Extraction et impact des connaissances sur les performances des systèmes de recherche d'information

Mohamed Hatem Haddad

UNIVERSITÉ JOSEPH FOURIER - GRENOBLE 1

THÈSE

pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ JOSEPH FOURIER

Discipline : Informatique

Mohamed Hatem HADDAD

24 Septembre 2002

TITRE

Extraction et Impact des connaissances sur les performances des Systèmes de Recherche d'Information

Composition du jury :

Président :	M. Jean-Pierre Giraudin
Rapporteurs :	M. Claude Chrisment
	M. Jean-Marie Pinon
Examinateurs :	M. Éric Gaussier
	Mme. Marie-France Bruandet
	M. Jean-Pierre Chevallet

Supposons par exemple qu'il n'ait pas existé jusqu'à notre époque de science de la géométrie et de l'Astronomie, et qu'un seul homme, par soi-même, prétende à connaître les dimensions des corps célestes (?), lui dirait-on que le soleil est environ cent cinquante ou cent soixante fois plus grand que la Terre, qu'il taxerait de folie celui qui tiendrait un tel propos, alors même qu'il s'agit là d'un fait établi en astronomie au moyen d'une démonstration qui ne soulève pas l'ombre d'un doute chez les savants en cette manière.

Averroès, Discours décisif, éd. Grarnier Flammarion, Paris 1996 trad. Geoffroy page 113

Remerciements

Je tiens à remercier :

M. Jean-Pierre Giraudin, professeur de l'Université Pierre Mendès-France, qui m'a fait l'honneur de présider ce jury;

Les rapporteurs sur ce travail : M. Claude Chrisment, professeur de l'Université P. Sabatier de Toulouse, et M. Jean-Marie Pinon, professeur de l'INSA de LYON, pour avoir accepté d'être rapporteur et pour l'intérêt qu'ils ont manifesté pour ce travail;

M. Éric Gaussier, ingénieur de recherche au Xerox Research Centre Europe à Meylan, pour son aimable participation au jury;

Mme Marie-France Bruandet, professeur de l'Université Joseph Fourier, et M. Jean-Pierre Chevallet, maître de conférence à l'Université Pierre Mendès-France, qui ont dirigé ce travail : leurs remarques attentives, leur encadrement de mon travail depuis le D.E.A et pendant ces années de thèse, la gentillesse et la patience qu'ils ont manifesté à mon égard durant cette thèse;

M. Philippe Mulhem, directeur du laboratoire IPAL à Singapour : ses conseils avisés, sa grande connaissance du domaine et son écoute attentive m'ont été très utiles;

Mon collègue et "pote" Mathias Géry; nous nous sommes serré les coudes quand la pente était trop raide;

Stéphanie, Anne et Dominique qui m'ont soutenu comme seuls savent le faire les amis;

Ma famille qui a su manifester son soutien et m'entourer d'affection malgré les kilomètres.

Résumé

Dans un contexte riche d'information, un système de recherche d'information doit être capable de trouver les meilleurs résultats possibles dans un océan d'information. Notre étude s'intéresse aux connaissances qui peuvent être extraites du contenu textuel des documents en associant la finesse d'analyse d'une approche linguistique (extraction et structuration) à la capacité d'une approche statistique de traiter de gros corpus. L'approche statistique se base sur la fouille de données textuelles et principalement la technique de règles d'association. L'approche linguistique se base sur les syntagmes nominaux que nous considérons comme des entités textuelles plus susceptibles de représenter l'information contenue dans le texte que les termes simples. Elle explicite les contraintes linguistiques nécessaires à l'extraction des syntagmes nominaux et défini les rapports syntagmatiques entre les composantes d'un syntagme nominal. Ces relations syntagmatiques sont exploitées pour la structuration des syntagmes nominaux. Une mesure, appelée "quantité d'information", est proposée pour évaluer le pouvoir évocateur de chaque syntagme nominal, filtrer et comparer les syntagmes nominaux. Le modèle proposé démontre que la combinaison d'une approche statistique et d'une approche linguistique affine les connaissances extraites et améliore les performances d'un système de recherche d'information.

Abstract

An information retrieval system is dedicated to find the best possible results in a rich information context. Our study is interested in the knowledge which can be extracted from textual documents contents by associating a linguistic approach to the capacity of a statistical approach to analyze big corpus. The statistical approach is based on Text Data Mining, more precisely on the association rule technique. The linguistic approach is based on noun phrases considered as more adequate to represent document content than single words. It clarifies the needed linguistic constraints for the extraction of noun phrases and explicits the syntagmatic relations between words in noun phrases. These phrasal relations are exploited to structure noun phrases. A measure, namely "information quantity", is proposed to estimate the suggestive power of every noun phrase, to filter and compare noun phrases. The proposed model demonstrates that the combination of a statistical approach and a linguistic approach refines the extracted knowledge and increases the performances of an information retrieval system.

Table des matières

Table des figures

Liste des tableaux

Première partie

Introduction, problèmatique et état de l’art

Chapitre 1

Introduction

Le commencement de toutes les sciences,
c'est l'étonnement de ce que les choses sont ce qu'elles sont.
ARISTOTE

Dès l'invention des ordinateurs les hommes sont à la recherche d'une manière efficace de gérer, de stocker, de diffuser et de rechercher l'information. Plusieurs méthodes et techniques de gestion et de traitement d'information ont été développées. Aujourd'hui, nous pouvons estimer que nous sommes à un haut niveau d'informatisation grâce au développement et à la maîtrise de la technologie (soit celle des matériels, soit celle de la communication, soit celle de la construction des logiciels ou soit celle de la gestion et du traitement de l'information) dont l'Internet est un exemple flagrant. Malgré cette évolution, la progression des moyens de recherche efficaces est encore insuffisante dans le domaine de l'information documentaire, plus spécifiquement dans celui du traitement et de la dissémination de l'information textuelle. Plusieurs recherches sont en cours de développement dans ce domaine, et pourtant, les problèmes d'indexation automatique et de recherche d'information sont encore très actuels. Le système qui s'occupe de cette tâche est connu sous le nom de *Système de Recherche d'Information* (SRI).

1.1 Définitions d'un SRI

Il y a plusieurs définitions d'un SRI, qui sont plus ou moins proches. Tomek Strzalkowski définit un SRI comme suit [Str93]:

> La tâche typique de la recherche d'information, est de sélectionner des documents dans une base de données, en réponse à une requête de l'usager, et leur

> rangement par ordre de pertinence [1].

Tandis que Alan Smeaton donne la définition suivante [Sme89]:

> Le but d'un système de recherche d'information est de retrouver des documents en réponse à une requête des usagers, de manière à ce que les contenus des documents soient pertinents au besoin initial d'information de l'usager [2].

Or, ces définitions restent à notre avis insatisfaisantes car elles n'explicitent pas les procédures de traitement de l'information ou de l'indexation automatique des documents. Ce sont des définitions qui prennent en compte ce que les utilisateurs perçoivent d'un SRI et ce qu'un SRI doit leur offrir. Or, il y a tout un ensemble de procédures pour que les usagers puissent accéder à l'information. Salton et Mcgill donnent une définition d'un SRI plus simple mais plus précise et complète [SM83]:

> Un SRI traite de la représentation, du stockage, de l'organisation et de l'accès aux éléments de l'information [3].

Nous définissons un SRI, illustré dans la Figure 1.1, comme étant un système composé d'une part par un module chargé du traitement, de l'indexation et du stockage de l'information : le module indexation. Ce module construit, à partir du traitement de l'information, une structure de données organisées de manière à permettre l'accès rapide à l'information. D'autre part, il est composé par un module qui sert à interagir avec les utilisateurs, doté des mécanismes de sélection d'information orientés par les requêtes des utilisateurs : le module interrogation. Enfin, un module de correspondance (fonction de correspondance) établit une association entre la requête de l'utilisateur et les documents traités.

1.2 Défis d'un SRI

On assiste aujourd'hui à une explosion de la quantité d'information disponible et accessible. En effet, l'évolution de la quantité d'information est exponentielle. D'après les dernières estimations, on parle de plus de 180 millions de serveurs hôtes sur Internet et on a sûrement dépassé 2 milliards de pages en 2000 [MM00]. Si on regarde les rapports

1. A typical information retrieval (IR) task is to select documents from a database in reponse to a user's query, and rank these documents according to relevance.

2. The aim of an information retrieval system is to retrieve documents in response to a users request in such a way that the content of the documents will be relevant to the user's original information need.

3. Information retrieval (IR) is concerned with the representation, storage, organization and accessing of information items.

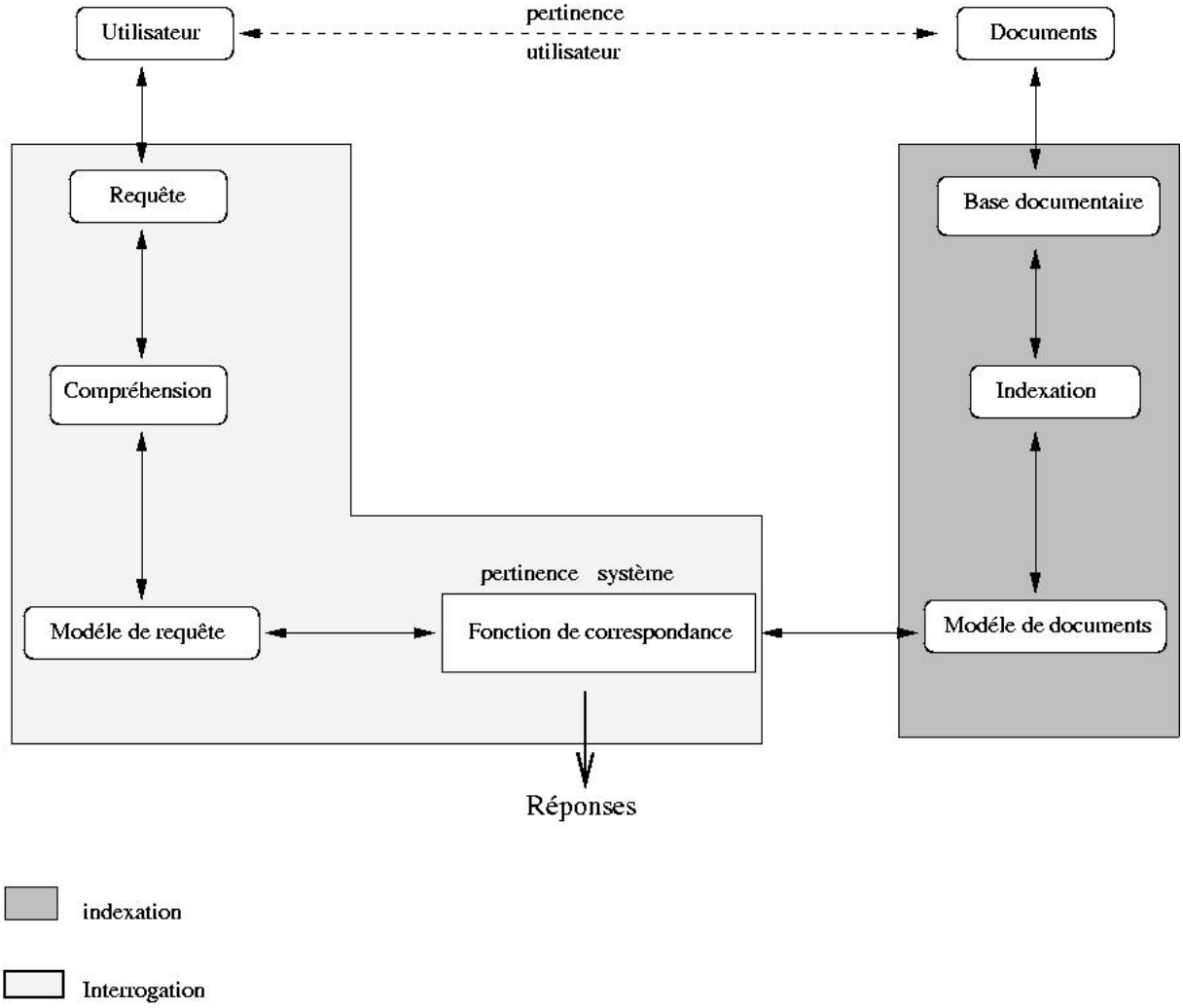

FIG. 1.1 – *Système de recherche d'information*

TREC[4], on remarque que la taille des corpus tests utilisés est passés de quelques mégaoctets durant les premières TREC à plusieurs Gigaoctets pour TREC 8[5].

Dans ce contexte riche d'information, un SRI doit être capable de trouver les meilleurs résultats possibles dans un océan d'information. Si le taux moyen de rappel et le taux moyen de précision dans une petite collection (quelques centaines de documents) peuvent

4. http://trec.nist.gov
5. http://trec.nist.gov/pubs/trec8/index.alpha.htm

être acceptables, ces taux sont loin d'être acceptables dans des grandes collections de données. En effet, avec les grandes collections de documents cités plus haut, les systèmes de recherche d'information que ce soit dans le cas des systèmes documentaires classiques ou dans le cas du Web ont pour objectif non seulement de trouver les documents pertinents mais aussi de classer ces documents parmi les premiers meilleurs résultats. Ceci est plus flagrant dans le cas du Web, car on sait que généralement l'utilisateur ne regarder pas au-delà des 20 premiers résultats. En effet, les utilisateurs d'un moteur de recherche sur le Web cherchent plutôt la précision dans les réponses. Ils préfèrent un nombre restreint de documents répondant à leur besoin qu'un grand nombre de documents contenant la réponse à leur besoin mais noyée dans l'ensemble des documents non pertinents [MBSC97].

Une solution qui a montré son efficacité est d'utiliser un ensemble de connaissances pour permettre d'augmenter les performances en terme de rappel et de précision. Ces connaissances peuvent être de différentes natures et sont classées d'après Bruandet selon deux classes [Bru90] : *les connaissances externes* qui sont indépendantes du modèle d'indexation ou d'interrogation telles que les connaissances du domaine, les connaissances de l'utilisateur, la connaissance de l'expert documentaliste, et *les connaissances internes* comme par exemple les connaissances du mécanisme d'indexation ou le modèle de documents.

Notre étude s'intéresse aux connaissances qui peuvent être extraites du contenu textuel des documents que nous intégrons dans la classe des connaissances internes. Nous nous intéressons à l'enrichissement de ces connaissances et particulièrement l'enrichissement par l'intégration de connaissances statistiques de cooccurrences entre termes, de connaissances syntaxiques et de l'utilisation de ces connaissances pour la représentation et l'interrogation des documents textuels ainsi que de l'impact de ces connaissances sur un SRI. Nous nous sommes posé les questions suivantes auxquels nous répondons tout au long de notre étude:

- Quels types de connaissances à prendre en compte dans le cadre de la recherche d'information textuelle et qui peuvent être qualifiées d'*utiles*?
- Comment les traitements de ces connaissances peuvent-ils s'effectuer?
- Sous quelles contraintes ces traitements peuvent être d'un coût raisonnable dans le contexte de collections volumineuses de documents?

1.3 Information et connaissance

D'un point de vue scientifique, l'information apparaît comme un sujet vague et incohérent. De ce fait, le mot information a des définitions multiples et ambiguës. La définition

du *Larousse* est tout à fait significative. Elle se décompose en plusieurs sous-définitions selon les critères suivant :

- Le critère Action : l'information est l'action d'informer, de se mettre au courant d'événements.
- Le critère Etat : l'information est une nouvelle, un renseignement que l'on communique ou que l'on obtient.
- Le critère Connaissance : l'information est un ensemble de connaissances acquises sur quelqu'un ou sur quelque chose.
- Le critère Contenu : l'information est le contenu proprement dit des messages transmis.
- Le critère Contenant : l'information est un signal par lequel un système donne connaissance de sa position à un autre.

L'Office de la Langue Française donne une définition plus précise de l'information:

> Élément de connaissance concernant un phénomène et qui, pris dans un contexte déterminé, a une signification particulière.

La note accompagnant cette définition est satisfaisante pour préciser notre point de vue de la définition de l'information:

> Parmi les phénomènes (ou les objets) pouvant donner naissance à une information, on peut mentionner : un fait, un événement, une chose, un processus, une idée, une notion. Dans le contexte plus précis du traitement des données, une information est une donnée qui a été interprétée (ou réinterprétée). Le cadre de référence qui détermine cette interprétation est constitué de la somme des connaissances et des expériences de la personne qui effectue l'interprétation. En anglais, dans l'usage courant, les termes data et information sont souvent utilisés comme des synonymes. En effet, on peut considérer que l'information, matérialisée par un support, devient une donnée pour un ordinateur. Il s'agit cependant de notions bien différentes [6]

Dans un contexte de recherche d'information, Paradis distingue, dans son modèle, le méta-contenu qui le définit comme étant *l'information qui se rattache au contenu en précisant sa nature ou en y ajoutant des informations telles que les attributs du contenu, les thèmes, les connaissances, etc.* ainsi que l'information qui se rattache à la structure des

6. Office de la langue française, 2002

documents [Par96]. Ces informations sont toutes, plus ou moins, intéressantes à intégrer dans un modèle de RI. Dans notre modèle, nous ne tenons pas compte de ces informations, bien que nous pensions qu'elles soient intéressantes et utiles dans le cadre de la recherche d'information, et nous ne nous focalisons que sur un seul type d'information qui est le *contenu textuel*. Ainsi, nous ne tenons pas compte par exemple des styles typographiques (italique, souligné, gras, etc.) et les informations de l'ordre de la logique qui définissent les éléments qui forment la structure du document (organisation en chapitre, sections, références, etc.).

Nous considérons alors comme *connaissance* toute information qui est utile pour un SRI en distinguant:

- *les connaissances internes* aux documents : l'ensemble des informations qui découlent ou qui sont dérivées du contenu.

- *les connaissances externes* aux documents : l'ensemble des informations qui sont injectées au contenu textuel et qui proviennent de sources diverses (thésaurus, base de connaissances externes, connaissances paradigmatiques, connaissances expertes d'un domaine, etc.) et qui sont du ressort de la modélisation d'un domaine.

Dans notre modèle, nous nous focalisons sur l'acquisition de connaissances internes que nous désignons *connaissances implicites* du fait qu'elles existent dans le texte mais noyées dans la masse d'information. Le refus délibéré de faire appel à des connaissances externes du type des traits sémantiques nous parait être essentiel pour tout système dont la vocation est d'appréhender des textes ouverts (sans restrictions sur les domaines). Ce refus est motivé aussi par la difficulté et la problématique liées à la construction de thésaurus ou ontologies et leur utilisation dans un contexte de recherche d'information.

1.4 Organisation du lexique mental

Lors de la production d'un discours, les mots sont employés selon un processus de sélection qui tient compte des relations entre les différents éléments du lexique. Des expériences en psycholinguistique sur l'organisation du lexique mental montrent que les relations entre les éléments du lexique sont de deux types. Il existe des relations intrinsèques, ou catégorielles, qui contiennent des informations linguistiques sur l'unité lexicale elle-même, et des relations d'association qui regroupent les unités dont la fréquence d'apparition dans un même contexte est importante (*ouvrier* avec *usine* ou *travail*) [Bog94, Iss97]. Les relations intrinsèques peuvent être décomposées en relations sémantiques (comme la synonymie ou l'antonymie), morphologiques (le domaine de la dérivation : *compétent, compétence, incompétence*), et phonologiques (les mots commençant ou se terminant par

les mêmes phonèmes). Les relations entre les éléments du lexique forment ainsi un réseau dans lequel les noeuds ne sont pas les mots eux-mêmes, mais leurs sens particuliers.

En recherche d'information, peu de travaux se sont intéressés à l'impact de ces relations sur les performances d'un SRI ainsi que le rôle que peuvent jouer ces relations. Dans ce domaine, Farradane se base sur l'hypothèse qu'une grande partie du sens d'une information est contenue dans les relations entre termes [Far80a, Far80b]. L'utilisation d'un vocabulaire contrôlé d'un domaine et de symboles qui indiquent des relations techniques entre des termes propres au domaine ne sont ni transportables ni applicables dans les autres domaines. Il y a donc un besoin de relations expressives entre les termes qui sont générales pour tous les domaines. L'indexation relationnelle semble alors être un moyen pour exprimer les relations en se basant sur le mécanisme de la pensée pour être converti directement en une notation d'indexation. Dans ce sens, Farradane dénombre 9 catégories de relations, qu'il qualifie d'intéressantes et suffisantes pour le domaine de la RI, que nous présentons dans l'annexe A, en se basant sur une étude des mécanismes de pensée psychologiques.

Les difficultés à ce niveau sont la complexité de la langue et l'imperfection des auteurs dans l'utilisation de la langue où il n'y a pas de standardisation des termes. En effet, dans les thesaurus ou les listes d'autorité, il n'y a pas de définition des termes et l'utilisateur est supposé connaître le sens des termes alors que l'utilisation d'un dictionnaire est trop encombrante.

L'utilisation de relations permet d'avoir des représentations du contenu du texte sémantiquement plus riches que ceux obtenus par les approches traditionnelles (utilisation de termes simples) [Oun98]. Dans ces expérimentations, Farradane construit ces relations manuellement ce qui présente un inconvénient majeur dans le contexte de grandes quantités d'information visée par notre modèle. En effet, il ne donne aucune spécification formelle indiquant comment procéder à l'indexation automatique des documents en fonction des relations qu'il présente.

Dans le chapitre 2, nous présentons des travaux qui traitent de ces relations. Nous détaillons le contexte de leur utilisation et leur impact dans un SRI. Dans notre modèle, nous exploitons, la relation d'association de Farradane présentée dans le tableau A.1.

1.5 Relation d'association

L'information obtenue par la structuration d'un corpus textuel (classification des documents, extraction de termes représentatifs du contenu des documents) ne constitue qu'une des facettes de la connaissance implicitement contenue dans un corpus. Pour cette raison, un des objectifs de la fouille de données textuelles [7] est également de proposer des tech-

7. en anglais *Text Mining*

niques permettant l'extraction d'informations implicites, présentes de façon diffuse dans la base documentaire (par exemple l'information distribuée sur plusieurs documents). Ce problème s'est aussi posé en intelligence artificielle dans le cas des bases de connaissances où l'objectif est de développer des formalismes de représentation suffisamment compatibles avec les techniques de raisonnement automatique pour permettre la dérivation automatique de l'ensemble des informations qui en sont logiquement déductibles. Cet objectif s'est avéré irréaliste pour des applications impliquant des volumes importants de connaissances.

Une partie des travaux d'extraction de connaissances dans le domaine de la fouille de données textuelles s'intéressent à des classes particulières de connaissances pour lesquelles des algorithmes de traitement, opérationnels dans le cas de volumes de données de taille importante, peuvent effectivement être proposés. Une des branches de la fouille de données textuelles s'intéresse aux implications qui décrivent d'une façon symbolique les différentes corrélations entre mots dans les documents en se basant sur la notion d'ensemble fréquent qui sous-entend tout ensemble de mots apparaissant dans une base de documents avec une fréquence supérieure à un seuil fixé a priori. L'extraction des implications permet l'extraction de corrélations entre ensembles de mots que l'on peut interpréter comme des implications probabilistes. Ces implications sont appelées **les règles d'association**.

1.6 Langage naturel

Le moyen naturel et habituel pour exprimer des informations est le langage naturel. Mais, pourquoi dit-t-on *langage naturel*? Le langage, parlé ou écrit, n'est-il pas toujours *naturel*? Ce terme, créé par les informaticiens, est utilisé par opposition aux langages formels, comme la logique ou les langages informatiques. A la différence de ces derniers qui sont des codes non ambigus, le langage humain ne se laisse pas facilement formaliser et on sait qu'il est extrêmement ambigu. Cette ambiguïté présente un obstacle à son utilisation pour le traitement de l'information. De fait, les systèmes informatiques éprouvent des difficultés en présence de la paraphrase ou de la construction de nouveaux concepts, omniprésents dans l'emploi de la langue. Ils ont tendance à buter sur l'ambiguïté de certains énoncés, pourtant clairs dans le contexte dans lequel ils ont été écrits, et ne peuvent de ce fait appréhender directement des textes.

Les notions de texte, de phrase et de mot sont des notions intuitives à la fois évidentes à comprendre dans le contexte de l'usage courant, et quasiment impossibles à définir de façon formelle. Plusieurs théories sont proposées pour décrire ces notions. Dans notre contexte, nous nous contenterons de dire qu'un texte est un ensemble d'énoncés (d'unités du discours) et que dans la plupart des textes écrits, les énoncés sont des phrases. La phrase

peut se définir intuitivement comme l'unité du langage supérieure au mot. Cette définition est celle de la grammaire traditionnelle, que nous retiendrons ici provisoirement pour sa valeur opératoire. A l'écrit, on peut considérer qu'une phrase est une séquence de mots qui se trouve entre deux points, définition empirique mais qui a le mérite de bien fonctionner pour l'identification dans le contexte d'un traitement informatique.

L'utilisateur d'un SRI exprime son besoin d'information par des mots et l'auteur d'un document textuel exprime l'information qu'il veut transmettre par des mots qui combinés forment des phrases dont l'ensemble forme le texte. Ce processus de communication entre l'utilisateur d'un SRI et le propriétaire du texte passe par le texte. Le texte est alors représenté par un ensemble de symboles dont la signification est un consensus entre l'auteur et l'utilisateur.

D'un autre côté, un SRI dispose d'une collection de documents et d'une requête d'un utilisateur. Il essaye alors de "piocher" dans les documents afin de trouver un texte dont le contenu informationnel correspond le mieux à la demande d'information de l'utilisateur. Mais un SRI, au contraire d'un être humain et des systèmes de compréhension de la langue naturelle ou des systèmes de questions/réponses, ne dispose pas de moyens pour comprendre et raisonner sur le sens comme par exemple un ensemble de connaissances pragmatiques sur le monde. Il n'a à sa disposition que le texte donc un ensemble de symboles qui traite son contenu. Ce texte à l'état brut (non traité ni passé par les phases de morphologique ni lexicales ni syntaxique) est la source d'information pour le SRI qui lui permet à partir du signal de s'approcher du sens véhiculé par le texte. La question qui se pose à ce niveau est la suivante : doit-il se préoccuper du contenu de ce qu'il transmet ou doit il être insensible au contenu et ne pas influencer la réception de l'information donc être une opération neutre et transparente ? Nous restons indécis dans un premier temps sur la réponse à cette question:

– Oui il doit s'occuper du contenu pour pouvoir représenter une information dans un formalisme qui soit le plus proche possible de l'information originale.

– Non il ne doit pas s'occuper du contenu d'une information afin de ne pas influencer le récepteur dans sa compréhension du sens de l'information transmise.

Pour résoudre cette problématique, un SRI veut être le plus fidèle possible au niveau de la représentation de l'information, au niveau de l'indexation ainsi que dans le cadre de l'interrogation. Cette fiabilité se traduit par la volonté des spécialistes de représenter l'information avec des formalismes plus complexes que l'utilisation des mots simples très utilisés jusque là en domaine de SRI.

D'un autre coté, un SRI doit répondre au besoin d'information d'un utilisateur dans un contexte où ce dernier exprime *mal* son besoin:

– l'utilisateur n'a pas un besoin précis et ne sait pas comment chercher ce qui est encore vague et flou dans sa tête.

– l'utilisateur interroge un corpus dont il ignore le vocabulaire et les termes qu'il emploie ne lui permettent pas d'avoir des réponses ou peu de réponses non pertinentes.

– l'utilisateur a un besoin très précis donc il emploie des mots très techniques ce qui ne lui permet pas d'avoir des réponses, ce qui se traduit par *le silence*.

– l'utilisateur emploie des mots trop généraux et le nombre de réponses est alors très important, ce qui se traduit par du *bruit*.

Parmis les types de besoins d'un utilisateur, deux types de besoins peuvent être distingués selon la demande d'information de l'utilisateur:

– la demande exploratoire naît quand l'utilisateur veut se faire une idée du contenu d'une collection donnée sans a priori. Il s'agit alors de lui proposer des extraits jugés représentatifs des thèmes des documents.

– la demande thématique est destinée à illustrer un thème. Le type de raisonnement alors suivi par l'utilisateur est un raisonnement par association d'idées stimulé par la visualisation des documents

S'il existe qu'une seule façon de compter et de délimiter les mots (en excluant les noms propres, les sigles, etc.), il existe une infinité de façons de regrouper ceux-ci pour constituer les thèmes. Chaque approche thématique, selon ses propres hypothèses a sa façon de regrouper les mots d'un texte afin de constituer des thèmes. Dans notre modèle, nous faisons appel, en plus des relations d'association, à l'axe syntagmatique pour construire les thèmes d'une collection de documents. Nous commençons d'abord par définir les notions fortes de notre modèle.

1.7 Uniterme et terme complexe

Les termes sont des objets linguistiques utilisés dans la littérature technique et scientifique et visent à faire référence à des concepts de façon non ambiguë. Les concepts sont des regroupements d'objets réels ou immatériels ayant des propriétés communes.

L'Office de la Langue Française [8] donne trois définitions de *terme*:

– Unité signifiante constituée d'un mot (terme simple) ou de plusieurs mots (terme complexe) et qui désigne une notion de façon univoque à l'intérieur d'un domaine.

– Mot ou groupe de mots employé pour représenter une notion.

8. Office de la langue française, 1985. http://www.olf.gouv.qc.ca/index.html

– Mot appartenant à un vocabulaire spécial notamment au vocabulaire scientifique.

La première définition place la notion et la face signifiante d'un mot en premier lieu. Le terme est ici considéré comme la traduction univoque d'une notion qui lui préexiste et qui est rattachée à un domaine. Cette vision se retrouve de façon prononcée dans le triangle sémiologique (Figure 1.2) et qui rejoint celle de Ferdinand de Saussure [Sau72]. En effet, pour Saussure, le mot qu'il désigne par *le signe linguistique* est formé de deux faces, le concept et l'image acoustique qui peuvent prêter à confusion. L'image acoustique seule pourrait subsister, absorbant le concept. Pour éviter cette fusion, il décide pour le signe de garder le mot et pour ses deux constituants il use des dérivés de signe. Il nomme *Signifié* le concept et *Signifiant* l'image acoustique. Le signifiant est alors la partie matérielle (physique) observable du signe (symbolique dans le contexte du texte) et le signifié est la partie conceptuelle du signe qui est la notion ou le sens à transmettre.

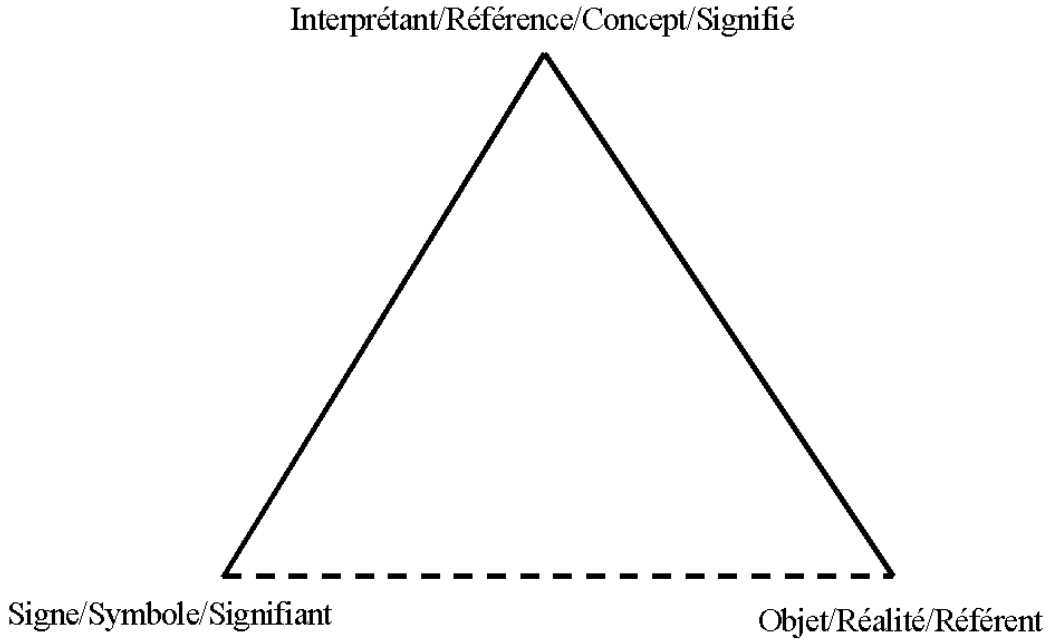

FIG. 1.2 – *Triangle sémiologique*

Le signe possède deux principes primordiaux selon Saussure : il est arbitraire, c'est à dire qu'il n'existe pas de lien motivé entre le signe et son référent et le signifiant possède un caractère linéaire. La langue, selon Saussure, est alors un système de signes et les rapports entre les signes d'une langue sont de deux types : syntagmatique (la langue comme

structure) ou paradigmatique (la langue comme système). Les rapports syntagmatiques sont organisés selon l'ordre de linéarisation du langage, dans une étendue spatiale et/ou temporelle. Cet ordre rend compte des relations positionnelles et des relations fonctionnelles. Ainsi, le rapport syntagmatique est le site des relations contextuelles. Tout système de signes qui a la possibilité de combiner des signes pour créer un nouveau texte, doit le faire en déployant ses signes sur un certain espace et selon certaines règles. C'est l'espace d'agencement des syntagmes : l'espace syntagmatique. Si les signes de la langue se combinent en syntagmes sur le plan de l'expression (axe syntagmatique) et si ces combinaisons sont identifiables en tant que telles, c'est parce que chacun des éléments de l'axe syntagmatique prend place dans une classe d'éléments qui pourraient virtuellement se substituer à lui : un paradigme. L'opposition syntagmatique/paradigmatique s'est imposée depuis Saussure dans toutes les conceptions structurales de la langue, et son pouvoir explicatif s'étend à plusieurs niveaux : phonologique, syntaxique, sémantique, etc.

Il existe différents points de vue pour la définition du sens d'un terme qui peut être abstraite tel que le sens stéréotype (le sentiment commun partagé) ou le sens d'un terme comme une intention (selon l'émetteur et le récepteur) ou encore comme une interprétation (le sens d'un mot n'est pas toujours donné par ce que *je dis* et par ce que *tu entends*, mais par ce que *je crois avoir dit* et par ce que *tu crois avoir entendu*).

Tant pour leur comportement en syntaxe que pour leur interprétation, il est utile de séparer les termes en deux catégories : les termes simples constitués d'un seul mot plein désignés par *unitermes* (termes simples) et les termes complexes contenant au moins deux mots pleins désignés par *multi-termes*. Les premiers sont fortement ambigus mais leur comportement en syntaxe est simple en raison de leur réduction à un atome syntaxique. Les seconds, en revanche, posent moins de problèmes de polysémie mais requièrent une étude fine de leur comportement en syntaxe.

1.8 Axe syntagmatique et axe paradigmatique

Saussure donne la définition suivante de l'axe syntagmatique [Sau72]:

> Dans le discours, les mots contractent entre eux, en vertu de leur enchaînement, des rapports fondés sur le caractère linéaire de la langue, qui exclut la possibilité de prononcer deux éléments à la fois. Ceux-ci se rangent les uns à la suite des autres sur la chaîne de la parole. Ces combinaisons qui ont pour support l'étendu peuvent être appelées syntagmes. Placé dans un syntagme, un terme n'acquiert sa valeur que parce qu'il est opposé à ce qui précède ou ce qui suit, ou à tous les deux.

Saussure enchaîne avec la définition de l'axe paradigmatique:

> D'autre part, en dehors du discours, les mots offrant quelque chose de commun s'associent dans la mémoire, et il se forme ainsi des groupes au sein desquels règnent des rapports très divers. Ainsi le mot enseignement fera surgir inconsciemment, à l'esprit une foule d'autres mots (enseigner, renseigner, etc., ou bien armement, changement, etc., ou bien éducation, apprentissage), par un côté ou un autre, tous ont quelque chose en commun entre eux.

La valeur d'une unité linguistique dans l'ordre syntagmatique découle de la linéarité donc due au contraste avec ce qui précède ou ce qui suit. Dans l'ordre paradigmatique (appelé aussi associatif ou sélectif), une unité linguistique s'oppose à celles avec lesquelles elle partage quelque chose par ressemblance ou dissemblance et qui ne peuvent pas apparaître dans une chaîne parce que cette unité existe déjà dans cette chaîne.

L'axe syntagmatique est alors relatif aux relations qu'entretiennent les unités lexicales entre elles dans une chaîne textuelle donc relatives à la succession des mots dans le discours ou dans un contexte, désignées par *relations linguistico-sémantiques* par Salton et Wong dans [SW75]. L'axe paradigmatique est relatif aux relations qui existent entre les unités susceptibles de commuter donc une relation entre des classes d'éléments linguistiques qui entretiennent entre elles des rapports de substitution, désignées par *relations logico-sémantiques* par Salton et Wong dans [SW75]. Chaque unité linguistique est donc liée par des rapports syntagmatiques et des rapports paradigmatiques à d'autres unités. Dans le cadre d'une indexation manuelle, l'indexeur choisit ou plutôt sélectionne les descripteurs qu'il juge représentatifs d'un document suivant l'axe paradigmatique pour les représenter dans un certain ordre selon l'axe syntagmatique.

Robins définit les relations syntagmatiques comme suit [Rob73]:

> Les relations syntagmatiques sont celles qui existent entre les éléments formant des structures sérielles à un niveau donné qui renvoient, mais bien sûr sans s'y identifier, au déroulement temporel du discours ou à des fragments linéaires d'écriture.

D'après le point de vue de Robins, une séquence de mots correspond à une transcription phonétique, une représentation phonologique abstraite et à un agencement morpho-syntaxique. Il s'agit de structures de composition, à différents niveaux, en relation syntagmatique.

Robins définit les relations paradigmatiques comme suit [Rob73]:

> Les relations paradigmatiques sont celles qui existent entre les éléments comparables placés à des endroits donnés dans des structures syntagmatiques.

Dans un contexte plus pratique, Debili désigne ces deux types de relations par : les relations lexicales sémantiques syntagmatiques (RLS syntagmatique) et les relations lexicales sémantiques paradigmatiques (RLS paradigmatique) [Deb82]. Il définit la RLS syntagmatique comme étant *une relation définie sur un n-uplet de mots pleins extraits d'un texte et vérifiant certaines conditions morpho-syntaxiques* et la RLS paradigmatique comme étant *une relation qui s'établit entre les unités lexicales qui sont susceptibles de commuter*. Une RLS syntagmatique est donc la cooccurrence de mots pleins dans une chaîne nominale ou verbale exprimant une réalité sémantique. Cette notion de réalité sémantique, évoquée aussi par Khoo [Kho95, Kho97], se base sur l'hypothèse qu'un ensemble de phénomènes syntagmatiques peuvent traduire des faits sémantiques ; hypothèse que nous retenons pour notre modèle.

1.9 Notre méthodologie

Le processus d'acquisition des connaissances est d'emblée reconnu dans la littérature comme un processus très complexe qui dépasse largement le fait d'extraire des termes d'un texte. Notre approche d'acquisition de connaissances textuelles consiste à prendre en considération les combinaisons des éléments textuels au niveau de l'analyse du texte ainsi que la relation d'association entre ces éléments. Contrairement à la plupart des travaux qui analyse le texte sous la forme de chaîne de caractères atomique ou individuelle, notre objectif est de traiter le texte en gardant l'information concernant les rapports qu'entretiennent ces éléments textuels que ce soit les rapports syntagmatiques au niveau de la syntaxe ou bien les rapports associatifs dans les documents. Cette méthodologie est motivée par le fait que ces connaissances existent bien dans un texte. Le fait de les ignorer au moment d'analyser un texte pour un SRI, nécessite l'ajout de la certitude au moment de restituer cette information. En effet, cette information est reconstituée généralement moyennant un thésaurus qui introduit une certaine certitude sur les relations entre les éléments textuels due à l'ignorance de leur contexte exact dans le texte [Rij79, XC00].

Si nous considérons le document [9] illustré dans la Figure 1.3, nous avons encadré, à titre d'exemple, des unités textuelles susceptibles de représenter le contenu du document. Dans le chapitre 7, nous montrons que ces unités sont en fait des syntagmes nominaux qui peuvent être détectés dans le texte en utilisant les rapports syntagmatiques entre les termes. D'autre part, nous avons représenté par des flèches des rapports d'association entre des unités textuelles. Par les rapports d'association, nous visons à extraire pour chaque unité textuelle son contexte d'utilisation dans les documents, que nous définissons dans le chapitre 5, et qui donne une information sur le sens d'une unité textuelle. Notre méthodologie d'acquisition de connaissances se base donc sur deux points de vue : un point de vue

9. Document Numéro : 2271490 de la collection OFIL

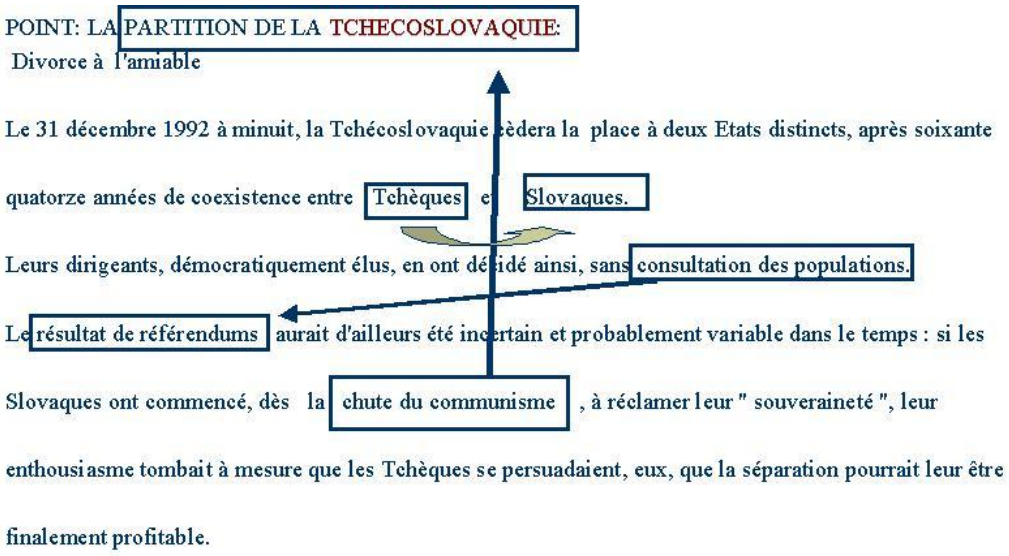

FIG. 1.3 – *Connaissances existantes dans le texte*

statistique et un point de vue linguistique.

Le point de vue statistique concerne la distribution des termes dans les documents. Il se base sur l'hypothèse que l'emploi de deux termes ensemble dans un document suggère une relation sémantique entre ces deux termes [Rij79, XC00]. Ce point de vue a montré des limites dues essentiellement à l'ignorance du contexte linguistique.

Le point de vue linguistique concerne les combinaisons des éléments textuels au niveau du texte. C'est un niveau qui est très proche de la syntaxe et qui prend en considération les rapports syntagmatiques entre les différentes unités textuelles. Il permet de mettre en évidence les combinaisons linguistiquement correctes donc qui sont susceptibles d'être sémantiquement plus riches. Notre objectif n'est pas une analyse en vue de la compréhension de la langue naturelle mais une reconnaissance de la structure linguistique des unités textuelles susceptibles de véhiculer le sens contenu dans un texte.

Le modèle d'acquisition de connaissances que nous présentons n'est pas une théorie, c'est un formalisme d'acquisition de connaissances à partir du texte qui s'inspire:

D'un point de vue théorique :

> Nous développons un formalisme linguistique basé sur le traitement automatique de la langue naturelle pour l'extraction d'unités textuelles représentatives du contenu textuel des documents. Dans ce formalisme, nous nous basons sur les rapports syn-

tagmatiques entre les éléments textuels et nous introduisons une nouvelle mesure, que nous désignons par *quantité d'information*, afin d'évaluer la représentativité des éléments extraits et de les comparer entre eux. Dans ce modèle, des considérations du niveau morphologique, lexicales et syntaxiques sont aussi traitées. En complément du modèle linguistique, un modèle statistique de l'analyse distributionnel est développé afin d'extraire les relations d'association entre les éléments extraits. Ces deux modèles sont assez génériques pour traiter la langue générale sans prendre en considération les propriétés et les caractéristiques des langues de spécialité.

