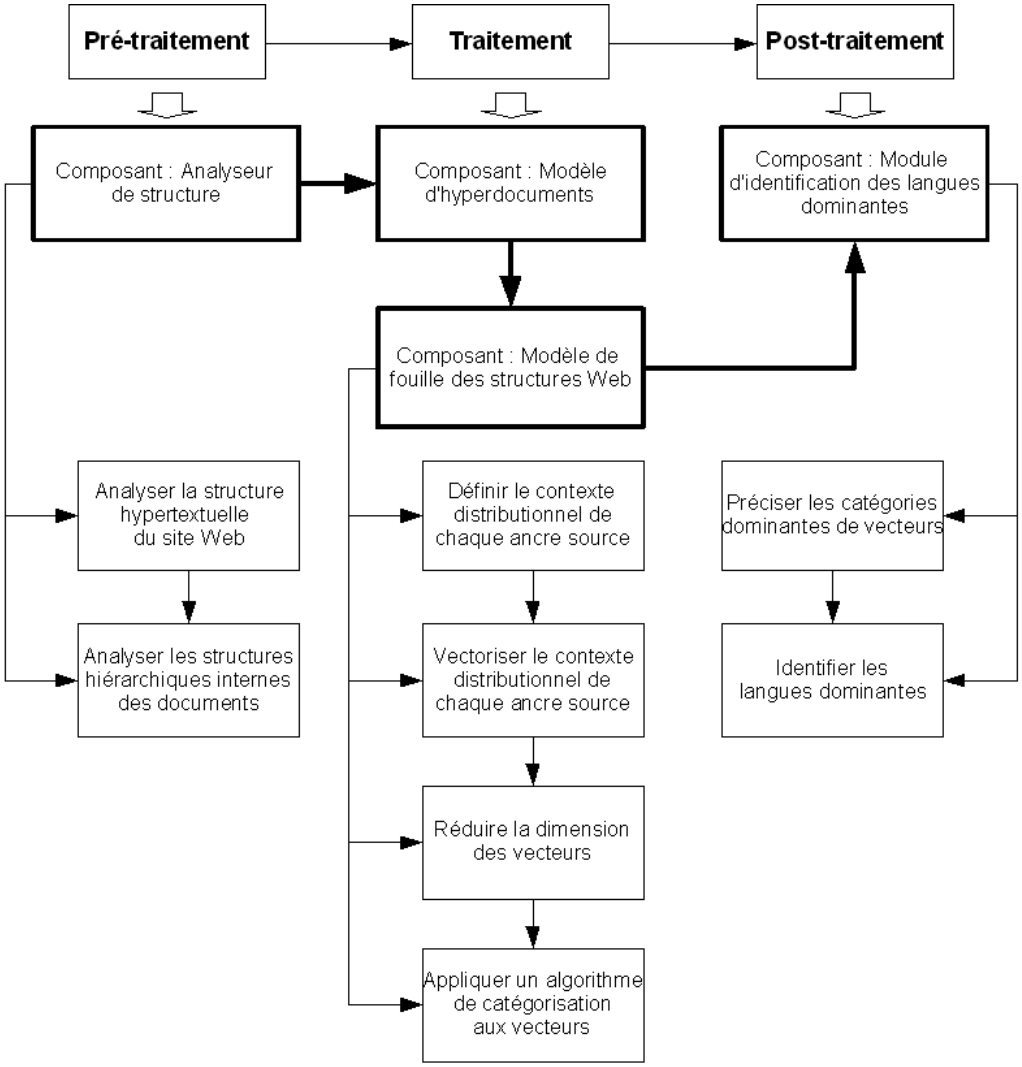


FIG. 6.1 – Architecture du module de reconnaissance des langues dominantes

Selon les différents niveaux de granularité, l'analyseur pourrait se trouver face à des composants élémentaires comme des images, des sons, des vidéos, des hyperliens, etc. Dans notre approche de modélisation, les composants élémentaires sont considérés comme des objets nominatifs et insécables ne pouvant pas contenir aucun autre composant.

Niveau de granularité	**Objet syntaxique**
Site	Nom du site
Documents	Page HTML
Sections	Balises HTML *séparateurs de sections*
Paragraphes	Balises HTML *séparateurs de blocs* Balises HTML *séparateurs de paragraphes*

TAB. 6.3 – Niveaux de granularité d'analyse

L'analyseur de structure enchaîne les opérations suivantes :

1. Fouiller la structure hypertexte du site Web. Cette opération est réalisée par un robot qui consiste à :
   – déterminer la page d'accueil du site Web,
   – parcourir les hyperliens : pour chaque hyperlien, déterminer le document source, le document destination et l'ancre source.
2. Analyser chaque document trouvé pour retenir l'information structurelle interne au document. Cette opération consiste à :
   – identifier des composants pré-spécifiés du document,
   – reconnaître la structure hiérarchique des composants.
3. Synthétiser des composants partagés entre plusieurs documents.

6.3.2 Modèle d'hyperdocuments

Dans le système Hyperling, nous proposons de représenter un site Web sous la forme d'un GDS (Graphe de Documents Structurés) (cf. section 3.4.3). Ce modèle permet de représenter la structure hypertextuelle du site Web, les structures hiérarchiques internes aux documents ainsi que d'élargir ultérieurement le modèle pour étudier les contenus textuels dans les paragraphes ou de décrire davantage les composants des documents par leurs caractéristiques quantitavies ou qualitatives. Selon le type d'application, tout le modèle ou des parties du modèle peuvent être exploités (Hyperling utilise le second niveau du modèle de représentation GDS).

Comme nous l'avons introduit dans la section 4.4.2 (cf. chapitre 4), ce module du système Hyperling consiste à travailler sur le graphe d'ancres sources, qui est le deuxième niveau de représentation du GDS.

6.3.3 Modèle de fouille des structures Web

La fouille des structures Web consiste à analyser la structure du Web, autrement dit l'architecture des sites et aussi des liens qui existent entre différents sites. La fouille des structures Web a pour but de produire des informations (de type synthétique ou résumé) sur la structure du site et des pages Web en essayant de découvrir la structure des hyperliens au niveau inter-documents.

La spécificité de ce modèle est qu'il travaille sur un graphe d'ancres sources au lieu d'un graphe de documents. Basé sur l'hypothèse de la convergence des ancres sources d'une même langue dans un graphe d'ancres sources représentant un site Web multilingue (cf. hypothèse 3), ce modèle de fouille des structures Web est construit en appliquant notre méthode d'identification des langues dominantes (cf. section 4.4.2).

Nous proposons également de déterminer le « contexte » de chaque ancre source sur un graphe d'ancres sources. Puis, le contexte de chaque ancre source sera vectorisé. Chaque ancre source est alors représentée par un vecteur de contexte.

Nous appliquons ensuite un algorithme de catégorisation travaillant avec les vecteurs de contexte pour permettre de trouver des catégories de vecteurs dominantes. A partir des catégories de vecteurs dominantes, nous proposons enfin une méthode d'identification (« hypothétique ») des langues dominantes.

Le modèle de fouille des structures Web est organisé en quatre phases :

Phase 1 : Déterminer le contexte distributionnel de chaque ancre source

Dans un graphe d'ancres sources, les ancres sources représentent les noeuds et les relations directes entre les couples d'ancres sources et sont représentées par des arcs. Cette représentation exige la prise en compte du contexte distributionnel pour chaque ancre source.

Définition 3 ***Contexte distributionnel d'une ancre source :*** *Etant donné une ancre source, qui est un noeud dans le graphe orienté, nous définissons son contexte distributionnel comme constitué par tous les ancres sources se trouvant sur les chemins partant de l'ancre source initiale du graphe orienté et passant par elle.*

Cette méthode consiste à déterminer un ensemble de chemins pour chaque ancre source, à l'exception des ancres sources éliminées. En générant ces chemins, Hyperling respecte les contraintes suivantes :

- tous les chemins commencent par l'ancre source initiale,
- un chemin ne contient pas un autre chemin.

Pour optimiser le processus et éviter les cycles nous considérons que :

- une ancre source peut apparaître une seule fois dans un chemin,
- une ancre source peut appartenir à plusieurs chemins.

Pour chaque ancre source, à partir d'un ensemble des chemins aboutissant à elle, nous déterminons un ensemble d'ancres sources qui forme le contexte distributionnel de cette ancre source.

Phase 2 : Vectoriser le contexte distributionnel de chaque ancre source

La majorité d'approches de classification ou de catégorisation opère sur des données qui sont prétraitées et représentées par des structures robustes et simplifiées, par exemple la représentation tabulaire (dans les arbres de décisions) ou la représentation vectorielle (pour la catégorisation conceptuelle, ou l'apprentissage paramétrique).

Hyperling, utilise des algorithmes de catégorisation de type SOM (Self-Organisation Map) et K-means dont l'exécution exige une suite de prétraitements des données extraites du site Web pour les transformer en structures homogènes calculables. Dans Hyperling nous optons pour une représentation numérique vectorielle. La numérisation d'une ancre source sera rendue sous forme de vecteur.

Définition 4 ***Vecteur de contexte distributionnel d'une ancre source :*** *L'information sur la distribution d'une ancre source (un noeud) dans les différents chemins d'un graphe d'ancres sources est représentée par un « vecteur de contexte distributionnel ».*

Dans notre approche, les relations entre ancres sources sont prises en comptes par leurs coexistences distribuées dans le modèle de représentation du site Web. Nous proposons donc une méthode qui consiste à « vectoriser » le contexte distributionnel de chaque ancre source.

Définition 5 ***Vectorisation du contexte distributionnel d'une ancre source :*** *Le processus de calcul des poids, qui sont des fréquences de co-existence d'une ancre source observée avec d'autres ancres sources dans son contextes distributionnel, et de les assigner aux éléments d'un vecteur est appelé la « vectorisation du contexte distributionnel d'une ancre source ».*

Le contexte d'une ancre source est défini de manière qu'il à ce reflète la fréquence et la densité de coexistences de cette ancre source avec les autres.

Phase 3 : Réduire la dimension des vecteurs

Les vecteurs ainsi générés constituent l'ensemble de l'apprentissage qui sera considéré par les algorithmes de catégorisation. Cependant, la grande taille des vecteurs pose toujours des problèmes d'optimisation (la mémoire, le temps de calcul) pour les algorithmes de catégorisation.

Les projections aléatoires sont considérées comme des méthodes puissantes pour la réduction de dimensions des vecteurs obtenus. Les résultats théoriques indiquent que la méthode de projection aléatoire préserve bien la similarité (du point de vu de distances) des vecteurs de données. De même, les expériences ont montré que l'application des projections aléatoires est moins coûteuse que d'autres méthodes, par exemple, celle de l'analyse de Composant Principal [BM01]. Pour ces deux raisons, Hyperling adopte la méthode Random Mapping, proposée par Kaski [Kas98].

Dans la méthode de projection aléatoire, proposée par Kaski, le vecteur de données originel, noté par $n \in \Re^N$, est multiplié par une matrix des valeurs aléatoires R. La projection $x = Rn$ résulte un vecteur de dimensions réduites $x \in \Re^d$. Les propriétés de cette projection ont été prouvées par Kaski [Kas98].

Phase 4 : Appliquer un algorithme de catégorisation aux vecteurs

Une fois que les vecteurs ont été extraits et optimisés formellement (une représentation ordonnées et des dimensions réduites), Hyperling dispose des méthodes de catégorisation dérivées à partir de K-means [Mac67], [BB95] et de SOM (Self-Organizing Map) [Koh95]. Hyperling propose d'en appliquer l'une ou l'autre ou les deux si on souhaite comparer la qualité des résultats obtenus. Par défaut Hyperling applique la méthode issue de K-Means. Quelque soit la méthode choisie, nous avons retenu la distance Euclidienne pour mesurer la distance et par conséquent la similarité entre les vecteurs.

K-means : créé par MacQueen [Mac67], est l'algorithme de catégorisation le plus connu et le plus utilisé, qui suit une procédure simple pour classifer un ensemble d'objets en un certain nombre K de clusters, avec K fixé à priori.

Soit un ensemble d'objets $D_n = (x_1, ..., x_n)$, avec pour tout i, x_i réel et soit μ_k, $1 < k < K$, les centres des K clusters, l'algorithme des K-means s'exécute en 4 étapes :

1. Choisir aléatoirement K objets qui forment ainsi les K clusters initiaux. Pour chaque cluster k, la valeur initiale du centre est $\mu_k = x_i$, avec x_i l'unique objet de D_n appartenant au cluster.
2. Ré-)Affecter les objets à un cluster. Pour chaque objet x, le prototype qui lui est assigné est celui qui est le plus proche de l'objet, selon une mesure de distance : $s = argmin_k \|\mu_k - x\|^2$.
3. Une fois tous les objets placés, recalculer les centres des clusters.
4. Répéter les étapes 2 et 3 jusqu'à ce que plus aucune réaffectation ne soit faite.

Self-Organizing Map (SOM) : introduit par Kohonen [Koh95], permet de convertir les relations statistiques complexes et non linéaires en relations géométriques simples dans une carte bidimensionnelle. De plus cet algorithme permet de garder les relations topologiques et métriques les plus pertinentes dans l'espace de données réel. Par conséquent, il représente un véritable outil de visualisation des données multidimensionnelles.

L'apprentissage de SOM dans les cartes topologiques procède en 3 étapes :

1. Les vecteurs référentiels sont initialisés par des valeurs aléatoires,
2. Pour chaque vecteur de données, un calcul de distance est effectué pour activer un neurone dit gagnant, c'est celui dont le vecteur référentiel est le plus proche du vecteur de données,

3. Le vecteur référentiel du neurone gagnant est localement ajusté, en minimisant la différence qui existe encore entre ce vecteur référentiel et le vecteur de données. En plus, cet ajustement se fait aussi pour les neurones voisins du neurone gagnant, selon un voisinage qui peut être une forme carrée, ronde ou hexagonale, ou qui est déterminée par une fonction de voisinage.

Nous avons implémenté et testé les algorithmes K-means et SOM dans [Ngu04], [NZ04], [NZ05], [ZN05], [NZ06].

6.3.4 Module d'identification des langues dominantes

Ce module est composé de deux phases pour reconnaître « hypothétiquement » des langues dominantes :
- préciser les catégories dominantes de vecteurs,
- identifier des langues dominantes.

Préciser les catégories dominantes de vecteurs

Les résultats obtenus par l'algorithme de catégorisation sont soumis à un certain nombre de règles « d'explication » dont le but est de retenir les catégories informationnelles, dites déterminantes.

Nous avons précisé les catégories de vecteurs par une méthode proposée par Vesanto [VA00]. Cette méthode consiste à catégoriser des vecteurs de prototypes acquis à partir des algorithmes de K-means ou SOM.

L'algorithme de catégorisation par agglomération des catégories :

1. Un ensemble des vecteurs de prototype est formé par l'algorithme K-means (ou SOM) à partir des vecteurs de données.
2. Calculer la distance entre deux vecteurs de prototype avec l'indice de Davies-Bouldin.
3. Si deux vecteurs de prototype sont proches, ils seront regroupés à une même catégorie.
4. Refaire l'étape 2 jusqu'au moment où aucun vecteur ne peut être regroupé avec aucun autre.
5. Regrouper les vecteurs de données dont les vecteurs de prototype sont de même catégorie.

L'indice de Davies-Bouldin est définis comme suivantes :

$$\frac{1}{C}\sum_{k=1}^{C} max_{l\neq k}\frac{S_C(Q_k)+S_C(Q_l)}{d_{ce}(Q_k,Q_l)}$$

où C : le nombre des catégories obtenues par K-Means ou SOM, $S_C = \frac{\sum_i \|x_i - c_k\|}{N_k}$: la distance inter-catégorie, et $d_{ce} = \|c_k - c_l\|$: la distance entre deux catégories, avec N_k : le nombre des vecteurs dans la catégorie Q_k, $x_i \in Q_k$, $c_k = \frac{1}{N_k \sum_{x_i \in Q_k} x_i}$.

La figure 6.2 illustre le processus de création des super-catégories à partir des catégories initiales [VA00]. L'algorithme permet aussi de former plusieurs niveaux d'abstraction des catégories (hiérarchie de catégories).

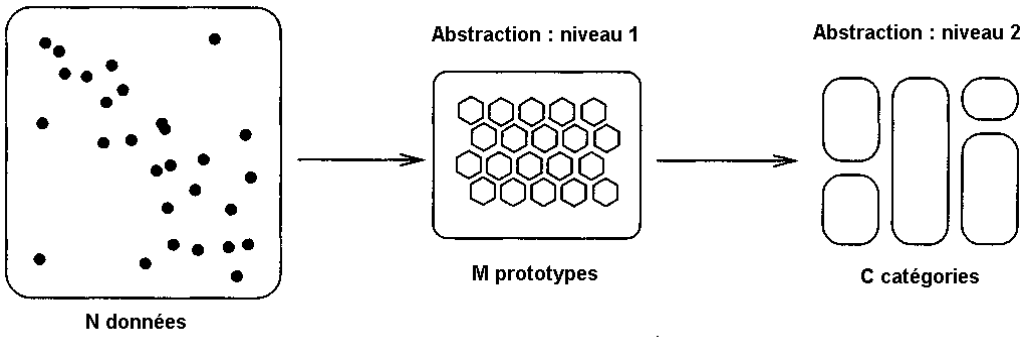


FIG. 6.2 – Processus de création des super-catégories

Pour renforcer l'indice Davies-Bouldin, Vesanto propose de comparer la somme des distances moyennes entre les vecteurs dans deux catégories avec la distance entre deux centres de catégorie [VA00].

Deux catégories se confondent quand un des deux critères suivants est satisfait :

- la valeur de l'indice Davies-Bouldin est supérieur de 1, ou
- l'écart entre deux catégories $d_s(Q_k, Q_l)$ est supérieur à la somme des distances moyennes entre les vecteurs dans ces deux catégories $S_{nn}(Q_k + S_{nn}(Q_l))$.

La figure 6.3 illustre les critères de regroupement des catégories [VA00]. La confusion est faite quand $d > S1 + S2$, où $S1$ et $S2$ sont les distances internes de deux catégories, et d est la distance entre deux catégories.

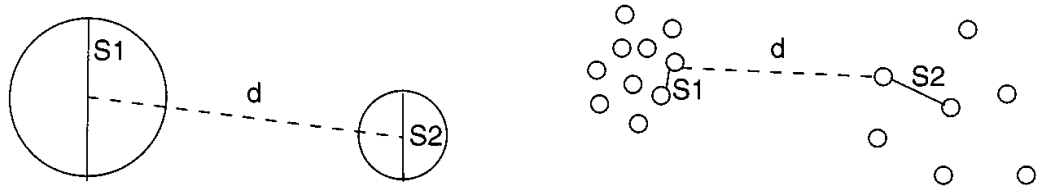


FIG. 6.3 – Critères de regroupement des catégories

Identifier les langues dominantes

Partant des catégories dominantes de vecteurs, nous pouvons déterminer les ancres sources vectorisées par ces vecteurs. Ces ancres sources sont considérées comme « importantes ».

Notion 8 ***Ancres sources importantes :*** *Les ancres sources importantes sont des ancres sources vectorisées par les vecteurs convergeant en catégories dominantes.*

Puis, nous déterminons les « documents importants » contenant ces ancres sources importantes.

Notion 9 ***Documents importants :*** *Les documents importants sont des documents contenant des ancres sources importantes.*

Les catégories de documents importants expriment la convergence des documents importants dans un espace multilingue (le site Web multilingue).

Pour identifier hypothétiquement les langues dominantes, nous proposons un algorithme qui se base sur une hypothèse supplémentaire. Cette hypothèse observe la langue principale de chaque catégorie de documents importants.

Hypothèse 5 *Dans un site Web multilingue, une langue est considérée comme dominante quand il existe une catégorie de documents importants dont la langue principale est cette langue.*

Donc, une langue est considérée comme dominante quand il y a au moins une catégorie de documents importants dont la plupart (définie par un seuil) est écrite en cette langue (autrement dit, cette langue est remarquée par la convergence des documents importants).

6.4 Expérimentations

Le module est implémenté pour permettre d'observer chaque étape du processus de fonctionnement. L'utilisateur peur également suivre les convergences des données qui sont montrées (en mode textuel sur l'écran ou dans les fichiers log), c'est à dire :

- le processus de classer des vecteurs en catégories (K-Means), ou l'adaptation des vecteurs référentiels du neurone gagnant et de ses neurones voisins (SOM),
- les changements des vecteurs de prototype,
- le processus de regroupement des catégories proches pour identifier les catégories dominantes.

Nous avons testé ce module sur les données de sites Web multilingues présentés dans le tableaux 5.4 (cf. chapitre 5). Le tableau 6.4 présente quelques caractéristiques structurelles obtenues par l'analyse des hyperliens dans ces sites.

Sites Web	**Liens inter-pages**	**Ancres sources distinctives**
IMF	480.121	20.815
UNDP	525.977	11.037
UNICEF	116.070	12.174
WTO	908.682	27.569
Ambassade de France en Espagne	59.850	3.892
Ambassade de France au Vietnam	20.642	1.247

Tab. 6.4 – Résultats d'analyse des hyperliens

Sur ce même jeux de sites nous avons expérimenté les deux types d'algorithmes K-Means et SOM [NZ04], [NZ05], [ZN05], [NZ06]. Les résultats obtenus sont très similaires et confirment la convergence des documents d'une même langue pour déterminer quelles langues (reconnues par le premier module d'Hyperling) sont dominantes. Ces résultats nous ont permis d'une part

de conserver notre architecture du module et d'autre part de renforcer la fiabilité de nos hypothèses proposées.

6.5 Conclusion

Les premiers résultats obtenus, sur des grands sites Web multilingues internationaux sont très favorables et ne montrent pas d'anomalie. La précision des catégories dominantes peut être mieux affinée en renforçant l'étape de post-traitement. Une approche de classification supervisée s'avère être assez appropriée.