D'un point de vue pratique :

Notre modèle définit les techniques à utiliser pour l'extraction d'unités textuelles et de relations d'association. Ces techniques, expérimentées sur des grandes collections de documents, ont montré leur efficacité à la fois au niveau de la qualité des résultats et au niveau des performances de traitement. Le modèle défini aussi le formalisme de structuration des connaissances extraites et d'exploitation dans un SRI.

Nous n'avons donc pas pour l'ambition de refaire une logique mais de bien définir un formalisme approprié pour l'acquisition de connaissances textuelles dans le cadre de la recherche d'information. L'idée d'associer un modèle linguistique à un modèle statistique est très prometteuse. Elle est également très pertinente en ce sens qu'elle associe la finesse d'analyse des méthodes linguistiques à la capacité des méthodes statistiques d'absorber de gros corpus.

1.10 Organisation de la thèse

Dans la suite de cette thèse, nous commençons par présenter les travaux relatifs à l'acquisition de connaissances à partir du texte, afin de positionner notre approche. Cette présentation ne veut pas être exhaustive mais elle cherche à préciser les éléments clés liés à notre approche. La deuxième partie de la thèse, après une introduction à l'approche de fouille de données, est consacrée à la fouille de données textuelles qui traite du point de vue statistique de notre méthodologie. Enfin, la troisième partie de la thèse traite le point de vue linguistique de notre méthodologie en développant la problématique de la représentation des documents avec des termes complexes ainsi que la proposition d'un modèle d'indexation qui prend en compte cette représentation dans un SRI.

La suite de cette thèse est organisée comme suit:

– le chapitre 2 présente une rétrospective des approches d'acquisition de connaissances à partir du texte. Nous détaillons certains travaux effectués dans ce domaine

en les classant d'abord par rapport à l'approche utilisée, ensuite par rapport aux ressources de connaissances utilisées et enfin selon leurs objectifs de classification et en discutant les avantages et les inconvénients de chacun d'eux.

– le chapitre 3 est une introduction à la problématique de la fouille de données dans les bases de données et particulièrement à la technique des règles d'association.

– le chapitre 4 présente la problématique de la fouille de données textuelles ainsi que quelques systèmes de fouille de données textuelles.

– le chapitre 5 présente la problématique des règles d'association dans le contexte des données textuelles. En précisons la signification et l'utilité de ces règles dans le cadre de la recherche d'information.

– le chapitre 6 présente le processus d'extraction des règles d'association et leur application pour l'expansion des requêtes.

– le chapitre 7 définit les syntagmes nominaux et leur capacité à représenter le contenu textuel des documents.

– le chapitre 8 présente notre méthodologie d'extraction des syntagmes nominaux ainsi qu'une expérimentation de l'évaluation de leur apport pour un SRI.

– le chapitre 9 présente la mesure de la quantité d'information et notre méthodologie de structuration des syntagmes nominaux.

– le chapitre 10 présente notre proposition d'un modèle d'indexation relationnelle basée sur les syntagmes nominaux.

Chapitre 2

Acquisition de connaissances à partir du texte

L'accumulation des connaissances n'est pas la connaissance.
ALBERTO MANGUEL, *Extrait d'Une histoire de la lecture*

Notre problématique est d'acquérir à partir de corpus de documents textuels un ensemble de connaissances utiles pour un SRI. Nous avons défini deux objectifs majeurs pour l'acquisition de connaissances dans des corpus textuels. Le premier objectif consiste à acquérir des termes significatifs et représentatifs du contenu informationnel du corpus. Le deuxième objectif est d'acquérir des relations entre ces termes.

Ces objectifs ne sont pas très éloignés des objectifs d'extraction de terminologie. En effet, la plupart des méthodes d'extraction de terminologie essayent de capturer la notion de concept à l'aide de classes, contenant des termes, utilisées pour préciser le concept. La tâche de construction de ressources terminologiques est en quelque sorte une activité d'interprétation de textes au cours de laquelle un analyste construit une description des termes existants dans le texte. Le terme est alors une unité de description résultant d'un travail d'interprétation et de modélisation mené à partir de l'analyse d'un corpus de référence et s'intégrant dans une ressource terminologique cible.

Nous pensons que, dans le domaine de la RI, nous utilisons certaines notions de la terminologie mais pas toutes. En effet, le but de la RI, lors de l'analyse de documents textuels, est d'extraire les termes représentatifs du contenu sémantique des documents et non pas de choisir les termes représentatifs des connaissances du domaine thématique traité dans le corpus. Bien que l'acquisition des connaissances (terminologie et classification) ne soit pas un domaine directement lié à notre problématique de recherche, certains travaux dans ce domaine peuvent cependant être intéressants et complémentaires.

Dans ce chapitre, nous présentons une rétrospective de différents travaux parallèles ou complémentaires à notre problématique. D'abord, nous classons ces travaux en 3 classes

selon l'approche utilisée : l'approche statistique, l'approche linguistique et l'approche hybride. Ensuite, nous présentons quelques travaux qui exploitent d'autres sources d'information autres que les corpus textuels et qui utilisent soit les connaissances acquises par l'utilisateur soit celles existantes dans des ressources lexicales tels que les dictionnaires ou les thesaurus. Nous présentons aussi des techniques de classification des connaissances extraites du texte qui peuvent être des techniques statistiques ou bien des techniques linguistiques. Enfin, nous présentons quelques techniques d'extraction d'information. Ce domaine connexe à la recherche d'information, a pour objectif d'identifier, réunir et normaliser l'information désirée par l'utilisateur. La différence avec le domaine de la RI est que cette information est définie au préalable, généralement dans des fiches [1]. Un système d'extraction d'information est donc généralement applicable à un domaine particulier afin de satisfaire le besoin en information des utilisateurs de ce domaine.

2.1 Approche statistique

L'approche statistique a commencé dans les années 60 avec les travaux de Salton [Sal68]. Elle est basée sur l'extraction des cooccurrences des termes dans un contexte particulier. Les outils statistiques sont bien adaptés à la détection des récurrences contextuelles contrairement aux analyseurs linguistiques qui n'observent pas les régularités sur des grosses masses de données. Cette information est exploitée en utilisant différentes mesures de similarité. Elle se base sur l'hypothèse que l'emploi de deux termes en cooccurrence est l'expression d'une relation sémantique entre ces termes [Rij79, JC94]. Cette relation s'exprime par des combinaisons de mots qui occurrent souvent dans un corpus et dont le statut linguistique (par exemple les catégories grammaticales des mots) peut varier. Trois approches statistiques fondées sur l'étude de cooccurrences entre termes peuvent être distinguées:

- la première, fondée sur l'hypothèse distributionnelle de Harris [Har71] (offrir une méthode formelle pour appréhender les corrélations entre les conditions de production des discours et leur forme textuelle) tente de découvrir les relations entre termes en fonction de leur régularité d'apparition commune.

- la seconde établit une mesure de proximité entre les contextes des termes. C'est *la ressemblance des contextes* qui fonde la détection de la relation de cooccurrence entre les termes. Les contextes associés aux termes sont construits à partir des cooccurrences [BRC99]. Le partage des contextes est exprimé par la notion de cooccurrence, c'est-à-dire la présence répétée des termes dans les mêmes contextes. Ces contextes représentent *l'environnement sémantique* d'un terme [Sim00]. L'étude

1. en anglais *templates*

du recouvrement de ses environnements sémantiques permet d'établir des relations entre termes y compris s'ils ne cooccurrent pas.

– la troisième approche, appelée *méthode des segments répétés* s'appuie sur la détection de chaînes constituées de morceaux existant plusieurs fois dans le même texte [Oue99]. Cette approche commence par stocker tous les mots du texte dans une table dont la valeur correspond soit à un terme, soit à une ponctuation, soit à un symbole de structure du texte (saut de paragraphe, chapitre, etc.) et une fréquence minimale d'apparition dans le texte est fixée afin d'éliminer les faibles occurrences. Pour chaque forme du texte, l'ensemble des suites dans le texte commençant par cette forme est répertorié. Le processus est réitéré pour chaque forme du texte.

Une autre branche de l'approche statistique est la statistique exploratoire ou l'analyse des données [Ben73]. Deerwester et al. ont exploité cette approche dans l'analyse des données textuelles sous le nom de *l'indexation sémantique latente*[2] [DDK+90]. Cette approche consiste à effectuer une décomposition en valeurs singulières de la matrice dont chaque colonne représente un document grâce à un vecteur des occurrences des termes qui le composent [BDO95, DDK+90, Hul94]. Cette matrice est projetée dans un espace de dimension plus petit. D'après les auteurs, avec cette méthode, les termes apparaissant ensemble sont projetés sur la même dimension. Cette représentation est censée résoudre partiellement le problème des synonymes et des termes polysémiques.

La littérature statistique est riche en mesures destinées à calculer la liaison ou la dépendance (ou contrairement l'indépendance) des termes. Dans l'annexe B, nous présentons certaines mesures utilisées dans la littérature pour calculer les dépendances entre les termes. La majorité de ces mesures, sinon toutes, se base sur la simple mesure de cooccurrences exprimée par la co-fréquence entre deux termes dont nous présentons les plus couramment utilisées [CH90, SHP95, YP97]:

– Information mutuelle.

  La mesure de l'information mutuelle (IM) permet de détecter des cooccurrences de deux mots en comparant la probabilité de les trouver simultanément avec la probabilité de les trouver indépendamment. C'est une mesure symétrique qui s'exprime comme suit:

$$IM(x,y) = log_2 \frac{P(x,y)}{P(x)P(y)}$$

  P(x) et P(y) sont les probabilités d'observer les mots x et y, et P(x,y) est la probabilité de les observer simultanément (sans notion d'ordre). Si les deux mots x et y

2. en anglais *latent semantic indexing (LSI)*

sont dépendants l'un de l'autre, l'IM a une forte valeur positive. S'ils ne sont pas en relation, la valeur est proche de 0. Si les deux mots sont en distributions complémentaires, la valeur est négative.

Une variante de cette mesure est proposée dans [CH90] sous le nom de *rapport d'information*[3] pour remédier au problème de la symétrie de la mesure de l'information mutuelle. La mesure du rapport d'information est donc une mesure non symétrique qui tient compte de la cooccurrence de deux termes dans une certaine fenêtre ainsi que de l'ordre d'apparition des mots dans cette fenêtre.

– Coefficient de Dice

La mesure du coefficient de dice est une mesure symétrique qui s'exprime comme suit:

$$Dice(x,y) = \frac{2P(x,y)}{P(x)+P(y)}$$

où P(x) et P(y) sont les probabilités d'observer les mots x et y, P(x,y) est la probabilité de les observer simultanément (sans notion d'ordre) et $0 \leq Dice(x,y) \leq 1$. Si les deux mots x et y sont dépendants l'un de l'autre le Coefficient de Dice a une valeur proche de 1. Dans le cas contraire, le Coefficient de Dice a une valeur proche de 0.

– Coefficient d'équivalence

C'est un indice statistique de classement des associations de termes qui est symétrique. Il relativise l'importance d'une association par rapport à la fréquence des deux termes qui le compose:

$$aij = \frac{f_{xy}^2}{f_x f_y} \text{ où}$$

f_{xy}^2 exprime la cooccurrence des termes x et y dans les documents. f_x et f_y expriment les occurrences des termes x et y dans ces même documents.

Les inconvénients de l'approche statistique est qu'elle doit disposer de corpus assez importants pour que les mesures statistiques puissent être validées et trouver des relations intéressantes entre les termes. Elle a l'avantage d'être simple à mettre en oeuvre et de prendre en compte des *graines documentaires* de tailles variables (document, paragraphe, phrase, etc.). Ces relations symétriques représentent des associations ou des dépendances entre les termes.

3. en anglais *information ratio* ou *association ratio*

2.2 Approche linguistique

L'approche linguistique se base sur l'étude de phénomènes langagiers propices à l'expression des relations entre termes. Elle exploite, la plupart du temps, des patrons syntaxiques ou de manière plus relâchée et souple des marqueurs linguistiques comme amorce à l'extraction de relations entre des termes.

Un patron syntaxique s'appuie sur un segment de texte, souvent des syntagmes nominaux simples ou complexes qui font référence à des candidats termes du domaine [Ril93], pour mettre en évidence des relations entre ces constituants selon leur enchâssement [Gre92, Rug97]. Un patron présuppose la présence d'indices linguistiques dans les segments de textes analysés pour déclencher l'extraction des relations sémantiques. Par exemple, l'indice linguistique *tel que* peut être un indice linguistique de la relation d'hyperonymie entre les deux mots DOS et Windows dans la séquence *ordinateur déjà équipé d'un système d'exploitation tel que DOS ou Windows*. Les patrons peuvent être établis en amont par un spécialiste ou un terminologue du domaine ou bien ils peuvent être obtenus à la suite d'une étude linguistique des structures présentes dans un corpus.

Un marqueur linguistique peut être défini comme étant une forme linguistique faisant partie de catégories prédéfinies (grammaticales, lexicales, syntaxiques ou sémantiques) dont l'interprétation définit régulièrement le même rapport de sens entre des termes. La différence entre l'approche s'appuyant sur les patrons syntaxiques et celle utilisant des marqueurs linguistiques est que ces derniers sont exploités comme des déclencheurs pour sélectionner le segment textuel alors que les patrons explicitent les contraintes syntaxiques et spécifient les séquences recherchées.

Une autre approche moins courante que celles exploitant les patrons syntaxiques et les marqueurs est celle basée sur l'utilisation de règles de transformation. Elle permet d'identifier dans un corpus des relations entre termes complexes à partir de connaissances extérieures servant de référence. Généralement, il s'agit d'identifier des variantes de termes contrôlés fournis par un thesaurus ou un vocabulaire du domaine [Jac97].

Dans la suite, nous détaillons quelques travaux utilisant l'approche linguistique.

2.2.1 Utilisation de patrons syntaxiques

Hearst dans [Hea92] propose d'exploiter, dans le cadre de l'acquisition de la relation d'hyperonymie, des patrons syntaxiques basés sur des indices linguistiques pertinents pour cette relation. Il utilise des paires de termes caractérisant la relation d'hyperonymie pour extraire de nouvelles instances par itération. Le principe de cette technique est le suivant:

1. Choisir la relation sémantique, par exemple l'hyperonymie.

2. Fournir une amorce, à partir d'un thésaurus ou manuellement, constituée de couples qui respecte la relation prédéfinie. Par exemple le couple *Angleterre-pays*[4] vérifie la relation d'hyperonymie.

3. Extraire du corpus les phrase qui contiennent les couples précédents (par exemple *"(...) most European countries, especially France, England and Spain"*) .

4. Trouver un environnement commun qui généralise les phrases extraites à l'étape 3 et révèle un candidat patron syntaxique tel que le patron suivant où NP est un syntagme nominal[5]:

   NP_0 *especially* $\{NP_i, \}^* \ (and|or) \ NP_n$

5. Faire valider par un expert les patrons syntaxiques.

6. Utiliser ces patrons pour trouver de nouveaux candidats couples de termes.

7. Faire valider par un expert les candidats couples de termes. Ces derniers sont ajoutés à la liste initiale puis le processus est réitérer à partir de l'étape 3.

Ainsi, à partir de la paire précédente *Angleterre-pays*, quatre schémas syntaxiques sont extraits [Hea92]:

- NP_0 *such as* $\{NP_i, \}^*(and|or)NP_n$
- $NP_0 \ \{, NP_i\}^*$ *or other* NP_n
- $NP_0 \ \{, NP_i\}^*$ *and other* NP_n
- NP_0 *especially* $\{NP_i, \}^* \ (and|or) \ NP_n$

L'avantage de cette approche, exploitée aussi dans [Lin98], est qu'elle est indépendante de la taille du corpus et qu'elle permet de mettre en évidence des relations étiquetées qui constituent une aide adaptée pour un expert en phase de modélisation des connaissances. Mais elle ne permet d'extraire qu'un seul type de relation qui est la relation d'hyperonymie. Un autre inconvénient est la description et la validation manuelle des patrons qui représente une forte contrainte. Cette approche manque d'exhaustivité dans le sens où elle ne permet d'acquérir que des relations du niveau local de la phrase.

Morin dans [Mor99] s'est basé sur les travaux de Hearst [Hea92] en intégrant des mesures statistiques pour le filtrage. Dans cette optique, le système PROMETHEE met en évidence des schémas syntaxiques de manière incrémentale à partir d'une liste de termes

4. *England-country*
5. en anglais *noun phrase*

liés par la relation d'hyperonymie. Une liste initiale de couples de termes qui respectent, dans le corpus, la relation d'hyponymie est d'abord fournie au système, par exemple *calcium EST-UN cation*. Les phrases du corpus qui contiennent les couples de la liste initiale sont extraites. Ainsi, le couple (cation, calcium) sélectionne la phrase *des cations tels que le sodium, le potassium, le calcium et le magnésium peuvent être dosés par une méthode de routine*. Le système examine dans toutes ces phrases, le contexte situé entre les deux éléments d'un couple et propose une liste de patrons syntaxiques à partir des séquences communes qui généralise les phrases sélectionnées. Dans la phrase précédente, le patron extrait est *SN tel que LIST*. Un expert examine ces patrons et ceux qui sont retenus sont regroupés dans des classes de manière à avoir dans une même classe les patrons syntaxiques qui partagent le même environnement. Pour chaque classe, un patron syntaxique qui subsume l'ensemble des patrons de la classe est généré [Mor98]. Les nouveaux patrons sont utilisés pour extraire de nouveaux candidats couples. L'expert retient les couples candidats pertinents qui sont ajoutés à la liste initiale. Le processus est réitéré avec la nouvelle liste et il s'arrête quand aucun nouveau patron n'est retenu. Une mesure statistique est utilisée pour calculer la similarité entre les patrons syntaxiques extraits. Après avoir testé un ensemble de mesures de similarité (Cosinus, Jaccard, etc.), Morin trouve que la mesure de Jaccard [6] est la plus satisfaisante. Par contre, cette mesure ne permet pas d'avoir le niveau de précision souhaité. Morin propose alors une alternative à la mesure de Jaccard qui repose sur l'identification de séquences contiguës au contraire de la mesure de Jaccard qui repose sur l'identification des éléments communs.

La qualité de la liste initiale est très importante pour le bon fonctionnement de cette méthode. Elle dépend entièrement du domaine du corpus traité et elle n'est pas facilement utilisable sur un autre domaine. L'intervention de l'expert est une phase importante de la méthode afin d'éliminer le bruit.

2.2.2 Utilisation de marqueurs

Dans cette partie, nous présentons deux systèmes utilisant des marqueurs linguistiques : le système COATIS et le système SEEK.

2.2.2.1 COATIS

D. Garcia a élaboré un outil d'aide à l'acquisition de connaissances causales (COATIS) qui permet de repérer à l'aide de marqueurs linguistiques les relations de causalité entre termes [Gar98]. Le système COATIS identifie dans les textes, l'expression de situations organisées par des rapports de cause à effet. Dans ce but, il y repère des verbes indicateurs de relations de causalité (environ 300 verbes du français tels que provoquer ou causer mais

6. soit X et Y deux séquences, la similarité de X et Y est $\frac{|X \cap Y|}{|X \cup Y|}$

également des verbes tels que gêner, modifier ou contribuer). L'occurrence d'un verbe, indicateur de la notion de causalité dans une phrase, déclenche l'analyse de celle-ci dans le but de vérifier la présence d'indices apportant une confirmation du caractère causal de la phrase tels que certains suffixes de nom et d'identifier les arguments de la relation qui jouent le rôle de la cause et ceux qui jouent le rôle de l'effet. Les marqueurs exploités par le système sont classés dans un modèle sémantique qui organise le lexique verbal de la notion de causalité en français. Ce modèle rend compte de vingt-cinq relations causales spécifiques (par exemple, créer, empêcher, faciliter ou pousser-à). D'après D. Garcia, ce traitement s'affranchit d'une analyse syntaxique exhaustive des phrases et il ne fait pas appel à des connaissances du domaine étudié (par exemple à des terminologies, à des thesaurus ou à des ontologies).

Un autre système dont l'objectif et d'acquérir des relations de causalité dans des documents textuels est le système KALIPSOS [Naz94]. Ce système de question-réponse procède par la compréhension en profondeur du contenu des documents afin d'accéder à l'information véhiculée dans le texte. Ainsi, il construit la représentation d'un texte sous forme de graphes conceptuels, en s'appuyant sur la structure syntaxique des phrases analysées, un ensemble de marqueurs linguistiques et une description conceptuelle des mots employés. Ce système donne une bonne représentation interne des documents pour un domaine donné. Cependant, la compréhension profonde d'un texte nécessite le recours à des connaissances sémantiques et paradigmatiques qu'on ne peut pas acquérir automatiquement.

2.2.2.2 SEEK

Le système SEEK (pour " Système Expert d'Exploration (K)contextuelle ") permet de détecter des relations statiques entre termes identifiés dans un texte (identification, incompatibilités, mesures, comparaison, inclusion, appartenance, localisation, partie/tout, possession et attribution) [Jou93, Pri00]. Il se base sur le modèle de Grammaire Applicative et Cognitive et la méthode d'exploration contextuelle. La méthodologie d'exploration contextuelle consiste à rechercher dans les textes des indices pertinents qui jouent le rôle de déclencheur. A partir de ces déclencheurs, une exploration contextuelle est effectuée en recherchant dans le contexte du déclencheur des indices complémentaires. SEEK utilise des listes de marqueurs (3300 marqueurs dans 240 listes) et des règles morphologiques (220 règles morphologiques) qui permettent l'inférence de relations sémantiques. Le langage de déclaration des règles d'exploration est du type :

SI <conditions> ALORS <actions> OU <conclusions>.

Les prémisses contiennent un ensemble d'éléments lexicaux combinés et certains schémas possédant des exceptions. Les conditions des règles expriment la co-présence

ou non d'unités linguistiques pertinentes dans le contexte.

Par exemple[7], les unités linguistiques *trouver, stationner, être localisé, être placé,* , etc., sont des déclencheurs de l'exploration contextuelle pour la relation de *localisation*. Ainsi l'existence du verbe *trouver* déclenche l'analyse de la phrase *le sabot de Vénus, une des rares orchidées de nos montagnes, se trouve dans les sous-bois humides* avec la règle d'exploration contextuelle qui correspond à la relation de localisation. Le résultat de cette analyse est l'identification de la relation de localisation entre *sabot de Vénus* et *sous-bois humides* : sabot de Vénus $\rightarrow^{\text{est localisé}}$ sous-bois humides.

Les auteurs dans [LCBD98, Pri00] ont procédé à un couplage de SEEK avec IOTA dans l'objectif de construire un réseau terminologique. Ce réseau contient des relations entre termes détectées par le système IOTA et étiquetées par SEEK.

2.2.3 Utilisation de règles de transformation

Le système FASTR (Filtrage et Acquisition Syntaxique de TERmes) est un outil de repérage de variations de termes à partir d'une liste de termes initiaux. L'auteur part de la constatation qu'un même terme peut apparaître sous des formes variées dans un corpus [Jac97]. FASTR prend en entrée un ensemble de termes simples. Il utilise des méta-règles pour repérer des variantes. Ces méta-règles se rapportent à la syntaxe, à la coordination de termes, à la modification et la substitution, à la permutation, à des règles morphosyntaxiques et sémantico-syntaxiques. Jacquemin répertorie trois familles de variations:

- les variantes syntaxiques : une expansion nominale remplacée par une conjonction (teneur en proreine $\rightarrow$ teneur en eau et proreine).
- les variantes morpho-syntaxiques : la tête ou l'expansion change de partie du discours (précipitation chimique $\rightarrow$ chimie des précipitations).
- les variantes sémantico-syntaxiques : la tête ou l'expansion est remplacée par un élément sémantiquement proche (fabricant d'autos $\rightarrow$ fabricant de voitures).

FASTR peut aussi servir à acquérir de nouveaux termes simples par un processus inverse.

2.3 Approche hybride

L'approche hybride associe des méthodes d'orientation statistique à des méthodes d'orientation linguistique. Elle exploite les avantages de l'approche statistique et de l'approche linguistique et tente de révéler les relations entre termes en repérant les candidats

7. exemple extrait de [Pri00]

termes à partir de schémas syntaxiques et de les filtrer à l'aide de méthodes statistiques [Gre93, Sim00].

2.3.1 SEXTANT

Grefenstette exploite les contextes syntaxiques des mots pour extraire des mots sémantiquement proche [Gre93]. Le système SEXTANT extrait un contexte pour chaque terme sous la forme de structures syntaxiques : adjectifs-noms, noms-noms et verbes-noms. L'hypothèse de cette approche est que les termes qui partagent les mêmes contextes sont des termes *sémantiquement proche* ce qui permet de construire automatiquement un thesaurus. Une analyse syntaxique permet de déterminer les termes qui modifient d'autres termes.

SEXTANT suit les étapes suivantes:

- une analyse morphologique du corpus dans le but de déterminer la catégorie grammaticale de chaque terme.
- une désambiguïsation syntaxique est effectuée pour associer à chaque mot une seule catégorie syntaxique.
- chaque phrase est découpée en syntagmes nominaux et verbaux. Dans la définition des syntagmes nominaux, Grefenstette inclus les prépositions.
- l'extraction des relations structurelles.

  Les syntagmes sont analysés pour extraire les articles, les adjectifs et les substantifs qui modifient les sujets des phrases ainsi que les substantifs connectés par les propositions. L'hypothèse est que les termes qui gouvernent les mêmes verbes et qui sont modifiés par les mêmes termes partagent une même sémantique contextuelle. Cela se produit par exemple:

  - quand un terme est modifié par un adjectif ou un autre terme.
  - quand un terme est modifié par un substantif via une préposition.
  - quand un terme apparaît comme sujet, complément d'objet direct ou complément d'objet indirect d'un verbe.

Nous montrons ci dessous un exemple extrait de [Gre93] pour le mot *CANCER*. La définition ci dessous est obtenue automatiquement à partir du corpus *MED*:

> CANCER:: [255 contexts, frequency rank: 29] MED Relat. lesion, tumor; tissue, disease; carcinoma,. Vbs. advance, disseminate. Exp. cancer patient

(cf. survival time, joint deformity), cancer chemotherapy (cf. survival time, intra-arterial infusion), cancer cell (cff. human cell, year period). Fam. cancer-specific, cancerous.

La structure de la définition est constituée des éléments suivants:

- le mot CANCER possède 255 contextes (attributs) qui sont utilisés pour mesurer sa similarité avec les autres mots. Les attributs sont les adjectifs qui modifient le mot CANCER, les verbes dont il est le sujet ou l'objet et les mots qui le modifient. La similarité entre les mots est calculée en utilisant la mesure de Jaccard.
- le mot CANCER est le $29^{\text{ème}}$ terme le plus fréquent.
- le nom du corpus est MED
- Relt désigne l'ensemble des mots reliés à CANCER c'est à dire les mots statistiquement les plus utilisés dans des contextes similaires et qui sont considérés comme ayant un certain degré de synonymie contextuelle avec le mot *CANCER*.
- Vbs désigne l'ensemble des verbes associés à CANCER c'est à dire les verbes pour lesquels CANCER est un sujet ou un objet au moins trois fois ou plus dans le corpus.
- Exp désigne l'ensemble des expressions selon le patron syntaxique *Substantif Substantif* de longueur 2 dont l'un des Substantifs est CANCER et qui apparaissent trois fois au moins dans le corpus. Chacun de ces composés est lui même accompagné d'un ou plusieurs composés utilisés dans des phrases similaires.
- Fam désigne l'ensemble des termes qui sont de la même famille que CANCER définis comme étant les termes possédant les même premiers caractères que CANCER.

La comparaison de cette définition avec une autre définition obtenue par le système SEXTANT à partir d'un autre corpus montre des similitudes et des différences. Les données associées aux entrées du thesaurus sont très dépendantes du corpus.

2.3.2 IOTA

L'approche adoptée par le système IOTA consiste à construire automatiquement un thesaurus à partir des textes sans une intervention humaine où un thesaurus comporte des classes de termes liés par le contexte dans lequel ils sont utilisés [Bru90]. La force de la liaison entre chaque couple de termes se base sur leur proximité contextuelle exprimée par la distance lexicale dans une même fenêtre textuelle (la phrase par exemple), de leur catégorie grammaticale ainsi que de leur fréquence de cooccurrence.

IOTA commence par une analyse morpho-syntaxique du texte où chaque terme est associé à une (ou plusieurs) catégorie syntaxique. A partir d'une matrice terme-terme où les valeurs de la mesure de la liaison entre termes sont enregistrées, les sous-graphes maximaux complets appelés *cliques* (sous-graphes dont les sommets sont tous interconnectés entre eux) sont extraits et sont censés représenter pour un terme donné ces contextes *forts* trouvés dans le texte [CGH00, Bru90]. Les cliques ont une structure proche des groupes nominaux, cependant les termes qu'elles contiennent dépassent le cadre des groupes nominaux. En effet, deux substantifs sont considérés comme reliés même si, dans une phrase, ils sont séparés par un verbe. Cependant, les liens entre les termes ne sont pas orientés.

2.3.3 ACABIT

Daille utilise une méthode mixte statistico-syntaxique qui effectue des calculs statistiques sur des composés repérés dans le corpus [Dai94]. Il s'agit de repérer des candidats-termes à partir de schémas syntaxiques puis de les filtrer à l'aide de méthodes statistiques. Elle établit une liste de types élémentaires de composés nominaux du domaine de télécommunication ainsi qu'une topologie des moyens dont dispose la langue pour engendrer des formes complexes à partir de formes élémentaires : la surcomposition, la modification et la coordination. Parmi ces divers types de composés, elle ne retient que certains composés de longueur 2 : *Nom Nom*, *Nom Adjectif*, et *Nom Préposition Nom* avec quelques possibilités d'insertion de modificateurs (Adjectif, Nom ou Adverbe) au sein des schémas retenus [Dai02].

ACABIT commence par cerner les spécifications linguistiques pour les termes (majoritairement des unités lexicales complexes de type nominal) et ainsi d'établir des filtres linguistiques. Ensuite, il sélectionne des candidats potentiels d'un corpus préalablement étiqueté et lemmatisé en utilisant les spécifications linguistiques exprimées en terme de patrons morpho-syntaxiques. Les termes extraits sont classés selon la mesure du coefficient de vraisemblance. ACABIT peut être appliqué aussi à des corpus monolingues ainsi qu'à des corpus bilingues alignés phrases à phrases et propose une liste de termes accompagnés de leur traduction.

2.3.4 Xtract

Smadja utilise une technique à base de fenêtre textuelle pour repérer des collocations en corpus par combinaison d'un filtrage statistique associé à une analyse linguistique [Sma93b, Sma93a]. Le filtrage statistique se base sur les fréquences des cooccurrences des mots qui composent des collocations. Cette approche utilisée dans le système Xtrac

repose sur les deux hypothèses suivantes pour l'acquisition de collocations:

– Les mots dans une collocation apparaissent ensemble plus fréquemment que par hasard.

– Les mots apparaissent dans une fenêtre de plus ou mois 5 mots correspondant à des contraintes syntaxiques particulières.

Xtract commence par repérer d'abord, à partir d'un corpus étiqueté les paires de mots statistiquement significatives. Les composants de ces paires peuvent être séparés l'un de l'autre par un maximum de 5 mots. Ensuite, il extrait, à partir des paires trouvées, des unités significatives plus longues. Enfin, il élimine des paires trouvées à la première étape toutes les combinaisons dans lesquelles les composants ne conservent pas toujours la même relation syntaxique.

2.4 Les sources d'acquisition de connaissances

2.4.1 Les techniques orientées utilisateurs

Il s'agit de l'acquisition de connaissances à partir du bouclage de pertinence ou bien la construction d'un pseudothesaurus dédié aux besoins d'un utilisateur particulier. Ce pseudothesaurus est construit à partir des requêtes d'un utilisateur. Guntzer et al dans [GJSS89] ont suggéré un système expert pour découvrir des relations entre termes en se basant sur des observations du comportement de l'utilisateur c'est à dire l'utilisation du feedback implicite et feedback explicite. Un ensemble de règles est établi pour vérifier les relations trouvées. Deux termes liés par une conjonction ou une disjonction dans une requête seront liés par une relation d'association dans le thesaurus en construction. Par exemple, si un utilisateur combine deux termes avec l'opérateur *OR* dans sa requête, ces deux termes sont probablement synonymes, c'est un cas de feedback implicite. Si l'utilisateur doit spécifier la relation entre deux termes qu'il a utilisé dans sa requête alors c'est le cas du feedback explicite. Le thesaurus obtenu permet de bien répondre aux besoins d'un utilisateur dans le cas d'une recherche ou un filtrage d'information mais ne peut pas être réutilisé par un autre utilisateur, sauf si le besoin d'information des deux utilisateurs est identique.

Naulleau se base sur un profil prédéfini du besoin de l'utilisateur pour extraire des syntagmes nominaux qualifiés alors de pertinents [Nau98, Nau99]. Un profil s'exprime par un ensemble de syntagmes nominaux existants donc un ensemble d'exemples des syntagmes que l'utilisateur juge pertinent ou non pertinent. L'approche consiste donc à prédire la pertinence en extrapolant les formes possibles à partir des formes observées. Un processus d'apprentissage essaye à partir des exemples des syntagmes de trouver les caractéristiques

et propriétés linguistiques de ces exemples. Naulleau s'est limité aux dépendances qui correspondent à des schémas du type *NOM Adjectif*, *NOM NOM* et *Nom Préposition NOM*. Même si l'approche n'a pas été testée dans le cadre de la RI, elle reste toutefois originale dans son principe car elle se base sur le besoin de l'utilisateur alors que la plupart des travaux se basent sur le contenu textuel ou l'intervention d'un expert linguistique.

2.4.2 Les techniques orientées ressources lexicales

2.4.2.1 Les dictionnaires

Les outils utilisant le dictionnaire repèrent des *variantes sémantiques* de termes et des relations entre eux dans des textes en établissant des régularités dans les définitions des termes [KF93, Ahl88]. Ainsi, le système SynoTerm est basé sur l'exploitation d'informations sémantiques extraites de ressources lexicales d'un dictionnaire de la langue générale [HNG98]. A partir de ces informations, des règles déduisent des relations entre des termes complexes validés ensuite par un terminologue. Les auteurs ont utilisé les liens de synonymie ou de *quasi-synonymie* fournis par un dictionnaire de la langue générale (Le Robert). Le principe consiste à considérer comme *synonymes* les termes complexes d'un corpus dont un des éléments au moins est donné comme synonyme d'un des éléments de l'autre terme complexe par le dictionnaire de la langue générale, et dont les autres éléments sont identiques ou synonymes. Ainsi, matériel et équipement étant donné comme synonymes par Le Robert, une relation de synonymie entre *matériel électrique* et *équipement électrique* est proposée par SynoTerm.

2.4.2.2 Les thesaurus

Un thesaurus peut être définit comme un réseau de références lexicales. Les thésaurus ont été utilisés comme source de connaissances dans le domaine de la RI ou bien combinés avec une base de connaissances spécifique à un domaine [Voo93, BS96].

Un thesaurus construit manuellement et fréquemment utilisé dans le domaine de la RI est Wordnet développé à l'université de Princeton et supposé traiter tous les mots anglais quels que soient leurs contextes. Sa conception est inspirée des théories de psycholinguistique sur la mémoire lexicale humaine (qu'est ce qui nous vient à l'esprit quand nous entendons le mot X). La version 1.6 de Wordnet [8] contient 109 377 classes de synonymes (synsets) regroupant 107 930 noms, 10 806 verbes, 21 365 adjectifs et 4 583 adverbes. Les substantifs, les verbes et les adjectifs sont organisés dans des ensembles appelés *synsets* dont chacun représente un concept.

8. source http://www.cogsci.princeton.edu/ wn/

Wordnet ressemble plus à un thesaurus qu'à un dictionnaire parce qu'il essaye d'organiser l'information lexicale en terme de sens des mots plutôt qu'en forme du mot. La conception de Wordnet pose un problème majeur dans le contexte de son utilisation pour la recherche d'information. Ce problème réside dans la redondance des entrées : si un mot x et un mot y sont synonymes alors les deux sens doivent exister; un pour x et l'autre pour y. Un mot peut alors appartenir à plusieurs synsets à la fois avec un sens différent pour chaque occurrence de ce mot [Voo93].

Les relations utilisées dans Wordnet sont la synonymie, l'antonymie, l'hyponymie, l'hypernymie, la meronymie et la holonymie [MBF$^+$90]:

– La synonymie:

  La synonymie est la relation la plus importante dans Wordnet. Deux expressions sont synonymes si la substitution de l'une par l'autre ne change rien au sens de la phrase dans laquelle la substitution est effectuée.

– L'antonymie

  L'antonymie d'un mot X est non-X mais ce n'est pas souvent le cas. Par exemple pour *riche* on a *pauvre*.

$$\text{grand} \Longleftrightarrow_{antonyme} \text{petit}$$

– L'hyponymie [9] et l'hyperonymie [10]:

  On peut ajouter aussi à cette catégorie les relations *est-un-type-de*, *est-un*, *espèce-genre*, etc.

  Un concept représenté par *x* est un hyponyme d'un concept représenté par *y* si un *x* est considéré comme un *type de y*.

  C'est une relation asymétrique et transitive. Cette représentation est fréquemment utilisée dans la RI dans les systèmes dits d'héritage. L'hyperonymie est la relation inverse de l'hyponymie.

$$\text{homme} \Longrightarrow_{hyponyme} \text{humain}$$

$$\text{humain} \Longrightarrow_{hyperonyme} \text{homme}$$

9. en anglais *hyponymy*
10. en anglais *hyperonymy*

– La méronymie[11] et l'holonymie [12]

  On peut ajouter aussi à cette catégorie les relations *possède* et *Est-une-Partie-De*.

  Un concept représenté par *x* est un *méronyme* d'un concept représenté par *y* si *x* est une *partie de* ou *fait partie de y*. C'est une relation transitive et asymétrique. L'holonymie est la relation inverse de la méronymie.

$$\text{ordinateur} \Longrightarrow_{holonyme} \text{processeur}$$

$$\text{processeur} \Longrightarrow_{meronyme} \text{ordinateur}$$

Wordnet est le thesaurus qui peut être qualifié du plus général et complet jusqu'à présent qui couvre plusieurs domaines mais il n'y a pas de séparation claire entre ces domaines ce qui provoque des ambiguïtés sémantiques [Nie97]. Il existe plusieurs expérimentation pour l'utilisation de Wordnet dans un SRI. Une base générale de connaissances terminologiques multilingues européenne a vu le jour en Janvier 1999 sous le nom de EuroWordNet [13].

Voorhees dans [Voo93, Voo94] utilise Wordnet comme outil de désambiguïsation pour l'expansion des requêtes. L'hypothèse sur la quelle se base ce travail est la suivante:

> Pour un ensemble de mots occurrant ensemble dans un contexte, le sens approprié d'un mot est déterminé par les autres mots même si ce mot est ambigu [14].

L'hypothèse sur laquelle se base Voorhees est qu'un ensemble de mots, même s'ils ont plusieurs sens, ils ont un sens unique s'ils sont regroupés ensemble dans le même contexte. Pour déterminer le sens d'un mot dans une phrase, l'auteur utilise la distance sémantique entre chaque synset possible du mot avec les synsets des autres mots de la phrase. Le synset qui est le plus proche des autres mots de la phrase est choisi. D'après cette hypothèse, le sens d'un mot dans un contexte est proche des sens des autres mots qui se trouvent dans le même contexte. La distance est calculée selon le nombre de liens entre les synsets. La désambiguïsation consiste d'abord à chercher pour chaque mot les synsets où il occurre. Si le mot n'existe dans aucun ensemble de WordNet, des transformations morphologiques simples sont effectuées sur le mot et les différentes variantes du mot sont utilisées pour une deuxième recherche.

11. en anglais *meronymy*
12. en anglais *holonymy*
13. http://www.hum.uva.nl/ ewn/
14. A set of words occurring together in context will determine appropriate senses for one another despite each individual word being multiply ambiguous

En comparant une indexation avec des mots sans désambiguïsation et une indexation avec des mots désambiguïsés avec Wordnet, les résultats montrent que la désambiguïsation en utilisant Wordnet détériore les performances du SRI [Voo98, Voo99].

Border et Song dans [BS96] utilisent deux bases de connaissances pour l'expansion des requêtes : une base de connaissances spécifique à un domaine et la base de connaissances générale Wordnet.

La base de connaissances spécifique à un domaine est construite automatiquement en se basant sur les deux hypothèses suivantes :

- les termes qui ont une fréquence élevée sont des termes généraux et ceux qui ont une fréquence faible sont des termes spécifiques.
- si la fonction de densité de deux termes est la même, le terme qui a une fréquence faible devient un descendant de celui qui a une fréquence plus élevée.

Le processus de construction de la base de connaissances spécifique à un domaine est la suivante:

1. tous les termes de la collection sont groupés dans des classes en se basant sur leur fréquence où les termes qui ont une fréquence élevée sont des racines et ceux qui ont une fréquence faible sont les feuilles.
2. des liens hiérarchiques de types IS-A sont calculés pour déterminer les descendants.
3. les termes qui n'ont pas de descendants sont des feuilles de la hiérarchie.
4. les étapes 2 et 3 sont itérer.

On obtient alors un réseau sémantique des termes qui est spécifique au domaine traité. La différence entre ce réseau sémantique et Wordnet est que Wordnet utilise plusieurs types de liens par contre la base de connaissances spécifique au domaine utilise un seul type de lien.

Les auteurs différencient entre deux stratégies de recherche:

- une stratégie de recherche isolée où on ne tient pas compte des relations entre les termes dans la requête qui peuvent exister dans la base de connaissances.

  On distingue trois types de recherches isolées:

  - la recherche des termes génériques où on ne cherche que les termes qui sont plus haut dans la hiérarchie dans l'objectif d'améliorer le rappel.
  - la recherche des termes spécifiques où on ne cherche que les termes feuilles dans la hiérarchie dans l'objectif d'améliorer la précision.

  - la recherche dans les deux sens de la hiérarchie qui semble la plus complète.

- une stratégie de recherche corrélée où on tient compte des relations entre les termes dans la requête. Cette stratégie consiste à chercher dans la base de connaissances si une relation existe entre deux termes de la requête.

Les stratégies de combinaison possibles entre les deux bases de connaissances sont les suivantes:

- l'union avec ou sans la pondération des termes communs aux deux bases de connaissances.

  La pondération des termes consiste à tenir compte du poids d'un terme lors de l'expansion et additionner les poids d'un terme dans le cas où le terme appartiendrait aux deux bases de connaissances ce qui donne un poids plus important pour les termes qui proviennent des deux bases de connaissances. Un union sans pondération ignore la possibilité que certains termes sont plus importants que d'autres. Le poids des termes ajoutés à la requête est calculé après l'expansion de la requête.

- le chaînage.

  Il s'agit d'utiliser une base de connaissances comme première source pour étendre la requête originale et le résultat est lui-même étendu avec la deuxième base de connaissances.

- la coordination.

  Une première source (la base de connaissances spécifiques au domaine) est utilisée pour étendre la requête originale et une deuxième source (la base de connaissances générales Wordnet) est utilisée pour assister dans la coordination des termes de la requête originale à ceux de la première source.

Les auteurs ont utilisé la collection de documents ADI [CY92]. Cette collection est, à notre avis, trop petite pour évaluer l'approche proposée. En effet, la collection ADI contient uniquement 82 documents et 822 termes. A l'image des travaux de Voorhees, les expérimentations dans [BS96] n'ont pas donné des résultats satisfaisants.

Nie dans [Nie97] a utilisé la base terminologique BTQ (La Banque de terminologie du Québec (BTQ) développée par l'Office de la langue française du Québec, avec plus de 800000 fiches terminologiques, plus de 3 millions de termes techniques en français et en anglais et plus de 2160 champs (160 spécialisés et 2000 généraux))[15]. Dans cette base, un terme est défini par une fiche contenant toute l'information sur le terme : son

15. http://www.olf.gouv.qc.ca/index.html

identifiant, son domaine, ses synonymes, les termes reliés (hypernyme, hoponyme, etc.), etc. Si un terme a plusieurs sens alors il a plusieurs fiches. Nie identifie la partie de la base terminologique utile pour la RI comme étant le terme lui-même et les termes qui lui sont reliés suivant les relations de synonymie, hyperonymie, holonymie, hyponymie et méronymie que Nie pondère respectivement avec les poids suivant : 0.08, 0.16, 0.32, 0.08 et 0.89. Ces même poids ont été attribués aussi à ces mêmes relations dans Wordnet.