Aussi, ces expérimentations nous ont permis de dresser quelques limitations et un certain nombre de points à étudier afin d'améliorer la performance d'Hyperling. Les outils proposés pour les différents modules sont de grande efficacité mais ils restent un peu limités sur le plan ergonomie d'interface. En effet ils ont été développés dans le cadre d'un prototype de recherche pour tester et valider nos hypothèses de regroupement structurel des hyperdocuments dans un site Web multilingue. Même le nombre d'itérations à effectuer par les méthodes de catégorisation peut être définit par défaut. Le temps d'exécution d'Hyperling est étroitement lié au volume du site, au nombre d'objets structurels et surtout au nombre d'itérations effectuées par les méthodes de catégorisation.

Nous tenons à rappeler le caractère expérimental de ce système. Les algorithmes de traitement, issus essentiellement des travaux en apprentissage automatique, demeurent approximatifs : Ils peuvent montrer la pertinence de la convergence des données statistiquement fréquents mais ils ne sont pas encore capables de démontrer ou de garantir la complétude et la certitude totales des résultats.

L'approche retenue est basée essentiellement sur le constat de la distribution statistique ne permettant que d'observer les objets qui sont très fréquents. Or, il serait possible de reprocher cette limitation à Hyperling. En effet l'élargissement du traitement, c'est-à-dire la non prise en compte de critères d'optimisation, pour couvrir l'ensemble des objets structurels, pourrait surmonter cette limitation. La prise d'une telle mesure serait au détriment du temps de calcul qui risquerait, selon la nature du site Web en question,

d'être plus coûteux. La question d'optimisation de la complexité des algorithmes utilisés devrait pouvoir clarifier l'ampleur d'un tel élargissement.

Finalement, l'ensemble de ces expérimentations et réalisations nous ont permis de mettre en avant l'importance des pré-traitement des données. Nous avons observé que dans un certain contexte l'algorithme de catégorisation n'est pas déterminant. Malgré que SOM et K-Means sont très différents nous avons obtenus des résultats très comparables mais aussi très satisfaisant. Ce fait, renforce vivement notre orientation qui consiste à donner davantage d'importance aux travaux de recherche en prétraitement des données d'une part et d'autre part à considérer l'importante valeur ajoutée que peut avoir l'information structurelle.

Quatrième partie

Conclusion générale

Synthèse et bilan de la thèse

Le multilinguisme sur le Web est un des événements les plus importants depuis la naissance de ce nouveau média d'information. La structure des ressources multilingues sur le Web représente un aspect particulier de la description de l'information. Différents travaux de recherches, dans plusieurs domaines, se sont intéressés à étudier et à exploiter les ressources multilingues. Ces travaux se sont focalisés, le plus souvent, sur les aspects linguistiques sans tenir compte suffisamment de l'importance et du potentiel informationnel que représente la structure de l'information.

Le premier objectif de cette thèse était axé sur la spécificité structurelle des hyperdocuments (des sites Web) multilingues. Nous avons étudié les spécificités des structures multilingues en précisant le rôle des ancres sources particulières (des ancres de changement de langue, des ancres partagées par plusieurs langues et des ancres de référence interlingues) et quelques phénomènes relatifs aux structures multilingues (des stratégies de changement de langue, des informations complémentaires, etc.). Dans ce contexte, nous considérons que la structure des hyperdocuments multilingues a été profondément analysée.

Notre deuxième objectif était de proposer un modèle de représentation de l'information structurelle en deux niveaux : le premier niveau représente la structure hypertextuelle du site Web ainsi que les structures hiérarchiques internes des documents, et le second niveau représente des relations entre des ancres sources. Nous avons également adopté un graphe constitué d'ancres sources pour mieux représenter des relations entre les documents de différentes langues dans la structure d'un hyperdocument multilingue. Nous avons observé qu'un graphe de documents Web n'est pas le mieux adapté pour déployer ces différents points.

Notre troisième objectif était d'élaborer une approche d'analyses structurelles d'extraction d'information. Ainsi nous avons développé le système « Hyperling » pour reconnaître les langues dominantes dans un site Web multilingue. Hyperling ne se limite pas à des conceptions traditionnelles du domaine de l'extraction d'information qui ont pour but d'extraire des motifs ou des types pré-spécifiés de données structurées. Hyperling est conçu selon des hypothèses « fortes » supposant la convergence des documents en fonction de leurs langues dans un site Web multilingue.

Le quatrième et dernier objectif de cette thèse était d'observer et de valider les limites de ces hypothèses par une série d'expérimentations effectuées sur plusieurs sites Web multilingues de volumes variées.

A travers la réalisation et le développement de ces objectifs, cette thèse nous a donné un bilan montrant l'originalité de notre approche basée sur les analyses structurelles dans le domaine d'extraction d'information multilingue. Nous pouvons résumé la contribution de nos travaux de recherche par les points suivants :

1. Le premier apport de cette thèse est la formalisation de l'information structurée et hypertextuelle du site Web concrétisée par la proposition d'un modèle exhaustif pour représenter l'hyperdocument.
2. Le deuxième apport réside dans l'élaboration, dans le cadre de la conception du système Hyperling, d'une méthode de vectorisation (numérisation) des contextes distributionnels des ancres sources et d'un algorithme d'interprétation des langues dominantes. Le système Hyperling intègre des mécanismes complémentaires optimisant son fonctionnement, tels que les modules de : réduction de la dimension des vecteurs obtenus (méthode Random Mapping [Kas98]), regroupement des catégories de vecteurs qui sont proches (méthodes proposées par Vesanto [VA00]) et de reconnaissance des langues naturelles dans les textes (dérivés des concepts n-grams [CT94]).
3. Le troisième apport de cette thèse est la validation des hypothèses portant sur la convergence « structurelle » des hyperdocuments issus de même langue dans un site Web multilingue.

Discussion

L'étude et l'analyse de la structure des sites Web multilingues nous ont permis d'entériner, à l'aide d'une approche de recherche expérimentale, les observations suivantes.

1. En tout état de cause, les sites Web multilingues représentent des organisations (logiques) d'information qui possèdent des structures spécifiques. Le point clé de ces structures multilingues se trouve dans les mécanismes, les méthodes et les moyens de changement de langue. Sur ce point de vue, les sites Web multilingues respectent, de plus en plus, des normes universelles (identifiables) dans l'organisation de leurs structures.

2. Chaque langue dans un site Web multilingue a une autonomie structurelle, elle entretien des relations avec d'autres langues qui sont traduites formellement par des ancres de changement de langue. La structure des documents dans chaque langue forme une structure de « communauté » monolingue dans une grande « communauté » multilingue du site Web. Les documents dans chaque communauté monolingue se relient très fortement.
3. Enfin, nous constatons que les documents complémentaires ainsi que les ancres sources de référence « interlingues » ou les ancres sources partagées entre plusieurs langues, etc. sont des phénomènes rencontrés très souvent dans les grands sites Web multilingues.

A partir de ces observations, nous avons élaboré un plateforme expérimental d'extraction d'information adoptant une approche d'analyse structurelle visant à déterminer les langues dominantes dans un site Web multilingue. Cette réalisation a été conduite comme suit :

1. Premièrement, regrouper les documents d'un site Web en catégories. Chaque catégorie devra, selon nos hypothèses, correspondre à une langue. Cette étape permet d'évaluer le caractère monolingue (un seul regroupement dominant) ou multilingue (plusieurs regroupements dominants) d'un site Web.
2. Deuxièmement, à partir des catégories retenues (ou langues détectées), nous cherchons à identifier, hypothétiquement, celles qui sont dominantes. Cette approche utilise des mesures quantitatives.

En somme, nos travaux de recherche ont permis la réalisation d'un système d'analyses statistico-structurelles (Hyperling). La qualité de fonctionnement de ce système dépend strictement de la qualité des données en entrée. Hyperling exige un grand volume des données, car comme nous l'avons signalé, Hyperling adopte un traitement statistique sur les données. Cependant, le volume des données ne pourra pas assurer leur qualité informationnelle qui est strictement dépendante de leur qualité structurelle. Aussi, nous rappelons que la notion des langues dominantes demeure très relative, et que nous ne prétendons pas d'avoir l'ambition de trouver une réponse absolue à ce point. Cependant, nous confirmons d'après nos expérimentations, que la détermination de certaines langues dominantes a un sens important dans la stratégie de construction des systèmes de recherche d'information ou d'extraction d'information multilingues.

Comme il s'agit d'une approche de recherche et développement expérimentale, les résultats obtenus jusqu'à lors valide et confirme nos hypothèses et nos réalisations. Bien évidement, nous n'excluons pas la possibilité (qui peut être non nulle) de rencontrer, dans un contexte particulier, un contre exemple (cas d'un site monolingue conçu par des structures de volumes équivalentes et fortement parallèles). Ceci ne peut être affirmé qu'après de nombreuses expérimentations

Quelques perspectives

Ce travail a rappelé encore une fois l'importance d'une approche de recherche expérimentale pour prouver et valider des hypothèses. Ainsi, et dans la perspective de cette recherche, nous souhaitons reprendre ces mêmes principes (traduits par le plateforme Hyperling) pour générer et tester de nouvelles hypothèses pouvant renforcer ou réfuter d'autres. Cependant, ceci exige d'une part qu'Hyperling soit fortement interactif et d'autre part de mener davantage d'expérimentations sur d'importants ressources de données (qui devraient être au moins équivalentes à celles traitées par l'équipe de Broder et Kumar [BK00], [KRR+00b] dans leurs recherches sur la structure macroscopique du Web).

Nous nous intéressons également à étudier la structure des sites Web monolingues ou des faux sites web multilingues (c'est-à-dire des sites monolingue utilisant des traducteurs automatiques pour simuler le multilinguisme). En effet l'analyse de ce dernier type de sites Web pose quelques sérieuses difficultés. D'une part ces sites disposent, généralement, des mécanismes de changement de langue (sachant qu'ils sont démunis des autres caractéristiques structurelles des sites Web multilingues), d'autre part les traducteurs automatiques ne peuvent pas encore assurer des traductions certaines et complètes des contenus en toutes les langues du site.

Finalement nous souhaitons étendre cette recherche pour savoir s'il existe d'autres critères structurels pouvant déterminer la langue la plus dominante, c'est-à-dire la langue mère (parmi les langues dominantes) dans un site Web multilingue. La réponse à cette question est susceptible de jouer un rôle pratique dans la sélection des stratégies de développement des systèmes multilingues dans les domaines de recherche d'information, d'extraction d'information ou d'autres.

Cinquième partie

Bibliographie

Bibliographie

[AB00] Fernando Aguiar and Michel Beigbeder. Des moteurs de recherche efficaces pour des systèmes hypertextes grâce aux contextes des noeuds. In *Colloque International : Technologies de l'Information et de la Communication dans les Enseignements d'ingénieurs et dans l'industrie (TICE'2000)*, 2000.