La valeur élevée de la pertinence des relations de méronymie est due, d'après Nie, au fait qu'il n'y a pas beaucoup de relation de méronymie dans Wordnet ce qui fait que les termes liés par une relation de méronymie sont fortement pertinents. La faible valeur des relations de synonymie est expliquée par les ambiguïtés que peut engendrer cette relation.

Les résultats des expériences de Nie ont montré le faible impact de BTQ, en comparaison avec Wordnet, sur les performances d'un SRI dû au fait de la rareté des relations entre termes dans BTQ.

L'inconvénient des techniques utilisant des thesaurus, ainsi que des ressources lexicales, est que l'exploitation de ces derniers et leur adaptation pour une tâche de recherche d'information ne sont pas claires. L'utilisation d'un thesaurus dans un système de recherche d'information nécessite une adaptation de sa structure à un besoin de recherche d'information et l'utilisation d'une stratégie de recherche d'information adéquate. Les types de relations entre termes utiles pour un besoin de recherche d'information ne sont pas identifiés.

La généralité des thesaurus tel que WordNet pose un problème dans un contexte de recherche d'information. En effet, avoir à disposition tous les sens d'un mot qui dans un contexte particulier n'en conserve qu'un seul pose des problèmes de désambiguïsation [Voo93]. De même, les mots dans un domaine technique possèdent souvent un sens particulier qui ne se retrouve pas forcément dans la langue dite générale.

L'avantage de l'utilisation d'un thesaurus est l'existence d'un nombre important et varié de relations typées. Un autre avantage est qu'il existe des thesaurus tel que BTQ qui sont composé de sous-parties relatives à des domaines particuliers.

2.5 La classification des connaissances textuelles

2.5.1 Les techniques statistiques de classification

La classification des termes consiste généralement à mettre en évidence les relations hiérarchiques entre termes à l'aide de méthodes statistiques. Ces méthodes peuvent se baser uniquement sur les classifications des termes ou bien sur les classifications des documents.

La relation de hiérarchie met en relation un terme générique et un terme spécifique (relation hyperonymie-hyponymie). Deux approches pour construire une hiérarchie des

termes sont distinguées [Sim00]:

– la définition d'un degré de hiérarchie.

  Le degré de hiérarchie d'un terme indique entre deux termes lequel est plus générique et lequel est plus spécifique à partir de la fréquence d'utilisation du terme. Un terme dont la fréquence est élevée est considéré comme étant un terme générique. A l'inverse, si un terme est peu fréquent alors il sera considéré comme spécifique.

– l'approche ensembliste.

  Fluhr dans [Flu77] propose une méthode de construction de relations entre termes fondée sur des opérations ensemblistes. Soit CS_t l'ensemble des termes qui cooccurrent avec t appelé champ sémantique de t. La mesure de Tanimoto pour la distance entre deux termes t_1 et t_2 est appliquée:

$$d(t_1, t_2) = |\frac{card(CS_{t_1} \cap CS_{t_2})}{card(CS_{t_1} \cup CS_{t_2})} - 1|.$$

  Si $d(t_1, t_2)$ est inférieur à un certain seuil s_1 alors t_1 et t_2 sont déclarés synonymes ou quasi-synonymes. Si $d(t_1, t_2)$ est comprise entre s_1 et s_2 et $card(CS_{t_1} \cap CS_{t_2}) << card(CS_{t_1})$ alors t_1 et t_2 sont déclarés liés par une relation associative. Si $card(CS_{t_1} \cap CS_{t_2}) << card(CS_{t_1} \cup CS_{t_2})$ avec $card(CS_{t_1}) << card(CS_{t_2})$ alors t_2 est déclaré générique de t_1.

Crouch et al, dans [Cro90], proposent une méthode de construction automatique d'un thesaurus fondé sur la valeur de discrimination d'un terme [Sal89]. La classification des documents consiste à regrouper les documents dans des clusters et de générer un thésaurus à partir des termes de faibles fréquences contenus dans les clusters. Une classification hiérarchique des documents est alors construite en utilisant l'algorithme de classification de lien complet, en utilisant une mesure de similarité qui considère que deux documents sont similaires s'ils partagent un grand nombre de termes de fréquences moyenne. Les classes sont générées en coupant les branches de la classification hiérarchique selon une valeur donnée de l'indice de hiérarchie et une valeur donnée du nombre de documents maximum d'une classe. Une classe de thesaurus est définit comme l'ensemble des termes discriminants communs à tous les documents d'une classe. Le résultat de cette approche sont des ensembles de termes dont l'intersection est vide.

L'inconvénient de cette approche est que la classification des documents est un processus plus coûteux que la classification des termes principalement dans le cas où le corpus serait dynamique. Un deuxième inconvénient est que les critères et les paramètres utilisés pour classer les documents sont difficiles à déterminer et particulièrement le seuil de similarité. En effet, un seuil très haut entraîne des classes qui contiennent un très petit nombre

de termes et, inversement, un seuil faible entraîne un nombre important de termes dans la même classe.

2.5.2 Les techniques linguistiques d'acquisition et de classification

LEXTER est un exemple de système utilisant une technique linguistique pour la construction d'une base de connaissances d'un domaine [Bou92, Ass98]. Il a été développé par Didier Bourigault à la Direction des Études et Recherches de EDF dans le cadre d'un projet de gestion de la documentation technique de l'entreprise et a pour objectif l'extraction de syntagmes (nominaux et adjectivaux) à partir de corpus spécialisés, dans une perspective d'acquisition terminologique.

LEXTER reçoit en entrée un corpus, en français, de textes techniques portant sur un domaine quelconque. Il propose comme résultat un réseau terminologique, c'est-à-dire un ensemble de groupes nominaux susceptibles d'être des termes complexes du domaine, organisé en réseau terminologique à l'aide de relations grammaticales en fonction de critères syntaxiques et non conceptuels. Ces relations relient chaque candidat terme à sa tête, à son expansion et aux candidats termes dont il est lui même tête ou expansion. Dans [BC99], Bourigault et al. donne l'exemple du candidat terme *modèle conceptuel* qui est la tête des candidats termes *modèle conceptuel de l'application, modèle conceptuel des données* et *modèle conceptuel des traitements*. Il est aussi l'expansion des candidats termes *construction d'un modèle conceptuel* et *validation d'un modèle conceptuel*.

Ce réseau de candidats termes, avec le corpus à partir duquel il a été extrait, est soumis, sous la forme d'un hypertexte, à un expert du domaine ou à un terminologue à des fins de validation.

Le système LEXTER est une chaîne logicielle composée de 3 maillons.

- Le module Catégorisation.

  Il s'agit d'un étiqueteur grammatical qui a pour fonction de déterminer pour chaque mot du texte sa catégorie grammaticale (nom, adjectif, verbe, participe passé, etc.). Il a recours à un dictionnaire informatique du français de taille importante, ainsi qu'à des règles de désambiguïsation, qui choisissent la bonne catégorie grammaticale d'un mot en fonction de son contexte.

- Le module extraction.

  LEXTER effectue une analyse grammaticale des textes catégorisés pour en extraire des groupes nominaux en fonction de leur position syntaxique et leur structure grammaticale qui sont désignés comme étant des unités terminologiques potentielles. Cette analyse grammaticale s'effectue en deux temps : un découpage puis une décomposition.

Pendant l'étape de découpage, LEXTER recherche des frontières entre groupes nominaux (les verbes conjugués, les articles indéfinis, les conjonctions, certains participes, etc.). Nous montrons ci dessous un exemple extrait de [BC99] où les frontières sont *assure, de sa* et *après une*.

Exemple 1 *le circuit d'aspersion de l'enceinte de confinement assure le maintien de sa température nominale de fonctionnement après une augmentation de pression.*

les candidats termes extraits sont alors:

- circuit d'aspersion de l'enceinte de confinement
- maintien
- température nominale de fonctionnement
- augmentation de pression

Pendant l'étape de décomposition, LEXTER analyse les groupes ainsi délimités pour, d'une part, en donner une décomposition en tête et expansion, et, d'autre part, en extraire des sous-groupes qui constituent eux-aussi de bons candidats termes. Un groupe nominal contient une seule tête, généralement un substantif (*noun*), qui est le concept central de l'expression et une expansion, appelée aussi argument ou modifieur, qui est un modifieur ajoutant de la précision à la tête.

A partir de la décomposition en tête et expansion de chacun des termes repérés, LEXTER organise la liste des termes candidats en un réseau, et effectue éventuellement un filtrage statistique.

- Le module Navigation

  Ce module est un générateur d'hypertexte qui organise le réseau des termes candidats et le corpus de textes dont il est issu sous la forme d'un réseau hypertextuel. Le terminologue ou l'expert du domaine qui a la charge de valider la terminologie proposée par LEXTER peut naviguer au sein du réseau des termes candidats, ainsi que des textes vers les termes candidats qui y ont été détectés.

Dans la littérature, l'utilisation de LEXTER est souvent accompagnée par d'autres systèmes, FASTR par exemple [Jac97], qui exploitent les résultats de LEXTER.

Les auteurs dans [IS99] proposent une méthode de classification fondée sur des relations linguistiques de dépendance entre les termes. Cette technique, appelée par l'auteur *condensation de données textuelles*, a les particularités suivantes:

- elle sous-entend, en amont de la phase de condensation, une analyse linguistique (morpho-syntaxique) des textes collectés afin de sélectionner les unités textuelles porteuses de sens. Ces unités sont identifiées comme étant les syntagmes nominaux.

– une analyse grammaticale. Sur le plan grammatical, un syntagme nominal (SN) se décompose en une tête et une suite d'expansion.

  Exemple 2 *root hair curling*

  Dans cet exemple, *curling* est le centre et *root hair* ses expansions.

Les auteurs identifient un certain nombre de relations de variation entre les unités textuelles qui se répartissent en deux axes:

– axe grammatical

  L'hypothèse est que les relations de variations impliquant un changement de la tête indique plus fortement le déplacement de thèmes (ou de sujets) au sein des objets d'étude.

  Exemple 3 *root hair curling et root hair deformation*

  Dans l'exemple précédent, on a un changement du centre *curling* en *deformation* alors que dans l'exemple suivant on a changement du modificateur *curling* en *deformation* qui n'est pas pris en compte.

  Exemple 4 *curling root hair et deformation root hair*

– axe "symétrie vs antisymétrie".

  Il s'agit de la modification (antisymétrie) ou non (symétrie) de la longueur d'un syntagme nominal donc du nombre de termes qui le composent .

La combinaison de ces deux axes donne lieu à quatre relations:

– une relation entre des syntagmes nominaux de longueur différente et partageant le même centre. Elle se décompose en deux sous-relations : l'insertion et l'expansion à gauche.

– une relation entre des syntagmes nominaux de longueur différente ayant des centres différents : l'expansion à droite.

– une relation de substitution de centre qui met en relation des syntagmes nominaux partageant les mêmes modificateurs et ayant la même longueur.

– une relation, qui modélise la substitution de modificateur, relie des syntagmes nominaux de même longueur.

Ces relations sont représentées sous forme d'un graphe. Une fonction de pondération permet d'indiquer la proportion d'une catégorie vis-à-vis d'une autre et d'empêcher que les informations portées par une catégorie minoritaire ne soient noyées. Le graphe initial est donc composé en composantes connexes moyennant un algorithme des sous-graphes complets (dont tous les éléments au sein d'une même composante partagent le même centre) avant d'agglomérer celles-ci en classes pour le besoin de classification.

2.5.3 Les techniques d'extraction d'information

Dans des domaines spécialisés et avec des informations plus spécifiques, les systèmes *d'extraction d'information* utilisent des structures syntaxiques locales pour extraire des relations propres au domaine d'étude. Ce sont donc des systèmes qui produisent une représentation de l'information textuelle pertinente dans le domaine en question, pour une application particulière. Ils obtiennent de bons résultats dans un domaine spécifique, mais sont relativement difficiles à adapter à un nouveau domaine. Ce type de système a été initié par la série des *Message Understanding Conferences* ou MUC qui est un programme d'étude visant à évaluer et comparer les performances d'outils d'extraction d'information des laboratoires de recherches soutenus par l'équivalent aux Etats-Unis du ministère français de la défense. Dans MUC-4 par exemple, la tâche à effectuer consistait à remplir les champs de patrons de description d'événements terroristes à partir de dépêches de presse.

L'extraction d'information peut être définie comme étant la tâche qui consiste à identifier de l'information bien précise d'un texte en langue naturelle et à la représenter sous forme structurée dans le but de réduire l'effort intellectuel humain. En effet, l'extraction d'information s'avère très pratique dans des applications où des opérations d'extraction sont quotidiennement effectuées à la main. Les entrées et les sorties d'un système d'extraction de connaissances sont définies a priori ce qui facilite l'évaluation des différents systèmes et approches [Mcc96].

Dans la suite, nous présentons deux systèmes d'extraction d'information : le système KEP et le système AutoSlog.

2.5.3.1 Le système KEP

Le système KEP [BHR96] se base sur la description d'un concept dans le domaine étudié. Les termes qui sont reliés à ce concept apparaissent dans des expressions où l'on est capable de reconnaître des locutions : *est un, définit comme, etc.*. Dans le domaine des sciences informatiques, il distingue trois relations. Soit, l'exemple suivant extrait de [BHR96] où le texte à analyser est :

> *We define a sorting routine to be a function which orders a list of items according to some criterion. An example of a sorting criterion is alphabetical*

order. The bubble sort and the quick sort are well-known examples of sorting routines. The four elements of a sorting routine are the input list, the output list, the sorting criterion, and the sorting algorithm.

Les relations de définition extraites sont:

- relation 1:
  - *Concept:* sorting criterion;
  - $Exemple_0$: alphabetical order.

  C'est l'expression *An example of* qui permet d'extraire cette relation.

- relation 2:
  - *Concept:* sorting routines;
  - $Exemple_0$: the bubble sort.
  - $Exemple_1$: the quick sort.

  Cette relation est extraite moyennant l'expression *are well-known examples*.

- relation 3:
  - *Concept:* sorting routine;
  - $Exemple_0$: input list.
  - $Exemple_1$: the output list.
  - $Exemple_2$: the sorting criterion.
  - $Exemple_3$: the sorting algorithm.

  L'expression *The four elements of* déclenche l'extraction de cette relation.

Le processus se déroule en deux étapes. Pendant la première étape, les expressions à extraire, sont déterminées manuellement. Pour chacune de ces expressions, un ensemble de déclencheurs est défini manuellement. Par exemple, les expressions *est un, définit comme, n'est pas un, etc.* sont des déclencheurs de la relation de définition. Les déclencheurs sont définis en deux catégories : les déclencheurs positifs et les déclencheurs négatifs. La deuxième étape consiste en une analyse linguistique du texte en utilisant les classes des déclencheurs positifs et des déclencheurs négatifs pour extraire les concepts et les relations entre concepts.

L'inconvénient de cette technique est qu'elle est dédiée à un domaine particulier où la définition des déclencheurs joue un rôle important et influence énormément les performances du système.

2.5.3.2 Le système AutoSlog

AutoSlog pratique un apprentissage supervisé [Ril93]. Il prend en entrée des textes d'entraînement et les réponses attendues pour ces textes, ainsi que des heuristiques permettant de reconnaître certaines constructions, comme par exemple *sujet verbe-au-passif*. En sortie, Autoslog fournit des entrées du dictionnaire conceptuel cible. Ainsi, sachant que pour la phrase *the diplomat was kidnapped*, il devra associer le patron, *KIDNAPPING (victim : diplomat)*, AutoSlog va défnir le mot *kidnapped* comme déclencheur du patron *KIDNAPPING* avec comme condition qu'il soit un verbe au passif et comme effet de remplir le champ *victim* par le sujet de la phrase. Après une phase d'acquisition, les entrées du dictionnaire extrait sont filtrées manuellement pour supprimer le bruit.

La conception d'AutoSlog se repose sur l'observation faite à partir du corpus d'entraînement et qui se traduit par le fait que le contexte syntaxique à proximité de termes cibles est un support pertinent pour mettre en relation ces termes et que ce contexte peut être réduit à une phrase ou un paragraphe. L'acquisition des noeuds de concepts repose sur un corpus d'entraînement qui identifie les syntagmes nominaux qui représentent des informations importantes. Ces syntagmes sont manuellement identifiés et typés.

Soit l'exemple du noeud de concept appelé *kidnap-passive* extrait des informations d'événements relatives au kidnapping. Ce noeud de concept est activé par le mot *kidnapped* et a des conditions qui lui permettent d'être activé seulement dans le contexte de construction passive : des expressions tel que *was kidnapped* ou *were kidnapped*. Le dictionnaire contient aussi un second noeud de concept appelé *kidnap-active* qui est aussi activé par le mot *kidnapped* mais qui a des conditions qui lui permettent d'être activer seulement dans un contexte de construction active tel que *terrorist kidnapped the mayor*. Chaque définition d'un noeud de concept contient un ensemble de champs. Nous montrons ci dessous un exemple de noeud de concept extrait de [Ril93]:

Exemple

> in la oroya, junin departement, in the central peruvian mountain range,, public building were bombed and a car-bomb was denoted.

Le système détecte dans la séquence le verbe *bombed* à la voix passive. Le noeud de concept qui peut être activé par exemple par le champ *public buildings* est le suivant:

> **Name:** target-subject-passive-verb-bombed
>
> **Trigger:** bombed
>
> **Variable Slots:** (target(*S*1))
>
> **Contraints:** (class phys-target *S*)
>
> **Constant Slots:** (type bombing)
>
> **Enabling Conditions:** ((passive))

Le verbe *bombed* active le noeud de concept *target-subject-passive-verb-bombed* ce qui permet d'extraire des informations pour remplir les différents champs du noeud.

AutoSlog exige une bonne définition de chaque noeud de concept. Ce procédé de recherche de schémas est souvent utilisé dans le cadre d'analyse de textes de presse où la première apparition du terme (ici concept) le place dans son contexte.

2.6 Conclusion

L'acquisition de connaissances consiste à extraire des connaissances sous la forme de termes et de relations entre les termes. Elle les normalise et les organise dans le but de les utiliser au cours de l'indexation ou l'interrogation de grands corpus en langue naturelle. Dans le tableau **??**, nous résumons les avantages et inconvénients des approches d'acquisition ainsi que les différentes techniques présentées dans ce chapitre.

On remarque que dans les derniers travaux faits dans ce domaine [Voo98, Voo99, BS96, Nie97], les auteurs se sont orientés vers l'utilisation des connaissances dans le but d'élargir les requêtes et plus spécialement les connaissances qui traitent des relations entre les termes. Cela s'explique par les observations suivantes:

- la représentation basée sur un mot simple n'est pas assez précise pour le contenu d'un document ou d'une requête à cause des ambiguïtés des mots d'où l'utilisation des thésaurus [Voo93, BS96, Nie97].

- un document pertinent ne partage pas toujours les mêmes mots-clés avec la requête d'où l'application de techniques qui cherchent le contexte d'utilisation des termes dans les documents [BRC99, Gre93].

Dans les travaux présentés dans ce chapitre, nous pouvons distinguer deux approches pour l'acquisition de connaissances textuelles que Morin dans [Mor99] qualifie d'approche ascendante et approche descendante.

L'approche ascendante cherche à faire émerger des informations du corpus textuel sans faire un *a priori* sur les informations à extraire [CH90, Gre93], c'est généralement le cas des techniques statistiques. L'approche descendante part de connaissances plus ou moins riches sur le fonctionnement de la langue. Les outils utilisant cette approche s'appuient souvent sur une analyse linguistique pour mettre en évidence des relations entre des termes.

On peut étudier le contexte des termes selon deux approches :

- On peut s'intéresser aux contextes dans lesquels apparaissent les termes (la phrase, les mots précédant ou suivant les termes, etc.), et on rapproche ceux qui partagent les mêmes contextes [Gre93];

Approche	*Avantages*	*Inconvénients*
Statistique	mise en oeuvre simple mise en évidence de relations exhaustives application sur des grandes unités textuelles	non typage de relations
Linguistique	typage de relations	relations entre termes du niveau local description manuelle des patrons
Hybride	exploitation des avantages de l'approche statistique et l'approche linguistique	
Extraction de connaissances	obtention de bons résultats dans un domaine	rôle important de la phase manuelle non transportable à un autre domaine
Les techniques orientées utilisateur	bien répondre au besoin d'un utilisateur	non applicable pour satisfaire le besoin d'un autre utilisateur
Les techniques orientées ressources lexicales	grand nombre de relation de relations	structure difficilement exploitable stratégie de recherche adéquate

TAB. 2.1 – *Avantages et inconvéients des approches et techniques d'acquisition de connaissances*

– On peut également s'intéresser à la structure même des termes et rapprocher ceux qui ont une structure semblable (par exemple, tous les termes dont un des éléments est le même) [Bou92].

Les systèmes d'extraction d'information ont eu des bons résultats dans des tâches précises mais souvent ils sont critiqués parce qu'ils dépendent d'un dictionnaire spécifique

d’un domaine particulier donc non transportable à un autre domaine [Ril93].

Comme Hearst, et au contraire de beaucoup d’autres, nous voulons traiter des corpus non-spécialisés. En revanche, nous ne pensons pas qu’une approche statistique seule ou bien une approche linguistique seule puisse cerner toutes les connaissances utiles à un SRI. Nous sommes convaincus que les connaissances à extraire du texte doivent être de nature statistique et linguistique.

Enfin, nous tenons à rappeler que les connaissances que nous visons à acquérir représentent une étape dans un processus complet qui mène à l’acquisition d’un niveau sémantique symbolique. Ce niveau symbolique très structuré doit être apte à représenter les connaissances du monde auxquelles un système de recherche d’information peut être confronté.

Deuxième partie

Fouille de données pour la recherche d’information

Chapitre 3

Fouille de données

Il n'y a pas une méthode unique pour étudier les choses.
ARISTOTE

3.1 Qu'est ce que la fouille de données?

La fouille de données ou Data Mining (DM) (appelée aussi paillage de données ou le forage de données)[1] est un terme utilisé pour décrire le processus de découverte automatique de modèles à partir de grandes quantités de données. Cette approche combine analyse et découverte et se justifie par le constat général qu'il y a beaucoup de données non exploitées. Elle est accompagnée de beaucoup de slogans de type "*Vos données travaillent pour vous*" ou "*Torturer vos données pour qu'elles se confessent*". Mis à part ces slogans, nous allons essayer de cerner la signification de cette problématique et en donner une définition plus précise.

3.1.1 Découverte de connaissances dans les bases de données

L'augmentation significative des informations au sein des organisations s'est accompagnée d'une prise de conscience de l'importance de développer des moyens informatiques plus efficaces pour traiter ces informations. En effet, les volumes astronomiques des bases de données, la diversité et l'hétérogénéité des sources de données nécessitent une nouvelle philosophie de traitement des données. Dans cet objectif, la découverte de connaissances dans les bases de données (KDD[2]) est dédiée à résoudre ces problèmes.

1. nous utilisons dans la suite le terme Data Mining (DM).
2. le terme KDD est l'abréviation de *Knowledge Discovery in Databases*

Elle est introduite dans [FPSS98] où les auteurs en donnent la définition suivante:

Définition 1 *Le processus d'identification d'une structure de données valide, nouvelle, potentiellement utile et finalement compréhensible. Ce processus fait intervenir la sélection de données, le prétraitement, la transformation, l'application du Data mining pour produire la structure et l'évaluation de la structure extraite.* [3]

Le terme **Structure** dans la définition signifie des modèles (une description des données) ou des patrons (une description d'un sous-ensemble de données) [BFM98].

Le KDD désigne le processus non-trivial de découverte d'informations implicites, précédemment inconnues et potentiellement utiles concernant les bases de données. Elle consiste donc à parcourir les immenses volumes de données contenus dans une base, à la recherche de connaissances. Cela s'applique notamment à des données en quantité trop importante pour qu'une étude visuelle soit possible. Si l'on considère par exemple l'ensemble des tickets de caisse d'un supermarché pendant 10 ans [4], il est aisé d'imaginer la quantité d'informations présentes, la diversité des champs, et donc la difficulté d'une exploitation de l'information, à part des données factuelles, telles que les quantités totales d'articles vendus, selon les mois, etc.

Les travaux dans ce domaine sont motivés par l'évolution très rapide des techniques de production, d'acquisition (telles que les techniques de lecture des codes barres des articles achetés en supermarché) et de stockage de données (augmentation de la capacité et diminution des coûts des disques durs par exemples) qui ont permis la création par les organismes de volumineuses bases de données concernant leurs activités.

Le KDD suppose qu'il est possible d'extraire des *informations cachées* dans ces masses de données pour l'aide à la décision, la gestion des informations et l'optimisation des requêtes, sous la forme de régularités, des anomalies et des modèles qui sont utiles, intéressants et compréhensibles du point de vue de l'utilisateur [FPSS98]. La connaissance, extraite de telles données, est exprimée par des liens entre différents champs d'information, mettant en évidence la dépendance entre les données.

Comme l'indique la Figure3.1, le KDD est un processus semi-automatique et itératif, constitué de plusieurs étapes allant de la sélection et la préparation des données jusqu'à l'interprétation des résultats [FPSS98].

3. The process of identifying valid, novel, potentially useful, and ultimately understandable structure in data. This process involves selecting or sampling data from a data warehouse, cleaning or preprocessing it, transforming or reducing if, applying a data mining component to produce structure, and the evaluating the derived structure.

4. Exemple classique pour définir la problématique du KDD

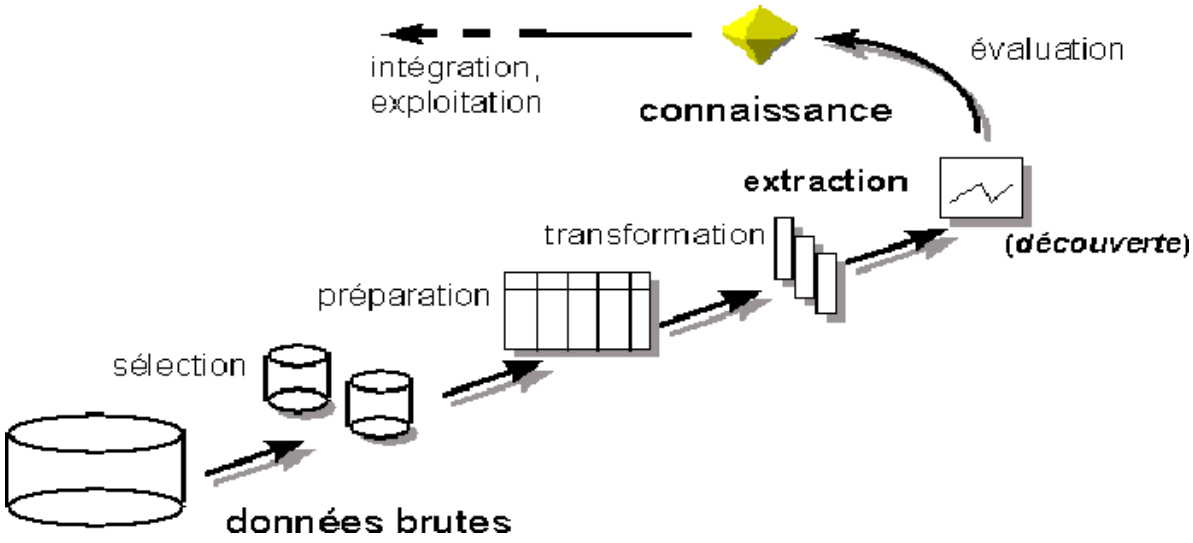

FIG. 3.1 – *Étapes de découverte de connaissances dans les bases de données*

3.1.2 Data Mining : définition et objectifs

Le Data Mining (DM) ne s'adresse qu'à la phase d'extraction et la phase de découverte du processus du KDD (Figure 3.1) [FPSS98]. S'il s'agit d'une part importante, il faut bien être conscient du fait que ce n'est pas la seule, et il est fondamental de ne pas négliger la sélection des données ainsi que la préparation et la transformation qui serviront pour les algorithmes de DM. Dans cette phase, des algorithmes spécifiques sont appliqués sur des données pour extraire des résultats utiles. Le Data Mining n'intègre pas donc toute la problématique de la découverte de connaissances.

Le terme Data Mining est souvent employé pour désigner l'ensemble des outils permettant d'accéder aux données et de les analyser. Nous restreindrons ici le terme de Data Mining aux outils et méthodes ayant pour objet de générer des informations intéressantes et de découvrir des modèles implicites dans les données. La définition la plus communément admise de Data Mining est celle de [FPSS98]:

> Le Data mining est un processus non trivial qui consiste à identifier, dans des données, des schémas nouveaux, valides, potentiellement utiles et surtout compréhensibles et utilisables.

Nous proposons de définir le Data Mining comme suit [Had98]:

> c'est le processus de découverte de connaissances, de dépendances endogènes et de règles dans des grandes quantités de données suivant des méthodologies

automatiques permettant de développer une représentation optimale de leurs structures.

Nous appellerons Data Mining l'ensemble des techniques qui permettent de transformer les données en connaissances. Son but est de remplir l'une des tâches suivantes : classification, estimation, prédiction, regroupement par similitudes, segmentation (ou clusterisation), description, etc. Selon les données sur lesquelles ces techniques s'appliquent, on désigne par des termes différents le DM : text mining, multimédia mining, Web mining, etc. Dans la suite de ce document, nous considérons que ces appellations font partie de l'ensemble Data mining car elles ne représentent pas de véritables innovations technologiques mais plutôt une adaptation des techniques à des besoins différents.

3.1.3 Les domaines d'application du Data Mining

Le Data Mining utilise le savoir-faire de plusieurs domaines : l'intelligence artificielle, l'apprentissage, la reconnaissance de modèle, l'acquisition de connaissances pour les systèmes experts, la visualisation des données, etc. La plupart des algorithmes du Data Mining sont dérivés de ces domaines. Le point commun entre tous ces algorithmes est l'extraction de modèles dans un contexte de grandes quantités de données.

Il y a donc différentes vues de Data Mining et différentes sortes d'applications qui dépendent des types des connaissances à découvrir, des types de bases de données à manipuler, des types de techniques à utiliser, etc. Chaque classe d'application est soutenue par un ensemble d'approches algorithmiques qui sont utilisées pour extraire les relations pertinentes dans les données : réseaux bayésiens, réseaux de neurones, clustering, règles d'association, etc. Ces approches se différencient suivant les problèmes qui sont les plus appropriés à résoudre. Certaines applications peuvent étudier les éléments qui ont un impact sur une variable donnée (par exemple, les caractéristiques démographiques des individus qui répondent à un mailing). D'autres applications permettent de segmenter la population en différentes classes, de regrouper les individus similaires ou de détecter les comportements atypiques. Néanmoins, les méthodes les plus novatrices concernent la recherche de règles d'association [AIS93] pouvant conduire à des observations de type *composition du panier d'achat du consommateur*, et l'étude des séquences fréquentes (un ensemble de champs de la base de données) permettant d'appréhender le comportement des clients dans le temps [AS95].

3.2 Les règles d'association

Les techniques d'extraction des règles d'association furent développées à l'origine pour l'analyse de bases de données composées des achats des consommateurs. Elles ont

pour but de découvrir des relations significatives entre une grande quantité de données dans une base. Soit une base de données avec un grand nombre de transactions où chaque transaction est constituée d'une liste d'articles achetés par un client. Il s'agit alors de trouver les règles qui associent la présence d'un ensemble d'articles avec la présence d'un autre ensemble d'articles. Une règle d'association est donc une relation d'implication $X \Rightarrow Y$ qui indique que les transactions qui contiennent X ont tendance à contenir les articles de l'ensemble Y. X est la prémisse de la règle et Y est la conclusion de la règle.

Si on considère par exemple la règle d'association suivante:

Example 1 *40% des transactions qui contiennent les articles pain et vin contiennent également l'article fromage et 5% des transactions contiennent ces trois articles.*

Dans cette règle, 40% est la *confiance* de la règle et 5% est le *support* de la règle. Ces deux mesures, inspirées des statistiques, permettent de déterminer l'importance et la signification de la règle : la confiance indique la proportion de clients qui ont acheté l'article *fromage* parmi ceux qui ont acheté les articles *pain* et *vin*, et le support indique la proportion de clients qui ont acheté les trois articles. Cette règle d'association est représentée comme suit:

$$\text{r: pain vin} \Rightarrow \text{fromage}$$

Ainsi, le problème de trouver des règles d'association est de trouver les règles dont les supports et les confiances sont supérieurs respectivement à un seuil de support minimum ($minsupport$) et à un seuil de confiance minimum ($minconfiance$), définis par l'utilisateur en fonction des ses objectifs et du type de données à traiter. Les règles satisfaisant ces deux critères sont qualifiées de règles *fortes*.

Nous présentons le formalisme des règles d'association proposé par Agrawal et al. dans [AS94] comme suit:

Soit $I = i_1, i_2, ..., i_n$ un ensemble de n littéraux, appelés *items*[5]. Soit $B = t_1, t_2, ..., t_n$ une base de données de n transactions et chaque transaction t_i est constituée d'un sous-ensemble $I \subseteq I$ d'items et identifiée par un identifiant unique appelé TID. Un sous-ensemble $I \subseteq I$ de taille k est appelé un k-itemset. Une transaction t_i contient un itemset I si et seulement si $I \subseteq t_i$. Le support d'un itemset I est le pourcentage de transactions de B qui contiennent I:

$$support(I) = \frac{|t \in B / I \subseteq t|}{|t \in B|} = P(I)$$

5. le terme "item" (traduction du terme anglais "article") a pour origine les base de donnée de transactions de ventes.

Un itemset dont le support est supérieur ou égal au seuil minimal de support défini par l'utilisateur est appelé *itemset fréquent*.

Une règle d'association est une implication de la forme $I_1 \Rightarrow I_2$ entre deux itemsets I_1,$I_2 \subseteq I$ telle que $I_1 \cap I_2 = \emptyset$. Cette règle a un support s dans la base des transactions B si $s\%$ des transactions T de B contiennet $I_1 \cup I_2$ [6]:

$$support(I_1 \cup I_2) = \frac{|T \in B / I_1 \cup I_2 \subseteq T|}{|B|} = P(I_1 \cup I_2)$$

La confiance de la règle d'association $r : I_1 \Rightarrow I_2$ est la probabilité conditionnelle qu'une transaction contienne I_2 sachant qu'elle contient I_1 :

$$confiance(r) = \frac{support(I_1 \cup I_2)}{support(I_1)} = \frac{P(I_1 \cup I_2)}{P(I)}$$

Le problème de l'extraction des règles d'association consiste à déterminer l'ensemble des règles d'association dont le support et la confiance sont au moins égaux à des seuils minimaux de support $minsupport$ et de confiance $minconfiance$ définis par l'utilisateur.

Il existe de nombreux algorithmes pour extraire des règles d'association dans des grandes quantités de données [Wu97]. Chaque type d'algorithmes est basé sur des propriétés et des méthodes qui prennent en compte principalement les structures des données à exploiter et les temps d'extraction des itemsets fréquents. Le premier algorithme d'extraction des itemsets fréquents est l'algorithme AIS [AIS93, AS94] qui procède par itérations : pour chaque itération k, un ensemble de k-itemsets candidat (les itemsets fréquents potentiels) est généré et les supports de ces candidats sont calculés lors d'un seul et même balayage, ce qui permet de limiter le nombre total de balayages réalisés. Plusieurs algorithmes optimisant l'efficacité de l'algorithme AIS ont été proposés tel que l'algorithme DHP (Direct Hashing and Pruning) proposé par Park et al. [PCY95] qui utilise des tables de hachage, l'utilisation des bitmaps hiérarchique proposé par Gardarin et al. dans [GPW98], la parallélisation du calcul proposé par Zaki dans [Zak99], etc.

3.3 Conclusion

Le Data mining s'intéresse à l'extraction d'informations auparavant inconnues et potentiellement utiles, généralement sous la forme de corrélations ou de tendances. Il englobe l'ensemble de processus utilisé pour optimiser le traitement de données et généralement implémenté comme une couche au-dessus des données. D'autres points qui restent encore assez peu traités sont:

– l'interface entre les utilisateurs et les techniques très complexes de traitement reste un problème majeur. Plusieurs méthodes courantes du Data Mining ne sont pas

6. $\cup$: union

réellement interactives et ne peuvent pas, pour une application d'un processus de Data Mining sur un ensemble de données, exploiter des connaissances découvertes antérieurement par ce même processus.

– le Data Mining est utilisé pour découvrir des structures dans des grandes bases de données mais la façon de représenter ces structures n'est pas souvent explicitée.

– la notion de sémantique est inexistante dans le Data Mining. En effet, traiter des données en se basant sur des statistiques et en donnant comme résultat des pourcentages ou des classes sans pour autant donner un sens est loin d'être satisfaisant.

– le Data Mining n'est pas très efficace pour les grandes bases de données dynamiques. En effet, les variables dans ce cas sont souvent modifiées. Par exemple, les valeurs des seuils de support et de confiance pour l'extraction des règles d'association sont souvent changées alors il faut réviser le support et la confiance minimaux pour éviter le risque que le Data Mining trouve des règles d'association trop fréquentes ou au contraire très peu fréquentes dans les données.

Au sein du Data mining, l'extraction des règles d'association dans les grandes bases de données de type *si condition alors conclusion* est devenue une technique essentielle en extraction des connaissances à partir des données. Cette technique a montré son efficacité dans le domaine des bases de données. Son application dans le domaine de la recherche d'information est promettant mais nécessite l'adaptation de sa problématique au traitement des données textuelles. En effet, en appliquant cette technique, notre objectif est d'extraire des associations entre des éléments textuelles, par définition, nouvelles et potentiellement utiles.

Dans le chapitre suivant, nous détaillons la problématique de l'application de la fouille de données dans le contexte de données textuelles et plus précisément dans le contexte de la recherche d'information.

Chapitre 4

La fouille de données textuelles

Ne craignez pas la perfection. Vous n'y parviendrez jamais.
SALVADOR DALI

Le problème avec la langue naturelle est qu'elle n'est pas conçue pour être traitée par les ordinateurs, à la différence des données stockées dans des bases de données. En effet, l'information est disponible à l'état brut et de ce fait est faiblement exploitable. Elle n'est pas explicite mais implicite, "enterrée dans le texte". La fouille de données textuelles associe des techniques d'analyse linguistique automatique aux techniques de fouille de données dans les bases de données en vue d'analyser le contenu des textes dans l'objectif de découvrir cette information implicite [GSEM97].

Or, la plus grande partie de l'information existante est sous une forme faiblement ou non structurée (documents, texte, vidéos, etc.). Il y a eu très peu de travail du style des outils de KDD pour l'analyse des collections de documents. L'objectif est d'exploiter les opérations de KDD utilisées pour les données structurées dans la découverte de connaissances textuelles (KDT)[1].

Pour une collection de documents, les besoins des utilisateurs sont très variés : les sujets traités, les types de documents existent, les liens entre les documents, etc. Une constatation générale de ces besoins est que l'utilisateur ne sait pas exactement ce qu'il cherche d'où une approche Data Mining semble adéquate à résoudre ces problèmes [AHKV97]. En effet, la structuration des données, la recherche d'association ou la sélection de schémas fréquents ont été largement utilisés dans le domaine des bases de données et leur application dans le cadre de données textuelles peut par exemple aider l'utilisateur à explorer le contenu des corpus dans une tâche de recherche d'information. Nous avons besoin d'outils qui permettent de manipuler les données textuelles. Ces outils doivent permettre d'examiner l'information et de la transformer en connaissances. La connaissance

1. le terme KDT est l'abréviation de *Knowledge Discovery from Text*

n'est alors qu'une information dans un certain contexte. Si on examine les moteurs de recherche traditionnels, ces derniers ne peuvent pas être qualifiés d'outils de fouille de données parce qu'ils trouvent uniquement les informations que l'utilisateur leur demande de chercher. Il y a un manque d'outils qui permettent d'acquérir des connaissances à partir d'un ensemble de documents. L'objectif est donc de trouver toute la connaissance qu'on peut déduire, explicitement ou implicitement, dans toutes les sources d'information disponibles. Ces outils sont connus sous le nom d'outils de **fouille de données textuelles** (désigné aussi **fouille de données documentaires**) et **fouille de données sur le Web**[2]. Le terme générique de ces outils est **découverte de connaissances dans le texte** (DCT). Nous proposons la définition suivante du DCT [Had98]:

Définition 2 *La DCT est le processus de découverte de modèles intéressants et utiles dans des corpus textuels non structurés. Le DCT combine les techniques d'extraction d'information, recherche d'information, traitement de la langue naturelle, etc. avec les méthodes du Data Mining. L'objectif de l'utilisation de la DCT est d'obtenir des connaissances précédemment inconnues et enfouies dans les textes.*

La fouille de texte aide à acquérir la connaissance latente (*cachée*) à partir du contenu des documents telles que des relations d'association entre les termes d'un corpus.

4.1 La fouille de données textuelles

Alors que le data mining agit sur des bases de données structurées, la fouille de données textuelles agit sur des textes individuels, des parties de textes ou des corpus. Dans la littérature, il n'y a pas une distinction nette entre la fouille de données textuelles et la fouille de données documentaires. Ces deux termes sont souvent confondus. La fouille de données textuelles est appliquée à l'analyse et l'accès à l'information en général qui se traduit par l'implémentation de fonctionnalités dans des moteurs de recherche Internet, des moyens de traitement de messages électroniques, des moyens de diffusion automatique de documents, des moyens d'interrogation de connaissances à partir de corpus documentaire, etc.

Un système de fouille de données textuelles reprend les étapes du processus du Data Mining et il en ajoute d'autres pour les adapter à son objectif. Quelques systèmes de fouille de données textuelles ont pour objectif de structurer le contenu des textes en découvrant des modèles pour les décrire. Ils se basent sur l'hypothèse d'une catégorisation a priori où il s'agit d'un prétraitement manuel des textes afin d'en extraire un certain nombre d'attributs comme les mots-clés ou les URLs. Une fois les attributs extraits, les méthodes classiques de Mining, telles que l'analyse statistique, association, etc., sont appliquées.

2. en anglais *Web mining*

D'autres systèmes appliquent l'analyse plein texte à des collections de documents pour construire une catégorisation de ces derniers. Deux types d'analyse sont possibles : l'analyse à but descriptif (fonctionnant en mode *non-supervisé* : l'outil analyse les documents sans référence à une classification prédéfinie) et l'analyse à but décisionnel (fonctionnant en mode *supervisé* : l'outil affecte automatiquement les documents selon une classification prédéfinie). Dans l'un et l'autre cas, le couplage de techniques linguistiques (une analyse linguistique rudimentaire du texte qui consiste à distinguer les catégories grammaticales des termes et enlever les mots vides) et statistiques (une analyse statistique des données qui va permettre de corréler les données entre elles pour en saisir les invariants et les règles qui les régissent) est utilisé.

Les techniques de fouille de données textuelles peuvent être classées en deux catégories [FFH+98]:

Techniques utilisant les mots-clés [FD95] :

> Les processus de découverte de connaissances sont appliqués à des mots-clés associés, généralement manuellement, à des documents. Des thèmes sont établis a priori et chacun de ces thèmes est représenté par un ensemble de mots-clés. Selon les mots-clés qu'il contient, un document appartient donc à un thème donné. Les inconvénients de cette approche est que les mots-clés sont extraits et associés à un document manuellement ou semi-automatiquement ce qui est difficilement faisable pour une grande collection de documents.

Techniques utilisant tout le texte [RB97, LAS97] :

> Un document est représenté par tous les mots qu'il contient. L'application des processus de la fouille de données textuelles a permis de découvrir des patrons de bas niveau c'est à dire au niveau des termes tels que les mots composés [RB97].

Dans ce qui suit, nous présentons quelques systèmes de fouille de données textuelles.

4.1.1 Le système PatentMiner

Le système d'extraction des textes décrit dans [LAS97] traite le problème de l'identification de phrases (séquence de termes) comme un modèle séquentiel dans le but de découvrir des tendances dans des bases de données textuelles. On dénote un terme par w et une phrase p par $< (w_1)(w_2)...(w_n) >$. Les auteurs définissent *une tendance* comme étant une sous-séquence d'une phrase qui satisfait la requête de l'utilisateur [3].

3. Trends are simply those k-phrases selected by the shape query with the additional information of the time periods in which the trend is supported [LAS97].

La méthodologie de cette approche est la suivante:

– identification des phrases fréquentes utilisant la technique de Data Mining : *traitement des patrons séquentiels*[4] et la mesure de support. La fenêtre de recherche est fixée à un paragraphe mais elle peut être plus petite. L'utilisateur peut aussi spécifier une taille de fenêtre pour préciser la distance entre les termes. La structure des documents est exploitée pour identifier les différentes sections.

– génération de l'historique des phrases.

  La granularité de partitionnement est spécifiée par l'utilisateur. Elle peut être l'année ou le mois (intéressante pour les documents sur le Web), etc. Pour chaque partition, une génération d'un ensemble de phrases fréquentes est calculée.

– recherche de phrases satisfaisantes pour une tendance.

  Les auteurs utilisent un langage spécifique appelé *SDL* [APWZ95] pour identifier les formes ("shape") intéressante pour un utilisateur. L'utilisateur interroge le système en spécifiant la tendance recherchée. Le résultat est exprimé par des courbes et des graphiques.

4.1.2 Le système des épisodes

Ahonen et al utilisent les notions d'*épisode* et *règle d'épisode*[5] [AHKV97] qui sont inspirées de la notion des règles d'association mais qui sont appliquées aux données séquentielles (où un certain ordre des données est pris en compte). Dans le cas du texte, les auteurs précisent que les épisodes sont des vecteurs composés de caractéristiques (un ensemble de caractéristiques ordonnées) et un index (contient la position d'un mot dans la séquence). Une caractéristique peut être un mot, une expression, un signe de ponctuation ou une étiquette (par exemple un tag SGML). Un épisode est donc une séquence de texte qui apparaît dans une fenêtre donnée avec un ensemble de caractéristiques et l'apparition de ces caractéristiques dans la séquence selon un certain ordre. La taille d'une fenêtre est mesurée en nombre de mots non vides.

Un épisode textuel est défini comme suit :

$$\alpha = (V, \leq) \text{ où}$$

– V est un ensemble de vecteurs de caractéristiques.

– $\leq$ est un ordre.

4. en anglais *Generalized Sequential Patterns [AS95]*

5. *en anglais episodes et Episode rule*

Pour une séquence de texte S, un épisode textuel occurre dans S si un vecteur de V est vérifié tout en respectant l'ordre $\leq$. Une fenêtre est utilisée pour déterminer des frontières où un épisode occurre dans le texte. Le support de α dans S respectivement à une taille de fenêtre W est le nombre d'occurrences de α dans S.

Une règle d'épisode donne la probabilité conditionnelle qu'un certain épisode occurre dans une certaine fenêtre. Elle est représentée sous la forme suivante:

$$\beta[win_1] \Rightarrow \alpha[win_2] \text{ où}$$

- α et β sont deux épisodes.
- β est un sous-épisode de α, donc une séquence incluse dans la séquence α.
- win_1 et win_2 sont deux tailles de fenêtres telles que $win_1 \leq win_2$.

La confiance de la règle est la probabilité conditionnelle que α occurre sachant que β occurre sous la contrainte de la taille de la fenêtre W.

Soit la règle d'épisode suivante extraite de [AHKV97]:

Exemple 5 *Knowledge, discovery, in [4] $\Rightarrow$ databases [5] (85%).*

Dans 85% des cas où les trois mots *Knowledge*, *discovery* et *in* occurrent dans une séquence de quatre mots alors le mot *databases* occurre dans la fenêtre de 5 mots.

Des seuils de support minimum et de confiance minimale doivent être spécifiés pour sélectionner les règles d'épisode.

Les auteurs soulignent l'importance de la phase de prétraitement des informations textuelles avant l'application des techniques du Data Mining. Ce prétraitement consiste à ne sélectionner que certaines catégories grammaticales (substantifs) ou à éliminer d'autres (préposition, articles, etc.) ou bien encore à lemmatiser les mots et permet, selon les auteurs, d'alléger le processus de traitement. Cependant, ils ont expérimenté leur approche que pour découvrir les phénomènes locaux dans quatorze documents et ne l'ont pas expérimenté pour découvrir des phénomènes dans des corpus volumineux. Dans une deuxième expérience, les auteurs se sont orientés vers la découverte des dépendances structurelles en ne traitant que les Tag SGML des documents.

4.1.3 Les systèmes Fact et KDT

Fact est un système qui est basé sur les cooccurrences des mots dans les documents [FH96]. Il génère un langage de requête grâce à une interface visuelle qui applique un processus de fouille de données textuelles sur les mot-clés associés aux documents par les auteurs. La collection utilisée est la collection-test Reuter-222173. Dans cette collection,

les mots-clés représentent 5 classes : les pays, les sujets des dépêches, les marchés, les personnes et les organisations. Le but du système est de trouver des règles d'association entre les termes, par exemple entre un pays et les organisations auxquelles il appartient. Les règles d'association extraites indiquent que la présence du terme (ou d'un ensemble de termes) X dans un document implique la présence du terme (ou d'un ensemble de termes) Y, avec deux probabilités (support et confiance présentées dans la section 3.2) calculées sur l'ensemble des documents du corpus. Les auteurs appliquent ce type d'extraction d'associations à des corpus textuels étiquetés par des mots-clefs.

Le système KDT[6] est inspiré du système Fact [FD95]. Les mots-clés dans le système KDT sont utilisés pour permettre à l'utilisateur de naviguer dans la collection des documents et de reconnaître les modèles des documents en utilisant la distribution des cooccurrences des mots-clés.

Dans leurs expériences, les auteurs ont utilisé une collection Reuter de 22000 articles de 25 megabytes. La hiérarchie des mots-clés est établie par le personnel de Reuter où différentes catégories et sous-catégories de termes sont déterminées a priori. La distance de Kullback-Leibler (la mesure d'entropie relative), détaillée dans l'annexe B, a été utilisée pour quantifier le degré d'intérêt d'une distribution.

Le système KDT calcule la distribution des catégories filles des termes par rapport à la catégorie mère. Par exemple, l'annotation des documents avec les filles de la catégorie *computer* peut être la distribution : $mainframe$= 0.1, $work - station$=0.4 et PC=0.5. Il s'agit d'attribuer à chaque noeud C de la hiérarchie une variable discrète aléatoire c dont la valeur est définie par les filles du noeud C et notée $P(C = c)$.

Le système KDT s'intéresse à la distribution conditionnelle des mots-clés sous la forme $P(C = c/x)$ où x est un événement conditionnel qui dénote une autre catégorie. Cette distribution dénote une cooccurrence de la catégorie x avec les autres filles de C.

4.2 Le Web mining

Berghel dans [Ber97a] considère les moteurs de recherche actuels comme des versions primitives des futurs moyens d'accès à l'information sur le Web, principalement en raison de leur incapacité à distinguer le *bon* du *mauvais* dans ce fouillis d'informations qu'est le Web. Selon l'auteur, cette évolution ne peut pas se faire par de simples améliorations des méthodes actuelles, mais nécessite au contraire le développement de nouvelles approches (agents logiciels intelligents, outils de personnalisation de l'information, outils de *push*, etc.) orientées vers un filtrage et une analyse plus fine des informations. Dans ce contexte, une nouvelle tendance dans le domaine de recherche sur le Web est connue sous le nom

6. Knowledge Discovery from Texts

de *découverte de connaissances du Web (Web mining)*. Le Web mining est l'application des techniques du Data Mining sur les ressources du Web.

Cooley et al. dans [CMS97] définissent le Web mining comme suit:

Définition 3 *La découverte et l'analyse d'informations intéressantes sur le Web incluant la recherche automatique de ressources d'information en ligne (la découverte de connaissances véhiculées par le contenu) et la découverte de modèles d'accès des utilisateur à partir des serveurs Web (la découverte de connaissances sur l'usage)*[7].

On distingue alors deux approches de Web mining : la découverte de connaissances véhiculées par le contenu (Web content mining) et la découverte de connaissances sur l'usage (Web usage mining). La découverte de connaissances véhiculées par le contenu est le processus de découverte de connaissances dans le contenu des documents et des textes ou la découverte de leurs descriptions. La découverte de connaissances sur l'usage est le processus d'extraction de modèle ou de patrons intéressants dans les *Web access logs* [MPC99, Mas02]. Les *Web access logs* sont les fichiers contenant l'activité des utilisateurs d'un site Web. On peut ajouter la découverte de connaissances sur la structure du Web qui est un processus d'inférence de connaissances à partir de l'organisation du Web et des liens entre les référents et les référencés sur le Web.

Les travaux sur la découverte de connaissance dans le contenu des pages Web sont généralement basés sur les agents logiciels. Plusieurs de ces travaux ont employé des techniques de clustering afin de filtrer, rechercher et classer des documents disponibles sur le Web par catégories. Theilmann et al. dans [TR98] proposent une approche appelée *domain expert*. Un domaine est un domaine de connaissances contenant des informations sémantiquement reliées. Il utilise un agent de filtrage mobile qui visite des sites spécifiques et examine la pertinence des documents qu'il contient par rapport au domaine. L'agent mobile utilise un ensemble de connaissances établies a priori et extraites des URL collectés par des moteurs de recherche existants. Les auteurs ont utilisé uniquement des connaissances qui décrivent les documents : mots-clés, métamots-clés, URL, date de modification, auteurs; et cette connaissance est représentée par des *facettes*.

4.3 Conclusion

Une architecture générale d'un système de fouille de données textuelles peut être la suivante (Figure 4.1):

– Sélection/Recherche.

7. the discovery and analysis of useful information from the World Wide Web including the automatic search of information resources available on-line, i.e. *Web content mining*, and the discovery of user access patterns from Web servers, i.e. *Web usage mining*

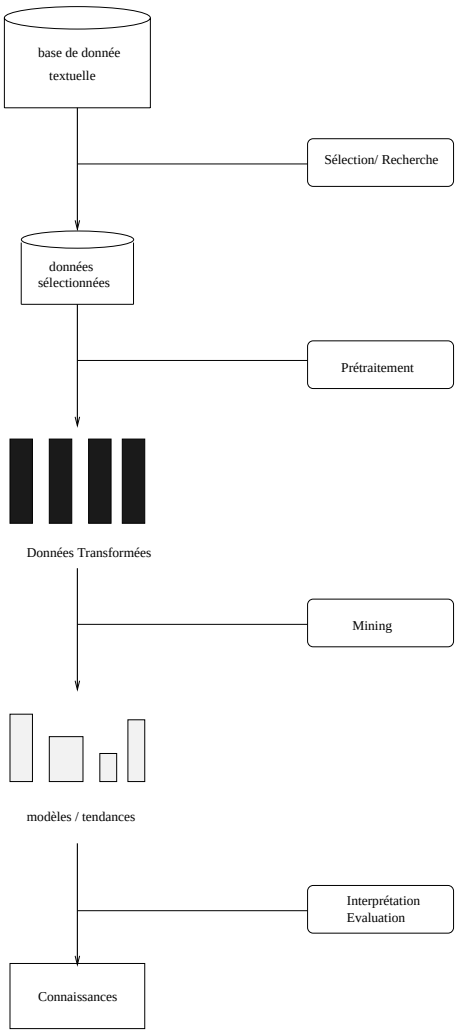

FIG. 4.1 – *Architecture d'un système de fouille de données textuelles*

A partir d'une base de données textuelles, des informations sont sélectionnées suivant certains critères ou les besoins des utilisateurs en informations.

– Prétraitement.

  Un prétraitement suivant les objectifs du système est appliqué aux données sélectionnées (élimination de mots vides, sélections de mots-clés, etc.).

– Mining

  Il s'agit d'appliquer une des techniques du Data Mining : épisode, classification, associations, etc.

– Interprétation/Évaluation

  Il s’agit d’interpréter les connaissances découvertes et de représenter les résultats à l’utilisateur dans une interface bien adaptée.

Dans le chapitre suivant, nous présentons notre méthodologie de la découverte des règles d’association dans le contexte de la recherche d’information.

Chapitre 5

Les règles d'association dans la recherche d'information

Nous ne connaissons pas le vrai si nous ignorons les causes.
ARISTOTE

Dans cette section, nous présentons une méthodologie basée sur la découverte de la connaissance sous la forme de termes et de liens entre les termes. Même si les liens ne sont pas typés, ils permettent à l'utilisateur de pouvoir préciser son besoin d'information en examinant le contexte d'utilisation des termes qu'il emploie. Par *contexte des termes*, nous désignons une indication sur la définition possible d'un terme ou de son environnement d'usage. Les approches présentées dans la section 4.1 exigent une modification de la structure de collection ou l'utilisation de la catégorisation. Notre méthodologie est établie sur l'hypothèse suivante:

Hypothèse 1 *Un document représente un ensemble de termes sémantiquement cohérent. Tous les termes participent à la signification globale du document. Chaque terme a une signification dans le contexte où il est utilisé dans le document.*

Notre hypothèse est qu'un document est représenté par un ensemble de termes et que ces derniers participent au sens global du document. Les termes sont alors des indicateurs d'une certaine connaissance qui révèle le contexte de l'utilisation des termes dans les documents. Par *termes*, nous désignons les mots qui ont un sens et nous excluons les mots vides. Ainsi, nous considérons que deux mots apparaissant dans le même document sont sémantiquement liés.

5.1 Définition du problème

Les règles d'association, décrites dans la section 3.2, sont basées, dans le contexte de données textuelles, sur les principes suivants:

- une transaction est alors une entité textuelle tel qu'un document, un paragraphe, une phrase, un message e-mail, un message dans les news, etc.
- les items sont alors les termes des entités textuelles.

Le formalisme des règles d'association dans le contexte de données textuelles est le suivant : soit $I = i_1, i_2, ..., i_n$ un ensemble de n termes, appelés *items*. Soit $B = t_1, t_2, ..., t_n$ un corpus documentaire de n entités textuelles, chaque entité textuelle t_i est constituée d'un sous-ensemble $I \subseteq I$ de termes, appelés *items*. Un sous-ensemble $I \subseteq I$ de taille k est appelé un k-itemset. Une entité textuelle t_i contient un itemset I si et seulement si $I \subseteq t_i$. Le support d'un itemset I est le pourcentage d'entités textuelles de B qui contiennent I. Une règle d'association est une implication de la forme $I_1 \Rightarrow I_2$ entre deux itemsets I_1,I_2 $\subseteq I$ telle que $I_1 \cap I_2 = \oslash$. Nous notons le support de I_1 et I_1 comme suit: $support(I_1 \cap I_2)$ et leur confiance comme suit $confiance(r) = \frac{support(I_1 \cap I_2)}{support(I_1)}$.

Un seuil minimal de support et un seuil maximal de support sont établis pour éliminer les règles très rares et celles très fréquentes qui ne sont pas utiles. En effet, si la mesure de support est très élevée cela veut dire que ces deux ensembles de termes cooccurrent trop souvent ensemble dans les entités textuelles de la collection. Cette connaissance est donc explicite (elle n'est pas nouvelle) et il n'est pas intéressant de l'extraire. Si la mesure de support est trop faible cela veut dire que les deux ensembles de termes cooccurrent rarement ensemble. Cette connaissance n'est pas vraiment utile.

La confiance d'une règle d'association $r : I_1 \Rightarrow I_2$ est la probabilité conditionnelle qu'une entité textuelle contienne I_2 sachant qu'elle contient I_1. Ainsi, c'est le pourcentage des entités textuelles qui contiennent tous I_1 et I_2 par rapport à celles qui contiennent seulement I_1. Donc, plus I_1 est fréquents dans B plus cette règle est inintéressante. Nous avons fixé un seuil minimal de confiance.

Les mesures destinées à calculer les dépendances entre les termes (Information mutuelle, Coefficient de Dice, Coefficient de cohérence, etc.)[1] prennent en compte les liaisons entre deux termes sans tenir compte du reste des termes dans le corpus. C'est généralement la fréquence des termes qui est utilisée pour calculer la dépendance. Ces mesures ne tiennent pas compte de la proportion que peut avoir cette dépendance par rapport à toutes les dépendances pouvant exister dans le corpus. En effet, l'objectif de la plupart de ces mesures est plutôt d'extraire des collocations ou des catégorisations des termes dans des ensembles homogènes alors que l'objectif des règles d'association est de trouver

1. section 2.1 et annexe B

un lien implicite entre des termes sans se soucier de classer ces derniers ni chercher des collocations. La sémantique d'une règle d'association $I_1 \Rightarrow I_2$ est alors différente de la sémantique d'une simple dépendance dans le sens où :

- elle exprime une probabilité de l'existence d'un terme par rapport à un autre.
- elle tient compte de l'ensemble des implications possibles dans le corpus d'où un classement de toutes les implications possibles dans le corpus en tenant compte de la dispersion de I_1 et de I_2 dans le corpus mais aussi de la dispersion des autres termes.

Le sens intuitive de $r : I_1 \Rightarrow I_2$ est que, dans un corpus donné, les entités textuelles qui contiennent I_1 contiennent aussi I_2. Les règles d'association candidates qui vérifient ces critères sont un ensemble de *règles d'association fortes* et qui représentent une base de connaissances exprimée sous la forme de termes et de liens entre des termes.

Nous montrons dans l'exemple ci dessous, le processus de sélection des règles d'association par rapport aux seuils du support et de la confiance.

Exemple 6 *Soient les règles d'association A $\Rightarrow$ B, C $\Rightarrow$ D et E $\Rightarrow$ F découvertes dans le corpus, on suppose que les seuils sont les suivants:*

- $10\% \leq s \leq 30\%$ *où* s *est la mesure du support*
- $50\% \leq c$ *où* c *est la mesure de la confiance*

Soient les mesures de support et confiance de chacune des règles d'association :

Règles d'association	*Support*	*Confiance*
$A \Rightarrow B$	20%	50%
$C \Rightarrow D$	25%	8%
$E \Rightarrow F$	35%	45%

TAB. 5.1 – *Exemple de règles d'association*

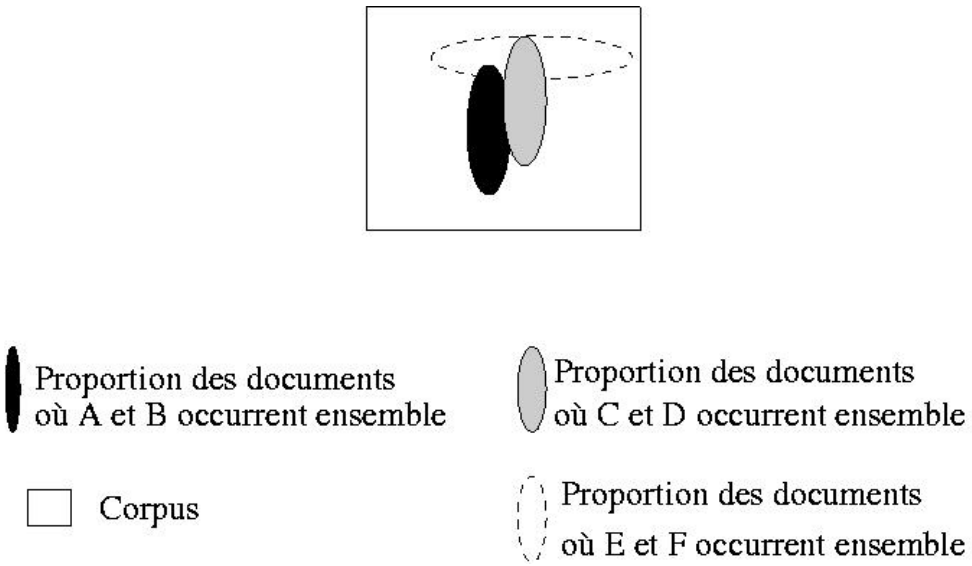

FIG. 5.1 – *Distribution des termes dans le corpus*

Les seuils de support sont d'abord utilisés pour une première sélection des règles d'association. La règle d'association $E \Rightarrow F$ a une mesure de support supérieure au seuil maximum de support. La fréquence des deux items E et F dans le corpus est jugée importante. Cette règle est alors éliminée. Les deux autres règles seront sélectionnées (Figure 5.1).

Le seuil de confiance est utilisé pour une deuxième sélection des règles d'association. La règle d'association $C \Rightarrow D$ a une mesure de confiance inférieure au seuil minimum de confiance. En effet, la proportion des documents contenant C et D est faible ce qui se reflète par une mesure de confiance faible (Figure 5.2). Cette règle est alors éliminée.

Dans la suite, nous utilisons la valeur absolue de la mesure de support. Cette mesure sera exprimée en nombre de documents et non plus en pourcentage.

5.2 Définition des transactions dans le contexte de la RI

Dans le contexte de la RI, une transaction dans la problématique des règles d'association peut s'identifier à une entité textuelle. Une entité textuelle peut être une phrase ou un paragraphe ou un document ou une fenêtre textuelle se déplaçant sur le texte, indépendamment de la structure logique du document. Elle permet le décomptage des fréquences et le calcul des règles d'association. La détermination de la fenêtre est donc un point important

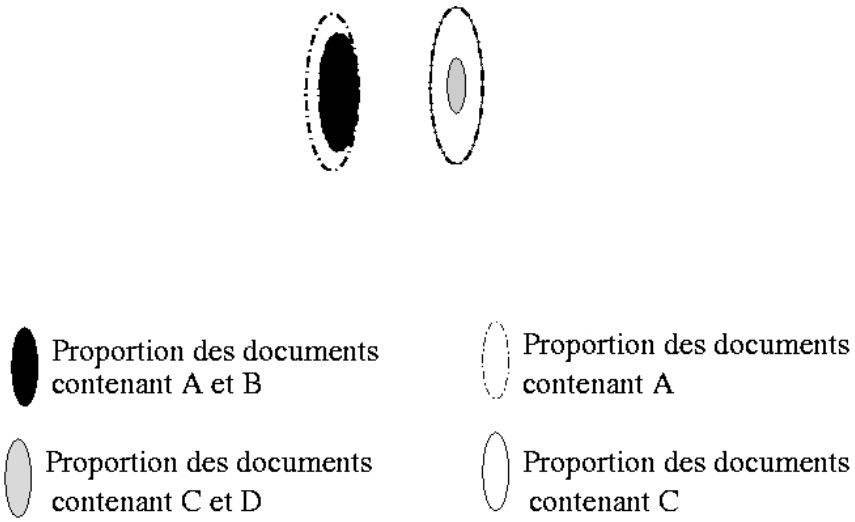

FIG. 5.2 – *Distribution conditionnelle des termes dans le corpus*

dans la phase de découverte de RA dans le texte. Dans cette section, nous discutons des cas où les transactions sont des phrases ou des paragraphes ou des documents.

5.2.1 Cas de la phrase

Dans le cas où une transaction est une phrase, les règles d'association permettent de découvrir des relations du niveau local (niveau de la phrase et de la construction de la phrase); ce qui revient à redécouvrir la structure syntagmatique de la phrase c'est-à-dire des relations entre des constituants des syntagmes mais avec beaucoup moins de qualité que l'application de patrons syntaxiques.

Exemple 7 *Collection OFIL1 : Document N 2271496*

POINT LA PARTITION DE LA TCHECOSLOVAQUIE : Les étapes de l'accord

En gestation dans les esprits depuis 1990, la partition a été opérée dans les six mois qui ont suivi les élections législatives de juin 1992.

Celles-ci ont amené au pouvoir, à Prague, le Parti libéral de M. Vaclav Klaus et, à Bratislava, celui du leader nationaliste Vladimir Meciar.

20 juin : MM. Klaus et Meciar signent un accord prévoyant la mise en route avant le 30 septembre du processus de division de la Tchécoslovaquie en deux Etats indépendants.

17 juillet : le Parlement de Bratislava proclame la souveraineté de la Slovaquie.

20 juillet : M. Vaclav Havel quitte son poste de président de la République tchécoslovaque.

Si on considère que la fenêtre est la phrase, les règles d'association découvertes seront entre des termes appartenant à une même phrase : par exemple une relation entre *Parti libéral de M. Vaclav Klaus* et *leader nationaliste Vladimir Meciar* ou bien *partition* et *TCHECOSLOVAQUIE* . On ne trouvera pas des relations entre des termes appartenant à des phrases différentes tel que:

- *élections législatives* et *PARTITION DE LA TCHECOSLOVAQUIE*
- *Parti libéral de M. Vaclav Klaus* et *Parlement de Bratislava*
- *M. Vaclav Havel* et *leader nationaliste Vladimir Meciar*

5.2.2 Cas du paragraphe

Un paragraphe peut être constitué d'une ou plusieurs phrases. Dans ce cas, nous retrouvons les règles d'association découvertes dans le cas de la phrase ainsi que des relations d'association interphrases. La plupart des relations d'association découvertes restent du niveau local mais on ne trouve pas les relations entre des termes appartenant à des paragraphes différents.

5.2.3 Cas du document

D'autres cas sont aussi envisageables telle que une fenêtre de n termes ou une fenêtre de m phrases où m est déterminé par rapport à la taille du document et/ou la moyenne des tailles des documents dans le corpus.

Dans le cas où la fenêtre est le document, le nombre des règles d'association découvertes est plus important que dans le cas de la fenêtre ou le paragraphe. En effet, pour la collection OFIL, le nombre de RA découvertes avec un support compris entre 20 et 500 documents et un seuil de confiance minimum de 20% est de 7030 relations d'association alors que pour les mêmes seuils, seulement 1343 relations d'association sont découvertes avec une fenêtre qui correspond à la phrase.

L'ensemble des règles d'association découvertes avec la phrase comme fenêtre est inclus dans l'ensemble des règles d'association découvertes avec le document comme fenêtre. Dans le tableau 5.2, nous illustrons l'exemple des règles d'association relatives au terme *chancelier* et découvertes à partir de la collection OFIL. Nous remarquons que toutes les règles d'association dans le cas de la phrase, sont aussi découvertes dans le cas du document. Dans ce dernier cas, d'autres règles d'association sont aussi découvertes. Nous pouvons ajouter que dans le cas de la phrase, la plupart des associations

Cas de la phrase (confiance, support)	Cas du document (confiance, support)
pacte ⇒ chancelier (20,2073% , 39)	pacte ⇒ chancelier (30% , 39)
unification ⇒ chancelier (22,4719%, 20)	unification ⇒ chancelier (32,7869%, 20)
chancelier ⇒ kohl (39,2694%, 86)	chancelier ⇒ kohl (63,7037%, 86)
kohl ⇒chancelier (49,1429%, 86)	kohl ⇒chancelier (843137%, 86)
chancelier ⇒ helmut (21,0046%, 46)	chancelier ⇒ helmut (34,0741%, 46)
helmut ⇒ chancelier (41,8182%, 46)	helmut ⇒ chancelier (47,9167%, 46)
bonn ⇒ chancelier (23,2044%, 42)	bonn ⇒ chancelier (35%, 42)
	chancelier ⇒ solidarité (34,074%,46)
	chancelier ⇒ déficit (20%, 27)
	chancelier ⇒ pacte (28,8889%, 39)
	chancelier ⇒ réunification (22,2222%, 30)
	réunification ⇒ chancelier (27,027%, 30)
	chancelier ⇒ bundesbank (22,2222%, 30)
	bundesbank ⇒ chancelier (20,6897%, 30)
	bonn ⇒ chancelier (31,1111%, 42)

TAB. 5.2 – *Règles d'association relatives au terme chancelier: cas de la phrase et cas du document*

découvertes sont des groupes nominaux ou des multi-termes tel que *chancelier helmut kohl*, ce que nous avons appelé des associations du niveau local. Les règles d'association $chancelier \Rightarrow kohl$, $kohl \Rightarrow chancelier$, $chancelier \Rightarrow helmut$ et $helmut \Rightarrow chancelier$ sont toutes relatives au syntagme nominal *chancelier helmut kohl*. Nous pensons que ces associations du niveau local ne sont pas d'un apport du point de vue connaissances, puisque l'utilisation de la syntaxe pour extraire les groupes nominaux donnerait une connaissance plus intéressante. Ce point est détaillé dans la section 5.3. Dans le chapitre 7, nous proposons l'utilisation des syntagmes nominaux ce qui permet de ne pas découvrir les règles d'association du niveau local. Nous considérons qu'une transaction est un document dans le contexte de la découverte de règles d'association dans des données textuelles.

5.3 Signification des Règles d'association

Si on considère deux règles d'association $X \Rightarrow Y$ et $Y \Rightarrow X$, ces deux règles ont la même mesure de support exprimée par le nombre de documents où X et Y occurrent. Cependant, les deux règles n'ont pas la même mesure de confiance. En effet, la mesure de confiance de la règle d'association $X \Rightarrow Y$ dépend de la dispersion de X alors que la

règle $Y \Rightarrow X$ dépend de la dispersion de Y. Une règle d'association $X \Rightarrow Y$ tient alors sa signification d'après la valeur de sa mesure de confiance et de l'existence ou non de la règle $Y \Rightarrow X$. Nous distinguons les différents cas suivants:

1. Existence de deux règles $X \Rightarrow Y$ et de $Y \Rightarrow X$ avec deux valeurs fortes de confiance presque identiques.

   Les deux valeurs fortes de confiance indiquent que:

   - $confiance(X \Rightarrow Y) = \frac{P(X \cap Y)}{P(X)} \approx 100\%$ donc $P(X \cap Y) \approx P(X)$.
     Cette règle d'association indique que chaque fois que X est dans un document alors il y a de forte chance que Y soit dans le document;
   - $confiance(Y \Rightarrow X) = \frac{P(X \cap Y)}{P(Y)} \approx 100\%$ donc $P(X \cap Y) \approx P(Y)$.
     Cette règle d'association indique que chaque fois que Y est dans un document alors il y a de forte chance que X soit dans le document.

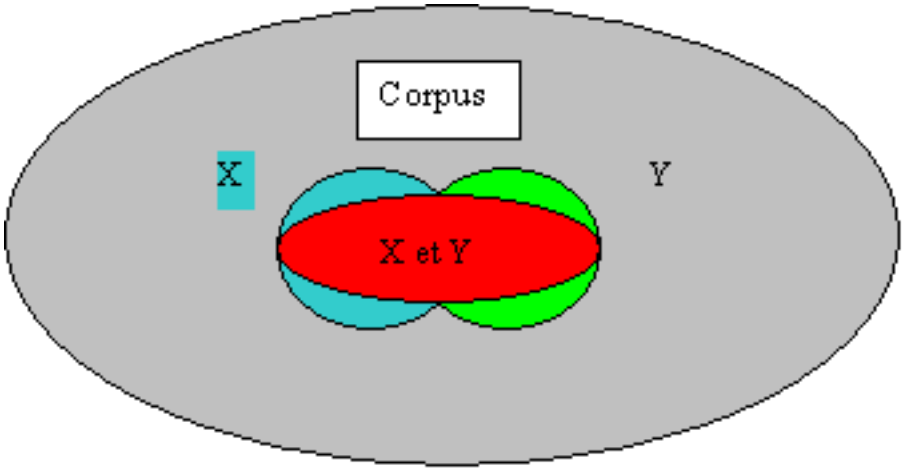

FIG. 5.3 – *Cas de deux valeurs fortes de confiance*

Ces règles d'association correspondent dans la plupart des cas observés à la découverte des associations du niveau local, c'est-à-dire des associations entre les composantes des groupes nominaux, des mots figés ou bien des noms propres et qui sont quasiment ensemble dans tous les documents (Figure 5.3). Ces règles d'association ne sont pas intéressantes puisqu'elles ne répondent pas aux objectifs visés par la découverte des règles d'association:

Règle d'association	mesure de confiance	mesure de support (nombre de documents)
ghali ⇒ boutros	99%	115
boutros ⇒ ghali	99%	115
end ⇒ week	91%	174
week ⇒ end	90%	174
boris ⇒ eltsine	87%	138
valéry ⇒ estaing	85%	128
eltsine ⇒ boris	81%	138

TAB. 5.3 – *Exemple de règles d'association découvertes dans la collection OFIL*

– découvrir des règles nouvelles :

  Les règles d'association ne sont pas nouvelles du fait que ces termes existent ensemble trop souvent.

– Exploration du corpus :

  Si on veut explorer l'environnement d'un terme, par exemple *Eltsine* (ou *Ghali*) de la collection OFIL illustré dans le tableau 5.3, le fait de proposer *Boris* (ou *Boutros*) ne rajoute pas une information intéressante.

– Extension des requêtes :

  Si on étend une requête contenant le terme *Ghali* par le terme *Boutros*, on ne va pas trouver de nouveaux documents ou très peu puisque à chaque fois qu'un document contient le terme *Ghali* il y a une forte probabilité qu'il contienne aussi le terme *Boutros*.

  Nous proposons de considérer les syntagmes nominaux, les mots figés et les noms propres comme étant des termes simples dont les composantes seront associées pour former un seul terme. Les règles d'association découvertes ne sont pas alors relatives aux termes *Ghali* ou *Boutros* mais au terme *Boutros Ghali*.

2. Existence de deux règles $X \Rightarrow Y$ et $Y \Rightarrow X$ avec deux valeurs de confiance sensiblement identiques:

   – $confiance(X \Rightarrow Y) \ll 100\%$ et $confiance(Y \Rightarrow X) \ll 100\%$ et

   – $confiance(X \Rightarrow Y) \approx confiance(Y \Rightarrow X)$

Ces règles signifient que:

- $P(X) \approx P(Y)$ donc X et Y ont presque la même dispersion dans le corpus et la probabilité de trouver X dans un document est la même que celle de trouver Y dans un document.
- Vu leurs dispersion dans le corpus, Figure 5.4, les deux termes sont associés avec la même valeur de confiance : la probabilité de trouver Y dans un document sachant que X existe est presque la même que celle de trouver X sachant que Y existe.

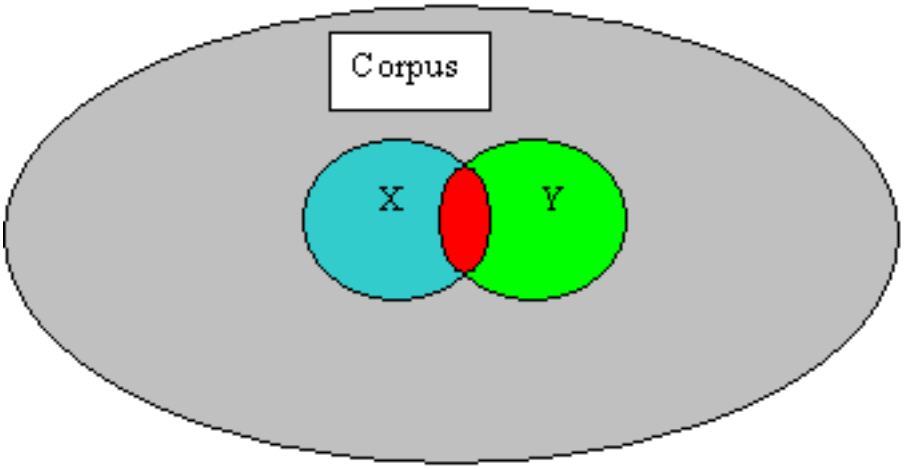

FIG. 5.4 – *Cas de deux valeurs de confiance sensiblement identiques*

3. Existence de deux règles $X \Rightarrow Y$ et $Y \Rightarrow X$ avec deux valeurs de confiance non identiques:
   - $confiance(X \Rightarrow Y) = a$
   - $confiance(Y \Rightarrow X) = b$
   - Avec $a > b$

   La valeur de confiance $b < a$ signifie que la P(Y) est plus forte de celle de P(X) : $P(Y) > P(X)$ ce qui revient à dire que Y est plus présent dans le corpus que X (Figure 5.5). C'est pourquoi la règle $X \Rightarrow Y$ a une plus forte valeur de confiance.

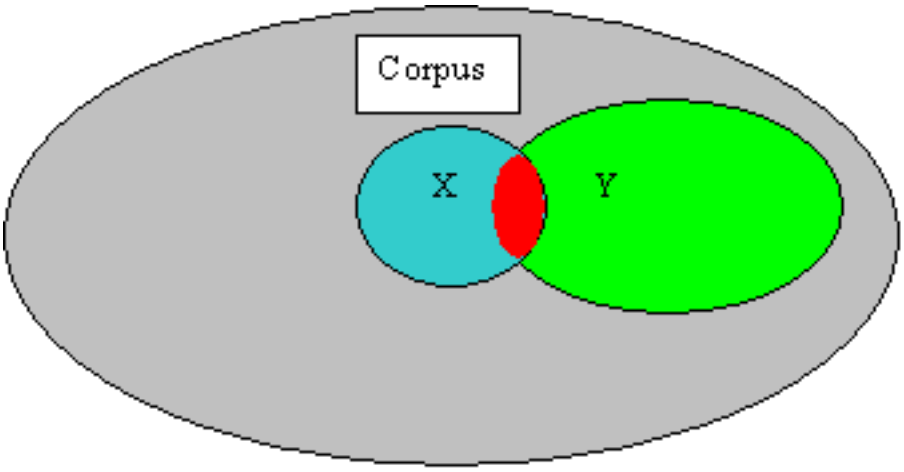

FIG. 5.5 – *Cas de deux valeurs de confiance non identiques*

Pratiquement, quand on extrait les règles d'association tous les termes qui cooccurrent dans un document sont associés par des règles d'association. Un filtrage élimine les règles qui:

- ne vérifient pas le seuil de support :

  Si deux termes X et Y ne cooccurrent pas assez souvent dans le corpus donc les deux règles $X \Rightarrow Y$ et $Y \Rightarrow X$ sont éliminées.

- ne vérifient pas le seuil de confiance :

  Si deux termes X et Y cooccurrent assez souvent dans le corpus mais que la confiance de la règle $X \Rightarrow Y$ ne vérifient pas le seuil de confiance cela signifie que X est trop présent par rapport à la présence de X et de Y dans un même document donc cette règle est éliminée.

Si on ajoute un autre filtrage pour éliminer les règles qui ont une grande valeur de confiance donc relatives à des syntagmes nominaux alors les règles sélectionnées sont celles qui:

- Vérifient le seuil de support : les termes X et Y cooccurrent assez fréquemment dans les documents.

- Vérifient le seuil de confiance donc X n'est pas trop présent et l'événement de sa cooccurrence avec Y n'est pas systématique

- N'ont pas une valeur de confiance importante donc $P(X \cap Y) \ll P(X)$: la probabilité d'avoir X et Y dans un document est inférieure à celle de trouver X

5.4 Utilité dans un contexte de RI

L'auteur d'un document exprime l'information qu'il veut passer par des mots qui combinés forment des phrases dont l'ensemble forme le texte. Ce processus de communication entre l'utilisateur d'un SRI et le propriétaire du texte passe par un document qui est le support de l'information. Nous considérons l'hypothèse suivante:

Hypothèse 2 *L'information véhiculée par un document dans sa globalité exprime le thème exprimé et discuté par l'auteur.*

A partir de cette hypothèse, nous pouvons considérer que:

- Un document représente un ensemble de termes sémantiquement cohérents qui forment le thème du document. Le thème du document est alors construit à partir de l'ensemble des termes contenu dans le document.
- Tous les termes participent à la signification globale du document où chaque terme joue un rôle dans la construction du thème d'un document mais les termes ne jouent pas le même rôle dans le sens où il y a:
  - des termes porteurs de plus d'information que d'autres
  - des termes plus précis que d'autres
  - l'introduction de termes nouveaux définis par rapport à des termes connus
- Chaque terme a une signification précise dans le contexte où il est utilisé.

  Ce contexte est généralement défini par les autres termes utilisés dans le document et qui lui donne un sens particulier parmi tous les sens possibles.

Le problème n'est pas aussi simple car des segments de texte peuvent aussi avoir des thèmes propres et que les documents n'ont pas nécessairement un thème unique. Dans un document on peut parler de plusieurs sujets : il peut y avoir un thème principal (générique) et plusieurs thèmes spécifiques pour des différentes parties du document. Cette division thématique ne correspond par forcément à une division physique ou logique d'un document (titre, paragraphe, chapitre, sous-section, etc.). Cette hypothèse est intéressante dans le cas où on voudrait chercher des parties d'un document ou des relations entre des termes appartenant à un sous-thème particulier d'un corpus.

On se restreint à la définition d'un thème d'un document comme étant le thème général où on considère le document dans sa globalité et que tous les termes jouent un rôle dans la construction de son thème. Le thème est donc exprimé par des termes qui sont proches ou éloignés dans le document. Des termes appartenant aux différentes parties du document participent à la signification du document et le sens d'un terme peut être défini par rapport

à l'ensemble des termes présents dans le document quelle que soit la distance qui le sépare de ces termes.

Plusieurs scénarios pour exploiter les associations entre termes dans un SRI peuvent être définis. En effet, elles peuvent être utilisées pour :

– L'expansion automatique de la requête. Les approches classiques de représentation du contenu textuel des documents et des requêtes sont basées sur l'utilisation de "mot-clé". Un mot-clé est supposé représenter une partie de contenu du document et de la requête. C'est raisonnable, en tenant compte de la représentativité des mots-clés pour le contenu et la simplicité de leur manipulation. Cependant, les opérations de recherche définies dans les modèles de RI pousse la notion de mot-clé plus loin, bien qu'implicitement. Il est implicitement supposé qu'un mot-clé soit l'unique représentant d'une signification unique. Autrement dit, il est supposé qu'il y ait une correspondance du type 1:1 entre les mots-clés et les sens ce qui est très rare voir impossible. En réalité, un mot peut avoir plusieurs sens, et un sens peut être exprimé par des mots différents.

  Afin de traiter cette correspondance n:n entre les mots-clés et les sens, on propose généralement de considérer des mots reliés pour étendre la requête. Ce traitement tente de considérer le fait qu'un sens puisse être véhiculé par des mots différents. Ce traitement est appelé l'enrichissement des requêtes [2].

  Une expansion de requête est alors vue comme un traitement pour *élargir* le champ de recherche pour cette requête. Une requête étendue va contenir plus de termes reliés. En utilisant le modèle vectoriel, par exemple, plus de documents seront repérés. Ainsi, ce traitement est souvent vu comme un moyen d'augmenter le taux de rappel.

  L'expansion de requête peut être effectuée à l'aide des règles d'association extraites. Pour chaque terme d'une requête, son profil relationnel dans le corpus est ajouté dans la requête d'origine. Par exemple, la neuvième requête de INIST contient les termes *système de scolarité*. Les termes *système* et *scolaire* sont associés à d'autres termes avec les règles d'association suivantes découvertes dans la collection INIST:

  – système $\Rightarrow$ structure
  – système $\Rightarrow$ infrastructure
  – scolarité $\Rightarrow$ collège
  – scolarité $\Rightarrow$ lycée

2. en anglais *query expansion*

Les termes *structure*, *infrastructure*, *collège* et *lycée* sont ajoutés à la requête. La requête enrichie, éloigne des premières réponses le sens *système de blanchiment* par exemple.

– une expansion interactive des requêtes (IQE) peut aider l'utilisateur à formuler sa requête [MR97] à l'image du module *Refine* d'*AltaVista* [Bou97], en utilisant le résultat d'une requête pour la reformuler, la filtrer et la réorienter en exploitant les termes liés aux termes de sa requête. En effet, l'utilisateur peut sélectionner des ensembles des termes ou des termes, suggérés, pour les ajouter à sa requête. Dans le cas d'un besoin d'information non précis c'est à dire que l'utilisateur a une idée vague de son besoin d'information, les contextes des termes de sa requête peuvent être suggérés et l'utilisateur choisit les termes à ajouter à sa requête. Dans le cas où le besoin d'information serait précis alors l'utilisateur peut choisir les termes d'un ensemble correspondant à son besoin d'information.

5.5 Conclusion

Pour conclure ce chapitre consacré à la définition des règles d'association dans le contexte de la recherche d'information, nous avons donné un exemple de l'utilisation de des règles d'association pour l'expansion des requêtes. Dans le chapitre suivant, nous appliquons une expansion automatique ainsi qu'une expansion interactive des requêtes. Dans cet objectif, nous allons nous focaliser sur la découverte de règles d'association entre deux termes dans des documents. Ainsi nous restreignons les définitions des ensembles I_1 et I_2, présentés dans le formalise des règles d'association dans le contexte de données textuelles présenté dans la section 5.1, à un seul terme : $I_1 = \{t_i\}$ et $I_2 = \{t_j\}$ où $t_i \in I$ and $t_j \in I$. Nous traitons une entité textuelle comme étant un document du corpus.

Chapitre 6

Extraction et exploitation de règles d'association

C'est par l'expérience que la science et l'art font leur progrès chez les hommes.
ARISTOTE

L'extraction des règles d'association à partir de corpus textuels, comme l'illustre la Figure 6.1 nécessite les étapes de sélection et de prétraitement (affinement linguistique) et extraction des règles d'association.