[ABD+95] J. Aberdeen, J. Burger, D. Day, L. Hirschman, P. Robinson, and M. Vilain. MITRE : Description of the alembic system used for MUC-6. In *Proceedings of the Sixth Message Understanding Conference (MUC-6)*. Morgan Kaufmann, California, 1995.

[ACT04] Bashir Ahmed, Sung-Hyuk Cha, and Charles Tappert. Language identification from text using n-gram based cumulative frequency addition. In *CSIS Research Day*, 2004.

[Ade98] Brad Adelberg. NoDoSE - a tool for semi-automatically extracting semi-structured data from text documents. In *SIGMOD Conference*, pages 283–294, 1998.

[Adr00] Mirna Adriani. Ambiguity problem in multilingual information retrieval. In *CLEF*, pages 156–165, 2000.

[AHB+95] E. R. Appelt, J.R. Hobbs, J. Bear, D. Israel, M. Kameyama, A. Kehler, D. Martin, K. Myers, and M. Tyson. SRI international FASTUS system : MUC-6 test results and analysis. In *Proceedings of the Sixth Message Understanding Conference (MUC-6)*. Morgan Kaufmann, California, 1995.

[AHG+99] S. Azzam, K. Humphreys, R. Gaizauskas, H. Cunningham, and Y. Wilks. Using a language independent domain model for multilingual information extraction. *Applied Artificial Intelligence*, 13(7), 1999. Special Issue on Multilinguality in the Software Industry : the AI Contribution (MULSAIC-97).

[AJB99] Reka Albert, Hawoong Jeong, and Albert-Lazlo Barabasi. The diameter of the World Wide Web. *CoRR*, cond-mat/9907038, 1999.

[AM99] Gustavo O. Arocena and Alberto O. Mendelzon. WebOQL : Restructuring documents, databases, and Webs. *TAPOS*, 5(3) :127–141, 1999.

[AMM97] Paolo Atzeni, Giansalvatore Mecca, and Paolo Merialdo. Semi-structured und structured data in the Web : Going back and forth. *SIGMOD Record*, 26(4) :16–23, 1997.

[App99] D. Appelt. An introduction to information extraction. *Artificial Intelligence Communications*, 12(3) :161–172, 1999.

[BA99] Albert-Laszlo Barabasi and Reka Albert. Emergence of scaling in random networks. *Science*, 286 :509–512, October 1999.

[Bar89] Jon Barwise. *The Situation in Logic*, volume 17 of *CSLI Lecture Notes*. Center for the Study of Language and Information Publications, 1989.

[Bat92] E. O. Batchelder. A learning experience : Training an artificial neural network to discriminate languages. Technical report, 1992.

[BB95] L. Bottou and Y. Bengio. Convergence properties of the K-means algorithm. *Tesauro, G., Touretzky, D., and Leen, T., editors, Advances in Neural Information Processing Systems 7. MIT Press, Cambridge, MA*, 1995.

[Bee88] K. R. Beesley. Language identifier : A computer program for automatic natural-language identification on on-line text. In *the 29th Annual Conference of the American Translators Association*, pages 47–54, 1988.

[BF03] Nicola Bertoldi and Marcello Federico. ITC-irst at CLEF 2003 : Monolingual, bilingual, and multilingual information retrieval. In *CLEF*, pages 140–151, 2003.

[BFF04] Romaric Besançon, Olivier Ferret, and Christian Fluhr. Integrating new languages in a multilingual search system based on a deep linguistic analysis. In *CLEF*, pages 83–89, 2004.

[BFG01] Robert Baumgartner, Sergio Flesca, and Georg Gottlob. Supervised wrapper generation with Lixto. In *VLDB*, pages 715–716, 2001.

[BFJ96] Justin Boyan, Dayne Freitag, and Thorsten Joachims. A machine learning architecture for optimizing Web search engine. In *Workshop on Internet-Based Information Systems (WAAAI'96)*, 1996.

[BH52] Y. Bar-Hillel. Semantic information and its measures. In *8ème Cybernetics - circular, causal and feedback mechanisms in biological and social systems (Cybernetics'52)*, 1952.

[BH98] Krishna Bharat and Monika R. Henzinger. Improved algorithms for topic distillation in a hyperlinked environment. In *Proceedings of SIGIR-98, 21st ACM International Conference on Research and Development in Information Retrieval*, pages 104–111, Melbourne, AU, 1998.

[BK00] Andrei Broder and Ravi Kumar. Graph structure in the Web. 2000.

[BKvH02] Jeen Broekstra, Arjohn Kampman, and Frank van Harmelen. Sesame : A generic architecture for storing and querying RDF and RDF Schema. In *International Semantic Web Conference*, pages 54–68, 2002.

[BM01] E. Bingham and H. Mannila. Random projection in dimensionality reduction : applications to image and text data. In *the seventh ACM SIGKDD International Conference on Knowledge discovery and Data mining*, 2001.

[Bra96] Tim Bray. Measuring the Web. *Computer Networks*, 28(7-11) :993–1005, 1996.

[Bre93] B. Breton. *Histoire de l'informatique.* La découverte, Paris, 1993.

[Bri98] Sergey Brin. Extracting patterns and relations from the World Wide Web. In *WebDB Workshop at 6th International Conference on Extending Database Technology, EDBT'98*, 1998.

[BRS92] Rodrigo A. Botafogo, Ehud Rivlin, and Ben Shneiderman. Structural analysis of hypertexts : Identifying hierarchies and useful metrics. *ACM Trans. Inf. Syst.*, 10(2) :142–180, 1992.

[BS98] Martin Braschler and Peter Schäuble. Multilingual information retrieval based on document alignment techniques. In *ECDL*, pages 183–197, 1998.

[Bus45] Vannevar Bush. As we may think. *The Atlantic Monthly*, 176 :101–108, Juillet 1945.

[BW04] K. Bontcheva and Y. Wilks. Automatic report generation from ontologies : the MIAKT approach. In *Nineth International Conference on Applications of Natural Language to Information Systems (NLDB'2004)*, 2004.

[Cal98] M. E. Califf. *Relational Learning Techniques for Natural Language Information Extraction*. PhD thesis, University of Texas at Austin, 1998.

[Car97] C. Cardie. Empirical methods in information extraction. *AI Magazine*, 18(4), 1997.

[CDK+98] Soumen Chakrabarti, Byron Dom, Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, Andrew Tomkins, David Gibson, and Jon Michael Kleinberg. Mining the Web's link structure. *IEEE Computer*, 32(8) :60–67, August 1998.

[CHJS94] Gavin Churcher, Judith Hayes, Stephen Johnson, and Clive Souter. Bigraph and trigraph models for language identification and character recognition. In *1994 AISB Workshop on Computational Linguistics for Speech and Handwriting Recognition*, 1994.

[CHS04] Philipp Cimiano, Siegfried Handschuh, and Steffen Staab. Towards the self-annotating Web. In *WWW '04 : Proceedings of the 13th international conference on World Wide Web*, pages 462–471, New York, NY, USA, 2004. ACM Press.

[CK97] S. Jeromy Carrière and Rick Kazman. WebQuery : Searching and visualizing the Web through connectivity. *Computer Networks*, 29(8-13) :1257–1267, 1997.

[CL96] J. Cowie and W. Lehnert. Information extraction. *Communications of the ACM*, 39(1) :80–91, 1996.

[CL01] Chia-Hui Chang and Shao-Chen Lui. IEPAD : information extraction based on pattern discovery. In *WWW*, pages 681–688, 2001.

[CLM00] Vincenza Carchiolo, Alessandro Longheu, and Michele Malgeri. Extracting logical schema from the web. In *PRICAI Workshop on Text and Web Mining*, pages 64–71, 2000.

[CM98] Valter Crescenzi and Giansalvatore Mecca. Grammars have exceptions. *Inf. Syst.*, 23(8) :539–565, 1998.

[CMBT02] H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan. GATE : A framework and graphical development environment for robust NLP tools and applications. In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics*, 2002.

[CMM01] Valter Crescenzi, Giansalvatore Mecca, and Paolo Merialdo. Automatic web information extraction in the ROADRUNNER system. In *ER (Workshops)*, pages 264–277, 2001.

[CR94] Vassilis Christophides and Antoine Rizk. Querying structured documents with hypertext links using OODBMS. In *ECHT*, pages 186–197, 1994.

[CS04] H. Cunningham and D. Scott. Introduction to the special issue on software architecture for language engineering. *Natural Language Engineering*, 2004. `http ://gate.ac.uk/sale/jnle-sale/intro/intro-main.pdf`.

[CT94] William B. Cavnar and John M. Trenkle. Ngram -based text categorization. In *In Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*, pages 161–175, 1994.

[CT98] S. Chen and J. Trenkle. An empirical study of smoothing techniques for language modeling. Technical report, Harvard University, 10 1998.

[Cun05] H. Cunningham. Information extraction, automatic. *Encyclopedia of Language and Linguistics, 2nd Edition*, 2005.

[Dam95] M. Damashek. Gauging similarity with n-grams : Language-independent categorization of text. *Science*, 267(10) :843–848, 1995.

[DC03] T. Declerck and C. Crispi. Multilingual linguistic modules for IE systems. In *Proceedings of Workshop on Information Extraction for Slavonic and other Central and Eastern European Languages (IESL'03)*, Borovets, Bulgaria, 2003.

[DCW03] A. Dingli, F. Ciravegna, and Y. Wilks. Automatic semantic annotation using unsupervised information extraction and integration. In *Workshop on Knowledge Markup and Semantic Annotation*, 2003.

[DD04] John Domingue and Martin Dzbor. Magpie : supporting browsing and navigation on the semantic web. In *IUI '04 : Proceedings of the 9th international conference on Intelligent user interface*, pages 191–197, New York, NY, USA, 2004. ACM Press.

[DEG+03] S. Dill, N. Eiron, D. Gibson, D. Gruhl, R. Guha, A. Jhingran, T. Kanungo, S. Rajagopalan, A. Tomkins, J.A. Tomlin, and J.Y. Zien. SemTag and Seeker : bootstrapping the Semantic Web via automated semantic annotation. In *Proceedings of the 12th International World Wide Web Conference*, pages 178–186. ACM Press, 2003.

[DeJ82] G. DeJong. An overview of the FRUMP system. In W. Lehnert and M.H. Ringle, editors, *Strategies for Natural Language Processing*, pages 149–176. Lawrence Erlbaum, 1982.

[DLM+05] Luca Dini, Doris Liebwald, Laurens Mommers, Wim Peters, Erich Peters, and Wim Voermans. Cross-lingual legal information retrieval using a WordNet architecture. In *ICAIL*, pages 163–167, 2005.

[DMS00] S. N. Dorogovtsev, J. F. Mendes, and A. N. Samukhin. Structure of growing networks with preferential linking. *Phys Rev Lett*, 85(21) :4633–4636, November 2000.

[Dun94] T. Dunning. Statistical identification of language. Technical Report Technical Report MCCS-94-273, Computing Research Laboratory, New Mexico State, 1994.