FIG. 6.1 – *Processus d'extraction des règles d'association*

6.1 Importance de l'étape de sélection et prétraitement

La première étape d'un processus de fouille de données textuelles débute par une phase de sélection et prétraitement des informations brutes afin d'en extraire des éléments pertinents (termes représentatifs des contenus dans notre contexte d'étude). Le problème avec le texte est que le nombre de mots différents qui peuvent exister dans des documents est largement plus grand que le nombre d'éléments (items) dans les bases de données transactionnelles. En effet, ces dernières peuvent avoir au mieux quelques centaines d'items alors

que, dans les collections textuelles OFIL et INIST, le nombre de mots différents est respectivement 119 434 et 174 659 mots. Cela se traduit par un nombre important de règles d'association extraites dans un processus de fouille de données textuelles. Le problème se situe alors sur la nature des mots à sélectionner.

Les critères de sélection peuvent être de nature statistique (fréquence des mots dans la collection), sémantique (utilisation de concepts et élimination des mots vides comme "la" ou "le") ou linguistique (filtrage selon les catégories grammaticales). Ce dernier critère semble très pertinent. En effet, comme l'illustre le tableau 6.1, des relations d'association faisant intervenir certaines catégories grammaticales semblent être vides de sens et plutôt inintéressantes dans notre contexte de découvertes de connaissances. La règle $r : sa \Rightarrow$ *orléans* par exemple, même si elle peut être statistiquement sélectionnable, reste non utile d'un point de vue sémantique et certainement pratique dans notre contexte d'étude. La règle d'association $r : vietnam \Rightarrow 0,5$ est une association entre le terme *vietnam* et le chiffre *0,5* qui n'est pas intéressante à découvrir car elle ne donne pas une précision sur le contexte du terme *vietnam*. La règle d'association $r : lever \Rightarrow il$ est une association entre le verbe *lever* et le pronom *il* qui ne donne pas une information intéressante sur le contenue de la collection. D'où l'importance de l'étape de sélection et de prétraitement qui est donc une phase cruciale de la fouille de données textuelles. Elle consiste à un prétraitement linguistiques et une sélection des mots représentatifs du contenu textuel des documents.

Règles d'association	Confiance
sa $\Rightarrow$ orléans	14.2%
vietnam $\Rightarrow$ 0,5	13.69%
brutalement $\Rightarrow$ escalier	21.73%
tgv $\Rightarrow$ symboliser	17.24%
panne $\Rightarrow$ lentement	12.4%
lever $\Rightarrow$ il	30.3%

TAB. 6.1 – *Exemple de règles d'association sans prétraitement*

6.2 Traitement linguistique

Nous avons choisi de ne traiter que certaines catégories grammaticales principalement les substantifs communs (SUBC). Ce choix se base sur notre hypothèse que la sémantique est portée par les SUBC (section 7.5.1). Les noms propres (SUBP) ont été aussi utilisés. La phase de prétraitement consiste donc en une analyse linguistique du texte afin d'affecter à chaque mot une catégorie grammaticale. Les mots sélectionnés sont alors ceux qui

correspondent aux catégories prédéfinies. Une liste de mots vides est utilisée pour éliminer les mots les plus communs (aujourd'hui, etc.).

L'analyseur morpho-syntaxique est un analyseur de surface qui utilise un dictionnaire associé à un modèle morphologique intégré dans le système IOTA [Pal90]. Un traitement particulier est appliqué aux formes non reconnues. Il permet en utilisant des schémas de résolution répertoriés manuellement, correspondant à des cas d'ambiguïté typiques et dont la résolution est connue, de leur attribuer une interprétation potentielle [CN97]. Le module donne en sortie une collection de textes étiquetés, c'est-à-dire dont tous les mots ont été catégoriés. La fréquence globale d'un terme dans une collection ainsi que sa fréquence selon une fenêtre sont calculées et permettent de calculer les valeurs de support et de confiance lors de l'extraction des règles d'association.

6.3 Extraction des règles d'association

Une règle d'association est donc une implication entre deux termes t et t' sous la forme $t \Rightarrow_c t'$ avec c la valeur de confiance relative à cette règle d'association. Une matrice de cooccurrences entre les termes est alors construite. L'ensemble des associations relatives à un terme, appelé profil relationnel d'un terme, est alors établi.

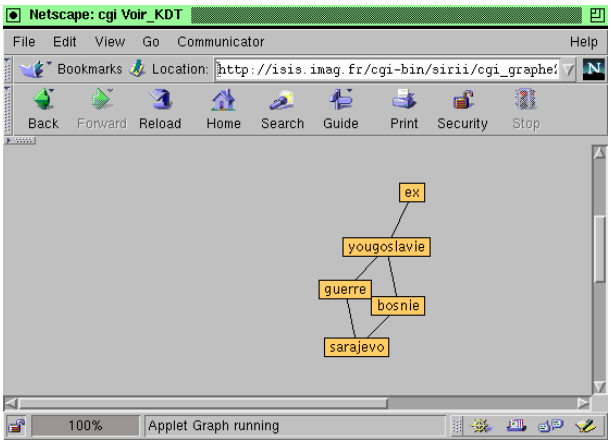

FIG. 6.2 – *Environnement du terme Sarajevo*

La Figure 6.2 illustre le profil relationnel du terme Sarajevo qui est formé par les termes *guerre* et *bosnie*.

Le profil relationnel du terme est alors défini par le vecteur des relations de ces termes dont le poids est la mesure de confiance relative aux deux termes qui forment la règle

d'association. Par exemple pour les termes t et $t^{'}$, la pondération de la règles d'association $t \Rightarrow_c t^{'}$ est notée $c_{tt'}$.

Notre algorithme d'extraction des règles d'association est inspiré de l'algorithme DHP (Direct Hashing and Pruning) proposé par Park et al. [PCY95]. En effet, il est basé sur une matrice de hachage où les entrées sont des termes. L'intersection d'une ligne X et d'une colonne Y représente les informations concernant les deux règles d'association $r : X \Rightarrow Y$ et $r : Y \Rightarrow X$. Etant donné, que le support des relations $r : X \Rightarrow Y$ et $r : Y \Rightarrow X$ est le même alors une seule valeur de support est calculée et stockée dans la matrice de hachage. La confiance de $r : X \Rightarrow Y$ ainsi que de $r : Y \Rightarrow X$ sont aussi stockée.

Le processus d'extraction des règles d'association se fait en un seul passage. En effet, l'algorithme commence par lire le contenu d'une collection. Pour chaque nouveau mot X trouvé, l'algorithme lui associe un identifiant ainsi que le support et la confiance des règles d'association potentielles qui peuvent l'associer à un autre mot dans la matrice. A chaque nouvelle occurrence de X, le support et la confiance de toutes les règles d'association qui contiennent X sont mises à jour. Vu que l'algorithme se base sur des fréquences d'occurrences des termes et que la phase de sélection et de prétraitement permettent d'éliminer une grande quantité de données, l'algorithme n'est pas coûteux en temps de calcul (douze secondes pour la collection OFIL).

6.4 Exploitation des règles d'association

A ce niveau, une base de connaissances exprimée sous la forme d'un réseau de termes connectés est construite. Ces associations sont antisymétriques. Notre approche pour la découverte d'associations entre termes se base sur les cooccurrences et la distribution des termes dans un corpus. L'interprétation des associations découvertes est donc complètement différente de celle des relations classiques dans un thésaurus. Il ne s'agit pas de connaissances encyclopédiques mais de connaissances qui se trouvent dans le texte lui-même.

Le tableau 6.2 présente les premières règles d'association découvertes à partir de la collection OFIL classées par ordre décroissant de la mesure de confiance. Les sept premières règles d'association représentent des mots composés ou bien des noms propres. Elles ont des probabilités de confiances importantes puisque les termes de chacune de ces règles cooccurrent souvent. Les autres règles sont plus intéressantes pour notre étude puisqu'elles représentent des bonnes dépendances sémantiques entre termes simples.

Règle d'association	mesure de confiance	mesure de support (nombre de documents)
ghali ⇒ boutros	99%	115
boutros ⇒ ghali	99%	115
end ⇒ week	91%	174
week ⇒ end	90%	174
boris ⇒ eltsine	87%	138
valéry ⇒ estaing	85%	128
eltsine ⇒ boris	81%	138
belgrade ⇒ serbie	61%	77
serbie ⇒ belgrade	44%	77
bundesbank ⇒ mark	43%	63
missile ⇒ irak	41%	47
album ⇒ chanson	40%	46
lycée ⇒ collège	36%	72
collège ⇒ lycée	36%	72
toile ⇒ peintre	36%	41
mur ⇒ berlin	36%	64

TAB. 6.2 – *Les 10 premières règles d'association de OFIL*

6.5 La campagne AMARYLLIS [Ama]

Amaryllis est une campagne d'évaluation destinée à favoriser l'émergence de nouveaux SRI. C'est une Action de Recherche Concertée (ARC) co-financée par l'Aupelf-Uref et le Ministère de l'Education Nationale, de l'Enseignement Supérieur et de la Recherche. Son objectif est de promouvoir l'élaboration de corpus et de procédures d'évaluation concernant le français, pour permettre à la recherche de progresser et au domaine de se doter d'instruments de mesure rendant possible une comparaison objective des différentes approches. Deux cycles ont déjà eu lieu, l'un en 1996-97, l'autre en 1998-99. La méthodologie employée dans les campagnes Amaryllis est très proche de celle des campagnes TREC [TRE]. Les équipes qui y participent ont accès à différents corpus documentaires, à des questions d'utilisateurs, et à des réponses "justes", c'est-à-dire constituées sur le principe des collections tests.

Pour la campagne d'évaluation Amaryllis'99, les fournisseurs des corpus sont :

– l'OFIL : Observatoire Français des Industries de la Langue

– l'INIST : Institut de l'Information Scientifique et Technique

– le LRSA : Laboratoire de Recherches Sémiographiques en Anthropologie -Université de Laval au Québec.

Les documents sont structurés selon une DTD ((Data Type Definition) issue de TEI en intégrant la gestion de la structure logique d'un ouvrage. Les thèmes (construits avec l'aide des documentalistes et des spécialistes des domaines concernés) des requêtes sont composés de cinq éléments :

– le domaine situant la thématique générale,

– le sujet : le titre du thème,

– la question proprement dite,

– des compléments d'information précisant le domaine de recherche et

– une liste de concepts (liste de termes délimitant la recherche).

L'exemple suivant illustre une requête de la collection OFIL [1]:

Exemple 8 *Domaine : International*
Sujet : La Guerre civile en Somalie
Question : Quelles sont les raisons de la poursuite des combats en Somalie ? Quel rôle l'ONU peut jouer en Somalie ?
Compléments : Les documents pertinents devront mettre en lumière les oppositions de clan et de chefs dans la poursuite des combats et ne devront pas ignorer que le comportement de l'ONU en Somalie n'obéit pas à une cohérence marquée.
Concepts: Etats-Unis, France, Négociations, Désarmement, Aide humanitaire, Guerilla.

Les évaluations décrites dans notre étude utilisent les trois jeux de test OT1 de OFIL, OD1 d'INIST et MD1 de LRSA.

6.6 Prétraitement des collections

La première étape de l'évaluation consiste à extraire le vocabulaire d'une collection et l'analyser en utilisant l'analyseur linguistique de IOTA [CN97]. Ensuite, la matrice de cooccurrence est construite. Les règles d'association sont extraites selon le formalisme présenté dans la section 5. Les expérimentations ont été menées sur les trois collections de la compagne Amaryllis : OFIL (plus de 35 Mo, 11,000 articles hétérogènes du journal "Le

1. requête numéro 7

Monde"), INIST (plus de 100 Mo, 165,431 d'articles scientifiques extraits des bases de données bibliographiques) et LRSA (plus de 3 Mo of 502 documents d'une monographe). Un ensemble de requêtes est associé à chaque collection et chaque requête est associée à un ensemble pertinent de documents : 26 requêtes pour la collection OFIL, 30 requêtes pour la collection INIST et 15 requêtes pour la collection LRSA. Les seuils de support et confiance utilisés ainsi que le nombre de règles d'association découvertes pour chaque collection sont présentés dans le tableau 6.3. Le seuil de support est le nombre minimum de documents où les termes doivent occurrer.

Nous avons utilisé le système de RI expérimental SMART [Sal71]. Ce système a été construit entre 1968 et 1970. Réécrit et réorganisé dans les années 1980, il utilise le modèle vectoriel.

Collection	support minimum (nombre de documents)	confiance minimum	nombre de règles d'association
LRSA	25	50%	3905
OFIL	110	20%	12774
INIST	1655	10%	93941

TAB. 6.3 – *Paramètres d'extraction des règles d'association*

On utilise deux critères classiques pour évaluer les performances d'un SRI [Sal71] : le rappel et la précision. Le rappel représente la capacité d'un système à retrouver des documents pertinents et la précision représente sa capacité à ne retrouver que des documents pertinents. Pour calculer ces deux critères, on utilise les ensembles suivant : R (ensemble de documents retrouvés par le système) et P (ensemble de documents pertinents pour l'utilisateur). On calcule:

$$\textbf{Rappel} = \frac{||R \cap P||}{||P||}$$

$$\textbf{Précision} = \frac{||R \cap P||}{||R||}$$

Pour faire une évaluation statistique de la qualité d'un système, il faut disposer d'une collection de tests : typiquement, un corpus de plusieurs milliers de documents et quelques dizaines ou centaines de requêtes, auxquelles sont associés des jugements de pertinence utilisateur, établis par des experts ayant une grande connaissance du corpus.

En général, pour chaque requête, on établit alors une **courbe de Rappel/Précision** (la précision en fonction du rappel). La moyenne de ces courbes permet d'établir un profil visuel de la qualité d'un système. Cela suppose que pour un certain nombre de requêtes, on est capable d'évaluer tous les documents dans le corpus pertinents à chaque requête

Dans la suite de cette section, nous présentons les résultats de nos expérimentations suivant les deux scénarios d'expansion automatique de la requête et de l'expansion interactive de la requête.

6.7 Expansion automatique des requêtes

Le déroulement du processus de l'expansion automatique des requêtes est le suivant:

– Trouver les meilleurs résultats du système SMART suivant les différents paramètres, présentés dans le tableau 6.4, de pondération, lemmatisation, etc. Ces résultats constituent nos résultats de référence.

Collection	pondération	lemmatisation	Liste de mots vides	Utilisation du titre
LRSA	ltc	oui	oui	non
OFIL	ltc	oui	oui	oui
INIST	ltc	oui	oui	oui

TAB. 6.4 – *Paramètres d'expérimentation de SMART*

Pour modéliser le contenu des documents, nous utilisons un des modèles les plus répandus en RI : le modèle vectoriel [Sal71] et le système SMART pour indexer les documents.

Un document est représenté par un vecteur dans un espace à n dimensions :

$$\vec{Contenu_i} = (w_{i1},\ w_{i2} \ldots w_{ij} \ldots w_{in})$$

Avec $w_{ij} \in [0, 1]$ qui représente le poids du terme t_j dans le document D_i.

La fonction de pondération ltc utilise la fréquence d'un terme tf_{ij}, le nombre de documents dans lesquels le terme t_j apparaît, la fréquence documentaire df_j et le nombre de documents N_{doc} du corpus comme suit:

$$w_{ij} = \frac{(log_2(tf_{ij})+1) * log_2(\frac{N_{doc}}{df_j})}{\sqrt{\sum_{j\in[1..n]} {w_{ij}}^2}}$$

– Expansion automatique des requêtes.

– Comparaison entre les meilleurs résultats de SMART et les résultats de notre approche.

Collection	*sans expansion*	*avec expansion (augmentation)*
OFIL	$31,64\%$	$32.73(+1.09)\%$
INIST	21.88%	$22.65(+0.77)\%$
LRSA	39.83%	$42.29\%(+2.46)$

TAB. 6.5 – *Résultats de l'expérimentation en précision moyenne en 11 points de rappel*

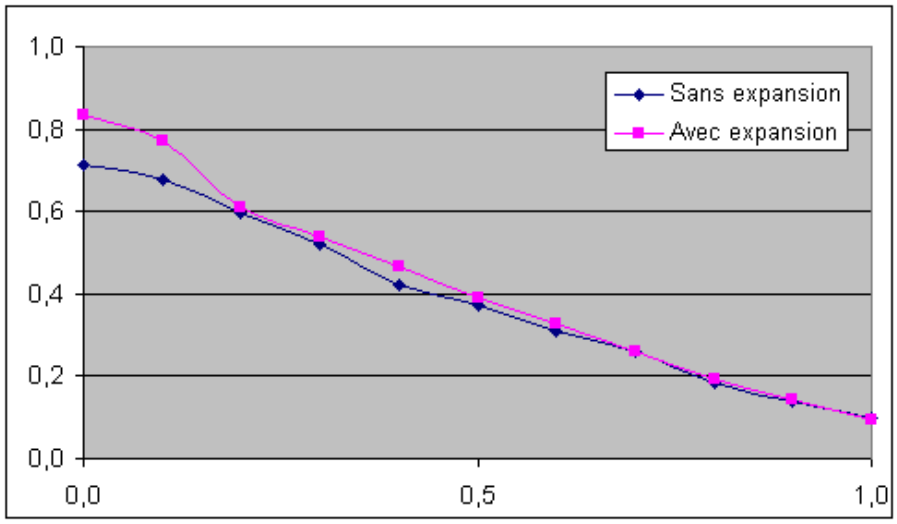

FIG. 6.3 – *Graphe Rappel/Précision de la collection LRSA*

Nous constatons que les mesures de la précision moyenne en 11 points de rappel dans le cas de l'expansion automatique des requêtes sont meilleures que ceux sans l'expansion pour les trois collections (tableau 6.5). Par contre, cette augmentation est plus significative pour OFIL (Figure 6.4) et LRSA (Figure 6.3). Ceci s'explique par le fait que INIST (Figure 6.5) est une collection scientifique où les termes ont de très faibles distributions et cooccurrences. Une deuxième raison est que l'analyseur linguistique de IOTA trouve beaucoup de difficultés à identifier les termes scientifiques de la collection INIST ce qui fait qu'une grande partie du vocabulaire n'est pas utilisée car elle n'est pas correctement analysée. Par exemple, les termes *diméthyl-2,4 et 2H-pyrindine-2* n'ont pas été correctement analysés. De même, ces termes ne sont pas pris en compte durant l'extraction des règles d'association.

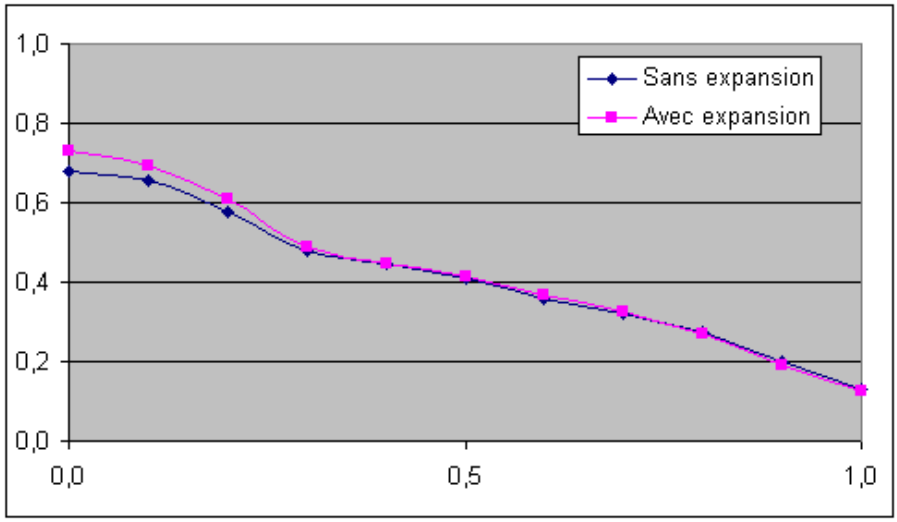

FIG. 6.4 – *Graphe Rappel/Précision de la collection OFIL*

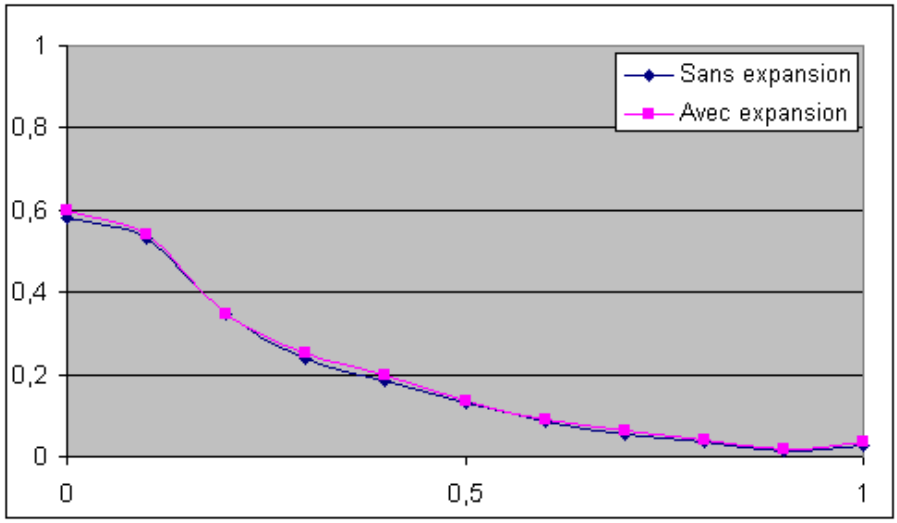

FIG. 6.5 – *Graphe Rappel/Précision de la collection INIST*

6.8 Expansion interactive des requêtes

Notre deuxième expérimentation consiste à une expansion interactive des 26 requêtes de la collection OFIL en utilisant la base de connaissances. Une expansion interactive, contrairement à l'expansion automatique utilisée lors de la première expérimentation, nécessite une implication de l'utilisateur dans la phase d'expansion. Ce dernier, selon sa

compréhension de la requête, ajoute un certain nombre de termes à la requête d'origine.

Pour cela, une interface a été développée. Les requêtes sont d'abord présentées aux utilisateurs (16 membres du laboratoire) comme illustré dans la Figure 6.6. En accédant à une des requêtes, les utilisateurs peuvent ajouter d'autres termes, extraits de la base de connaissances, à la requête originale. La Figure 6.7 présente une requête de la collection OFIL.

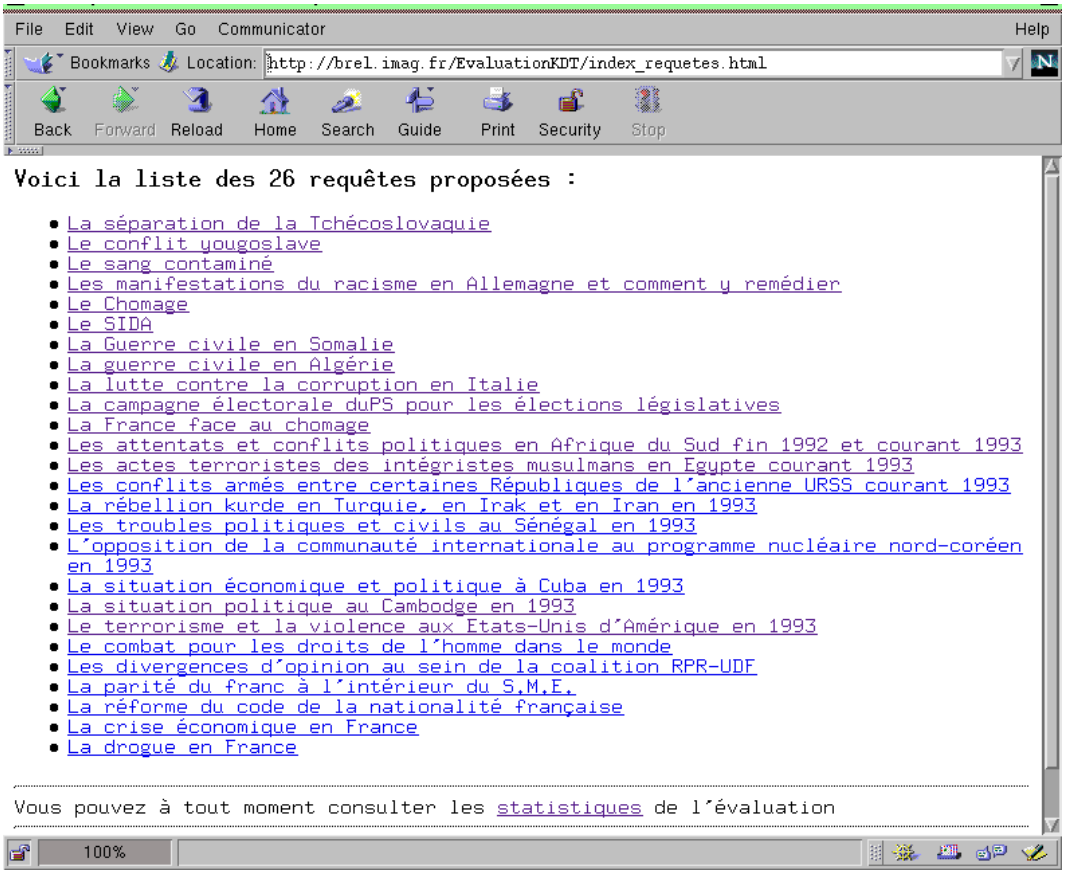

FIG. 6.6 – *Requêtes de la collection OFIL*

Les termes soulignés de la requête possèdent des liens dans la base de connaissances. En cliquant sur un de ces termes, la liste des liens relatifs à ce terme s'affiche dans la frame *visualisation des liens*. En consultant cette liste, l'utilisateur peut ajouter, suivant sa compréhension de la requête, un ou plusieurs termes, qu'il juge pertinents, à la requête.

Etant donné que les requêtes peuvent être évaluées individuellement (évaluation du rappel et la précision pour une requête particulière), les utilisateurs peuvent étendre une,

plusieurs ou toutes les requêtes. Pour les 74 expansions interactives effectuées par les utilisateurs, 6.79 termes, en moyenne, ont été ajoutés à chaque requête. En gardant les mêmes paramètres de SMART du tableau 6.4 ainsi que les paramètres d'extraction des règles d'association du tableau 6.3, nous avons comparé les performances des requêtes étendues par rapport aux requêtes originales. Nous avons constaté une augmentation des performances, en terme de rappel et précision, pour 51 requêtes étendues. Cette augmentation est en moyenne de 2.1%.

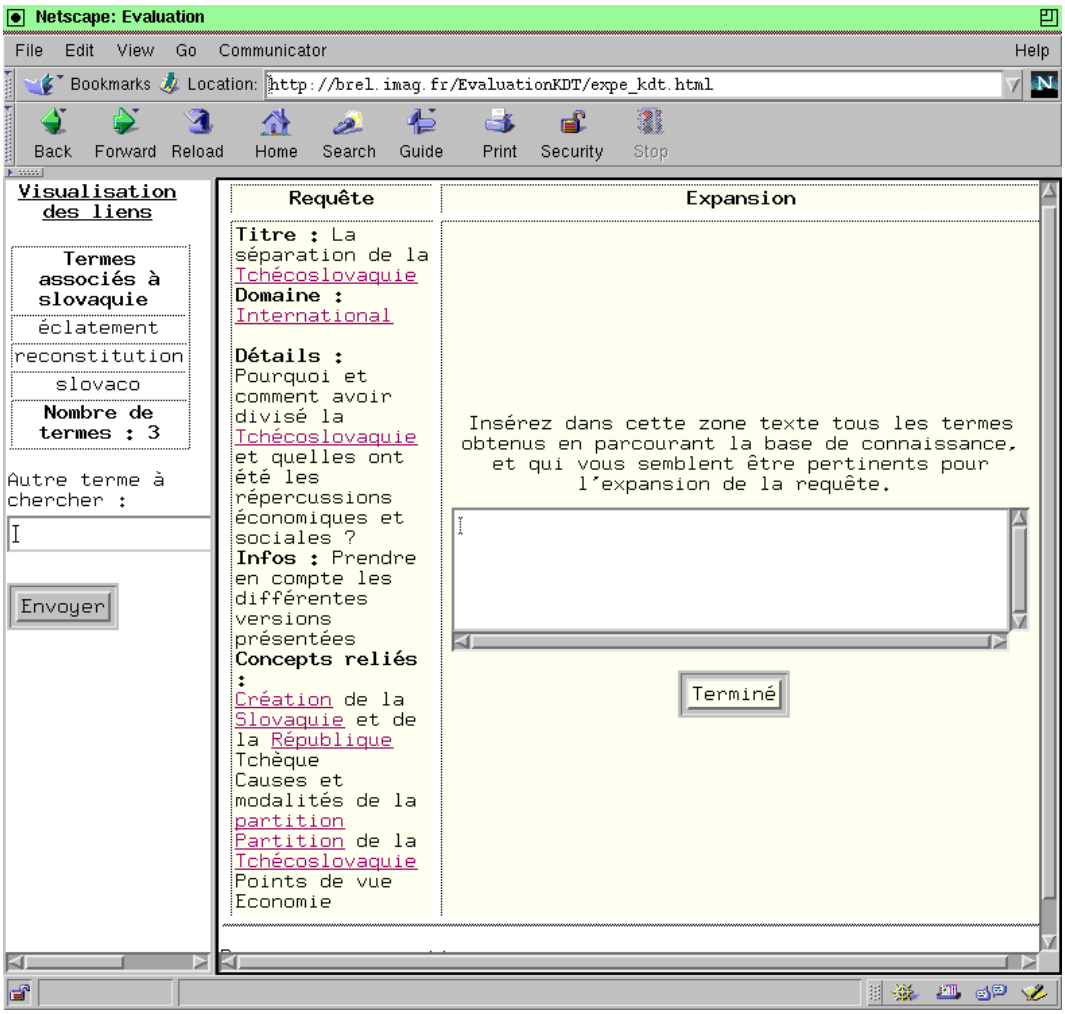

FIG. 6.7 – *Interface d'expansion*

Par cette expérimentation, nous avons voulu évaluer notre système dans un cadre autre

que les collections-test et qui ressemble plus à un contexte de recherche d'information où un utilisateur interroge un SRI et utilise les règles d'association pour enrichir sa requête. Ce contexte peut entre comparé à l'interrogation sur le Web qui est le sujet de notre section suivante.

6.9 Application au Web

Dans cette section, nous présentons une expérimentation de l'application des règles d'association dans le contexte du Web. L'objectif est de montrer l'impact de l'expansion des requêtes en utilisant des règles d'association en comparant les résultats obtenus aux résultats de deux moteurs de recherche : Altavista et Hotbot. Cette expérimentation, nous a amené à des réflexions intéressantes telle que la structure du Web. Nous commençons d'abord par présenter le type d'évaluation à utiliser dans le contexte du Web. Ensuite, nous présentons les étapes de notre expérimentation ainsi que les résultats obtenus.

6.9.1 Évaluation d'un SRI sur le Web

Dans le contexte du Web, une évaluation en terme de Rappel et Précision pose de nombreux problèmes :

- La quantité de documents présents sur le Web ne permet pas de déterminer quels sont les documents pertinents pour une requête donnée; il est impossible de porter un jugement sur chacun des documents d'une collection de plusieurs millions de pages HTML.

- L'hétérogénéité des documents du Web, aussi bien du point de vue de son contenu que de sa présentation, ne permet pas d'établir un jugement de pertinence *universel*, valable pour tous les documents.

- Le Web est dynamique et évolutif : le nombre de documents est en croissance constante. Les documents subissent une modification régulière pour la plupart d'entre eux.

- La pertinence d'une page Web n'est pas due uniquement à son contenu thématique propre mais aussi à l'espace informationnel auquel elle permet d'accéder. En effet, une page contenant essentiellement des liens peut se révéler très pertinente si ces liens sont des pointeurs vers les pages les plus intéressantes traitant du thème de la requête [Gér99].

- Une critique porte sur le jugement de pertinence binaire qui est généralement utilisé dans les collections de tests classiques [GS99]. Cette critique peut être faite pour

toute évaluation d'un SRI. Cela nous semble d'autant plus vrai dans le contexte hétérogène du Web. Il est intéressant d'utiliser un jugement de pertinence non binaire.

Dans ces conditions, établir une collection de tests (c'est-à-dire un ensemble de document, un ensemble de requêtes et des jugements de pertinence associés) n'est pas faisable. En particulier, le calcul du rappel est impossible : il n'existe pas de moyen permettant de déterminer TOUS les documents pertinents pour une requête. Par contre, il est envisageable de calculer la précision à n documents : un SRI propose n pages Web en réponse à une requête, un juge consulte chacune de ces pages pour émettre un jugement de pertinence et détermine les p documents pertinents pour la requête.

$$\textbf{Précision à } n \textbf{ documents} = \frac{||R_p||}{||R||}$$

Il existe une telle collection de test pour le Web [HCTH99], créée dans le cadre de la piste Web de la conférence TREC (Text REtrieval Conference).

Deux échantillons de 1 Go et 10 Go respectivement ont été extraits d'un sous-ensemble du Web de 100 Go. Un jeu de requêtes inspiré de ceux utilisé dans un contexte plus classique a été utilisé sur ces collections avec plusieurs SRI de la conférence TREC. Puis, des jugements de pertinence ont été faits sur ces réponses. Un des critères de qualité de ces SRI est la précision à 20 documents. .

Une critique importante à cette approche est le jugement de pertinence utilisé. Ce jugement est binaire, et ne tient pas compte de l'aspect hypertextuel du corpus comme nous l'avons évoqué plus haut et qui nous semble primordial.

Une approche similaire est présentée par Chakrabarti et al. dans [CDG+98]. Pour comparer son système Clever avec Altavista [Alt] et Yahoo! [Yah], il a défini une méthode basée sur la comparaison de la précision entre des systèmes. Il appelle cette méthode la *précision comparative*[2].

Pour cela, 26 requêtes sont utilisées : les 10 meilleures pages répondant à chacune de ces requêtes selon Clever, Altavista et Yahoo! sont présentées à des juges. Le juge peut considérer tous les paramètres qu'il pense nécessaires : contenu du document, présentation, espace d'information accessible, etc.

Le principal inconvénient de la précision comparative est que l'on ne peut pas différencier les améliorations/dégradations de performances dues :

– à la manière de constituer le corpus de pages Web, c'est-à-dire si l'indexation réalisée par le moteur a une bonne couverture du Web, si les pages indexées ne sont pas redondantes, etc.

2. en anglais *comparative precision*

– à l'indexation des pages, c'est-à-dire la manière d'extraire le contenu sémantique des documents.

– à la manière d'effectuer le classement des documents, c'est-à-dire de déterminer les documents les plus pertinents en appliquant une fonction de correspondance entre la requête et les documents.

En dépit de ces inconvénients, nous avons choisi d'évaluer notre système de cette manière.

6.9.2 Précision comparative

Pour une l'évaluation de notre système, nous avons défini un ensemble de requêtes portant sur des thèmes abordés dans bon nombre de pages du corpus tel que : interaction multimodale, réseaux de neurones pour la reconnaissance de la parole, recherche information multimédia.

Pour bien préciser le besoin, une description de chaque requête détaillant son sens est fournie aux juges. Huit juges ont participé à l'évaluation (professeurs et thésards du laboratoire CLIPS). Trois systèmes ont été évalués : Altavista, Hotbot et notre système. Les juges ont donné un score de pertinence pour chaque page : 0 pour non pertinent, 1 pour faiblement pertinente, 2 pour pertinente et 3 pour très pertinente.

Pour calculer la précision comparative, nous définissons la formule suivante pour évaluer la pertinence non binaire :

$$Score_s^r = \frac{\sum_{d=1}^{nb}(\sum_{j=1}^{nb^j} \frac{jug_j^d}{3})}{nb}$$

– $Score_s^r$ est le score du moteur de recherche s pour la requête r.

– d est le numéro d'un document.

– nb est le nombre de documents trouvés par le moteur de recherche s pour la requête r.

– j est le numéro d'un juge.

– nb^j est le nombre de juges qui participent à l'évaluation.

– jug_j^d est le score attribué par le juge j au document d.

Une moyenne pour chaque système a été calculée sur l'ensemble des requêtes.

6.9.3 Expérimentation

Pour évaluer l'apport de l'utilisation des règles d'association dans le cadre d'une interrogation sur le Web, nous nous sommes basés sur la *précision comparative*. Cette évaluation a nécessité plusieurs étapes :

- Collecte de pages Web : il est nécessaire de disposer d'un robot afin de constituer un corpus de pages HTML.
- Indexation : nous avons utilisé le SRI SMART [Sal71], basé sur un modèle vectoriel pour l'indexation classique des documents.
- Extraction des règles d'association : une base de connaissances constituée de règles d'association extraites du corpus Web est construite. Cette base sert à étendre les requêtes.
- Interrogation : il est nécessaire d'interfacer les différents moteurs de recherche afin de pouvoir soumettre des requêtes à chacun d'entre eux et de récupérer les références des pages Web retrouvés.
- Évaluation : ces références doivent être fusionnées pour les présenter aux juges qui vont associer à chacune un jugement de pertinence. Le fusionnement consiste à unifier les références en enlevant les doublons dans une seule liste afin de rendre la provenance d'une référence anonyme aux juges.

Dans la suite, nous allons détailler ces différentes étapes:

Création d'un corpus Web de pages HTML :

Dans le cadre de cette évaluation, CLIPS-Index (détaillé dans Géry et al [GH99]), a été utilisé pour construire localement un corpus de pages Web. Le corpus choisi est celui des pages Web de la fédération de laboratoires de l'IMAG (Institut d'Informatique et de Mathématiques Appliquées de Grenoble) accessibles à partir de l'URL *http://www.imag.fr*[3]. Les principales caractéristiques de ce corpus sont les suivantes :

- Taille en format HTML : environ 60.000 pages HTML identifiées par leurs URLs, pour un volume de plus de 415 Mo. Les serveurs concernés sont géographiquement proches les uns des autres, le rapatriement s'effectue en environ 1 heure.

3. La construction du corpus a été effectuée en Février 1999

- Taille en format textuel : après l'analyse de ces pages, il reste environ 190 Mo de données purement textuelles.
- Serveurs Host : ces pages résident sur 37 serveurs.
- Type de pages : la grande majorité des pages sont au format HTML. Toutefois, nous avons conservé d'autres formats purement textuels.
- Thèmes abordés : un grand nombre de thèmes sont abordés dans ces pages, mais on note évidemment une prépondérance des documents scientifiques, plus particulièrement dans le domaine de l'informatique.
- Taille du lexique : plus de 150.000 termes différents.

Indexation des documents :

Nous avons suivi la même approche d'indexation décrite dans la section 6.7.

Extraction des règles d'association :

Une phase de prétraitement des données, détaillée dans la section 6.6, a été effectuée sur le corpus (élimination des mots vides, certaines catégories grammaticales, etc.). Après cette phase de prétraitement, le corpus filtré obtenu est de taille réduite (environ 100 Mo) et il ne reste qu'environ 70 000 termes distincts sur les 150 000 initiaux. Les seuils de support et confiance utilisés pour l'extraction des règles d'association ainsi que le nombre de règles d'association découvertes sont présentés dans le tableau 6.6.

seuil de support minimum (nombre de documents)	seuil de confiance minimum	nombre de règles d'association
20	10%	24650

TAB. 6.6 – *Paramètres d'extraction des règles d'association*

Dans la base des règles d'association découvertes, un terme est, en moyenne, relié à quatre autres termes par des règles d'association. Pour chaque terme d'une requête, les termes reliés à ce dernier dans la base des règles d'association sont ajoutés dans la requête originale (expansion automatique).

Interrogation :

Dans l'objectif de comparer notre système à des moteurs de recherche sur le Web, une interface d'interrogation a été développée pour l'interrogation de SMART ainsi

que différents moteurs de recherche (Altavista, Yahoo, Hotbot, etc.) avec les mêmes requêtes. Les étapes de l'interrogation sont les suivantes:

- Une ou plusieurs requêtes sont spécifiées, soit dans un fichier soit directement en ligne. Dans ce dernier cas, l'interrogation se rapproche d'une interrogation d'un meta-chercheur comme savvysearch [Sav] ou metacrawler [Met].
- Envoi d'une requête HTTP précisant l'ensemble des informations nécessaires : termes de la requête, domaine de restriction, etc. L'expression de ces informations est différente pour chaque moteur : notre système est actuellement capable d'interroger les moteurs les plus connus et/ou les meilleurs du Web. Malheureusement, tous ces systèmes n'offrent pas les mêmes fonctionnalités : par exemple, plusieurs d'entre eux ne permettent pas de restreindre une interrogation à un sous-domaine du Web. Dans le cadre de cette expérimentation, nous avons restreint l'ensemble des moteurs de recherche à ceux qui permettent cette dernière fonctionnalité. Ainsi, nous pouvons restreindre l'interrogation des moteurs au domaine *http://www.imag.fr*. Particulièrement, nous avons sélectionné les moteurs de recherche *Altavista* et *Hotbot*.
- Analyse de la page de résultats fournis par le système. L'analyse est différente pour chaque moteur.
- Production automatique d'une page HTML présentant les résultats aux juges. Pour ne pas avantager tel ou tel moteur de recherche, les références proposées sont identifiées uniquement par leur URL et par leur titre. Le reste des informations habituellement proposées par les moteurs de recherche (résumé, taille, date, etc.) est éliminé. L'ordre dans lequel apparaissent les références est aléatoirement choisi. Quand plusieurs moteurs proposent une même page, elle n'apparaît qu'une seule fois dans nos résultats.

Résultats :

Les résultats de l'évaluation sont présentés dans le tableau 6.7.

	Altavista	Hotbot	notre système
Score	24.99	45.83	34.16

TAB. 6.7 – *Résultats de l'évaluation*

Une analyse détaillée de ces résultats montre que dans le cas d'une requête imprécise, notre système donne les meilleurs résultats. Dans les autres cas, c'est Hotbot qui donne les meilleurs résultats. Par exemple, les dix documents trouvés par notre

système pour la requête "*recherche d'informations multimédia*", sept de ces documents ont un score de 3, deux ont un score de 2, et le dernier un score 0. Alors que pour la même requête, les dix pages d'Altavista ont un score de 0 tandis que quatre pages de *Hotbot* ont un score de 2 et les autres ont un score de 1.

6.9.4 Conclusion

Cette approche, relativement facile à implanter, est capable d'utiliser des moteurs de recherche existants évitant une réindexation de tout le Web. Notre expérimentation a montré que l'utilisation des connaissances supplémentaire obtenues grâce à l'utilisation de la phase d'extraction des règles d'association peut augmenter l'efficacité des moteurs de recherche existants pour certaines requêtes. On a montré que les relations entre termes exprimées sous la forme de règles d'association peuvent améliorer les performances d'un système de recherche d'information sans avoir recours à des connaissances prédéfinies. Cette expérimentation a montré aussi l'impact des traitements linguistiques sur les résultats d'un SRI.

6.10 Fouille de Données Images

Dans les sections précédentes, nous avons montré que les règles d'association permettent d'acquérir des associations entre des termes d'un corpus. Que ce soit dans le contexte de collections d'évaluation classiques de recherche d'information ou bien dans le contexte du Web, nous avons montré que les règles d'association peuvent améliorer les performances d'un SRI en terme de rappel et précision. L'approche des règles d'association permet aussi de traiter des éléments de natures différentes. En effet, elle peut être appliquée aux éléments d'une base de données, aux termes d'un corpus mais aussi elle peut être appliquée à un ensemble de données hétérogènes. Dans ce chapitre, nous appliquons l'approche des règles d'association à une combinaison d'éléments de deux types : des éléments multimédia et des éléments textuelles. La souplesse de l'approche des règles d'association permet de l'adapter selon l'objectif visé et la nature des éléments à traiter. Les algorithmes d'extraction des règles d'association sont alors adaptés au traitement souhaité.

Nous introduisons dans cette section un champ d'application de la fouille de données aux collections d'images, que nous appelons *Fouille de Données Images* (FDI). Tous les aspects de la fouille de données peuvent être intéressants pour la FDI mais nous nous concentrons sur le problème de découverte de connaissances à partir du contenu des images sous la forme d'associations entre des éléments symboliques (concepts) et des éléments du signal (caractéristiques ou propriétés des images).