[ECJ+99] David W. Embley, Douglas M. Campbell, Y. S. Jiang, Stephen W. Liddle, Yiu-Kai Ng, Dallan Quass, and Randy D. Smith. Conceptual-model-based data extraction from multiple-record Web pages. *Data Knowl. Eng.*, 31(3) :227–251, 1999.

[Eik99] L. Eikvil. Information extraction from World Wide Web - a survey. Technical Report 945, Norweigan Computing Center, Oslo, Norway, 1999.

[EM81] Robert Estival and Jean Meyriat. La dialectique de l'écrit et du document. un effort de synthèse. *Schéma et schématisation*, pages 82–91, 1981.

[ER59] P. Erdos and A. Renyi. On random graphs. *Publicationes Mathematicae*, 6 :290–297, 1959.

[ERC+00] K. Efe, V. Raghavan, C. Chu, A. Broadwater, L. Bolelli, and S. Ertekin. The shape of the Web and its implications for searching the Web. 2000.

[Etz96] Oren Etzioni. The World-Wide Web : Quagmire or gold mine ? *Commun. ACM*, 39(11) :65–68, 1996.

[Fel98] Christiane Fellbaum. *WordNet - An Electronic Lexical Database*. 1998.

[FLGC02] Gary William Flake, Steve Lawrence, C. Lee Giles, and Frans Coetzee. Self-organization and identification of Web communities. *IEEE Computer*, 35(3) :66–71, 2002.

[Flu05] Christian Fluhr. Systèmes multilingue recherche interlingue. In *Conférence Internationale sur le Document Electronique (CiDE.8)*, 2005.

[FMSDW93] Michael Fuller, Eric Mackie, Ron Sacks-Davis, and Ross Wilkinson. Structured answers for a large structured document collection. In *16ème ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'93)*, pages 204–213, 1993.

[Für99] Johannes Fürnkranz. Exploiting structural information for text classification on the WWW. In *IDA*, pages 487–498, 1999.

[Fre98] D. Freitag. Information extraction from HTML : Application of a general machine learning approach. In *Proceesings of Fifteenth National Conference on Artificial Intelligence (AAAI-98)*, 1998.

[Gig95] Emmanuel Giguet. Multilingual sentence categorization according to language. *CoRR*, cmp-lg/9502039, 1995.

[Gig98] Emmanuel Giguet. *Méthode pour l'analyse automatique de structures formelles sur documents multilingues*. PhD thesis, Université de Caen, France, 1998.

[GKR98] David Gibson, Jon M. Kleinberg, and Prabhakar Raghavan. Inferring Web communities from link topology. In *Hypertext*, pages 225–234, 1998.

[GL02] Jean-Loup Guillaume and Matthieu Latapy. The Web graph : an overview. In *Quatrièmes Rencontres francophones sur les aspects algorithmiques des télécommunications (ALGOTEL'02)*, 2002.

[GL03a] Jean-Loup Guillaume and Matthieu Latapy. Modèles pour les topologies réalistes. In *Cinquièmes Rencontres francophones sur les aspects algorithmiques des télécommunications (ALGOTEL'03)*, 2003.

[GL03b] Jean-Loup Guillaume and Matthieu Latapy. Topologie d'Internet et du Web : mesure et modélisation. In *Premier colloque Mesures de l'Internet*, 2003.

[GL04a] Jean-Loup Guillaume and Matthieu Latapy. Bipartite graphs as models of complex networks. In *CAAN*, pages 127–139, 2004.

[GL04b] Jean-Loup Guillaume and Matthieu Latapy. Bipartite structure of all complex networks. *Inf. Process. Lett.*, 90(5) :215–221, 2004.

[Gér02] Mathias Géry. *Indexation et interrogation de chemins de lecture en contexte pour la Recherche d'Information Structurée sur le Web*. PhD thesis, Université Joseph Fourier - Grenoble I, France, 2002.

[Gri95] R. Grishman. TIPSTER architecture design document version 2.0 (tinman architecture). Technical report, Department of Computer Science, New York University, 1995.

[Gri01] R. Grishman. Adaptive information extraction and sublanguage analysis. In *Proceedings of Workshop on Adaptive Text Extraction and Mining at Seventeenth International Joint Conference on Artificial Intelligence*, Seattle, USA, 2001.

[Gru93] T. R. Gruber. A translation approach to portable ontologies. *Knowledge Acquisition*, 5(2) :199–220, 1993.

[GS93] R. Grishman and J. Sterling. Description of the Proteus system as used for MUC-5. In *Proceedings of the Fifth Message Understanding Conference (MUC-5)*, pages 181–194. Morgan Kaufmann, California, 1993.

[GS96] R. Grishman and B. Sundheim. Message Understanding Conference - 6 : A brief history. In *Proceedings of the 16th International Conference on Computational Linguistics*, Copenhagen, June 1996.

[GS99] X. Gao and L. Sterling. Autowrapper : automatic wrapper generation for multiple online services. In *The Asia Pacic Web Conference*, 1999.

[GW98] R. Gaizauskas and Y. Wilks. Information extraction : Beyond document retrieval. *Journal of Documentation*, 54(1) :70–105, 1998.

[GWH+95] R. Gaizauskas, T. Wakao, K Humphreys, H. Cunningham, and Y. Wilks. Description of the lasie system as used for muc-6. In *Proceedings of the Sixth Message Understanding Conference (MUC-6)*. Morgan Kaufmann, California, 1995.

[GY96] Georges Gardarin and Shim Yoon. Hyweb : Un système d'interrogation orienté objet pour le web. In *BDA*, pages 205–224, 1996.

[Hab04] Benjamin Habegger. *Extraction d'informations à partir du Web*. PhD thesis, Université de Nantes, 2004.

[HAT+92] J.R. Hobbs, D. Appelt, M. Tyson, J. Bear, and D. Israel. Description of the FASTUS system as used for MUC-4. In *Proceedings of the Fourth Message Understanding Conference MUC-4*, pages 268–275. Morgan Kaufmann, California, 1992.

[Hay93] Judith Hayes. Language recognition using two-and three-letter clusters. Technical report, School of Computer Studies, University of Leeds, 1993.

[HD98] C. N. Hsu and M. T. Dung. Generating finite-state transducers for semistructured data extraction from the Web. *Information Systems 23(8), Special Issue on Semistructured Data*, 1998.

[HGMN+97] Joachim Hammer, Hector Garcia-Molina, Svetlozar Nestorov, Ramana Yerneni, Markus M. Breunig, and Vasilis Vassalos. Template-based wrappers in the TSIMMIS system. In *SIGMOD Conference*, pages 532–535, 1997.

[HM01] Md Maruf Hasan and Yuji Matsumoto. Multilingual document alignment - a study with chinese and japanese. In *NLPRS*, pages 617–623, 2001.

[Hob91] J.R. Hobbs. Description of the TACITUS system as used for MUC-3. In *Proceedings of the Third Message Understanding Conference MUC-3*, pages 200–206. Morgan Kaufmann, California, 1991.

[HSC02] S. Handschuh, S. Staab, and F. Ciravegna. S-CREAM - semi-automatic creation of metadata. In *13th International Conference on Knowledge Engineering and Knowledge Management (EKAW02)*, pages 358–372, Siguenza, Spain, 2002.

[Huf95] S. Huffman. Learning information extraction patterns from examples. *Workshop on new approaches to learning for natural language processing (IJCAI-95)*, pages 127–142, 1995.

[Jak63] Roman Jakobson. *Essais de linguistique générale*. 1963.

[jHtY97] Jane Yung jen Hsu and Wen tau Yih. Template-based information mining from HTML documents. In *AAAI/IAAI*, pages 256–262, 1997.

[JR90] P. S. Jacobs and L. F. Rau. Scisor : Extracting information from on-line news. *Communications of the ACM*, 33(11) :88–97, 1990.

[Kas98] S. Kaski. Dimensionality reduction by random mapping : Fast similarity computation for clustering. In *International Joint Conference on Neural Networks (IJCNN'98)*, 1998.

[KB00] R. Kosala and H. Blockeel. Web mining research : A survey. *SIGKDD Explorations*, 2(1) :1–15, 2000.

[Kes63] M. M. Kessler. Bibliographic coupling between scientific papers. *American Documentation*, 14 :10–25, Janvier 1963.

[KH01] P. Kogut and W. Holmes. AeroDAML : Applying information extraction to Generate DAML Annotations from Web pages.

In *First International Conference on Knowledge Capture (K-CAP 2001), Workshop on Knowledge Markup and Semantic Annotation*, Victoria, B.C., 2001.

[KKR+99] Jon M. Kleinberg, Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, and Andrew Tomkins. The Web as a graph : Measurements, models, and methods. In *COCOON*, pages 1–17, 1999.

[KL01] J. Kleinberg and S. Lawrence. The structure of the Web. *Science*, 294 :1849–1850, 2001.

[Kle99a] Jon M. Kleinberg. Authoritative sources in a hyperlinked environment. *J. ACM*, 46(5) :604–632, 1999.

[Kle99b] Jon M. Kleinberg. Hubs, authorities, and communities. *ACM Comput. Surv.*, 31(4es) :5, 1999.

[KM95] J. Kim and D. Moldovan. Acquisition of linguistic patterns for knowledge-based information extraction. *IEEE Transactiops on Knowledge and Data Engineering*, 7(5) :713–724, 1995.

[KO80] Catherine Kerbrat-Orecchioni. *L'énonciation de la subjectivité dans le langage*. 1980.

[Koh95] T. Kohonen. Self-organizing maps. *Berlin, Heidelberg, New-York : Springer*, 1995.

[KRR+00a] Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, D. Sivakumar, Andrew Tomkins, and Eli Upfal. Random graph models for the Web graph. In *FOCS*, pages 57–65, 2000.

[KRR+00b] Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, D. Sivakumar, Andrew Tomkins, and Eli Upfal. The Web as a graph. In *PODS*, pages 1–10, 2000.

[Kru95] G.R. Krupka. Description of the SRA system as used for MUC-6. In *Proceedings of the Sixth Message Understanding Conference (MUC-6)*, pages 221–236. Morgan Kaufmann, California, 1995.

[KSG04] Satoshi Sekine Kiyoshi Sudo and Ralph Grishman. Cross-lingual information extraction system evaluation. In *In Proceedings of COLING*, 2004.

[KSS97] Henry Kautz, Bart Selman, and Mehul Shah. Referral Web : combining social networks and collaborative filtering. *Commun. ACM*, 40(3) :63–65, March 1997.

[Kus97] Nicholas Kushmerick. *Wrapper induction for information extraction*. PhD thesis, 1997. Chairperson-Daniel S. Weld.

[Lan01] Stefan Langer. Natural languages on the Word Wide Web. *Bulag. Revue annuelle. Presses Universitaires Franc-Comtoises*, pages 89–100, 2001.

[LCD+05] M. Lyckova, I. Charon, L. Denoeud, O. Hudry, and A. Lobstein. Optimisation et modélisation du graphe du Web. Rapport sur le projet WEB-MOPT, Ecole Natinale Supérieure des Télécommunications, août 2005.

[LG86] S.L. Lytinen and A. Gershman. Atrans : Automatic processing of money transfer messages. In *Proceedings of the Fifth National Conference on Artificial Intelligence (AAAI-86)*, pages 1089–1093, 1986.

[LG94] L. Lamel and J. Gauvain. Language identification using phone-based acoustic likelihoods. In *the IEEE International Conference on Accoustics, Speech, and Signal Processing (ICA94)*, 1994.

[Lin95] D. Lin. Description of the PIE system as used for MUC-6. In *Proceedings of the Sixth Message Understanding Conference (MUC-6)*, pages 113–126, San Francisco, 1995. Morgan Kaufmann, California.

[LO99] B. F. Lavoie and E. T. O'Neill. How "World Wide" is the Web ? trend in internationalization of Web sites. *Annual Review of OCLC Research*, 1999.

[LPH00] Ling Liu, Calton Pu, and Wei Han. XWRAP : An XML-enabled wrapper construction system for Web information sources. In *ICDE*, pages 611–621, 2000.

[LRNdS02] Alberto H. F. Laender, Berthier A. Ribeiro-Neto, and Altigran Soares da Silva. DEByE - data extraction by example. *Data Knowl. Eng.*, 40(2) :121–154, 2002.

[LRNdST02] A.H.F. Laender, B.A. Ribeiro-Neto, A.S. da Silva, and J.S. Teixeira. A brief survey of Web data extraction tools. *SIGMOD Record*, 31(2) :84–93, 2002.

[Mac67] J. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297, 1967.

[Mas04] Philipp Masche. Multilingual information extraction. Master's thesis, Master's Thesis, University of Helsinki, Faculty of Science, Department of Computer Science, April 2004.

[May03] D. Maynard. Multi-source and multilingual information extraction. *Expert Update*, 2003.

[MdMLD02] Adilson E. Motter, Alessandro P. S. de Moura, Ying-Cheng Lai, and Partha Dasgupta. Topology of the conceptual network of language. *Physical Review E*, 65(065102), 2002.

[MMK98] I. Muslea, S. Minton, and Craig Knoblock. Stalker : Learning extraction rules for semistructured. In *Proceedings of AAAI-98 Workshop on AI and Information Integration*, 1998.

[MMK99] Ion Muslea, Steven Minton, and Craig A. Knoblock. Active learning for hierarchical wrapper induction. In *AAAI/IAAI*, page 975, 1999.

[MMM96] Alberto O. Mendelzon, George A. Mihaila, and Tova Milo. Querying the World Wide Web. In *PDIS*, pages 80–91, 1996.

[MTB+03] D. Maynard, V. Tablan, K. Bontcheva, H. Cunningham, and Y. Wilks. MUSE : a multi-source entity recognition system. *Submitted to Computers and the Humanities*, 2003.

[Mus65] S. Mustonen. Multiple discriminant analysis in linguistic problems. *Statistical Methods in Linguistics*, 4, 1965.

[Nel72] Ted Nelson. As we will think. In *Proceedings of Online 72 Conference*, Uxbridge, England, 1972. Brunel University.

[New87] P. Newman. Foreign language identification : First step in the translation process. In *the 28th Annual Conference of the American Translators Accociation*, pages 509–516, 1987.

[Ngu04] Dang Tuan Nguyen. Nouvelle méthode syntagmatique de vectorisation appliquée au Self-organizing map des textes vietnamiens. In *POSTER, RECITAL (Rencontre des Etudiants Chercheurs en Informatique pour le Traitement Automatique des Langues)*, 2004.

[NZ04] Dang Tuan Nguyen and Khaldoun Zreik. Multilingual hyperdocument recognition : a document mining approach. In *International Conference on Information and Communication Technologies : from Theory to Applications (ICTTA'04)*, 2004.

[NZ05] Dang Tuan Nguyen and Khaldoun Zreik. Hyperling : Système de reconnaissance et de classification des hyperdocuments multilingues. In *International Conference in Computer Science « Research, Innovation and Vision of the Future» (RIVF'05)*, 2005.

[NZ06] Dang Tuan Nguyen and Khaldoun Zreik. Multilingual Web documents : the system Hyperling. In *International Conference on Information and Communication Technologies : from Theory to Applications (ICTTA'06)*, 2006.

[OLB03] E. T. O'Neill, B. F. Lavoie, and R. Bennett. Trends in the evolution of the public Web. *D-Lib Magazine*, Volume 9 Number 4, April 2003.

[PKK+03] Borislav Popov, Atanas Kiryakov, Angel Kirilov, Dimitar Manov, Damyan Ognyanoff, and Miroslav Goranov. KIM - semantic annotation platform. In *International Semantic Web Conference*, pages 834–849, 2003.

[PKK+04] B. Popov, A. Kiryakov, A. Kirilov, D. Manov, D. Ognyanoff, and M. Goranov. KIM - semantic annotation platform. *Natural Language Engineering*, 2004.

[Poi99] Thierry Poibeau. Mixing technologies for intelligent information extraction. In *Actes du workshop Intelligent Information Integration (III), 16th International Joint Conference on Artificial Intelligence (IJCAI'99)*, 1999.

[PP04] Muntsa Padro and Lluis Padro. Comparing methods for language identification. *Procesamiento del Lenguaje Natural*, 33 :155–162, 2004.

[PPR96] Peter Pirolli, James Pitkow, and Ramana Rao. Silk from a Sow's Ear : Extracting usable structures from the Web. In *Proc. ACM Conf. Human Factors in Computing Systems, CHI*, pages 118–125, New York, NY, 1996. ACM Press.

[PS03] Fuchun Peng and Dale Schuurmans. Combining Naive Bayes and n-gram language models for text classification. In *ECIR*, pages 335–350, 2003.

[RH05] Lawrence Reeve and Hyoil Han. Survey of semantic annotation platforms. In *SAC*, pages 1634–1638, 2005.

[Ria98] Farshad Riahi. Elaboration automatique d'une base de données à partir d'informations semi-structurées issues du Web. In *INFORSID*, pages 327–341, 1998.

[Ril93] E. Riloff. Automatically constructing a dictionary for information extraction tasks. *Proceedings of the Eleventh Annual Conference on Artificial Intelligence*, pages 811–816, 1993.

[RM00] Davood Rafiei and Alberto O. Mendelzon. What is this page known for ? Computing web page reputations. *Computer Networks*, 33(1-6) :823–835, 2000.

[RNLdS99] Berthier A. Ribeiro-Neto, Alberto H. F. Laender, and Altigran Soares da Silva. Extracting semi-structured data through examples. In *CIKM*, pages 94–101, 1999.

[RSY02] Ellen Riloff, Charles Schafer, and David Yarowsky. Inducing information extraction systems for new languages via cross-language projection. In *COLING*, 2002.

[RTPG04] Filippo Ricca, Paolo Tonella, Emanuele Pianta, and Christian Girardi. Experimental results on the alignment of multilingual web sites. In *CSMR*, pages 288–295, 2004.

[RtY02] Dan Roth and Wen tau Yih. Probabilistic reasoning for entity and relation recognition. In *Proceedings of the 19th international conference on Computational linguistics*, pages 1–7, Morristown, NJ, USA, 2002. Association for Computational Linguistics.

[SA99] Arnaud Sahuguet and Fabien Azavant. Web ecology : Recycling HTML pages as XML documents using W4F. In *WebDB (Informal Proceedings)*, pages 31–36, 1999.

[SAS96] Gerard Salton, James Allan, and Amit Singhal. Automatic text decomposition and structuring. *Inf. Process. Manage.*, 32(2) :127–138, 1996.

[Sch00] B. F. Schloman. Breaking through the foreign language barrier : Resources on the web. *Online Journal of Issues in Nursing*, 2000.

[SFAL95] Stephen Soderland, David Fisher, Jonathan Aseltine, and Wendy G. Lehnert. Issues in inductive learning of domain-specific text extraction rules. In *Learning for Natural Language Processing*, pages 290–301, 1995.

[SHSM03] Jun Suzuki, Tsutomu Hirao, Yutaka Sasaki, and Eisaku Maeda. Hierarchical directed acyclic graph kernel : Methods for structured natural language data. In *ACL*, pages 32–39, 2003.

[Sma74] Henry Small. Co-citation in the scientific literature : A new measure of the relationship between two documents. *Essays of an Information Scientist*, 2 :28–31, Février 1974.

[SO00] Ruth Sperer and Douglas W. Oard. Structured translation for cross-language information retrieval. In *SIGIR*, pages 120–127, 2000.

[Sod97] Stephen Soderland. Learning to extract text-based information from the World Wide Web. In *KDD*, pages 251–254, 1997.

[Sod99] S. Soderland. Learning information extraction rules for semi-structures and free text. *Machine Learning, Special Issue on Natural Language Learning*, 34(1-3), Feb. 1999.

[Spe97] Ellen Spertus. ParaSite : Mining structural information on the Web. *Computer Networks*, 29(8-13) :1205–1215, 1997.

[Str01] S. H. Strogatz. Exploring complex networks. *Nature*, 410(6825) :268–276, March 2001.

[SYC01] Heekyoung Seo, Jaeyoung Yang, and Joongmin Choi. Knowledge-based wrapper generation by using XML. In *In IJCAI-2001 Workshop on Adaptive Text Extraction and Mining*, 2001.

[The01] Mike Thelwall. Extracting macroscopic information from web links. *JASIST*, 52(13) :1157–1168, 2001.

[UM05] A. Ultsch and F. Moerchen. ESOM-Maps : tools for clustering, visualization, and classification with Emergent SOM. Technical Report 46, Technical Report Dept. of Mathematics and Computer Science, University of Marburg, Germany, 2005.

[UN90] Yoshio Ueda and Seiichi Nakagawa. Prediction for phoneme/syllable/word-category and identification of language using hmm. In *the 1990 International Conference on Spoken Language Processing*, 1990.

[VA00] J. Vesanto and E. Alhoniemi. Clustering of the Self-Organizing Map. In *Student Member, IEEE*, 2000.

[VVMD+02] Maria Vargas-Vera, Enrico Motta, John Domingue, Mattia Lanzoni, Arthur Stutt, and Fabio Ciravegna. MnM : Ontology driven semi-automatic and automatic support for semantic markup. In *EKAW '02 : Proceedings of the 13th International Conference on Knowledge Engineering and Knowledge Management. Ontologies and the Semantic Web*, pages 379–391, London, UK, 2002. Springer-Verlag.

[Wei95] R. Weischedel. Description of the PLUM system as used for MUC-6. In *Proceedings of the Sixth Message Understanding Conference (MUC-6)*, pages 55–70, San Francisco, 1995. Morgan Kaufmann, California.

[WF94] Stanley Wasserman and Katherine Faust. *Social Network Analysis : Methods and Applications*. Cambridge University Press, 1994.

[WM89] H. D. White and K. W. McCain. Bibliometrics. *Annual Review of Information Science Technology*, 24 :119–165, 1989.

[WS98] Duncan J. Watts and Steven Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393 :440–442, June 1998.

[XNS00] F. Xu, K. Netter, and H. Stenzhorn. MIETTA - a framework for uniform and multilingual access to structured database and Web information. In *IRAL2000*, 2000.