L'accroissement des données multimédia oblige à aller au-delà d'une indexation manuelle et à prendre en compte ces données au cours du processus d'interprétation et d'indexation. Notre but est alors d'exploiter l'information contenue dans ces données pour un besoin de recherche d'information (RI) et plus particulièrement pour une indexation automatique des images dans un système de recherche d'information multimédia (SRIM). Nous n'allons pas aborder dans cet article les problèmes d'indexation multimédia mais notre objectif est d'étudier la combinaison de données multimédia et des données textuelles. Pour cela, les techniques de la fouille de données nous semblent intéressantes à appliquer dans ce contexte. Plus particulièrement, nous allons adapter la technique des règles d'association pour découvrir des associations entre des éléments du signal et des éléments symboliques. Notre objectif est alors d'indexer des images avec des éléments du signal en leurs associant automatiquement des éléments symboliques. Les règles d'association visées sont celles où un ensemble d'éléments symbol (la prémisse de la règle d'association) implique un ensemble d'éléments symboliques (la conclusion de la règles d'association):

$$\text{un ensemble d'éléments symbol} \Rightarrow \text{un ensemble d'éléments symboliques}$$

Notre approche est alors basée sur l'utilisation d'une collection d'apprentissage de photographies décrites à la fois par une annotation manuelle et par des régions segmentées automatiquement. Cette collection permet de découvrir les règles d'association utilisées pour indexer ultérieurement de nouvelles images. Chacun de ces éléments, la segmentation, l'annotation manuelle, la définition des règles d'association considérées ainsi que les évaluations son décrites dans la suite.

6.10.1 Processus de segmentation et d'extraction des caractéristiques des images

Le processus de segmentation a pour but de regrouper les pixels en régions d'après des caractéristiques propres aux pixels et à leur pixels voisins. Nous avons détaillé le processus de segmentation dans [HM01a].

Le résultat de ce processus est, pour chaque pixel de l'image traitée, une association à un et un seul cluster. Dans un second temps, nous déterminons les régions de pixels connexes appartenant au même cluster. Les régions obtenues sont donc disjointes, i.e. ne partagent aucun pixel. La Figure 6.8 montre le résultat de la segmentation pour la photographie de la Figure 6.9.

Une fois le processus de segmentation achevé, chaque région est associée à un ensemble de caractéristiques. Les caractéristiques retenues sont les suivantes:

– la couleur principale : la couleur la plus dominante de la région. Ces couleurs sont représentées dans l'espace RGB sous-échantillonné pour obtenir 64 couleurs.

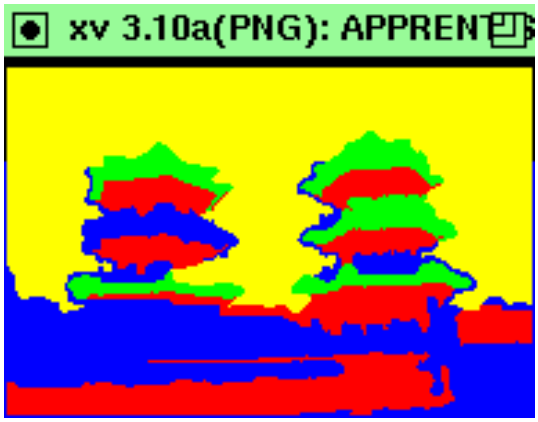

FIG. 6.8 – *Segmentation de l'image "le jardin chinois"*

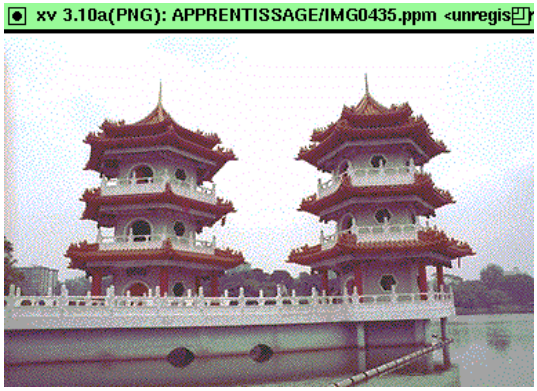

FIG. 6.9 – *Exemple d'image d'apprentissage : le jardin chinois*

– la couleur secondaire : la seconde couleur la plus dominante de la région. Ces couleurs sont représentées dans l'espace RGB sous-échantillonné pour obtenir 64 couleurs.

– la direction principale et la direction secondaire : nous déterminons les directions d'une région en passant sur tous les pixels de la région un filtre de gradient (opérateur de Solbel) pour déterminer les pixels qui sont sur des zones de forts gradients d'énergie. L'intérêt de l'opérateur de Sobel est de fournir également la direction de ce gradient. Pour chaque pixel de la région, nous pouvons donc déterminer la

direction de gradient la plus dominante, appelée direction principale (36 directions principales possibles), et la direction de gradient la plus dominante après la direction principale, appelée direction secondaire (36 directions secondaires possibles)

– la texture ou la non texture : en nous basant sur les résultats obtenus sur les coefficients DCT obtenus pour chaque pixel d'une région, nous sommes en mesure de définir si une zone est texturée ou non. Plus précisément, nous nous basons sur les moyennes des coefficients DCT des pixels. On a alors 2 choix de textures possibles.

L'utilisation des couleurs et des directions dominantes est inspirée de l'une des propositions faites par Yihong Gong dans [Gon99]. Le nombre de dimensions des caractéristiques provenant du signal est important (64 couleurs principales * 64 couleurs secondaires * 36 directions principales * 36 directions secondaires * 2 choix de texture). Ces caractéristiques sont utilisées par la suite lors des opérations de la fouille de données.

6.10.2 Processus d'annotation manuelle des images

L'annotation manuelle des images est réalisée par l'intermédiaire d'une interface présentée Figure 6.10. La personne qui annote détermine le contour intérieur approximatif d'une région et associe cette région à un symbole pris parmi une liste prédéfinie. Dans les travaux reportés ici, nous avons choisi de nous intéresser à des images d'extérieur comprenant des paysages et des bâtiments. La Figure 6.9 présente l'une de ces images. La liste de concepts déterminée est composée de 26 symboles.

La Figure 6.10 montre le processus d'annotation manuelle de la photographie de la Figure 6.9. Les termes d'annotations de cette photographie sont les suivants : Ciel, Façade Immeuble, Rivière, Autre Construction et Groupe Arbre.

6.10.3 Collections d'apprentissage

Le corpus d'apprentissage est composé de 67 photographies d'extérieur (cf. section 6.10.2). En moyenne, le nombre de régions par image est de 9.8 et le nombre de concepts manuellement associés à une image est de 7,6.

Notre objectif est de découvrir des règles d'association entre des caractéristiques signal extraites de la segmentation et des caractéristiques symboliques venant d'une annotation manuelle.

Nous étudions l'influence de trois éléments indépendants sur la qualité des règles d'association découvertes :

– l'indépendance des caractéristiques signal extraites,

– les associations entre caractéristiques et symboles dans une image ou une région,

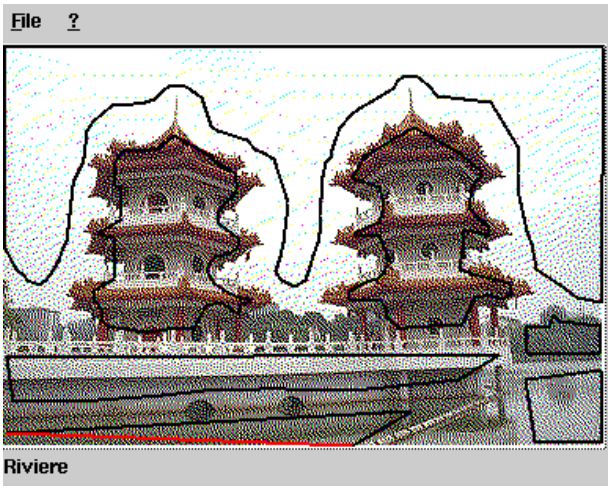

FIG. 6.10 – *Annotation de l'image "le jardin chinois"*

– la dimension des caractéristiques signal potentielles.

Ces éléments sont détaillés ci-dessous:

Paramètre d'indépendance des caractéristiques signal :

Les caractéristiques signal extraites peuvent être utilisées de deux manières :

– l'ensemble des caractéristiques d'une région est considéré comme atomique.
– chacune des caractéristiques d'une région est atomique donc les caractéristiques sont considérées comme indépendantes.

Prenons l'exemple d'une région segmentée, elle possède comme couleur principale (40, 40, 29), comme couleur secondaire (30, 10, 0), comme direction principale 280 degrés, comme direction secondaire 190 degrés, et n'est pas texturée. Le premier choix considère tous ces paramètres comme fournissant une caractéristique de l'image. Le second choix considère que l'image contient une région possédant une région de couleur (40, 40, 29), une région de couleur secondaire (30, 10, 0) qui peut être la même ou une autre, etc. Nous avons choisi ce second choix pour étudier dans quelle mesure l'affranchissement des contraintes inter-caractéristiques signal influe sur les résultats.

Paramètres liés à une image ou une région :

Ce paramètre détermine si les liaisons entre signal et symboles fournies au processus d'apprentissage sont basées sur les descriptions de l'image complète ou bien sur chaque région:

- soit on applique l'apprentissage sur des descriptions n'utilisant pas de liens entre les caractéristiques signal et symboliques d'une image, dans ce cas il s'agit d'un apprentissage non supervisé. Notre espoir est que l'apprentissage va permettre de retrouver les liaisons avec les symboles pertinents.
- soit on utilise l'apprentissage sur des descriptions utilisant une simple union des caractéristiques signal et symboliques d'une image. Nous sommes donc dans le cas d'un apprentissage supervisé.

Nous allons montrer que, comme nous pouvons nous y attendre, ce paramètre a une grande influence sur la qualité des règles générées.

Paramètres liés à la dimmensionnalité des caractéristiques signal :

Vu le nombre important de dimensions pour les caractéristiques signal de l'image par rapport aux nombres de concepts, trois stratégies ont été appliquées pour diminuer le nombre de dimensions signal et équilibrer les dimensions signal et symboliques : la réduction des couleurs, la réduction des directions et la réduction des couleurs et les directions.

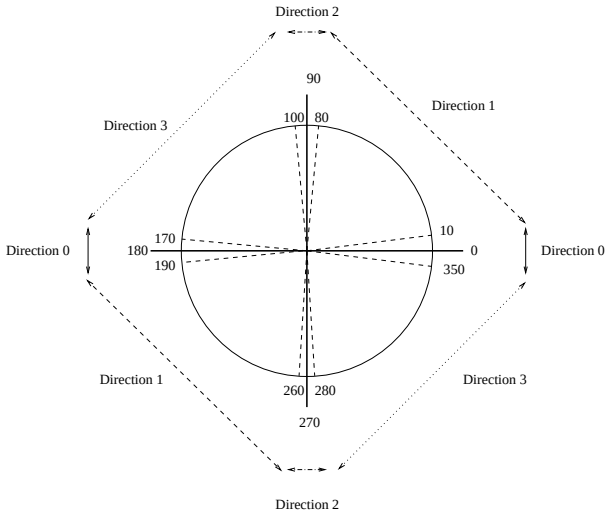

FIG. 6.11 – *Réduction des directions à 4 directions*

- La réduction des couleurs.

  Nous avons choisi de réduire les dimensions des couleurs à 8 couleurs possibles au lieu de 64, aussi bien pour les couleurs principales que pour les couleurs secondaires. Les combinaisons des caractéristiques dans ce cas sont réduites à 165888 combinaisons possibles.

- la réduction des directions.

  Il s'agit de réduire les 36 directions à 4 directions (Figure 6.11). Nous avons choisi de regrouper les directions verticales en une seule direction (direction 0), les directions horizontales en une seule direction (direction 2) et les directions obliques en deux directions (direction 1 et direction 3). Les combinaisons des caractéristiques dans ce cas sont réduites à 131072.

- la réduction des couleurs et des directions.

  Il s'agit d'appliquer les deux processus de réduction décrits ci-dessus. Les combinaisons des caractéristiques dans ce cas sont limitées à 2048.

L'ensemble des collections d'apprentissage :

En utilisant les différents choix décrits plus haut, nous obtenons 16 collections d'apprentissage. Dans la Figure 6.12, qui présente 12 de ces collections, nous utilisons le suffixe _C8 pour les collections où l'on a réduit le nombre les couleurs, le suffixe _D4 pour les collections où l'on a réduit le nombre de directions et le suffixe _A pour les collections où l'on a appliqué un apprentissage supervisé. Le processus de fouille de données est appliqué sur chacune de ces collections.

Nous avons choisi de nous limiter à l'étude des résultats obtenus pour les collections COL1_A_C8,COL2_A_C8, COL1_C8_D4 et COL2_C8_D4. Ce choix est dû au fait que ces collections nous donnent les résultats les plus significatifs pour notre étude.

Les règles d'association :

Afin d'extraire des règles d'association dont la prémisse est formée d'éléments symbol et la conclusion est formée d'éléments symboliques, l'algorithme utilisé dans le contexte de données textuelles, présenté dans la section 6.3, n'est pas adapté à ce besoin. En effet, ce dernier ne permet pas d'avoir plusieurs éléments dans la prémisse ainsi que plusieurs éléments dans la conclusion. Ainsi, nous utilisons l'algorithme APRIORI [AS94], détaillé dans l'annexe C, pour extraire les règles d'association que nous modifions pour l'adapter à notre objectif. L'application de l'algorithme APRIORI pour la base d'images suppose que l'on définisse dans ce cadre la notion de transaction et d'items. Pour cette application, on considère que les caractéristiques signal et les concepts sont des items. Dans le cas des collections COL1_C8_D4 et COL2_C8_D4, les transactions sont représentées par les images

Collection	Niveau atomique	type de réduction	type de l'apprentissage
COL1	région		non supervisé
COL2	caractéristique		non supervisé
COL1_C8	région	couleurs	non supervisé
COL2_C8	caractéristique	couleurs	non supervisé
COL1_A	région		supervisé
COL2_A	caractéristique		supervisé
COL1_A_C8	région	couleurs	supervisé
COL2_A_C8	caractéristique	couleurs	supervisé
COL1_D4	région	directions	non supervisé
COL2_D4	caractéristique	directions	non supervisé
COL1_C8_D4	région	couleurs et directions	non supervisé
COL2_C8_D4	caractéristique	couleurs et directions	non supervisé

FIG. 6.12 – *Caractéristiques des collections d'apprentissage*

alors que dans le cas des collections COL1_A_C8 et COL2_A_C8, les transactions sont représentées par le couple caractéristique signal et symbole.

6.10.4 Processus d'évaluation

Nous avons envisagé deux types d'évaluations qualitatives des règles d'association : d'une part du point de vue de l'étiquetage sémantique des régions et d'autre part du point de vue de l'indexation automatique des images.

Filtrage des règles d'association :

La problématique des règles d'association est de découvrir s'il existe des associations fortes entre les caractéristiques physiques de l'image, que se soient les régions ou les caractéristiques, et l'annotation symbolique. Pour cette raison, on a ajouté la contrainte qui consiste à avoir exclusivement une ou plusieurs caractéristiques de l'image dans la partie gauche d'une règle d'association et exclusivement un seul concept dans la partie droite d'une règle d'association.

Etiquetage sémantique des régions :

Pour une région C donnée d'une image, s'il existe une règle d'association R qui associe cette région à un concept S sous la forme $C \Rightarrow S$ alors on évalue si cette région correspond bien au concept S. Dans le cas où plusieurs règles d'association

seraient sélectionnées pour une région donnée, la règle qui possède la plus grande valeur de confiance est utilisée pour associer un concept à une région.

Deux mesures ont été définies pour évaluer les résultats. La première mesure, le Rappel des régions, notée Rappel_régions, évalue la capacité du système à associer un concept avec une région.

$$\text{Rappel\_régions} = \frac{\text{nombre de régions associées à un concept}}{\text{nombre de régions total de la collection}}$$

La deuxième mesure, notée Précision_régions, évalue la capacité du système à associer le bon concept à une région.

$$\text{Précision\_régions} = \frac{\text{nombre de régions correctement associées à un concept}}{\text{nombre de régions associées à un concept}}$$

Plus les valeurs de Rappel_région et de Précision_région sont importantes plus le système est jugé performant pour étiqueter sémantiquement les régions des photographies.

Évaluation de l'indexation automatique :

Nous nous plaçons ici dans le cadre où les règles d'association seront utilisées comme base d'indexation des photographies. Les concepts associés aux régions d'une image sont alors évalués par rapport aux concepts affectés grâce à l'annotation manuelle des photographies.

Nous avons définit d'une part la complétude d'une image qui évalue l'association d'un concept d'indexation correctement à une image :

$$\text{Complétude d'une image} = \frac{\text{nombre de concepts correctement associés à une image}}{\text{nombre de concepts associés manuellement à une image}}$$

La complétude moyenne des images est alors la moyenne des complétudes des images.

Nous avons définit d'autre part la complétude d'un concept qui évalue l'association correcte d'un concept aux images :

$$\text{Complétude d'un concept} = \frac{\text{nombre d'association correcte du concept aux images}}{\text{nombre d'associations manuelles du concept aux images}}$$

La complétude moyenne des concepts est alors la moyenne des complétudes des concepts.

Ces deux mesures de complétude sont inspirées de [Ber97b].

Collection	Nombre de transactions	Nombre de items dans la collection	support minimum	confiance minimale	nombre de 1-item fréquents	nombre de règles découvertes
COL1_A_C8	673	265	0.2%	5%	200	33
COL2_A_C8	2866	158	0.1%	5%	197	112
COL1_C8_D4	67	228	2%	10%	59	996
COL2_C8_D4	67	120	6%	20%	52	41557

TAB. 6.8 – *Mesures utilisées lors du traitement des collections d'apprentissage*

6.10.5 Expérimentations

Processus d'apprentissage :

Nous avons appliqué l'algorithme APRIORI aux 12 collections d'apprentissage. Les seuils de support minimum et de confiance minimum utilisés pour les meilleures 4 collections d'apprentissage ainsi que les nombres de transactions, d'items, de 1-items fréquents et de règles d'association sont présentées dans le tableau 6.8.

Processus d'évaluation :

Nous avons sélectionné 100 images, qui ne font pas parties de la collection d'apprentissage, pour constituer notre collection test. A partir de cette collection, nous avons obtenu 12 collections tests en suivant la même approche que dans la section 6.10.3 appliquée aux collections d'apprentissage.

Pour les collections COL1_A_C8 et COL1_C8_D4, le nombre de règles d'association est très faible. Les valeurs des mesures des rappels des régions et des concepts sont faibles, par conséquent la complétude des images et la complétude des concepts le sont aussi. Ceci s'explique par le fait qu'il n'y a pas des répétitions fréquentes dans ces collections. Cette approche de représentation des caractéristiques des images semble alors trop rigide pour permettre des découvertes d'associations entre des données de l'image et des données symboliques.

L'utilisation des caractéristiques indépendantes dans le cas des collections COL2_A_C8 et COL2_C8_D4 permet d'indexer toutes les images. Ceci s'explique par le fait que cette approche est plus souple que la première. En effet, à la différence des collections mentionnées précédemment où il faut que toutes les caractéristiques d'une région soient associées à un concept dans une règle d'association pour que la région soit associée à ce concept, une région est associée à un concept même si quelques caractéristiques ne sont pas associées à ce concept.

L'approche non supervisée a permis d'avoir des meilleurs résultats que celle supervisée. En effet, cette dernière semble restrictive du fait qu'elle ne permet pas de construire des règles d'association qui n'existent pas déjà dans les données d'apprentissage. Avec les nouvelles données tests, elle n'est pas capable de se prononcer sur les associations latentes probables. Par contre, l'approche non supervisée est capable de découvrir des associations nouvelles entre les données de l'apprentissage. Cet avantage lui permet de découvrir un grand nombre de règles d'association et donc d'être plus performante sur les données à tester.

L'influence de la réduction des couleurs ainsi que la réduction des directions a été détaillée dans [HM01b]. Nous présentons dans le tableau 6.9, les résultats des quatre collections jugées les meilleures pour notre objectif d'indexation automatique. La collection test COL2_C8_D4 est celle qui permet d'avoir les meilleures performances.

	COL1_A_C8	COL2_A_C8	COL1_C8_D4	COL2_C8_D4
Complétude moyenne des images	5.29%	44.72%	29.47%	59.3%
complétude moyenne des concepts	1.10%	12.08%	6.87%	19.33%
Rappel_régions	7.35%	78.36%	26.69%	80.88%
Précision_ régions	40%	66.33%	50.26%	87.27%

TAB. 6.9 – *Résultats des évaluations*

On remarque que la complétude moyenne des images est supérieure à la complétude moyenne des concepts. Ceci s'explique par le fait que le nombre d'occurrences de certains concepts, tels que Ciel ou Façade Immeuble, qui sont présents dans presque toutes les images, est très important par rapport au nombre d'occurrence des autres concepts tels que Piscine ou Rocher qui occurrent dans une ou deux images ce qui fait qu'ils ne sont pas des 1-items fréquents et donc ces concepts ne peuvent pas être associés à une région.

6.10.6 Conclusion

Notre objectif est de combiner les données textuelles avec celles qui proviennent du signal dans un but d'indexation automatique des images. Il s'agit d'améliorer la qualité d'un SRIM en vue d'obtenir une indexation sémantique. On a été emmené à réfléchir sur l'approche à suivre pour aboutir à cet objectif. Pour l'atteindre, nous avons considéré des associations entre des caractéristiques signal et des caractéristiques symboliques. La technique des règles d'association permet de répondre précisément à ce besoin. Nous avons

choisi une approche par apprentissage dans un contexte de collection d'images homogènes. Les résultats de nos expérimentations montrent qu'un processus de fouille peut être utilisé pour combiner deux sources de données différentes, dans notre contexte des données signal et des données symboliques. La FDI est alors une perspective intéressante pour l'indexation automatique des images.

6.11 Conclusion

Les résultats présentés dans ce chapitre montrent bien l'intérêt, du point de vue de la recherche d'information, des règles d'association. Nous avons montré que l'utilisation des règles d'association, que ce soit dans le contexte de l'expansion automatique des requêtes ou dans le contexte de l'expansion interactive des requêtes, permet d'augmenter les performances d'un système de recherche d'information en terme de rappel et précision. Nous pouvons maintenant affirmer que les règles d'association permettent de découvrir des liens entre les termes qui reflètent le contexte d'utilisation d'un terme dans un corpus ainsi que le contenu thématique des documents du corpus.

Nous avons également montré l'intérêt, dans le contexte de la fouille de données images, des règles d'association de traiter des données de natures différentes (signal et symbole). Cette souplesse des règles d'association est une spécificité précieuse pour le traitement de données hétérogènes ce qui est le cas du Web par exemple.

Cependant, nous avons mis l'accent sur l'importance de la phase de prétraitement des données, principalement le traitement linguistique. Cette dernière évite la redécouverte de la structure syntagmatique des phrases par l'utilisation des syntagmes nominaux. Dans la partie suivante, nous présentons le point de vue linguistique de notre formalisme qui prend en compte les rapports syntagmatiques entre les unités textuelles.

Troisième partie

Les syntagmes nominaux pour représenter le sens

Chapitre 7

Représentation du contenu textuel

La définition fait connaître ce qu'est la chose.
ARISTOTE

On peut noter en examinant les recherches récentes dans le domaine de la recherche d'information un regain d'intérêt des approches linguistiques. Ce réveil est mis en évidence par l'importance soudaine qu'a pris la technologie de la RI dans les applications industrielles, et particulièrement sur Internet comme le mentionne les rapports de la piste traitement automatique du langage naturel [1] (TALN) de TREC. Les conférences annuelles TREC (Text REtrieval Conference) ont l'objectif d'évaluer les SRI manipulant des corpus volumineux textuels. Démarrées en 1992, ces conférences visent à fournir aux concepteurs de SRI des pistes de développement dont la piste TALN qui a commencé lors de TREC-4. Les résultats de ces rapports soulignent à l'unanimité le rôle important et l'impact que les traitements linguistiques peuvent avoir sur un Système de Recherche d'Information (SRI) [SC94, SC95, SFLG+97].

Dès l'origine, il a été proposé un minimum de traitements linguistiques dans un SRI. Ces traitements se limitaient à la troncature des mots extraits du corpus et à l'élimination des mots outils de la langue. Ces traitements, bien que rudimentaires, sont toujours appliqués dans les travaux sur l'indexation car ils sont aisés à mettre en oeuvre sur des grands corpus. Ces travaux ne tiennent pas compte des phénomènes linguistiques tels que la variation morphologique, lexicale, syntaxique ou sémantique [Jac97, CH01].

7.1 Traitement automatique de la langue naturelle

Le traitement automatique du langage naturel (TALN) peut être défini comme étant le domaine de l'ingénierie linguistique qui a comme objectif la conception de logiciels

1. en anglais *Natural Language Processing*

ou programmes, capables de traiter de façon automatique des données linguistiques. Ces données linguistiques peuvent être des textes écrits ou bien des dialogues oraux ou encore des unités linguistiques de taille inférieure à ce que l'on appelle habituellement des textes (par exemple : des phrases, des énoncés, des groupes de mots ou simplement des mots). C'est donc un champ de la technologie informatique qui permet d'analyser et de représenter des données textuelles à un ou plusieurs niveaux de compréhension (morphologique, syntaxique, etc.). On parle également de linguistique informatique ou linguistique computationnelle pour la partie de la linguistique qui concerne le traitement automatique; le terme d'informatique linguistique désigne, quant à lui, plutôt la discipline de l'informatique qui s'intéresse au langage.

On distingue 6 niveaux pour l'analyse linguistique qui reflètent 6 niveaux de compréhension de la langue [Jou93, AWKB00, Ama00, Sau72] :

– Niveau phonologique :

  La phonétique est l'étude scientifique des sons du langage humain. Ce niveau réfère à la façon dont les mots sont prononcés. Il n'est pas important en ce qui concerne le repérage de textes écrits, mais s'avère crucial pour la compréhension du langage oral et dans les systèmes de reconnaissance vocale. A ce niveau, la plus petite unité de traitement est le *phonème*.

– Niveau morphologique :

  L'analyse morphologique permet de traiter les variations de surface de chaque mot du texte (chaîne de caractères séparés par des espaces) en prenant en compte les formes fléchies ou variations apparentes du mot. Un analyseur morphologique permet par exemple de : traiter les formes du pluriel d'un mot, identifier les caractères minuscules ou majuscules, les abréviations, reconnaître les locutions, les expressions et les noms composés, isoler une seule forme canonique pour toutes les formes rencontrées d'un mot, etc. L'unité minimale d'une forme signifiante est le *morphème*.

– Niveau lexical:

  L'analyse lexicale permet d'une part de rechercher l'existence des mots et des expressions du texte dans un dictionnaire linguistique. D'autre part, elle permet de confirmer ou d'infirmer l'existence des morphèmes identifiés par l'analyse morphologique. Par opposition aux morphèmes, nous parlerons ici de *lexème* pour désigner un mot canonisé et signifiant.

– Niveau syntaxique:

  L'objectif de l'analyse syntaxique est d'exploiter toutes les indications provenant de la structure du texte et permettant d'en construire une représentation sémantique

la plus exacte et complète possible. La syntaxe nous dit, en particulier, quelles associations entre prédicats et arguments peuvent être exprimées d'après les propriétés syntaxiques et l'arrangement des mots dans une phrase donnée. Un analyseur syntaxique analyse dans un premier temps les groupes de mots de la phrase qui forment des unités fonctionnelles (principalement les syntagmes) et génère dans un deuxième temps un arbre syntaxique de la phrase. Une des difficultés de l'analyse syntaxique est la détection par exemple de syntagmes nominaux ou encore la désambiguîsation syntaxique d'un mot.

– Niveau sémantique:

  L'analyse sémantique est l'étude linguistique du sens et son objectif est donc de déterminer le sens des mots et des phrases. Les mots et les structures des phrases identifiées lors des analyses morphologiques, lexicales et syntaxiques, constituent autant d'indices pour le calcul du sens.

– Niveau discursif:

  Le niveau discursif exploite la structure documentaire des différents types de documents et de requêtes en vue d'une extraction du thème. On peut ainsi tirer parti, par exemple, des traits structurels caractéristiques d'un article de journal, d'un article scientifique, d'un roman policier, etc. En profitant de cette structure prévisible, le TALN peut déterminer le rôle d'un fragment d'information spécifique dans un document (opinion, fait, prédiction, conclusion, etc.). La résolution des anaphores peut se faire également à ce niveau du TALN.

– Niveau pragmatique:

  L'analyse pragmatique permet d'utiliser les connaissances pragmatiques (par exemple les connaissances qui découlent du sens commun) afin d'interpréter les situations du monde réel.

La limite entre ces différents niveaux d'analyse n'est pas toujours très nette. Du fait de la très grande interaction entre les niveaux, il est parfois difficile de définir à quel niveau se situe l'analyse (interaction syntaxe-sémantique, morphologie-syntaxe, etc.). Les difficultés rencontrées par le traitement automatique augmentent au fur et à mesure que l'on avance vers les niveaux supérieurs de compréhension. Bien entendu, tous les systèmes de TALN n'opèrent pas sur l'ensemble de ces niveaux. Le TALN peut être intégré à une ou plusieurs composantes d'un SRI et correspond à plusieurs niveaux de compréhension des documents :

– Pour l'indexation des documents et le traitement des requêtes : identification de termes complexes et de *bonnes unités textuelles* pour représenter le contenu.

– Pour la formulation des requêtes : une analyse de la requête de l'utilisateur et un processus de dialogue permettant à l'utilisateur de mieux formuler son besoin.

L'utilisation du TALN pour la recherche d'information soulève un certain nombre de questions:

– Le TALN est un vaste domaine où il existe différentes approches et techniques destinées à différents phénomènes linguistiques. Le choix de traiter un phénomène langagier plutôt qu'un autre n'est pas une question facile à aborder.

– A quel(s) niveau(x) de traitement et quelle(s) composante(s) du SRI ces phénomènes langagiers doivent être utilisés ?

  Un traitement linguistique peut être appliqué à une ou plusieurs composantes d'un SRI (indexation et/ou interrogation). Le choix de l'appliquer à une composante particulière plutôt qu'une autre n'est pas évident.

– Quel impact qualitatif et quantitatif va avoir l'ajout du TALN sur un SRI ?

  L'utilisation du TALN dans un SRI a pour but d'augmenter les performances de ce dernier. Par performances, nous faisons référence aux performances quantitatives et qualitatives. Les performances quantitatives sont exprimées généralement par les mesures de Rappel et de Précision. Ces mesures sont trop limitatives pour permettre d'évaluer l'apport du TALN. C'est pourquoi une étude qualitative qui s'exprime généralement par une évaluation subjective (effectuée généralement manuellement par un linguiste) est effectuée en parallèle de l'étude quantitative.

Tout au long de notre étude, nous essayerons de répondre à ces questions rarement évoquées dans la littérature.

7.2 Unités linguistiques

Les niveaux de l'analyse linguistique permettent d'identifier et de décrire des unités linguistiques. Les plus petites unités linguistiques sont les phonèmes. Les phonèmes se combinent pour former des morphèmes, puis des mots, puis des phrases. Entre le mot et la phrase, il existe deux autres unités grammaticales : le syntagme et la proposition. Nous pouvons distinguer le syntagme et la proposition de la manière suivante : tout groupe de mots qui est grammaticalement équivalent à un seul mot et qui n'a pas son propre sujet et son propre prédicat est un syntagme, au contraire, un groupe de mots qui a son propre sujet et son propre prédicat est une proposition s'il est inclus dans une phrase plus grande [Lyo70]. Excluant les phonèmes, puisqu'on se place dans le cadre d'unités textuelles, les unités linguistiques sont donc le morphème, le mot, le syntagme, la proposition et la

phrase. La relation, qui existe entre les cinq unités est une relation de composition. En effet, on peut dire, d'un point de vue grammaticale, que les phrases sont composées de propositions qui sont composées de syntagmes eux même sont composés de mots et ces derniers sont composés de morphèmes. Il y a donc une hiérarchie entre ces unités. Une unité linguistique peut donc s'analyser en unités de niveau inférieur [Lyo70]. Nous allons exploiter cette caractéristique de la langue dans l'extraction d'unités textuelles (section 7.5.2 et 7.5.3) et principalement dans leur structuration (section 9.1).

7.3 Représentation simple vs représentation complexe

Les fondements du domaine de la RI depuis plusieurs dizaines d'années, ont été définis principalement par Gerald Salton [Sal71]. La plupart des SRIs actuels se basent toujours sur l'hypothèse initiale qu'un document doit partager les termes d'une requête pour être identifié comme pertinent. Bien entendu, la force de cette relation de pertinence [2] est proportionnelle à l'intersection des termes entre le document et la requête. Un poids affecté à un mot clé précise l'importance de ce dernier dans le document. Que le modèle soit vectoriel ou probabiliste ou logique, ce poids est une fonction du nombre d'occurrences du terme dans le document. Le problème de la RI semble alors se résumer à un simple calcul de correspondance entre un ensemble de mots clés de la requête de l'utilisateur avec l'ensemble des mots clés représentant le document. Cette représentation souffre d'un sérieux inconvénient qui est le fait que les termes simples sont souvent ambigus et peuvent, suivant les contextes, se référer à des concepts différents:

– Dans le contexte d'unité lexical atomique [3].

  Si l'on considère le mot composé *pomme de terre*, les mots simples *pomme* et *terre* ne gardent pas leur propre sens que dans l'expression *pomme de terre* et si on les utilise séparément ils deviennent une source d'ambiguïté.

– Dans le contexte de termes complexes.

  Si l'on considère les deux termes *voiture* et *marque*, ils ne sont pas assez spécifiques pour qu'une distinction existe entre *voiture de marque* et *marque de voiture*.

Dans le premier cas, il est souhaitable d'utiliser l'expression pour l'indexation et ne pas inclure les mots simples dans l'index. En effet, l'inclusion des mots simples permet de retourner le document dans le cas d'une interrogation avec *pomme terre* alors qu'il n'est

2. appelée pertinence système par opposition à la pertinence utilisateur que le système tente d'approcher
3. une unité lexicale atomique fait référence à une expression où au moins le sens d'un des constituants est différent s'il est utilisé seul ou dans une expression

pas pertinent pour cette interrogation. Dans le second cas, il est souhaitable de remplacer les mots simples par l'expression plus spécifique et de fait plus discriminatoire.

Les méthodes de représentation du contenu à partir de mots-cléfs simples sont alors biaisées. Si pertinentes soient-elles, les combinaisons de mots-clefs associées aux documents ne sont que des vues très partielles, *des trous de serrure* [Jac97], par lesquels le texte est imparfaitement représenté. La faiblesse de la méthode réside donc dans sa phase initiale qui consiste à substituer à un texte plein un ensemble de mots. On peut également le comprendre en examinant le texte : *des quartiers avec une architecture moderne ont vu le jour à côté de la vieille ville* qui ne traite pas de *l'architecture moderne de la vieille ville* même si les même mots apparaissent ensembles dans le document.

Nous considérons donc qu'un texte n'est pas seulement un *sac de mots*[4], mais il est bien véritablement un ensemble fortement structuré de termes qui permettent de communiquer des informations d'une grande précision. Les mots simples ne peuvent pas être considérés comme un langage de représentation expressif et précis du contenu sémantique

Plusieurs travaux ont souligné le problème de représentation du contenu en RI, principalement la représentation des termes et des relations entre les termes dans une phrase ou un morceau de phrase. La question qui se pose est de savoir si la représentation de cette information est importante et si on doit en tenir compte lors de l'analyse du texte. L'utilisation d'une représentation complexe revient à laisser les mots dans le contexte dans lequel les auteurs les ont écrits, en opposition à l'indexation classique, où les mots sont détachés de leurs contextes. Une première étape consiste à utiliser les syntagmes nominaux, pour la représentation du contenu sémantique d'un document. Nous sommes conscients que les performances et la qualité de l'extraction et de la représentation dépendront de la qualité des syntagmes nominaux extraits. En effet, le repérage et le filtrage de syntagmes nominaux corrects du point de vue linguistique est une phase cruciale du modèle. Nous avons opté pour une stratégie d'implémentation du modèle par étapes où il s'agit d'intégrer les éléments proposés dans le modèle un par un et d'évaluer l'impact effectif que va avoir chacun d'eux sur le SRI. L'évaluation consiste à comparer les performances de notre approche à une indexation classique à base de termes simples. Cette dernière utilise des traitements linguistiques rudimentaires qui ont une faible complexité et nécessitent des ressources linguistiques limitées mais permet pourtant d'avoir des résultats assez satisfaisants. L'objectif de cette comparaison est de s'assurer que l'intégration d'un traitement linguistique plus élaboré ne doit pas détériorer les performances mais plutôt les augmenter.

4. Expression caricaturale souvent utilisée pour la représentation avec des termes simples ; en anglais *bag of words*

7.4 Termes complexes en RI

Dans cette section, nous incluons dans la catégorie des termes complexes les groupes de mots extraits statistiquement même si ces derniers n'ont pas un comportement syntaxique correct mais pour le besoin de notre étude et de l'abondance des travaux qui lui sont dédiés nous les intégrons aux termes complexes définies précédemment. Les termes complexes peuvent être sélectionnés statistiquement, linguistiquement ou en combinant les deux approches[5]. Les techniques statistiques permettent de découvrir des séries de mots ou de combinaisons de mots qui occurrent fréquemment dans un corpus. Les techniques linguistiques permettent de découvrir des groupes de mots qui sont proches de la notion de syntagmes nominaux. L'avantage de l'approche statistique est qu'elle permet de couvrir de manière exhaustive toutes les combinaisons possibles des termes dans une fenêtre textuelle allant d'un bigramme (en terme de mots) au document tout entier. Mais c'est aussi son inconvénient majeur du fait que certaines combinaisons, valables d'un point de vue statistique, peuvent ne pas être justifiées linguistiquement et aussi sémantiquement. Ceci est du principalement à l'ignorance du contexte linguistique de ces termes. Les approches linguistiques permettent d'extraire des combinaisons de termes dont la structure syntagmatique est valable (ordre syntagmatique). Son inconvénient est le nombre important de combinaisons possibles et de candidats comme termes d'indexation obtenus à partir du texte.

7.4.1 Problème d'évaluation

Une étude comparative des résultats des approches d'extraction et d'indexation avec des termes complexes (statistique, linguistique et hybride) n'a pas abouti à des conclusions claires en ce qui concerne leur utilité en RI [Fag87, MBSC97, KP98, CB97]. La plupart des travaux montrent que l'utilisation des syntagmes offre un avantage pour un SRI [WBH+00]. Certains travaux ont trouvé de moins bons résultats et justifient leur échec par le fait que la fonction de correspondance utilisée n'était pas adaptée [SOK94] ou que l'étude des termes complexes n'a été appliquée qu'aux requêtes [CTL91, SR88] alors que les auteurs dans [HGS+97, MBSC97] ont montré que l'indexation avec des syntagmes nominaux extraits linguistiquement affecte plus positivement les résultats d'un SRI que celle avec des groupes de mots extraits statistiquement.

L'évaluation et la comparaison des différents travaux ne sont pas évidentes. Une difficulté majeure pour évaluer l'effet du TALN sur un SRI est le fait que la composante TALN ait été intégrée pratiquement à tous les niveaux du SRI que se soit au niveau de l'extraction terminologique, au niveau de l'indexation, dans le traitement de la requête, etc. Ce qui rend très difficile une évaluation objective et détaillée de l'effet du TALN sur

5. Chapitre 2

un SRI. Une autre difficulté, non moins importante que la première, est la différence entre les SRI utilisés, que ce soit au niveau du modèle utilisé ou bien au niveau des fonctions de pondération utilisées, ou bien encore à quel niveau la composante TALN a été intégrée ce qui rend la comparaison entre les différents résultats des travaux effectués dans ce sens très subjective. Il est difficile alors de tirer des conclusions sur l'efficacité des méthodes d'indexation et d'interrogation qui ont été testées par différentes équipes de recherche, à des époques différentes dans des conditions d'expérimentation complètement différentes [6]. L'indexation, l'interrogation, les procédures d'évaluation et surtout les tailles et la nature des collections et les langues des documents ne sont pas donc les mêmes.

Néanmoins, il reste instructif de comparer des résultats d'expérimentation dans l'objectif d'avoir une vue sur l'efficacité atteinte par les méthodes d'indexation et d'interrogation. En absence d'une approche formelle d'évaluation, la capacité d'un SRI est évaluée en tant qu'un ensemble de composantes diverses permettant une bonne recherche d'information. Le peu qu'on puisse dire sur l'utilisation du TALN dans un SRI est *Est ce que l'ajout du TALN a permis d'avoir de meilleures performances ou non ?*. Les conférences TREC est le meilleur exemple de ce type de comparaison où différents systèmes rentrent en compétition et sont jugés d'après leurs performances respectives.

7.4.2 Indexation avec des termes complexes

Le besoin de représentation des documents avec des termes complexes [7] s'est réveillé dès la prise de conscience des limites de la représentation avec un terme simple mais les tentatives ont été limitées et difficiles à mettre en place du fait de la puissance limitée des ordinateurs. Aujourd'hui, il existe des analyseurs morphologiques et lexicaux ainsi que des analyseurs syntaxiques qui nous permettent d'aborder l'application du TALN au RI avec beaucoup de motivation et d'optimisme.

L'hypothèse sous-jacente à l'utilisation des termes complexes, est que ces derniers sont plus aptes à désigner des entités sémantiques (concepts) que les mots simples et constituent alors une meilleure représentation du contenu sémantique des documents [MBSC97]. Dans la suite, nous détaillons quelques travaux et systèmes qui ont utilisé des termes complexes pour l'indexation des documents.

Le système PRISE représente les documents avec des expressions linguistiques en utilisant l'analyseur TTP (Tagged text Parser) [SLPC97]. Pour extraire ces expressions, le système utilise des patrons syntaxiques tel que *Adjectif+Substantif, Substantif+Substantif, Substantif+X* avec X un argument du Substantif, etc. A noter ici que la relation *tête expansion* est exprimée avec un + entre les paires de termes. Par exemple, les expressions : *information retrieval, retrieval of information, retrieve more information* et *information*

6. la puissance des ordinateurs des années 80 n'est pas la même que celle des années 2000
7. syntagmes, mots-composés, noms propres, etc.

that is retrieved, sont normalisées avec la paire *retrieve+information* où *retrieve* est la tête et *information* est l'expansion.

Cette représentation n'est qu'une première étape, certes importante et significative, vers une représentation du sens du texte mais elle souffre à notre avis de deux inconvénients majeurs. Le premier inconvénient est lié au fait que cette approche ne tient pas compte de l'ordre des mots dans le texte. En effet, si on considère l'exemple précédent où toutes les paires de termes sont représentées par *retrieve+information*, dans certains cas cette représentation est très ambiguë et peut introduire du bruit comme par exemple dans *system of investigation* et *investigation of system* s'ils sont normalisés en *system+ investigation*. Le deuxième inconvénient concerne certains patrons syntaxiques utilisés tel que *verbe+objet* ou *verbe+sujet*, etc. Nous ne voyons pas l'intérêt d'indexer des documents avec ce type d'information vu qu'elle est ambiguë. En effet, si on considère l'exemple donné dans [SLPC97]:

> *While serving in South Vietnam, a number of U.S. Soldiers were reported as having been exposed to defoliant Agent Orange. The issue is veterans entitlement, or the awarding of monetary compensation and/or medical assistance for physical damages caused by Agent Orange.*

Les paires de termes *tête+expansions* extraites de ce paragraphe sont les suivantes : *damage+physical, cause+damage, award+assist, award+compensate, compensate+monetary, assist+medical, entitle+veterans*.

Si nous considérons la phrase suivante *Fire damages cause major replanting*, cette phrase sera aussi indexée par *cause+damage*. Si un besoin d'information est les documents qui parlent des causes des dommages subis par les soldats US, le document qui traite les *major replanting* causé par le *Fire damages* sera aussi retrouvé comme réponse alors qu'il n'est pas pertinent. Ces simples exemples montrent l'intérêt d'aller plus en profondeur dans la structuration des expressions et de ne pas négliger les rôles des mots ainsi que les relations qu'ils entretiennent entre eux dans le texte.

Nous retiendrons dans notre étude le formalisme de représentation *tête expansion*, utilisé aussi dans le système Zellig [HBGN+97], mais nous tiendrons compte de l'ordre des mots donc de la structure linguistique des syntagmes.

Les résultats donnés dans [SLPC97] est un exemple pertinent de la difficulté à évaluer l'impact du TALN sur un SRI. En effet, les auteurs ont utilisé la technique de *tête expansion* ainsi que les noms propres (nom de personne, nom d'organisation, nom de localisation, etc.) avec des méthodes d'indexation parallèles. Pour cela, ils ont conçu un métaSRI qui exploite les résultats de plusieurs SRIs à la fois (SMART, INQUERY, NIST's Prise, etc.). Les performances obtenues n'ont pas été significatives mais un jugement objectif de l'impact du TALN ne peut pas être tiré vu que les différents systèmes utilisés ont des caractéristiques propres particulières (modèle de recherche d'information, procédure

de lemmatisation ou de troncature, etc.) qui peuvent influencer positivement ou négativement sur les résultats obtenus.