[YG98] Roman Yangarber and Ralph Grishman. NYU : Description of the Proteus/PET system as used for MUC-7. In *Proceedings of the Seventh Message Understanding Conference (MUC-7)*, 1998.

[Zie91] D. V. Ziegler. *The automatic identification of languages using linguistic recognition signals*. PhD thesis, SUNY Buffalo, 1991.

[ZN05] Khaldoun Zreik and Dang Tuan Nguyen. Catégorisation des hyperdocuments multilingues : système Hyperling. In *Conférence Internationale sur le Document Electronique (CiDE.8)*, 2005.

[ZS94] Marc A. Zissman and Elliot Singer. Automatic language identification of telephone speech messages using phoneme recognition and n-gram modeling. In *the IEEE International Conference on Accoustics, Speech, and Signal Processing (ICA94)*, 1994.

Sixième partie

Annexes

A. Liste des sites Web étudiés

Organisation	Nom complet	URL
EUROPA	Gateway to the European Union	http ://europa.eu/
FAO	Food and Agriculture Organization of the United Nations	http ://www.fao.org/
ILO	International Labour Organization	http ://www.ilo.org/
IMF	International Monetary Fund	http ://www.imf.org/
UN	United Nations	http ://www.un.org/
UNDP	United Nations Development Program	http ://www.undp.org/
UNFPA	United Nations Population Fund	http ://www.unfpa.org/
UNICEF	United Nations Children's Fund	http ://www.unicef.org/
WB	World Bank	http ://www.worldbank.org/
WTO	World Trade Organization	http ://www.wto.org/
	Ambassade de France en Espagne	http ://www.ambafrance-es.org/
	Ambassade de France au Vietnam	http ://www.ambafrance-vn.org/

B. Formats du fichier des vecteurs de données

```
# comment
#
% n
% m
% s1          s2                ..          sm
% var_name1 var_name2           ..          var_namem
x11          x12               ..          x1m
x21          x22               ..          x2m
.            .                 .           .
.            .                 .           .
xn1          xn2               ..          xnm
```

```
Notes :

n      Nombre de vecteurs.
m      Nombre de colonnes.
si     Type du colonne: 1 pour la valeur, 0 pour l'ignorance.
var_namei    Nom du i-ème colonne.
xij    Valeur du élément (ligne i, colonne j).
```

C. Formats du fichier des vecteurs de prototype

```
% k l
% m
w001          w002          ..     w00m
w011          w012          ..     w01m
w021          w022          ..     w02m
.             .             ..      .
w0(l-1)1      w0(l-1)2      ..     w0(l-1)m
w101          w102          ..     w10m
w111          w112          ..     w11m
.             .             ..      .
w1(l-1)1      w1(l-1)2      ..     w1(l-1)m
.             .             ..      .
.             .             ..      .
w(k-1)(l-1)1      w(k-1)(l-1)2      ..     w(k-1)(l-1)m


Notes:

k      Nombre de lignes.
l      Nombre de colonnes.
m      Nombre de dimensions du vecteur de prototype.
wijh   h-ème valeur du vecteur de prototype (ligne i, colonne j).
```

D. Exemple (extrait) du processus de convergence des vecteurs

Dans cette annexe, des exemples du processus de convergence des vecteurs, obtenus par K-means et SOM sur les données du site Web de l'Ambassade de France au Vietnam, sont présentés comme.

k-means.log

```
Cycle d'apprentissage: 1
        Vecteur 1       Distance minimum: 0.347676332397685     Centre: 3-2
        Vecteur 2       Distance minimum: 0.346860783807277     Centre: 2-5
        Vecteur 3       Distance minimum: 0.355660839750708     Centre: 4-4
        Vecteur 4       Distance minimum: 0.35798726923873      Centre: 5-3
        Vecteur 5       Distance minimum: 0.3474837714329 35    Centre: 1-1
        Vecteur 6       Distance minimum: 0.347483771429034     Centre: 5-3
        Vecteur 7       Distance minimum: 0.347483771488094     Centre: 4-5
        Vecteur 8       Distance minimum: 0.342962140021731     Centre: 2-5
        Vecteur 9       Distance minimum: 0.34686078389959      Centre: 2-2
        Vecteur 10      Distance minimum: 0.707106779982256     Centre: 4-3
        Vecteur 11      Distance minimum: 0.362598354435532     Centre: 3-2
        Vecteur 12      Distance minimum: 0.362598354517415     Centre: 1-3
        Vecteur 13      Distance minimum: 0.362598354578344     Centre: 5-3
        Vecteur 14      Distance minimum: 0.356963797763598     Centre: 4-4
        Vecteur 15      Distance minimum: 0.354435916829769     Centre: 2-5
        Vecteur 16      Distance minimum: 0.35443591670215      Centre: 4-4
        Vecteur 17      Distance minimum: 0.355540541325343     Centre: 5-5
        Vecteur 18      Distance minimum: 0.999999999003662     Centre: 5-1
        Vecteur 19      Distance minimum: 0.35554054126999      Centre: 3-1
        Vecteur 20      Distance minimum: 0.354435916772259     Centre: 1-3
        Vecteur 21      Distance minimum: 0.355540541318327     Centre: 1-3
        Vecteur 22      Distance minimum: 0.35683501135189      Centre: 1-3
        Vecteur 23      Distance minimum: 0.355540541159192     Centre: 2-2
        Vecteur 24      Distance minimum: 0.355540541275586     Centre: 2-5
        Vecteur 25      Distance minimum: 0.362598354605356     Centre: 2-1
        Vecteur 26      Distance minimum: 0.342962140141487     Centre: 4-3
        Vecteur 27      Distance minimum: 0.342962140124159     Centre: 1-2
        Vecteur 28      Distance minimum: 0.35564346435696      Centre: 2-1
        Vecteur 29      Distance minimum: 0.349641858194428     Centre: 2-2
        Vecteur 30      Distance minimum: 0.354496047242075     Centre: 5-2
        Vecteur 31      Distance minimum: 0.342962139949884     Centre: 2-1
        Vecteur 32      Distance minimum: 0.342962140161199     Centre: 2-1
        Vecteur 33      Distance minimum: 0.999999999026611     Centre: 5-3
        Vecteur 34      Distance minimum: 0.342962140103016     Centre: 3-1
        Vecteur 35      Distance minimum: 0.342962139996403     Centre: 2-5
        Vecteur 36      Distance minimum: 0.355643464359204     Centre: 5-2
        Vecteur 37      Distance minimum: 0.355643464240258     Centre: 4-4
        Vecteur 38      Distance minimum: 0.35564346428202      Centre: 3-1
        Vecteur 39      Distance minimum: 0.36259835459966      Centre: 4-4
        Vecteur 40      Distance minimum: 0.999999999031128     Centre: 3-3
        Vecteur 41      Distance minimum: 0.350809192185725     Centre: 5-4
        Vecteur 42      Distance minimum: 0.350809192280347     Centre: 5-2
        Vecteur 43      Distance minimum: 0.355540541174042     Centre: 2-2
        Vecteur 44      Distance minimum: 0.355540541161368     Centre: 4-4
        Vecteur 45      Distance minimum: 0.999999999019684     Centre: 2-1
        Vecteur 46      Distance minimum: 0.339383836514559     Centre: 1-3
        Vecteur 47      Distance minimum: 0.355540541354868     Centre: 2-5
        Vecteur 48      Distance minimum: 0.355540541176842     Centre: 4-4
        Vecteur 49      Distance minimum: 0.354538018183663     Centre: 4-4
        Vecteur 50      Distance minimum: 0.355540541213556     Centre: 4-4
        Vecteur 51      Distance minimum: 0.355540541379621     Centre: 4-3
        Vecteur 52      Distance minimum: 0.355540541251075     Centre: 5-2
        Vecteur 53      Distance minimum: 0.355540541365485     Centre: 4-4
        Vecteur 54      Distance minimum: 0.999999999012238     Centre: 4-4
        Vecteur 55      Distance minimum: 0.355540541359758     Centre: 2-1
        Vecteur 56      Distance minimum: 0.355540541370622     Centre: 3-1
        Vecteur 57      Distance minimum: 0.355540541079053     Centre: 2-2
        Vecteur 58      Distance minimum: 0.355540541323854     Centre: 4-3
        Vecteur 59      Distance minimum: 0.355540541355157     Centre: 1-4
        Vecteur 60      Distance minimum: 0.355643464282303     Centre: 4-4
        Vecteur 61      Distance minimum: 0.355643464301626     Centre: 2-5
        Vecteur 62      Distance minimum: 0.355643464155875     Centre: 2-2
        Vecteur 63      Distance minimum: 0.355643464351353     Centre: 4-4
        Vecteur 64      Distance minimum: 0.999999999017273     Centre: 2-1
        Vecteur 65      Distance minimum: 0.342962140198766     Centre: 2-1
        Vecteur 66      Distance minimum: 0.342962140374607     Centre: 4-1
        Vecteur 67      Distance minimum: 0.362598354537697     Centre: 4-5
        Vecteur 68      Distance minimum: 0.355540541291268     Centre: 3-1
        Vecteur 69      Distance minimum: 0.342962139952763     Centre: 4-5
        Vecteur 70      Distance minimum: 0.362598354451338     Centre: 4-5
        Vecteur 71      Distance minimum: 0.342962140064243     Centre: 4-5
        Vecteur 72      Distance minimum: 0.362598354582601     Centre: 1-2
        Vecteur 73      Distance minimum: 0.355643464312581     Centre: 5-2
        Vecteur 74      Distance minimum: 0.999999999167358     Centre: 4-3
        Vecteur 75      Distance minimum: 0.342962140173891     Centre: 4-3
```