Le système CLARIT, un des systèmes les plus cité dans ce domaine, se base sur un filtrage syntaxique suivi d'un filtrage statistique. C'est le système qui, à notre avis, a tiré le plus d'avantages du TALN comme le montre le net gain de performances d'une nouvelle version du système par rapport à la version qui la précède durant les différentes sessions de TREC. Ces améliorations sont expliquées par les concepteurs du système par l'enrichissement de l'extracteur de syntagmes nominaux de CLARIT d'un module de génération d'abord de noms propres et ensuite de syntagmes syntaxiques employés comme index complexes et qui a permis d'obtenir des meilleurs résultats.

Durant TREC-5, les auteurs dans [ZTME96] se sont basés sur les deux hypothèses suivantes :

– L'utilisation d'expressions figées tel que *hot dog* pour remplacer les mots simples dans l'index peut augmenter la précision et le rappel.

– L'utilisation d'expression syntaxique tel que *junior college* pour remplacer les mots simples peut augmenter la précision sans pour autant affecter le rappel.

La première hypothèse n'a pas montré des améliorations nettes des performances en terme de rappel et précision alors que la deuxième a été largement vérifiée dans certains thèmes de la piste TREC-5 (entre 9% et 139% d'augmentation de la précision moyenne).

Les auteurs dans [MBSC97] ont appliqué le TALN aux réponses des requêtes. En utilisant le système SMART, les documents sont indexés avec des termes simples suivant une pondération *tf.idf*. Pour une requête donnée lors de l'interrogation, les 100 premiers documents obtenus sont analysés pour extraire les expressions syntagmatiques. Etant donné que les expressions syntagmatiques ne sont extraites qu'à partir de 100 documents, il est alors impossible de calculer l'*idf* de ces expressions. Cette valeur est alors remplacée par une estimation qui prend en compte la fréquence d'apparition des constituants des expressions dans le corpus. Ces expressions syntagmatiques sont utilisées pour indexer les documents réponse de la requête. Pour un document donné, son index est alors constitué de deux champs : un champ pour l'indexation avec des termes simples et un champ pour l'indexation avec des expressions syntagmatiques. La fonction de correspondance est une nouvelle fois appliquée en tenant compte cette fois du score des documents lors de la première interrogation, donc avec des index constitués seulement de termes simples, et d'un paramètre qui reflète l'importance donnée au champ index syntagmatique. Il s'agit alors d'une reclassification des 100 documents afin d'avoir plus de documents pertinents en tête. Cette approche n'a pas montré des améliorations nettes des résultats même en examinant la précision à 20 documents.

L'équipe XEROX durant TREC-5 [HGS$^+$97] a testé l'impact de la reconnaissance de la dépendance syntaxique des mots pour éliminer le bruit dans les couples de mots

extraits statistiquement et réduire le silence par la reconnaissance de paires de termes reliés syntaxiquement. Pour cela, un ensemble d'expérimentations a été réalisé avec des comparaisons entre les différents résultats:

- Une expérimentation purement statistique qui consiste à extraire statistiquement des couples de termes et à les lemmatiser. Les mots vides ont été éliminés. En utilisant le système Smart, les auteurs ont ajouté un deuxième champ dans l'index des documents qui représente les couples de termes extraits.

- Une expérimentation avec le TALN qui consiste à utiliser des patrons syntaxiques : *sujet-verbe, substantif substantif, adjectif substantif,* etc. Les SNs extraits du texte sont alors utilisés à l'indexation.

- Une expérimentation avec le TALN et l'approche statistique; dans ce cas l'index est composé de trois champs différents : un champ pour les termes simples, un champ pour les couples de termes extraits statistiquement et un champ pour les syntagmes nominaux extraits avec le TALN.

- Une expérimentation avec le TALN, l'approche statistique et un filtrage manuel des couples qui sont jugés non corrects.

Les résultats de ces expérimentations montrent clairement que l'indexation avec des termes complexes extraits syntaxiquement affecte plus positivement les résultats d'un SRI (augmentation de 15% de la précision moyenne en 11 points de rappel) que les groupes de mots extraits statistiquement dans les cas où les requêtes sont longues. Cette constatation est très pertinente à notre avis dans le sens où une indexation des documents avec des syntagmes doit être accompagnée aussi par une indexation des requêtes avec des syntagmes.

Arampatzis et al [ATKW98] ont proposé un modèle de RI basé sur l'exploitation des propriétés linguistiques des termes extraits des documents. Les auteurs mettent l'accent sur le besoin d'étudier et d'exploiter les variations morphologiques, lexicales, sémantiques et syntaxiques des termes. Ils proposent de compléter les termes simples par des expressions nominales[8]. Les expressions verbales sont transformées en expressions nominales et des normalisations morphologiques, lexicales, sémantiques et syntaxiques sont appliquées. Ces expressions sont représentées sous la forme de têtes et expansions.

Cette approche souffre à notre avis de deux lacunes:

- Les expressions sont pondérées selon une fonction de pondération qui tient compte d'une part du rôle joué par une expression (tête ou expansion) et d'autre part, dans le cas où une expression est une expansion, de la distance qui la sépare du concept central c'est-à-dire la tête. Cette fonction de pondération ne tient pas compte des catégories grammaticales des composantes d'une expression ni du nombre de constituants

8. en anglais *noun phrases*

de cette dernière. La tête d'une expression est considérée par les auteurs comme étant l'élément central d'une expression mais cette caractéristique n'a pas été prise en compte dans la fonction de pondération ni celle de la correspondance. En effet, les propriétés linguistiques sur lesquelles les auteurs ont insisté et basé leur modèle ne sont pas utilisées pour valuer l'importance des expressions dans les documents ni dans l'évaluation de la correspondance entre un document et une requête.

– Dans la mise en oeuvre du modèle, les auteurs ont utilisé le modèle vectoriel pour évaluer leur proposition. Ils avouent que ceci va à l'encontre de l'hypothèse de dépendance entre les termes sur laquelle se base leur proposition d'autant plus qu'ils ont utilisé le même schéma de pondération pour les expressions que pour les termes simples. Dans la mise en oeuvre du modèle, les auteurs se sont donc éloignés de leur proposition théorique.

Les évaluations du modèle proposé ont été conduites dans un environnement de filtrage d'information et de bouclage de pertinence avec le système IRENA [9] [ATK97]. Elles ont montré que l'ajout d'expressions nominales dans l'index permet d'augmenter les performances du système IRENA (+5% de la précision moyenne en 11 points de rappel) [AvdWK+00].

7.4.3 Conclusion

L'analyse des travaux précédents nous a conduits à un ensemble de constats. Tout d'abord, dans le cas des SRIs qui utilisent la statistique, la notion de terme complexe désigne dans ce contexte deux mots qui cooccurrent fréquemment dans une fenêtre textuelle donnée. Les termes complexes sont clairement une approximation des termes complexes dans le sens TALN même s'ils n'ont pas un comportement syntaxique défini mais qui sont souvent ambiguës. En effet, certains de ces termes complexes s'avèrent être un groupement au hasard ou accidentel d'un ensemble de mots. Dans la plupart de ces approches, les termes complexes de plus de 2 mots sont généralement ignorés et la structure interne n'est pas considérée. C'est pourquoi nous nous sommes orientés vers une approche linguistique. Adopter une approche linguistique, c'est s'autoriser a priori une extension vers la prise en compte de phénomènes linguistiques. Il est cependant clair que l'exploitation du potentiel de ces traitements linguistiques est très délicate à mettre en oeuvre, car les phénomènes linguistiques sont variés et complexes. Par ailleurs, ces traitements doivent s'appliquer à des corpus en croissance permanente [10]. Il est alors difficile de choisir a

9. Information Retrieval Engine based on Naturel language Analysis

10. pour illustration, les corpus de test sont passés de 2 mégabits à plusieurs Gigaoctet pour la conférence TREC, le corpus du Web est lui estimé à plus de 2 milliards de pages en 2000 [MM00]

priori un ensemble de traitements linguistiques applicable aux corpus actuels et capables d'assurer une augmentation significative de la qualité des réponses du système.

Amar fait remarquer que la linguistique est pertinente à l'indexation à deux niveaux [Ama00]:

– au niveau des descriptions des faits de la langue:

  les propriétés que la linguistique dégage à partir de l'étude de ces propres objets peuvent, moyennant une décontextualisation et une généralisation, expliquer le fonctionnement d'un certain nombre de mécanismes à l'oeuvre dans l'indexation, mécanismes peu visibles hors du prisme de l'analyse linguistique;

– au niveau des postulats sur la langue et le langage :

  la théorie linguistique, si elle n'est pas la seule à ternir des propositions générales sur la langue et le langage, fournit l'un des cadres possibles pour discuter la conception du langage sous-tendue par la pratique d'indexation ; à ce titre, on pourra confronter la conception du langage telle que la véhicule l'indexation à la conception du langage postulée par les modèles linguistiques.

Amar donne des exemples de caractéristiques qui sont issues de la pratique d'indexation et qui peuvent être appréhender, au premier ou au deuxième point présentés au-dessus, dans le cadre d'une approche linguistique telles que les formes nouvelles de composition nominale, la représentation de la relation entre les mots et les choses, la question du sens, etc.

La plupart des travaux attribuent les faibles résultats obtenus au manque de robustesse et aux faibles performances des analyseurs utilisés. Nous pensons qu'il est important de bien mesurer ce qu'apporte l'usage d'un traitement linguistique plus précis de la langue par rapport à un traitement plus frustre. Il s'agit en fait, de déterminer la couverture des phénomènes linguistiques (et même sémantiques sous jacents) de ce nouveau traitement, par rapport à l'ancien, puis d'évaluer l'impact effectif qu'il va avoir sur le SRI dans sa globalité.

Au contraire de la plupart des travaux cités, notre approche se veut la plus généraliste possible afin de traiter n'importe quel domaine d'application, et plus particulièrement le Web. Pour cela, elle se base sur les patrons syntaxiques les plus utilisés dans une langue (la langue française dans le contexte de notre étude). Etant donnée la masse d'information, un traitement linguistique approprié s'impose. La mise en oeuvre d'un traitement linguistique en profondeur repose sur des analyseurs robustes et exhaustifs de la langue, trop complexes pour l'objectif visé et contraignants dans le cadre d'un SRI. C'est pourquoi nous adoptons une analyse de surface qui élimine la détermination de la structure linguistique profonde et ne tient compte que de l'extraction des syntagmes nominaux (SNs). Nous accordons aussi beaucoup d'attention à la phase d'extraction car nous sommes conscients

de son importance et l'effet qu'elle peut avoir sur les résultats d'un SRI. Nous avons donc consacré une grande partie de nos travaux à l'extraction de syntagmes nominaux de bonne qualité. Dans notre approche, les requêtes sont analysées linguistiquement de la même façon que les documents. Les requêtes utilisées pour nos expérimentations dans le cadre du projet Amaryllis sont des requêtes longues où des syntagmes nominaux peuvent être extraits.

7.5 Les syntagmes nominaux

7.5.1 Les Syntagmes pour représenter le thème

L'une des hypothèses forte dans notre approche est l'importance des termes complexes ayant une fonction référentielle forte à un concept, c'est-à-dire faisant référence à un objet de l'univers de discours. D'après la littérature dans ce domaine, cette fonction référentielle est majoritairement assumée par les groupes nominaux et plus précisément, par ce que les linguistes appellent les *syntagmes nominaux*. En effet, plusieurs travaux menées par des linguistiques ont montré le lien entre syntagmes nominaux et thèmes (*ce dont on parle* ou *ce dont il est question*) d'une part, et d'autre part entre syntagmes verbaux et rhèmes (*ce qu'on en dit* ou *le propos*) [Ama00, Feu88, Mar88, Gue84]. Plus précisément, ils s'accordent sur le fait que seuls les groupes nominaux peuvent être des référents [Ama00]. Dans un processus de communication, le thème est considéré comme étant le point de départ de la communication et son support alors que le rhème constitue le but de la communication [Feu88]. C'est pourquoi dans le domaine de la RI les syntagmes nominaux ont eu plus d'attention puisque c'est le thème qui est intéressant plus que le rhème. Sur ce point [Ama00] présente le point de vue de Marandin dans son hypothèse du descripteur comme thème du discours [Mar88] et celui de Le Guern dans son modèle logico-sémantique [Gue84] qui précisent qu'un thème est toujours un nom, et plus précisément, un individu linguistique de type syntagme nominal.

Nous avons choisi de nous intéresser aux informations exprimées dans un syntagme défini sommairement comme un ensemble de termes respectant des lois de la morphologie et de la syntaxe, et possédant une signification propre. Plus précisément, nous nous intéressons aux syntagmes nominaux en tant que thèmes. Le fait que le mot, pris isolément, soit un signe sans référence de même que l'idée que les descripteurs devraient être un signe avec références renforce la validité de notre approche, à savoir l'utilisation des syntagmes nominaux comme descripteur au lieu d'utiliser les mots isolés [MBSC97].

Les traitements linguistiques nécessaires à la mise en oeuvre du formalisme d'extraction et de représentation par des SNs doivent permettre non seulement de focaliser sur l'analyse sur des SNs mais surtout ces traitements doivent en proposer une structure. C'est justement cette structure qui doit supporter une partie de la signification des SNs.

Il reste néanmoins très difficile de placer les SNs réellement à un niveau sémantique. De manière pratique, c'est la structure syntaxique qui sert de passerelle vers le niveau sémantique [Par96, Kho97]. En effet, Strzalkowski et al dans [SSWB00] indiquent qu'un traitement linguistique, pour une représentation des documents avec des termes complexes, peut couvrir, contrairement à une représentation avec des mots simples, certains aspects sémantiques du contenu des documents. Nous nous intéressons alors aux SNs au niveau syntagmatique de l'analyse linguistique sans prendre en considération les niveaux sémantique et paradigmatique. La question pratique que l'on doit finalement se poser sur l'extraction des SNs concerne la profondeur de l'analyse à mettre en oeuvre. Une analyse de surface avec des patrons syntaxiques semble suffisante comme l'atteste les travaux de Daille [Dai94] et de Debili [Deb82]. Ces patrons ne traitent que les relations homosyntagmatiques c'est-à-dire les relations s'établissant entre éléments appartenant à une même phrase.

7.5.2 Description linguistique générale des syntagmes nominaux

Dans le cadre de l'analyse syntaxique d'une phrase, on parle de segmentation en unités fonctionnelles appelées syntagmes. Les syntagmes peuvent avoir la même fonction qu'un mot seul et ils peuvent également inclure un ou plusieurs autres syntagmes. Il existe en français plusieurs types de syntagmes, y compris le syntagme nominal, le syntagme verbal, le syntagme prépositionnel, le syntagme adjectival et le syntagme adverbial.

Tous les syntagmes partagent un certain nombre de caractéristiques, mais l'essentiel est sans doute le fait que tous ont une tête, c'est-à-dire, un élément central qui contrôle les autres et qui donne son nom au syntagme. Le contrôle exercé par la tête peut se manifester par exemple par l'accord en nombre ou en genre. La tête est donc considérée comme le noyau du syntagme dont dépendent éventuellement d'autres éléments [Gen94].

Linguistiquement, un syntagme nominal peut être caractérisé d'une part par les catégories grammaticales de ces composantes et d'autre part par les règles syntaxiques de l'agencement de ces composantes. Les catégories grammaticales des éléments d'un syntagme nominal sont : substantif, préposition, conjonction, article, adjectif, verbe à l'infinitif, participe passé et adverbe. L'ordre d'enchaînement de ces catégories dans un syntagme nominal respecte des règles linguistiques qui permettent d'avoir des syntagmes nominaux corrects.

A partir de ces deux caractéristiques des syntagmes nominaux, des patrons syntaxiques peuvent être construit comme en témoigne les travaux de Debili [Deb82]. Ces patrons décrivent les catégories grammaticales et l'ordre dans le quel les éléments d'un syntagme nominal doivent apparaître. Nous utilisons donc les syntagmes nominaux pour représenter le contenu des documents ainsi que le contenu des requêtes. Pour cela, nous exploitons les caractéristiques grammaticales des syntagmes nominaux ainsi que l'ordre syntagmatique

de l'enchâssement de leurs composantes.

7.5.3 Patrons syntaxiques

Nous définissions un patron syntaxique comme étant une règle qui, à partir d'un vocabulaire extrait du corpus, définit les catégories lexicales qui peuvent être utilisées dans un syntagme nominal ainsi que le lexique et l'ordre d'enchaînement des catégories. Nous désignons par :

- V : le vocabulaire extrait du corpus
- C : un ensemble de catégories lexicales
- L : le lexique $\subset$ V x C

Un patron syntaxique est une règle de la forme:

$$\text{X} := Y_1\ Y_2\ ...\ Y_k\ Y_{k+1}\ Y_n$$

avec $Yi \in C$ et X un syntagme nominal.

Un ensemble des syntagmes nominaux les plus couramment utilisés dans la langue française et les plus susceptibles de contenir le maximum d'information du texte a été décrit dans la littérature [Deb82, Jac97, Bou92]. Dans la section 8.2, nous établissons d'après une expérimentation, les catégories grammaticales qui sont les plus utilisées dans la langue française et qui nous guident dans nos choix de catégories grammaticales. Dans l'annexe E, nous présentons des exemples de patrons syntaxiques.

7.6 Conclusion

Dans ce chapitre, nous avons montré que l'utilisation des termes simples n'est pas suffisante pour représenter le contenu textuel des documents. Les unités textuelles les plus susceptibles de représenter les thèmes des documents sont les syntagmes nominaux. Nous avons présenté les caractéristiques linguistiques des SNs. Nous nous sommes basés sur ces caractéristiques afin de définir des patrons syntaxiques. Nous utilisons ces patrons pour extraire des syntagmes nominaux dont la structure syntagmatique est correcte. Notre approche d'extraction de SNs se base sur une analyse de surface qui doit être efficace sur des grandes collections que se soit au niveau du temps de traitement qu'au niveau de la qualité d'extraction, objectif que nous nous sommes fixé dès le départ. Dans le chapitre suivant, nous présentons notre méthodologie d'extraction ainsi que des évaluations quantitatives et qualitatives de la qualité des syntagmes extraits et l'impact de leur utilisation dans la représentation des documents dans le cadre d'un SRI.

Chapitre 8

Méthodologie d'extraction des syntagmes nominaux

C'est par l'expérience que la science et l'art font leur progrès chez les hommes.
ARISTOTE

Comme nous l'avons mentionné dans le chapitre précédent, la nature du traitement linguistique et ses performances pour extraire des SNs (qualité et temps d'exécution) sont des contraintes fortes. Nous adoptons une méthodologie qui se base sur les patrons syntaxiques car elle présente un bon compromis traitement/résultat, comme il a été montré dans la littérature et que nous démontrerons dans nos expérimentations.

8.1 Extraction des syntagmes nominaux

L'extracteur est composé d'un vocabulaire et d'une liste de patrons. Le traitement est réalisé par un algorithme à pile qui permet de traiter le texte en un seul passage. Il permet d'extraire les SNs, de calculer leurs fréquences et de filtrer la collection pour générer un nouveau vocabulaire contenant les unitermes et les SNs.

Après l'analyse linguistique effectuée par le système IOTA, présentée dans le chapitre 3, qui génère une collection étiquetée, la deuxième étape utilise cette collection étiquetée et en extrait un ensemble de syntagmes nominaux. Les syntagmes nominaux candidats sont extraits par repérage de patrons syntaxiques. Un filtre syntaxique permet de ne garder que les syntagmes nominaux valides par rapport à l'ensemble de patrons syntaxiques préétablis. Le filtrage favorise les syntagmes nominaux longs afin d'avoir des syntagmes nominaux les plus riches possibles d'un point de vue informationnel et de ne pas avoir un grand nombre de syntagmes nominaux extraits. Dans le processus de traitement, si plusieurs patrons peuvent être appliqués sur un segment de texte, c'est le patron le plus long

qui est utilisé pour extraire un syntagme nominal et les autres patrons sont ignorés. Par exemple, si dans le texte on rencontre la séquence suivante *aide humanitaire internationale* alors le syntagme nominal extrait est celui composé par ces trois mots même si *aide humanitaire* correspond à un patron syntaxique (*Substantif Substantif*) mais il ne sera pas extrait.

8.2 Application et Évaluation de l'extraction

Par cette expérimentation, notre objectif est d'étudier la faisabilité et l'efficacité de notre méthodologie d'extraction et ceci sur plusieurs points. Tout d'abord, notre approche d'extraction de SNs doit être efficace sur des grandes collections que se soit au niveau du temps de traitement qu'au niveau de la qualité d'extraction. Pour cela, outre les collections-tests d'Amaryllis, nous avons besoin d'autres collections plus volumineuses. Le Web représente un cadre idéal pour répondre à notre besoin puisqu'il représente une grande masse de données [VG01]. Cette expérimentation nous permet aussi d'observer le comportement de nos paramètres linguistiques dans un contexte de traitement de grandes collections.

Nous commençons par présenter la construction de corpus textuels à partir du Web. Ensuite, nous détaillerons la chaîne de traitement complète qui permet, à partir du Web, d'extraire des corpus textuels, de les traiter et d'en extraire des SNs. Enfin, nous présentons les caractéristiques générales des corpus collectés et nous analyserons les résultats obtenus par nos expérimentations d'extraction de SNs.

8.2.1 Chaîne de traitement

Nous présentons dans cette section la chaîne de traitement qui permet de collecter des données brutes sur le Web, de les analyser pour en extraire des corpus textuels normalisés, et d'utiliser ensuite ces corpus pour l'extraction de SNs [GV02]. Le schéma général simplifié de la chaîne de traitement est représentée dans la Figure 8.1.

Nous avons utilisé le robot CLIPS-Index [1] pour parcourir le Web, collecter et stocker des pages Web [VG01, GC01]. La phase suivante consiste à normaliser les données brutes collectées sur le Web. Les étapes les plus importantes étant les suivantes :

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

1. http://brel.imag.fr:8000/CLIPS-Index/

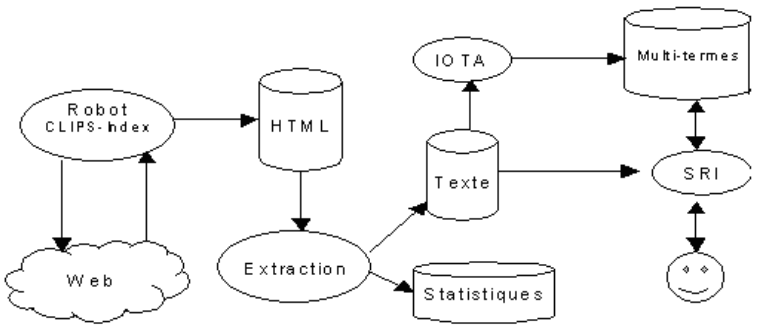

FIG. 8.1 – *Chaîne de traitement*

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

– L'extraction du lexique et le calcul de la couverture lexicale du corpus en calculant le pourcentage de formes lexicales d'un lexique témoin qui apparaissent dans chaque corpus.

– L'extraction de statistiques diverses, par exemple la langue des documents ou des informations ayant trait à la structure des pages Web [GC01].

Nous avons utilisé le système IOTA pour l'analyse morpho-syntaxique et l'extraction des SNs. La première étape de l'analyse, présentée dans le chapitre 3 permet d'avoir une collection de textes étiquetés où tous les mots ont été catégorisés. La fréquence globale d'un terme dans une collection ainsi que sa fréquence selon une fenêtre sont calculées. La deuxième étape utilise cette collection étiquetée et en extrait un ensemble de SNs comme présenté dans la section 8.1.

8.2.2 Constitution des corpus extraits du Web

Nous avons choisi d'étudier deux corpus extraits du Web:

– *Tunisie* : un corpus de pages collectées sur le domaine ".tn", pour obtenir un corpus qui contienne une majorité de documents francophones, et qui soit représentatif d'un domaine.

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

Un filtre, paramètre de CLIPS-Index, permet de sélectionner les sites Web à collecter. Il est exprimé à l'aide d'une expression régulière sur le nom du site. Celui utilisé pour "Tunisie" permet de se restreindre aux noms de sites se terminant par ".tn". Dans le cas des collections "Journaux", le filtre parcoure des parties de la hiérarchie de l'annuaire Yahoo! [2] et extrait les sites existants dans ces parties.

8.2.2.1 Analyse des corpus extraits du Web:

Nous avons collecté les corpus "Tunisie" et "Journaux" à différentes dates, pour analyser leur évolution dans le temps. Le Tableau 8.1 montre les statistiques générales de ces collectes, qui se sont déroulées jusqu'à épuisement des URLs sur les domaines recherchés.

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

TAB. 8.1 – *Caractéristiques des collectes.*

8.2.2.2 Caractéristiques des corpus utilisés:

Chaque corpus HTML est analysé pour en extraire un corpus de textes distincts, les documents doublons étant détectés et éliminés. Le Tableau 8.2 montre les caractéristiques générales des corpus textuels.

Le rapport entre la taille du corpus HTML et celle du corpus textuel va de 4,9 à 7,17. Ce phénomène vient principalement de l'utilisation de plus en plus importante de balises HTML et autres, pour la présentation des pages Web, mais est aussi lié aux outils de production de pages HTML qui insèrent de plus en plus de données dans une page.

2. par exemple $/Actualites-et-medias/Journaux-et-magazines/$

Corpus	Nombre de documents	Taille HTML/TEI Collection	Taille HTML/TEI en Ko/page	Taille Texte Collection	Taille Texte en Ko/page
INIST	163 308	100 Mo	0,63	79 Mo	0,50
OFIL	11 016	33 Mo	3,06	32 Mo	2,93
Tunisie	27 959	161 Mo	5,90	27 Mo	1,00
Tunisie2	43 651	397 Mo	9,31	55 Mo	1,30
Tunisie3	79 361	863 Mo	11,13	165 Mo	2,13
Journaux	198 158	4’391 Mo	22,69	896 Mo	4,63
Journaux2	345 860	7’728 Mo	22,88	1’491 Mo	4,41

TAB. 8.2 – *Caractéristiques générales des corpus textuels*

8.2.2.3 Répartition des langues:

L’extraction de la langue d’un document, est basée sur les fréquences des mots les plus courants de chaque langue (anglais, français, italien, allemand, espagnol, hollandais, danois, etc.). Par exemple, si le document comporte une plus grande proportion de *le, la, les, un, une, des, dans*, etc. alors il est de langue française par contre s’il comporte une plus grande proportion de *and, any, but, by, for, not, of, the, this, to*, etc. alors il est de langue anglaise. Pour chaque langue, ces mots fréquents sont extraits d’un corpus de textes de référence. Pour la langue française, par exemple, nous avons utilisé un lexique extrait du CD-ROM du Monde Diplomatique (1987-1997) [3]. Pour d’autres langues tels que le danois, le hongrois et le hollandais, nous avons utilisé la bible. On observe une très grande majorité de pages francophones, en particulier dans les corpus *Journaux* et *Journaux 2* (Figure 8.2). La grande proportion de pages dont la langue n’a pas été extraite (inconnue) s’explique, pour les corpus tunisiens, par un nombre important de pages vides d’éléments textuels, souvent remplacés par des images, alors que les pages de *Journaux* et *Journaux 2* contiennent presque toujours du texte.

3. par Jean Véronis (Université de Provence) http://www.up.univ-mrs.fr/ veronis

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

TAB. 8.3 – *Caractéristiques du contenu des corpus textuels*

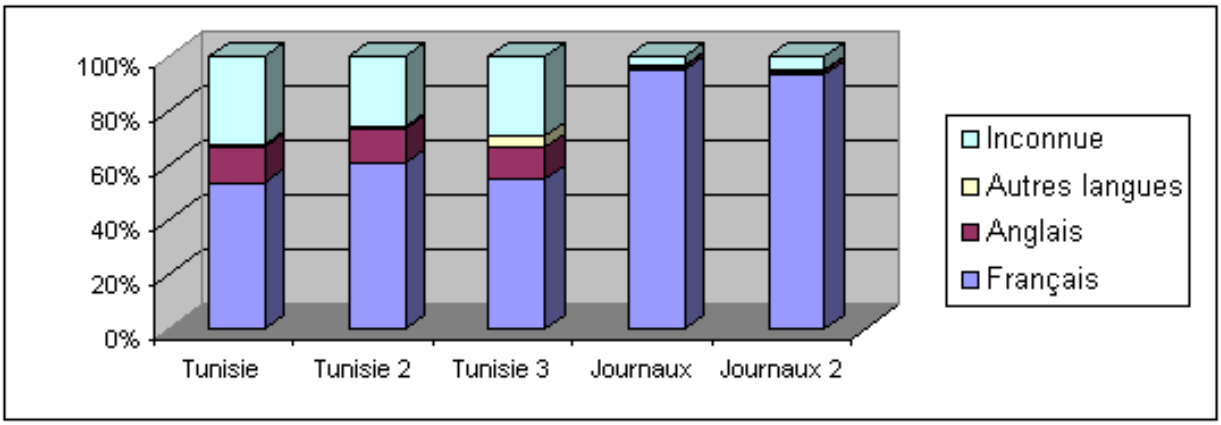

FIG. 8.2 – *Répartition des langues*

8.2.2.4 Lexique et couverture du français:

Le Tableau 8.3 récapitule le nombre de termes distincts apparaissant dans chaque collection, ainsi que le nombre total de termes par collection et par document.

Nous obtenons des corpus volumineux, jusqu'à 30 fois plus grand que ceux d'Amaryllis pour "Journaux 2". Les documents des collections de journaux ("OFIL", "Journaux", "Journaux 2") sont en moyenne de plus grande taille.

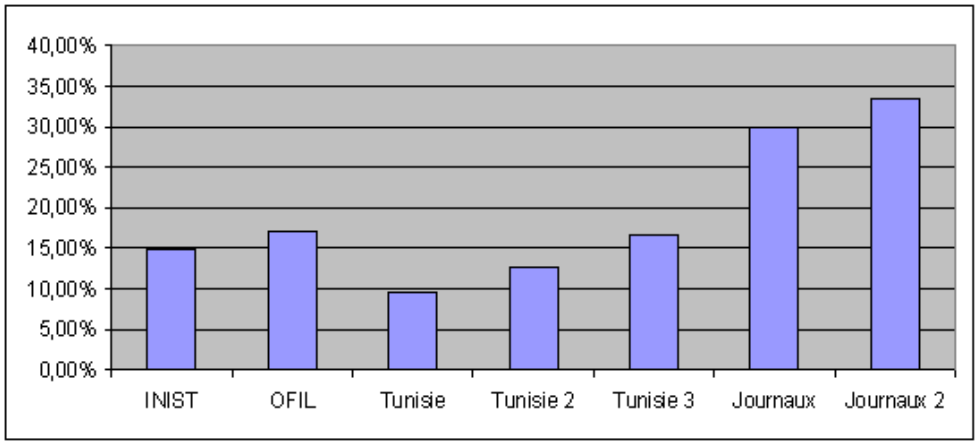

FIG. 8.3 – *Couverture du français*

Afin d'évaluer la diversité des corpus, nous avons estimé leur couverture lexicale de la langue française, en calculant le pourcentage de formes lexicales d'un lexique témoin qui apparaissent dans chaque corpus. Ce lexique de 400 000 formes lexicales a été construit à partir du lexique de l'Association des Bibliophiles Universels [Ass] (comportant plus de 270 000 formes lexicales), et du lexique BDLex [CP98] duquel nous avons dérivé plus de 300 000 formes lexicales. La Figure 8.3 montre la couverture lexicale de chacun des corpus étudiés, qui est beaucoup plus importante pour les corpus "Journaux" et "Journaux 2" que pour les corpus classiques.

8.2.3 Extraction des SNs

Distribution des catégories grammaticales :

La distribution des catégories grammaticales des termes est presque identique pour les différents corpus. Comme le montre le Tableau 8.4, les catégories dominantes sont les substantifs, les adjectifs et les noms propres.

Proportion des SNs extraits :

L'extraction des SNs a donné un nombre plus important de SNs dans les collections Web, par rapport aux collections Amaryllis, comme illustré dans la Figure 8.4. Par

Catégories	OFIL	INIST	Tunisie	Tunisie2	Tunisie3	Journaux	Journaux2
Substantif	30,60%	33,62%	29,21%	29,30%	28,14%	28,88%	28,52%
Adjectif	27,25%	31,55%	26,48%	27,00%	26,15%	27,84%	27,71%
Nom propre	18,38%	11,75%	12,96%	12,89%	13,79%	16,37%	16,42%
Reste	23,76%	23,09%	31,35%	30,80%	31,91%	26,91%	27,35%

TAB. 8.4 – *Distribution des catégories grammaticales des termes*

contre, le nombre moyen de SNs dans un document est plus important dans le cas de la collection OFIL, et très faible dans le cas de la collection INIST comme illustré dans la Figure 8.5.

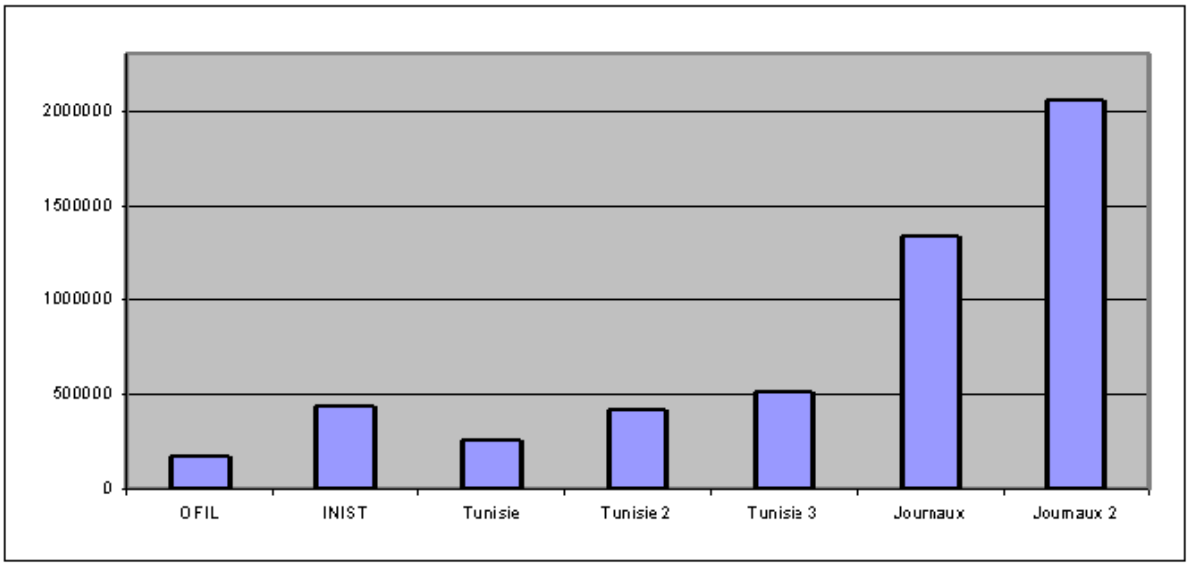

FIG. 8.4 – *Nombre de SNs extraits*

La première constatation s'explique par le fait qu'il y a un certain nombre de documents dans les collections Web, principalement les collections tunisiennes, qui ne contiennent pas ou très peu de texte. La deuxième constatation s'explique par le

fait que la collection INIST est une collection scientifique avec des documents de petite taille où les phrases sont très courtes avec un style descriptif simple. Il est intéressant de constater que pour une même collection, les fréquences de certains SNs ont largement augmentées. Par exemple, le SN ”enseignement supérieur” était pratiquement inexistant dans la collection ”Tunisie” et passe à une fréquence de 1773 dans la collection *Tunisie 3* contre seulement 992 dans ”Tunisie 2”. Des nouveaux SNs apparaissent d'une collecte à une autre, et reflètent par exemple un événement médiatique, tel que le SN *jeux méditerranéens* dont la fréquence est de 972 dans *Tunisie 2* et 178 dans *Tunisie*.

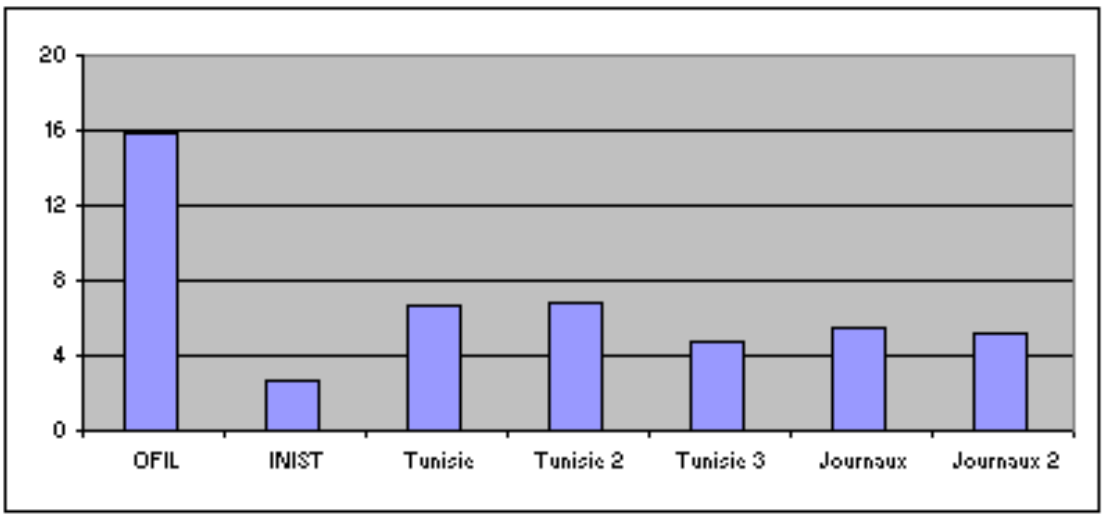

FIG. 8.5 – *Nombre de SNs extraits par document*

Répartition des patrons :

La répartition des patrons syntaxiques varie d'une collection à une autre. Par contre, nous avons constaté la dominance du patron syntaxique ” substantif adjectif ” pour toutes les collections étudiées. Dans la Figure 8.6, nous avons illustré la répartition des patrons syntaxiques dans les collections OFIL et INIST. Nous remarquons que les patrons syntaxiques de taille 3 sont plus rares que ceux de la taille 2. En effet, les SNs de taille 3 représente que 16% des SNs extraits de la collections OFIL et que 22 % dans le cas de la collection INIST.

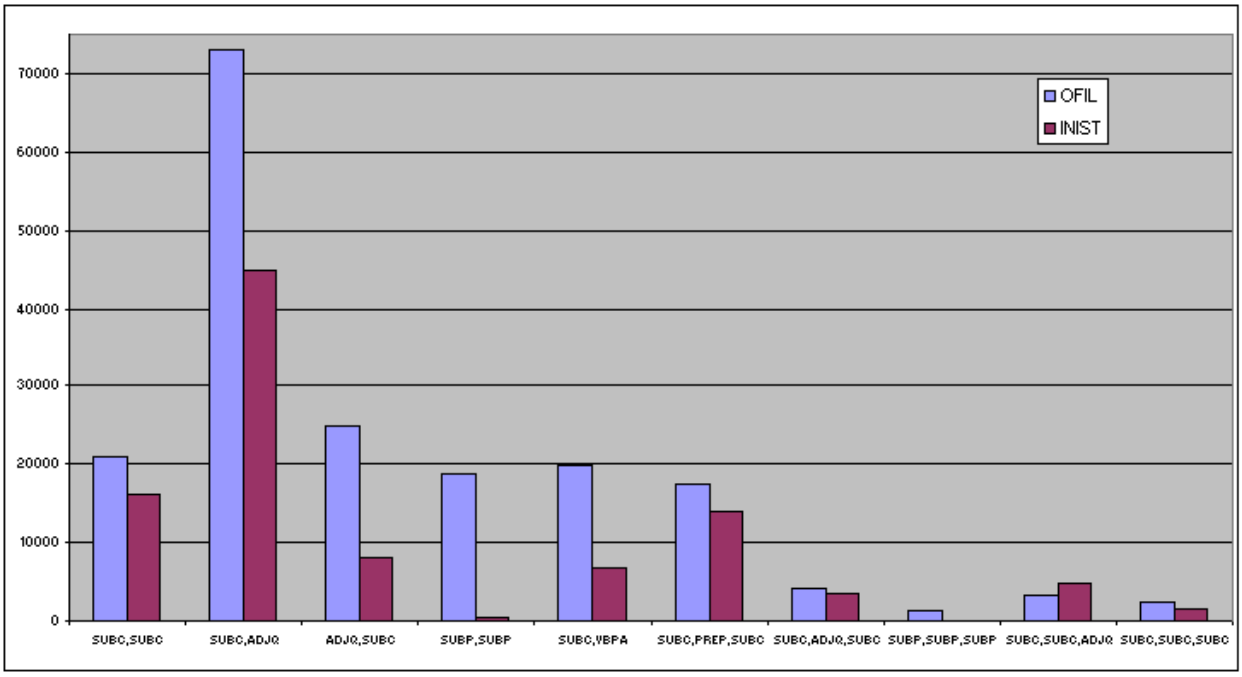

FIG. 8.6 – *Nombre de patrons syntaxiques extraits des collections OFIL et INIST*

8.3 Indexation avec des SNs

Dans le cadre du projet Amaryllis, nous avons expérimenté une indexation par des cliques définis comme étant des contextes forts dans les phrases [Bru90]. Les résultats obtenus n'ont pas été satisfaisants. En effet, une indexation avec des cliques n'a pas permis d'améliorer les résultats d'une indexation avec des unitermes, et pour certaines collections les résultats ont été détériorés. Ceci s'explique par le fait que l'analyse linguistique utilisée dans cette expérimentation et dans le cadre d'extraction des cliques s'est révélée peu performante. L'analyse linguistique a été incapable d'extraire des cliques des requêtes. Dans cette partie, nous reproduisons la même expérimentation effectuée dans [CGH00] mais avec notre approche d'extraction des SNs.

Après l'analyse linguistique et l'extraction des SNs, les documents et les requêtes sont indexés avec le système SMART selon la même approche que dans la section 6.5. Outre l'indexation classique avec des unitermes, nous avons testé les stratégies d'indexation suivantes [Had02]:

– Stratégie 1 : indexer les unitermes et les syntagmes nominaux ensemble dans un

même vecteur.

– Stratégie 2 : indexer les unitermes et les syntagmes nominaux séparément.

Le nombre de syntagmes nominaux extraits étant négligeable par rapport au nombre de unitermes, le temps d'indexation et d'interrogation pour les différentes stratégies sont sensiblement les mêmes.

8.3.1 Indexer les unitermes

Comme indiqué dans la section 6.5, après avoir fait varier bon nombre de paramètres, nous avons défini la base qui permet d'avoir les meilleures performances et qui consiste à utiliser une troncature, un anti-dictionnaire, et une pondération ltc (section 6.4). Ces performances sont comparées aux résultats obtenus avec l'intégration des syntagmes nominaux pour juger leur impact sur un SRI.

8.3.2 Indexer les unitermes et les SNs ensemble

Les syntagmes nominaux extraits sont ajoutés dans les documents et les requêtes comme étant des unitermes simples. Par exemple, si les deux unitermes *éducation* et *nationale* forme un syntagme nominal alors ils sont remplacés par le syntagme nominal *éducation nationale*. Les unitermes et les syntagmes nominaux sont utilisés ensemble dans un index.

8.3.3 Indexer les SNs indépendamment des unitermes

Pour chaque document ou requête, un nouveau index est créé où les syntagmes nominaux extraits sont ajoutés. Ces syntagmes nominaux sont alors indexés indépendamment des unitermes. Cela crée deux sous-vecteurs dans SMART : le premier correspond aux unitermes et le deuxième aux syntagmes nominaux. Au cours de l'interrogation, nous avons fait varier le poids des deux vecteurs. Les meilleures performances ont été trouvées quand le poids du vecteur syntagmes nominaux est sensiblement inférieur à celui des unitermes.

Le nombre moyen de syntagmes nominaux dans un document est plus important dans le cas de la collection OFIL que dans le cas de la collection INIST comme illustré dans la Figure 8.7, comme nous l'avons indiqué dans la partie 8.2.3. Par contre, le nombre moyen de syntagmes nominaux par requête est plus important dans le cas des requêtes INIST où les termes scientifiques sont plutôt des termes complexes (*structure chimique*, *approche sociologique*, etc.).