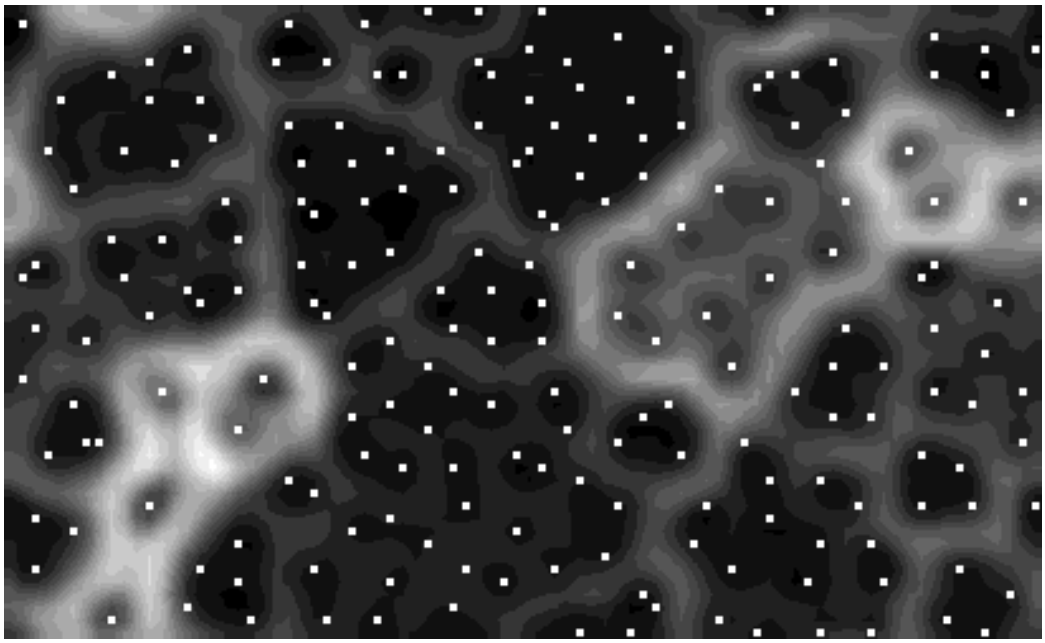
som.log

```
Traitement: 1 eme etape d'apprentissage ...
        Vecteur:13      Neurone gagnant: (2,3)  Dmin = 801.382840437023
Traitement: 2 eme etape d'apprentissage ...
        Vecteur:32      Neurone gagnant: (2,3)  Dmin = 799.127228940913
Traitement: 3 eme etape d'apprentissage ...
        Vecteur:19      Neurone gagnant: (2,3)  Dmin = 794.604649724756
Traitement: 4 eme etape d'apprentissage ...
        Vecteur:147     Neurone gagnant: (2,3)  Dmin = 787.6614082009
Traitement: 5 eme etape d'apprentissage ...
        Vecteur:190     Neurone gagnant: (2,3)  Dmin = 778.399045644949
Traitement: 6 eme etape d'apprentissage ...
        Vecteur:114     Neurone gagnant: (2,3)  Dmin = 766.957712408136
Traitement: 7 eme etape d'apprentissage ...
        Vecteur:5       Neurone gagnant: (2,3)  Dmin = 753.505824111861
Traitement: 8 eme etape d'apprentissage ...
        Vecteur:116     Neurone gagnant: (2,3)  Dmin = 737.947720817965
Traitement: 9 eme etape d'apprentissage ...
        Vecteur:77      Neurone gagnant: (2,3)  Dmin = 720.66300062898
Traitement: 10 eme etape d'apprentissage ...
        Vecteur:66      Neurone gagnant: (2,3)  Dmin = 701.535048741628
Traitement: 11 eme etape d'apprentissage ...
        Vecteur:179     Neurone gagnant: (2,3)  Dmin = 680.905073966206
Traitement: 12 eme etape d'apprentissage ...
        Vecteur:11      Neurone gagnant: (2,3)  Dmin = 658.813479026299
Traitement: 13 eme etape d'apprentissage ...
        Vecteur:121     Neurone gagnant: (2,3)  Dmin = 635.556684686432
Traitement: 14 eme etape d'apprentissage ...
        Vecteur:92      Neurone gagnant: (2,3)  Dmin = 611.286218640222
Traitement: 15 eme etape d'apprentissage ...
        Vecteur:158     Neurone gagnant: (2,3)  Dmin = 586.108400902514
Traitement: 16 eme etape d'apprentissage ...
        Vecteur:173     Neurone gagnant: (2,3)  Dmin = 560.328236849434
Traitement: 17 eme etape d'apprentissage ...
        Vecteur:164     Neurone gagnant: (2,3)  Dmin = 534.033143511905
Traitement: 18 eme etape d'apprentissage ...
        Vecteur:22      Neurone gagnant: (2,3)  Dmin = 507.511756780828
Traitement: 19 eme etape d'apprentissage ...
        Vecteur:1       Neurone gagnant: (2,3)  Dmin = 480.722274741938
Traitement: 20 eme etape d'apprentissage ...
        Vecteur:35      Neurone gagnant: (2,3)  Dmin = 454.026211230439
Traitement: 21 eme etape d'apprentissage ...
        Vecteur:85      Neurone gagnant: (2,3)  Dmin = 427.527058461898
Traitement: 22 eme etape d'apprentissage ...
        Vecteur:7       Neurone gagnant: (2,3)  Dmin = 401.335031417448
Traitement: 23 eme etape d'apprentissage ...
        Vecteur:203     Neurone gagnant: (2,3)  Dmin = 375.605373118317
Traitement: 24 eme etape d'apprentissage ...
        Vecteur:130     Neurone gagnant: (2,3)  Dmin = 350.45894829233
Traitement: 25 eme etape d'apprentissage ...
        Vecteur:203     Neurone gagnant: (2,3)  Dmin = 325.990687533723
Traitement: 26 eme etape d'apprentissage ...
        Vecteur:58      Neurone gagnant: (2,3)  Dmin = 302.334425157282
Traitement: 27 eme etape d'apprentissage ...
        Vecteur:7       Neurone gagnant: (2,3)  Dmin = 279.500787252824
Traitement: 28 eme etape d'apprentissage ...
        Vecteur:33      Neurone gagnant: (2,3)  Dmin = 257.608165998975
Traitement: 29 eme etape d'apprentissage ...
        Vecteur:32      Neurone gagnant: (2,3)  Dmin = 236.897544176813
Traitement: 30 eme etape d'apprentissage ...
        Vecteur:178     Neurone gagnant: (2,3)  Dmin = 217.137678604563
Traitement: 31 eme etape d'apprentissage ...
        Vecteur:131     Neurone gagnant: (2,3)  Dmin = 198.365373052461
Traitement: 32 eme etape d'apprentissage ...
        Vecteur:50      Neurone gagnant: (2,3)  Dmin = 180.55763043733
Traitement: 33 eme etape d'apprentissage ...
        Vecteur:80      Neurone gagnant: (2,3)  Dmin = 163.885992370519
Traitement: 34 eme etape d'apprentissage ...
        Vecteur:76      Neurone gagnant: (2,3)  Dmin = 148.40564281297
Traitement: 35 eme etape d'apprentissage ...
        Vecteur:154     Neurone gagnant: (2,3)  Dmin = 133.967331097917
Traitement: 36 eme etape d'apprentissage ...
        Vecteur:80      Neurone gagnant: (2,3)  Dmin = 120.580518603655
Traitement: 37 eme etape d'apprentissage ...
        Vecteur:91      Neurone gagnant: (2,3)  Dmin = 108.263004512223
Traitement: 38 eme etape d'apprentissage ...
        Vecteur:205     Neurone gagnant: (2,3)  Dmin = 96.8060974246342
Traitement: 39 eme etape d'apprentissage ...
        Vecteur:161     Neurone gagnant: (2,3)  Dmin = 86.4353159484689
Traitement: 40 eme etape d'apprentissage ...
        Vecteur:180     Neurone gagnant: (2,3)  Dmin = 76.9125770193177
Traitement: 41 eme etape d'apprentissage ...
        Vecteur:90      Neurone gagnant: (2,3)  Dmin = 68.2612653003094
```

Page 1

E. Exemples de convergence des vecteurs en catégories

La convergence des vecteurs avec SOM sur les données du site Web de l'Ambassade de France au Vietnam est visualisée par un outil d'E-SOM [UM05][1].



Voici l'exemple des catégories obtenues par K-means et SOM sur les données du site Web de l'Ambassade de France au Vietnam.

[1] http ://databionic-esom.sourceforge.net/

k-means

1-1:	162	5										
1-2:	115	137	138	25	26	27	29					
1-3:	142	46										
1-4:	59											
1-5:	120											
2-1:	139	31	32	35	45	64	8					
2-2:	143											
2-3:	109	111	112	165	166	167	168	169	170	171	172	173
174	175	177	178	179	180	181	182	183	184	185	187	188
189	24											
2-4:	54											
2-5:	204	205	206	207	61							
3-1:	19	34	38	56	68	94	98					
3-2:	103	11	12	13	141	144	145	147	164	176	186	191
202	87	88	89	91								
3-3:	40											
4-1:	201	66	80									
4-3:	10	135	196	197	198	199	200	51	74			
4-4:	1	104	136	14	148	149	15	150	151	152	153	154
155	156	157	158	159	16	160	161	163	17	192	193	194
195	2	20	203	21	22	23	28	3	30	36	37	43
44	47	48	49	50	52	53	55	57	58	60	62	63
73	77	82	83	9	90							
4-5:	100	101	102	105	106	107	108	110	113	114	116	146
190	65	67	69	70	71	72	75	76	78	79	81	84
85	86	92	93	95	96	99						
5-1:	18											
5-2:	140	39	41	42								
5-3:	134	33	4	6	7							
5-4:	117	121	124	128	130	133						
5-5:	118	119	122	123	125	126	127	129	131	132	97	

som

1-1:	101	107	146	149	152	160	170	172	177	189	190	197
2	204	41	52	58	76	99						
1-2:	1	109	178	188	24	91						
1-3:	104	147	155	171	180	194	28	3	5	7	72	78
1-4:	12	161	31	4								
1-5:	145	151	16	182	198	23	36	48	6			
2-1:	110	113	148	156	157	169	173	174	186	38	73	77
2-2:	106	116	141	167	195	206	8					
2-3:	13	134	137	164	192	202	207	21	50	83	87	96
2-4:	138	9										
2-5:	118	126	27	47	53	59	64	88				
3-1:	135	140	184	19	22	49	60					
3-2:	108	112	114	115	150	158	18	196	203	74	90	
3-3:	122	154	32									
3-4:	11	123	199	200	201	43	62					
3-5:	165	37										
4-1:	136	144	15	187	20	61	82					
4-2:	117	139	179	56	71	92	93	98				
4-3:	44	55	68	70	79	81	89					
4-4:	10	125	168	181	191							
4-5:	105	124	193	25	67							
5-1:	100	111	127	128	131	14	142	143	159	163	166	17
183	26	57	63	75	80							
5-2:	103	130	29	33	34	85	86	95	97			
5-3:	102	120	129	132	176	46	84	94				
5-4:	205	40	42	65	66							

F. Exemple du regroupement des catégories

Dans cette annexe, nous présentons l'exemple des catégories de vecteurs obtenues par K-means sur les données du site Web de l'Ambassade de France au Vietnam, puis regroupées en super-catégories.

davies-bouldin.log

1-1:	120	54	40	18								
1-2:	59	143	103	11	12	13	141	144	145	147	164	176
186	191	202	87	88	89	91	100	101	102	105	106	107
108	110	113	114	116	146	190	65	67	69	70	71	72
75	76	78	79	81	84	85	86	92	93	95	96	99
2-1:	19	34	38	56	68	94	98					
2-3:	134	33	4	6	7	117	121	124	128	130	133	
1-3:	142	46										
2-5:	204	205	206	207	61							
4-1:	201	66	80									
4-3:	10	135	196	197	198	199	200	51	74			
4-4:	1	104	136	14	148	149	15	150	151	152	153	154
155	156	157	158	159	16	160	161	163	17	192	193	194
195	2	20	203	21	22	23	28	3	30	36	37	43
44	47	48	49	50	52	53	55	57	58	60	62	63
73	77	82	83	9	90							
5-2:	140	39	41	42								
5-5:	118	119	122	123	125	126	127	129	131	132	97	

Page 1