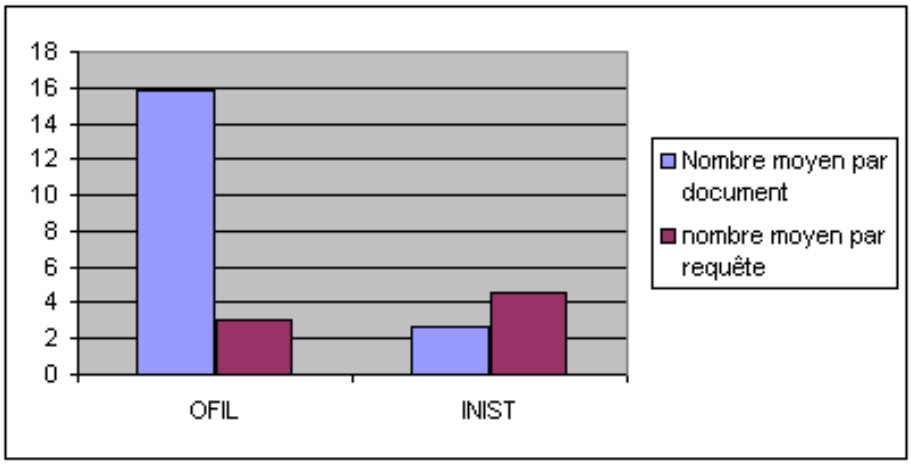

FIG. 8.7 – *Distribution des syntagmes nominaux*

8.3.4 Évaluation Rappel/Précision

En comparant les courbes de rappel et précision de l'INIST, présentées dans la Figure 8.8, et les courbes de rappel et précision de l'OFIL, présentées dans la Figure 8.9, nous pouvons constater qu'en intégrant des syntagmes nominaux dans l'indexation nous pouvons obtenir de meilleures performances par rapport à l'utilisation des unitermes (les performances sont exprimées en terme de précision moyenne en 11 points de rappel).

En particulier, la séparation des unitermes et des SNs dans deux sous-vecteurs différents donne les meilleurs résultats. Cette amélioration est plus nette dans le cas de la collection OFIL où la stratégie 1 a permis d'augmenter les performances de 3.64% (+1.18% pour l'INIST) alors que cette augmentation est de 6.05% (+3.29% pour l'INIST) dans le cas de la stratégie 2. La séparation des unitermes et des syntagmes nominaux donne les meilleures performances principalement dans le cas où le vecteur des syntagmes nominaux est faiblement pondéré par rapport au vecteur des unitermes (entre 10% et 30%). Dans le cadre de cette expérimentation, le vecteur des syntagmes nominaux est pondéré de 30%.

En examinant les résultats des deux stratégies, nous constatons que le nombre de documents retrouvés et pertinents est presque identique pour les deux stratégies. Ce qui diffère est le classement des documents trouvés. En effet, la stratégie 2 permet de favoriser le classement des documents pertinents en les mettant en tête de la liste des documents trouvés. D'après cette stratégie, les documents pertinents ont une similarité avec la requête plus importante que dans le cas de la stratégie 1 ce qui permet d'augmenter la précision

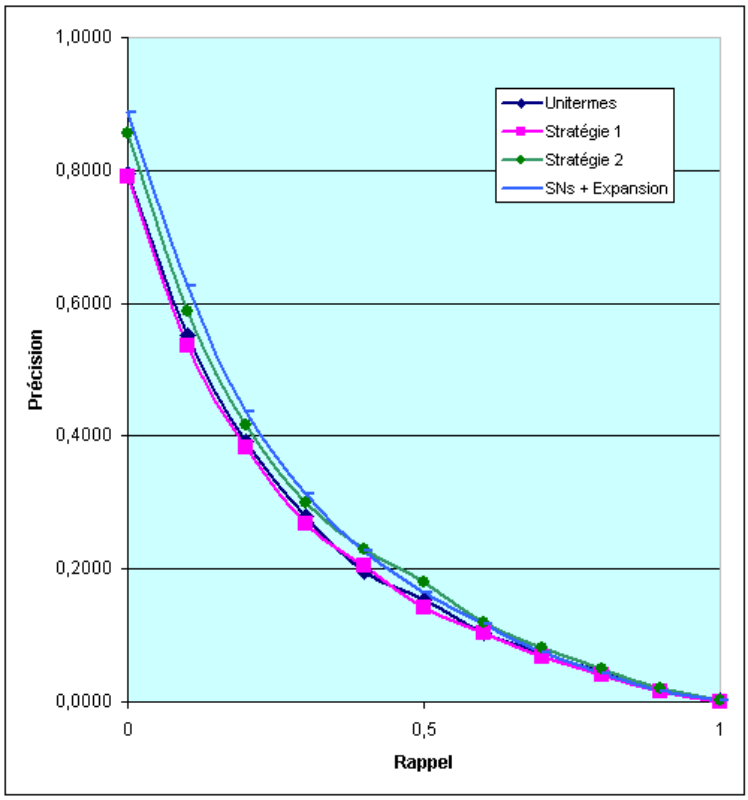

FIG. 8.8 – *Courbes de Rappel-Précision de la collection INIST*

des réponses. Ceci se reflète dans les résultats de la précision pour les premiers documents (précision à 5, 10, 15 et 30 documents) qui sont supérieurs en moyenne de 4% par rapport à ceux de la stratégie 1, ce qui confirme notre hypothèse que les syntagmes nominaux aident à augmenter la précision d'un SRI. Cette remarque se confirme dans le cas d'une indexation avec des unitermes et des syntagmes nominaux séparément et avec une expansion automatique des unitermes et des syntagmes nominaux des requêtes. Cette stratégie d'indexation, désignée par *SNs + Expansion*, reprend une indexation selon la stratégie 2 et applique une expansion automatique des requêtes selon le processus présenté dans la section 6.7. Elle permet d'augmenter les performances de $4,37\%$ pour la collection INIST et de $8,36\%$ pour la collection OFIL (Tableau 8.5). Cependant, cette augmentation est due principalement à une augmentation de la précision. En effet, même si cette stratégie permet de trouver de nouveaux documents pertinents qui ne sont pas trouvés avec les autres stratégies, elle permet de mieux classer les documents pertinents à une requête de façon à

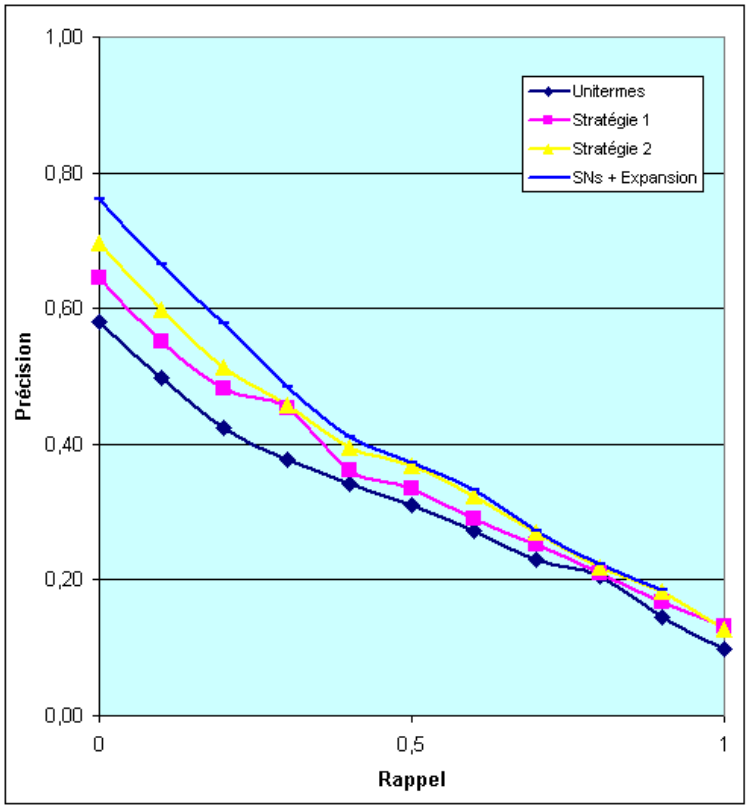

FIG. 8.9 – *Courbes de Rappel-Précision de la collection OFIL*

les mettre dans les premiers classés de la liste des documents trouvés.

Si nous reprenons l'exemple suivant de la requête de la collection OFIL présentée dans la section 6.5:

Exemple 9 *Domaine : International*

Sujet : La Guerre civile en Somalie

Question : Quelles sont les raisons de la poursuite des combats en Somalie ? Quel rôle l'ONU peut jouer en Somalie ?

Compléments : Les documents pertinents devront mettre en lumière les oppositions de clan et de chefs dans la poursuite des combats et ne devront pas ignorer que le comportement de l'ONU en Somalie n'obéit pas à une cohérence marquée.

Concepts: Etats-Unis, France, Négociations, Désarmement, Aide humanitaire, Guerilla.

Dans le cas de la stratégie 2, un ensemble de syntagmes nominaux est extrait de la requête et ajouté à cette dernière dans le sous-vecteur nommé *.SN* comme suit:

Exemple 10 *Domaine : International*
Sujet : La Guerre civile en Somalie
Question : Quelles sont les raisons de la poursuite des combats en Somalie ? Quel rôle l'ONU peut jouer en Somalie ?
Compléments : Les documents pertinents devront mettre en lumière les oppositions de clan et de chefs dans la poursuite des combats et ne devront pas ignorer que le comportement de l'ONU en Somalie n'obéit pas à une cohérence marquée.
Concepts: Etats-Unis, France, Négociations, Désarmement, Aide humanitaire, Guerilla.
.SN
guerre_civile
combat_en_somalie
aide_humanitaire

Une requête est alors composée de deux sous-vecteurs : le premier est composé d'unitermes et le deuxième est composé de syntagmes nominaux.

L'expansion de cette requête dans le cadre de la stratégie *SNs + Expansion*, consiste à ajouter, dans le premier sous-vecteur de la requête, un ensemble d'unitermes qui ont un lien d'association avec les unitermes de la requête et à ajouter un ensemble de syntagmes nominaux, dans le deuxième sous-vecteur de la requête, qui ont un lien d'association avec les SNs de la requête.

Collection	Unitermes	Stratégie 1	Stratégie 2	SNs + Expansion
OFIL	$31,64\%$	$35,28\%$ (+ $3,64\%$)	$37,69\%$ (+ 6.05%)	40% (+ $8,36\%$)
INIST	$22,08\%$	$23,26\%$ (+ $1,18\%$)	$25,37\%$ (+ $3,29\%$)	$26,45\%$ (+ $4,37\%$)

TAB. 8.5 – *Résultats de l'indexation avec des SNs en précision moyenne en 11 points de rappel*

Exemple 11 *Domaine : International*
Sujet : La Guerre civile en Somalie

Question : Quelles sont les raisons de la poursuite des combats en Somalie ? Quel rôle l'ONU peut jouer en Somalie ?

Compléments : Les documents pertinents devront mettre en lumière les oppositions de clan et de chefs dans la poursuite des combats et ne devront pas ignorer que le comportement de l'ONU en Somalie n'obéit pas à une cohérence marquée.

Concepts: Etats-Unis, France, Négociations, Désarmement, Aide humanitaire, Guerilla.

discussion conférence réunion administration communauté nation représentant position conflit territoire paix solution

.SN

guerre_civile
combat_en_somalie
aide_humanitaire
sanglant_affrontement
volonté_collectif
rapprochement_régional
propre_indépendance
communauté_mondiale
force_militaire
population_somalienne
mohamed_farah
alliance_nationale_somalienne

Cette requête permet de retrouver 15 documents pertinents pour chaque stratégie d'indexation. Cependant, ces documents n'ont pas le même classement selon la stratégie appliquée. Dans le cas de la stratégie *SNs + Expansion*, quatre documents pertinents sont dans les dix premiers documents trouvés alors que ces derniers sont classés au delà de la dixième position avec les autres stratégies (Tableau 8.6). Par conséquent, la précision moyenne en 11 points de rappel de cette requête dans le cas de la stratégie *SNs + Expansion* ($37,99\%$) est supérieure à celles des autres stratégies ($32,56\%$ dans le cas de la stratégie 2 et $30,48\%$ dans le cas des unitermes).

Documents	Unitermes	Stratégie 2	SNs + Expansion
294	28	27	5
597	27	26	7
979	44	41	6
1820	16	14	4

TAB. 8.6 – *Classement des documents pertinents trouvés*

8.4 Conclusion

Les expérimentations présentées dans ce chapitre nous ont permis de montrer l'efficacité de notre méthodologie d'extraction de SNs sur des collections hétérogènes dont quelque unes ont une taille particulièrement grande. L'étude de la distribution des catégories grammaticales dans les différents corpus a montré la dominance de certaines catégories qui sont en fait les composantes des patrons syntaxiques que nous avons établis. Ceci renforce notre hypothèse selon laquelle les patrons syntaxiques préétablis permettent d'extraire la plus grande partie d'information contenue dans les collections. L'analyse linguistique de surface du système IOTA s'est montrée suffisante pour notre besoin et efficace en terme de temps de traitement sur les grandes collections qui s'est effectuée avec une vitesse quasi linéaire pour toutes les collections de 5,5 Mo par minute. De même, la vitesse d'extraction des SNs est quasi linéaire de 8 Mo par minute [4]. Une étude qualitative des SNs ne peut pas être envisageable dans ce contexte vu le nombre de SNs extraits. Une étude qualitative sur un échantillon de chaque collection a montré la bonne qualité des SNs extraits. Par bonne qualité, nous faisons référence à la bonne structure linguistique.

L'utilisation des syntagmes nominaux extraits dans l'indexation des documents a permis d'augmenter les performances du SRI. Cette augmentation se présente essentiellement sous la forme d'une augmentation de la précision dans les performances du SRI. En effet, les résultats de nos expérimentations ont montré que l'utilisation des syntagmes nominaux dans l'indexation permet de classer les documents pertinents trouvés dans les premiers rangs de la liste des documents trouvés. Cependant, cette indexation est réalisée dans le cadre du modèle vectoriel qui n'est pas adapté à une indexation avec des syntagmes nominaux. En effet, même si cette indexation permet d'augmenter les performances du SRI, elle ne permet pas de prendre en considération les caractéristiques des syntagmes nominaux et principalement les dépendances entre les syntagmes nominaux définis dans la section 9.2. Dans le chapitre 10, nous présentons un modèle d'indexation qui permet de tenir compte des particularités des syntagmes nominaux.

Une étude qualitative a été effectuée dans le cadre de l'action AUPELF ARC A3. L'ARC A3 est une Action de Recherche Concertée financée par l'AUF (Association des Universités Francophones). Ce projet, dont l'intitulé est "ÉVALUATION DES SYSTEMES DE CONSTRUCTION DE TERMINOLOGIE ET DE RELATIONS SEMANTIQUES ENTRE TERMES", cherche à promouvoir l'élaboration de corpus et de procédures d'évaluation concernant le français. L'évaluation a été essentiellement qualitative, et a été effectuée par les experts (indexeurs, terminologues travaillant dans le domaine) sur la base de l'analyse de l'utilisabilité de l'information procurée par le système en s'appuyant sur

4. les expérimentations sont effectuées sur un serveur Intel Pentium III Xeon 500 MHz

des thésaurus de référence. Deux corpus sont proposés:

– Le corpus INRA proposé par l'Institut National de Recherche en Agronomie. Il est constitué d'articles dans le domaine des biotechnologies.

– Le corpus SPIRALE (Revue de recherche en éducation).

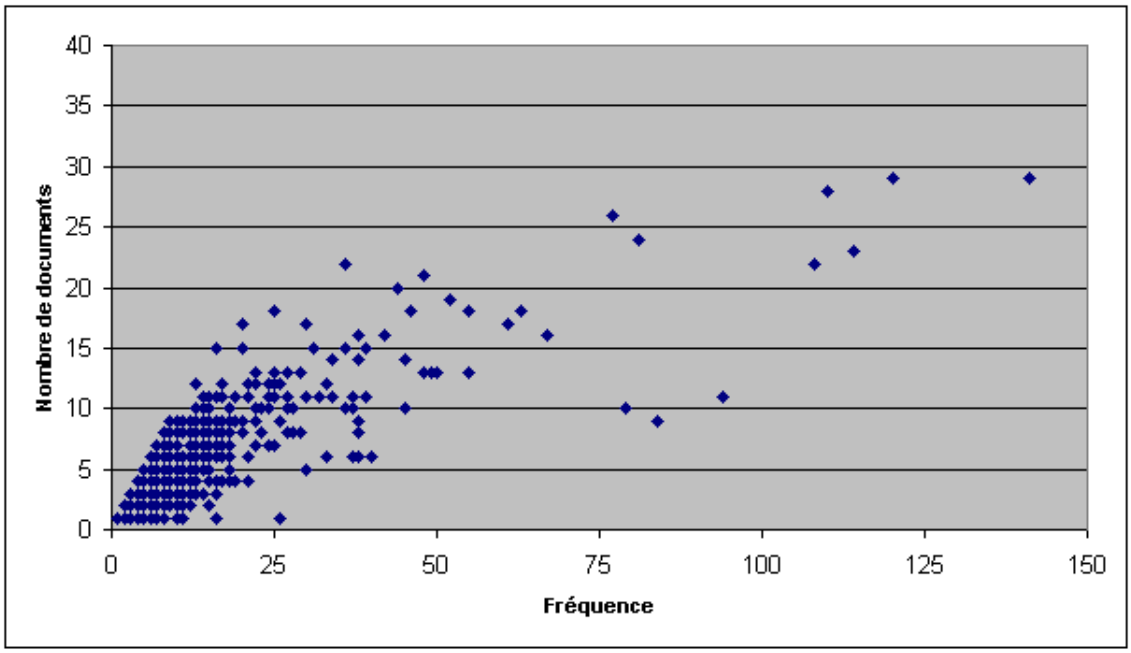

FIG. 8.10 – *Répartition des SNs selon leurs fréquences globales et documentaires dans le corpus INRA*

Les syntagmes nominaux ont une très faible fréquence documentaire. En effet, comme l'illustre la Figure 8.10, plus de 80% des SNs de la collection INRA apparaissent seulement dans un très peu de documents ce qui leur donne un pouvoir de discrimination important.

L'évaluation par un expert des nos résultats dont une partie est illustrée dans le Tableau 8.7, où nous avons présenté d'une part des syntagmes nominaux extraits du corpus SPIRALE ainsi que des associations extraites entre SNs par des règles d'association, montre que la précision de cette extraction est très bonne car sur les SNs que nous avons proposé, seul 9% doivent être considérées comme du bruit. Les règles d'association extraites à partir de ces SNs ont été aussi évaluées et seulement 27% doivent être considérés comme non valides.

Syntagme Nominal	Fréquence totale	Fréquence documentaire	Association
transfert de noyau	45	21	fécondation in vitro transfert de gène
stade morula	44	28	Zone pellucide
fécondation in vitro	162	58	Cycle ovarien Concentration de spermatozoïde
taux de gestation	114	43	Nombre d'embyon
transfert de embryon	120	52	Morulas compactées

TAB. 8.7 – *Exemples de SNs extraits du corpus INRA*

Dans le chapitre suivant, nous présentons notre approche de structuration des syntagmes nominaux extraits dans une base de connaissances.

Chapitre 9

Structuration des SNs

La science consiste à passer d'un étonnement à un autre.
ARISTOTE

9.1 Besoin de structuration

Une simple liste de termes (unitermes et syntagmes nominaux) et de relations d'association entre eux n'est pas suffisante pour une application de recherche d'information. La description des connaissances d'un corpus est généralement un ensemble structuré composé de termes et de relations unissant ces termes. Les connaissances extraites doivent être structurées afin d'en faciliter la consultation de manière interactive par un utilisateur ou de fournir lors du processus d'interrogation des suggestions pour élargir ou focaliser les stratégies de recherche. Pour cela, il peut y avoir plusieurs possibilités de structuration:

– Le regroupement par fréquences:

  Pour identifier les grandes thématiques, il est possible de s'appuyer sur la fréquence des termes : les termes fréquents sont alors considérés comme des mots-thèmes.

– Le regroupement par suffixes :

  Les suffixes peuvent servir à détecter des formes dont les membres partagent une caractéristique commune. Par exemple, les formes qui finissent par -ité expriment souvent des qualités, celles en -ation correspondent souvent à des actions ou à des processus.

– Le regroupement par familles de mots :

  Pour contrer la dispersion de l'information due à la variété de la nature grammaticale des mots beaucoup plus grande dans un corpus textuel que dans une liste de mots

clés contrôlés où ne sont conservés en général que les noms, il est commode de procéder à des regroupements autour d'un même radical. Cette opération s'effectue à partir de la troncature à droite.

– L'utilisation des relations hiérarchiques :

  Les relations hiérarchiques constituent l'ossature du thésaurus. Leur dépistage s'effectue au moyen d'expressions comme : est un(e), sont, tel(le)(s) le, la, te(le)(s) que, etc.

Dans la section 7.4, nous avons vu que la plupart des travaux cités se basent sur une structuration en un réseau de dépendance syntaxique (tête et expansion). Nous appliquons la même stratégie de structuration pour notre base de connaissances. Cette structuration en dépendance syntaxique est bien entendue différente d'une structuration basée sur des dépendances sémantiques (réseaux sémantiques) ou celle d'un thésaurus. Notre analyse n'aborde pas le niveau sémantique de la compréhension de la langue. Le découpage opéré en tête-expansion et les regroupements par têtes et par expansions offrent une vision synthétique du fonctionnement syntagmatique et paradigmatique des termes d'un corpus. Nous avons indiqué dans la section 7.5.2[1] que la tête d'un SN est le concept central (le thème). Notre hypothèse est la suivante:

Hypothèse 3 *Les SNs qui partagent la même tête représentent le même thème.*

Cette structuration a un double objectif : d'une part elle permet d'ordonner la multiplicité des unités lexicales recueillies et d'autre part de fournir une image cohérente du corpus traité. Elle est sous la forme d'un réseau global de dépendance construit pour l'ensemble du corpus où chaque unité est liée à sa tête et à son expansion syntaxique. D'autre part, chaque unité est liée à d'autres unités par des règles d'association comme indiqué dans le chapitre 6[2].

9.2 Structuration de dépendance *tête expansion*

La structuration selon les dépendances syntaxiques des SNs situe les relations à l'intérieur d'un syntagme nominal et porte sur des relations entre expansions. Chacun des SNs est alors structuré en tête et expansions. Pour cela nous utilisons des règles sur les patrons syntaxiques. Une règle de structuration analyse un syntagme nominal de façon binaire et produit un résultat binaire : une tête et une expansion. Ces relations de dépendance syntaxique ne correspondent donc pas à de simples proximités textuelles, mais à des relations

1. Description linguistique générale des syntagmes nominaux
2. Extraction et exploitation de règles d'association

de dépendance vérifiées dans l'analyse linguistique. L'intérêt de ce type de décomposition est de permettre les regroupements paradigmatiques pour rapprocher les SNs sur la base des contextes dans lesquels ils apparaissent et dégager certaines des relations sémantiques qu'ils entretiennent. Ainsi, comme le précise Bourigault dans [Bou93], cette structuration permet de mettre en évidence des liens de co-hyponymie et des liens d'hyperonymie. Un liens de co-hyponymie est une relation entre deux SNs qui partagent la même tête mais qui ont des expansions différentes. Par exemple, le terme *structure* est la tête d'un ensemble de SNs qui partagent un lien de co-hyponymie tel que [3] *structure de soutènement, structure anisotrope, structure d'évaluation, structure spatial, structure électronique, structure de fluorite, structure spiralé, structure du tricyanovinyl, structure du trifluorométhyl, structure de développement, structure géométrique, structure à bande, structure d'empilement,* etc. Un lien d'hyperonymie est un liens entre un SN et un autre qui le prolonge comme par exemple *assurance multirisque* et *assurance multirisque habitation*.

Comme pour l'extraction des SNs, les patrons syntaxiques sont utilisés pour structurer les SNs extraits sous la forme de *tête expansion*. Cette structuration se base alors sur le fait que pour un patron syntaxique donné, une règle précise le rôle que joue chaque composante du patron. Dans le tableau 9.1, nous donnons quelques exemples de patrons syntaxiques et leur structuration en têtes et expansions.

Patrons syntaxique	Tête	Expansion
SUBC1 SUBC2	SUBC1	SUBC2
SUBC ADJQ	SUBC	ADJQ
ADJQ SUBC	SUBC	ADJQ
SUBC1 PREP SUBC2	SUBC1	SUBC2

TAB. 9.1 – *Exemples de structuration de patrons syntaxiques*

Le cas des noms propres est un cas particulier où les composants ne peuvent pas être décomposés en tête et expansion. Nous considérons alors que la dernière composante du syntagme est la tête et le reste des composantes sont les expansions.

9.3 Quantité d'information

Dans les systèmes de recherche d'information, des poids sont affectés aux termes selon leurs occurrences dans le corpus et dans un document. Le *tf*idf* est la fonction la plus utilisée. Par cette fonction de poids, les termes complexes sont désavantagés par rapport

3. Exemples extraits de la collection INIST

aux termes simples par le fait qu'ils sont moins fréquents dans un document et dans le corpus à cause de:

– L'effet de style : un auteur utilise des anaphores ou des références elliptiques.

– Les phénomènes linguistiques : la variance linguistique par exemple est un phénomène qui fait que le comptage des termes complexes ne se fait pas correctement. Par exemple, la variation lexicale concerne le glissement d'un terme vers un autre sémantiquement proche comme, par exemple, *voiture* et *automobile* alors que la variation syntaxique concerne les différentes constructions syntaxiques ayant un sens voisin comme *pollution de l'air du fait des moteurs diesel* avec *pollution de l'air par des moteurs diesel*.

– Les termes complexes sont moins fréquents par rapport aux termes simples ; cette propriété est intrinsèque aux termes complexes.

Par exemple, la pondération [4] des termes du premier document de la collection OFIL dont le titre est *Les plumes de l'ange,* illustrée dans le tableau 9.2, montre que le poids de la plupart des termes complexes est inférieur à celui des termes simples.

Le poids des termes *texte* et *écrit* est supérieur à celui du syntagme nominal *texte_écrit* car ce dernier a une fréquence plus faible que les deux termes simples. Pour résoudre ce problème et dans le souci de donner plus d'importance aux termes complexes, on peut envisager de résoudre la majorité des phénomènes linguistiques (comme l'anaphore par exemple). Cette solution a l'inconvénient de nécessiter une analyse en profondeur du texte et l'emploi de connaissances d'ordre sémantiques et paradigmatiques. Or, un SRI n'a pas comme tâche le traitement en profondeur de la langue qui est une analyse lourde et ne possède pas (sauf exception et dans des domaines fort spécialisés où des connaissances d'ordre sémantique et paradigmatique ont été injectées dans le système) les connaissances nécessaires à la compréhension de la langue.

A défaut de résoudre les problèmes cités précédemment, la solution retenue est celle d'augmenter *artificiellement* le poids des termes complexes. " Artificiellement " dans le sens où la fréquence des termes ne suffit pas à mesurer l'importance des termes dans un document ou dans le corpus ; l'importance reste une mesure subjective qui dépend de plusieurs critères, et non seulement du nombre de fois où un terme apparaît. Un de ces critères que nous proposons est la *qualité d'information*.

Si nous tenons à l'hypothèse que les syntagmes nominaux sont les unités linguistiques les plus susceptibles de représenter le contenu des documents, la question que nous nous posons est comment mesurer cette représentativité ? En d'autres termes, pourquoi un SN est plus représentatif qu'un autre SN ?

4. Pondération ltc

La représentativité d'un SN est sa capacité à encapsuler le contenu textuel et sa contribution à spécifier une information, donc son apport informationnel, pour un utilisateur. Il nous faut alors une mesure qui permette de calculer l'apport informationnel d'un SN. L'information contenue dans la phrase et, plus généralement, exprimée par le document, doit se refléter dans le SN. Cette grandeur, a priori, indépendante de sa pertinence au document, exprime la quantité d'information *empruntée* au document tout entier. Cependant, on peut remarquer que les SNs porteurs de beaucoup d'information, pourront probablement être jugés fortement pertinents pour le document.

Terme	Pondération ltc
harmonique	0.30438
théorème	0.30438
assonance	0.30438
baudoin	0.30438
plume	0.30438
dessinée	0.30438
pasolini	0.28171
invente	0.25905
belle	0.25905
écrit	0.24579
texte	0.24074
ange	0.23638
intervention_forte	0.16365
laura_betti	0.16365
massimo_girotti	0.16365
dessin_achevé	0.16365
pier_palo_pasolini	0.16365
improbable_exploit	0.16365
demeure_milanaise	0.16365
langage_scientifique	0.16365
film_homonyme	0.15146
texte_écrit	0.14433
jeu_de_miroir	0.14433

TAB. 9.2 – *Poids des termes du document Les plumes de l'ange*

Nous voulons alors que la quantité d'information(Qinf(SN)) d'un SN permette d'exprimer le pouvoir évocateur de ce SN. Un SN avec une quantité d'information non nulle, reflète alors une partie de l'information exprimée par le document dans lequel il se trouve.

Si on considère l'exemple de la *séparation de la république tchèque*, on peut noter que ce syntagme est plus *riche* et plus *précis* du point de vue informatif que le syntagme *séparation de la république* ou bien celui de *république tchèque*.

Si nous reprenons les catégories grammaticales des éléments d'un syntagme nominal, nous considérons que certaines catégories ont un comportement informatif quasi nul c'est-à-dire n'apportent pas une information importante (par exemple les conjonctions) au SN alors que d'autres apportent une quantité d'information élevée (par exemple les substantifs).

En outre, la tête d'un SN est considérée comme une entité prépondérante et il est nécessaire de la privilégier par rapport aux autres entités. La tête d'un SN ne peut avoir que la catégorie grammaticale Substantif [5]. Cette catégorie grammaticale se voit alors associée une quantité d'information plus importante que les autres catégories. On propose alors de classer les catégories grammaticales des SN selon la hiérarchie suivante :

- La catégorie des substantifs est la catégorie la plus porteuse d'information.
- Les catégories d'adjectif, verbe à l'infinitif, participe passé et adverbe ont une quantité d'information moyenne.
- Les catégories de proposition, conjonction et article ont une quantité d'information nulle.

Nous déterminons empiriquement deux valeurs α et β où

- α exprime la quantité d'information de la catégorie des substantifs.
- β exprime la quantité d'information des catégories d'adjectif, verbe à l'infinitif, participe passé et adverbe.

Avec $\alpha \gg \beta$ [6] ce qui reflète l'importance accordée à la catégorie des substantifs par rapport aux autres catégories.

Pour prendre en compte tous ces éléments, nous utilisons la formule heuristique suivante pour représenter la quantité d'information:

$$Qinf(SN) = \sum_{a \in SN} Qinf(Cat(a))$$

5. Substantif commun et Substantif propre
6. $\gg$: *bien supérieur à*

Où *a* représente une composante de SN et Cat(a) la catégorie grammaticale de cet atome.

Si nous reprenons l'exemple du SN *séparation de la république fédérale tchèque*, ce SN étant constitué de deux Substantifs (séparation et république) et de deux adjectifs (fédérale et tchèque) sa quantité d'information est calculée comme suit :

$$Qinf(SN) = \sum_{a \in SN} Qinf(Cat(a)) =$$

$$Qinf(Cat(\text{séparation})) + Qinf(Cat(\text{république}))+$$

$$Qinf(Cat(\text{fédérale})) + Qinf(Cat(\text{tchèque})) =$$

$$Qinf(Substantif) + Qinf(Substantif) + Qinf(Adjectif)+$$

$$Qinf(Adjectif) = 2\alpha + 2\beta$$

La mesure de la quantité d'information peut être considérée de deux point de vue:

1. Indépendante du contexte du terme (un point de vue cognitif) :

   la mesure de la quantité d'information est considérée indépendamment du contexte où le terme se trouve et elle est la même quelque soit le contexte où ce terme est employé donc elle est basée seulement sur la longueur et les catégories grammaticales de ces composantes. Ce point de vue répond à la question suivante : dans le schéma cognitif du lecteur combien ce terme est discriminant et permet de préciser le signifié référencé par ce terme?

   Si on prend l'exemple du syntagme *ballon bleu*, ce syntagme peut évoquer un certain nombre d'entités pour le lecteur. Cet ensemble d'entités est susceptible d'être très large vue que ce terme dénote un ensemble relativement grand d'entités dont chacun peut être un *ballon bleu*. Par contre, le terme *ballon de foot* dénote un type particulier de ballon donc dans le schéma cognitif du lecteur il évoque moins d'entités mais l'ensemble évoqué est moins large donc plus précis.

2. Dépendante du contexte du terme:

   Ce point de vue tient compte de la fréquence du terme dans le document et dans le corpus ainsi que de la quantité d'information du terme. Le poids du terme n'est plus en fonction de la fréquence du terme (tf*idf) mais aussi de la quantité d'information du terme.

9.4 Comparaison des SNs

La quantité d'information permet de comparer les unités lexicales entre elles. Les SNs comparables sont alors ceux qui partagent de l'information. Dans le cadre de la structuration par dépendances syntaxiques, les SNs qui partagent la même information sont ceux qui partagent la même tête. Si nous considérons deux syntagmes nominaux SN et SN', nous pouvons calculer la différence entre eux en terme de quantité d'information. Cette différence consiste à une dérivation de SN en SN' représentée par le coût de cette transformation.

Comme nous ne faisons pas appel à des connaissances sémantiques prédéfinies, les transformations qui s'appliquent sont d'ordre syntaxique. L'ensemble de ces connaissances, noté K, s'intéresse aux catégories grammaticales des composantes d'un SN, les règles syntaxiques de l'agencement de ces composantes, les interdictions et les possibilités d'insertion d'un nouvel élément ou de la suppression d'un élément existant du syntagme nominal. Ainsi, une dérivation d'un SN en SN' consiste à ajouter ou supprimer une de ses extensions tel que SN' vérifie un des patrons syntaxiques définis. Par exemple, le SN *livre d'images*, où l'expansion *images* est dominée par la tête *livre*, peut être transformé par simplification en *livre* mais pas en *images*. En effet, comme nous l'avons souligné, nous considérons la tête d'un SN comme étant une entité prépondérante qui exprime le concept central du SN. Il est donc évident que cette entité ne peut pas être supprimée au cours d'une transformation au risque de perdre le sens exprimée par le SN. Par contre, les expansions peuvent faire l'objet d'une transformation soit en les éliminant soit en ajoutant de nouvelles expansions au SN original. Ainsi *livre d'images* peut être transformé en *petit livre d'images*.

Les transformations d'un syntagme nominal correspondent à des dérivations utilisant des schémas de transformation. Ainsi, un syntagme nominal SN peut être transformé en d'autres SNs. Soient P l'ensemble des patrons syntaxiques et P le patron syntaxique de SN, une insertion d'une extension à un syntagme est définie comme suit:

> une insertion est une transformation de SN en SN' qui associe une expansion à la tête de SN ou à l'une de ces expansions et le patron syntaxique de SN' vérifie un des patrons de P.

Si nous reprenons l'exemple de *la république tchèque*, ce syntagme correspond au patron syntaxique suivant:

$P :=$ Substantif Adjectif

Une dérivation de ce syntagme en *république fédérale tchèque* consiste à insérer l'adjectif *fédérale*. Le patron syntaxique du nouveau syntagme est le suivant:

$P' :=$ Substantif Adjectif Adjectif

Si $P' \in P$ alors cette insertion est valide et SN peut être transformé en SN'.
De même, une suppression est définie comme suit:

> une suppression est une transformation de SN en SN' qui élimine une expansion de la tête de SN ou de l'une de ces expansions et le patron syntaxique de SN' vérifie un des patrons de P.

L'ensemble K correspond donc à l'ensemble des insertions et des suppressions possibles sur un syntagme nominal.

Nous considérons que la transformation de SN en un autre syntagme nominal SN' est une implication. La certitude de cette implication $SN \rightarrow SN'$ est déterminée par les transformations nécessaires effectuées sur (SN) pour le ramener à SN'. La mesure de ces transformations est utilisée pour donner une mesure à $SN \rightarrow SN'$ désignée par le coût d'une dérivation notée $T(SN \rightarrow SN')$.

La fonction de calcul du coût d'une dérivation est définie comme étant la mesure de la perte d'information (coût de transformation positif) ou de gain d'information (coût de transformation négatif) due à une transformation. S'il s'agit d'une transformation importante qui met en jeu des substantifs alors le coût de la dérivation est important. Si les transformations portent sur des adjectifs, adverbes, verbe à l'infinitif ou participe passé alors le coût de la dérivation est moins important.

Un gain d'information indique que le nouveau SN contient plus d'information donc c'est une transformation vers un SN plus spécifique. En revanche, une perte d'information indique que le nouveau SN contient moins d'information donc c'est une transformation vers un SN plus générique.

Soit par exemple la dérivation du syntagme nominal I_1 en un syntagme nominal I_2 où I_1 = *séparation de la république tchèque* et I_2 = *séparation de la république fédérale tchèque*. Nous avons alors le coût de dérivation de I_1 en I_2, noté $T(I1 \rightarrow I2)$, suivant :

$$T(I_1 \rightarrow I_2) = Qinf(I_1) - Qinf(I_2) = (2\alpha + 1\beta) - (2\alpha + 2\beta) = -\beta.$$

La coût de la dérivation de I_1 en I_2 est inférieur à 0 ce qui reflète une transformation d'un SN vers un syntagme plus spécifique.

Si on considère l'exemple suivant où le même syntagme nominal I_1 de l'exemple précédent est dérivé en I_3 avec I_3 = *séparation de la république*, le coût de cette dérivation est calculé comme suit :

$$T(I_1 \rightarrow I_3) = |Qinf(I_1) - Qinf(I_3)| = (2\alpha + 1\beta) - (2\alpha) = \beta$$

La coût de la dérivation de I_1 en I_3 est supérieur à 0 ce qui reflète une transformation d'un SN vers un syntagme plus générique.

Dans la section 10.2, nous montrons que ce processus de dérivation des SNs peut être considéré comme étant un processus de calcul de correspondance incertain de déduction.

9.5 Filtrage des syntagmes nominaux

Pour la structuration des SNs dans la base de connaissances, il n'est pas envisageable de garder tous les SNs extraits d'un corpus. Un filtrage des SNs s'impose avant de les intégrer dans la base de connaissances. Plusieurs approches ont été utilisées pour le filtrage d'unité lexicale d'une collection comme par exemple, la valeur de discrimination ou un schéma de pondération (tf*idf). L'approche la plus utilisée est celle qui défini deux seuils de fréquence et ne garde que les termes dont la fréquence est incluse dans cet intervalle. Ce filtrage se base alors sur l'hypothèse que plus le terme est fréquent plus il est important dans une collection. Il est clair que cette approche favorise plus les unitermes que les séquences de termes et dans notre cas les SNs. Ces derniers sont très peu fréquents par rapport aux unitermes donc moins avantagés dans un processus de filtrage qui se base sur les fréquences. Pour cette raison, nous utilisons aussi la quantité d'information dans notre approche de filtrage.

Dans la partie 9.3, nous avons défini la quantité d'information comme étant la mesure de représentativité des unités lexicales indépendamment de leur fréquence dans les documents. La quantité d'information d'une unité lexicale augmente quand ce dernier se spécialise c'est-à-dire augmente en longueur mais sa fréquence devient moins importante (Figure 9.1).

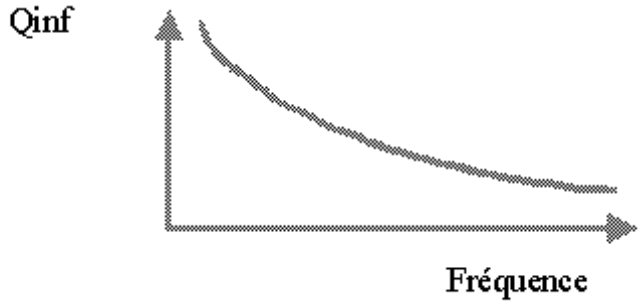

FIG. 9.1 – *Comportement de la quantité d'information par rapport à la fréquence*

L'utilisation de la quantité d'information dans le filtrage peut alors pallier au fait que les SNs ont une faible fréquence dans les documents. Nous définissons alors une fonction de filtrage F qui prend en compte de la façon suivante la fréquence ainsi que de la quantité d'information:

$$F(SN) = Qinf(SN) \times Frquence(SN)$$

L'introduction de la quantité d'information dans la fonction de filtrage permet d'augmenter l'importance des SNs artificiellement.

Si on considère les paramètres de filtrage suivant:

- La quantité d'information de la catégorie des substantifs = α = 4
- la quantité d'information des catégories d'adjectif, verbe à l'infinitif, participe passé et adverbe = β = 2
- Le seuil de filtrage = 15

SN	Patron syntaxique	Fréquence	Qinf	Fréquence *Qinf
système	SUBC	1824	4	7296
système monétaire	SUBC ADJQ	134	6	804
système éducatif	SUBC ADJQ	68	6	408
système de soins	SUBC PREP SUBC	5	8	40
système démocratique	SUBC ADJQ	5	6	30
système multimédia	SUBC ADJQ	2	6	12
système téléphonique interne	SUBC ADJQ ADJQ	1	8	8

TAB. 9.3 – *Exemple de SNs dont la tête est système*

L'application de ce filtrage dans le cas de la collection OFIL élimine pratiquement 96% des SNs. En effet, le nombre de SNs est passé de 174950 à 7642 après filtrage. Dans le tableau 9.3, nous illustrons un exemple d'un ensemble de SNs dont la tête est *système* ainsi que les mesures de la fréquence, la quantité d'information et la Fréquence * Qinf relatives à chaque SN. Si nous considérons la fréquence comme étant le seul paramètre de filtrage et que le seuil de filtrage est fixé à 15, les SNs qui ne seront pas sélectionnés sont : *système téléphonique interne, système multimédia, système démocratique* et *système de soins*. Par contre si on considère la quantité d'information aussi dans les paramètres de filtrage alors les SN éliminés sont *système multimédia* et *système téléphonique interne*

Pour les mêmes paramètres de filtrage, le nombre de substantifs passe de 61758 à 15533 substantifs. La proportion des substantifs éliminés est d'environ 75%.

Les deux syntagmes *système de soins* et *système démocratique* ont la même fréquence dans la collection OFIL où ils occurrent 5 fois chacun. Par contre, leurs mesures de quantité d'information ne sont pas les mêmes. En effet, la valeur de la quantité d'information de *système de soins* est de 40 alors que celle de *système démocratique* est de 30. La différence entre ces deux valeurs est du fait que *système de soins* est jugé plus informatif que *système démocratique*.

9.6 Méthodologie de structuration

La structuration des SNs extraits se base alors sur une structure de dépendance syntaxique des SNs où un SN est défini par rapport à sa tête, sa quantité d'information et sa fréquence :

SN : [Qinf, freq, X]

où :

- Qinf : quantité d'information
- Freq : fréquence
- X : tête

Nous utilisons une analyse ascendante qui permet de regrouper les SNs en allant des plus petits aux plus grands.

Dans la Figure 9.2, nous illustrons la structuration de quelques SNs dont la tête est le terme *système* et que nous détaillons dans l'annexe G:

- système [4, 1824, ∅]
- système monétaire [6, 134, système]
- système de soins [8, 6, système]
- système monétaire de l'Europe [10, 2, système monétaire]
- système monétaire international [8, 13, système monétaire]
- système de soins spécialisé [10, 2, système de soins]
- système de soins Français [10, 2, système de soins]

Les SNs sont alors structurés du plus court, donc le terme *système* au quel seront reliés les SNs *système monétaire* et *système de soins*, au plus long. Chaque SN dans la structure est accompagné de ces caractéristiques : quantité d'information, fréquence et la tête dont il est l'expansion. Le symbole ∅ relatif à la tête de *système* signifie que ce dernier n'est pas l'expansion d'un SN. Nous remarquons que la valeur de la quantité d'information augmente en allant des SNs courts vers les SNs longs et, inversement, la fréquence diminue.

Dans la structure, les SNs sont aussi liés entre eux par des règles d'association, présentées dans la Figure 9.3 avec des lignes orientées et pointillées, découvertes à partir du corpus traité et que nous détaillons dans l'annexe F. Une règle d'association est alors une lien orienté dont le poids est la valeur de la confiance de cette règle.

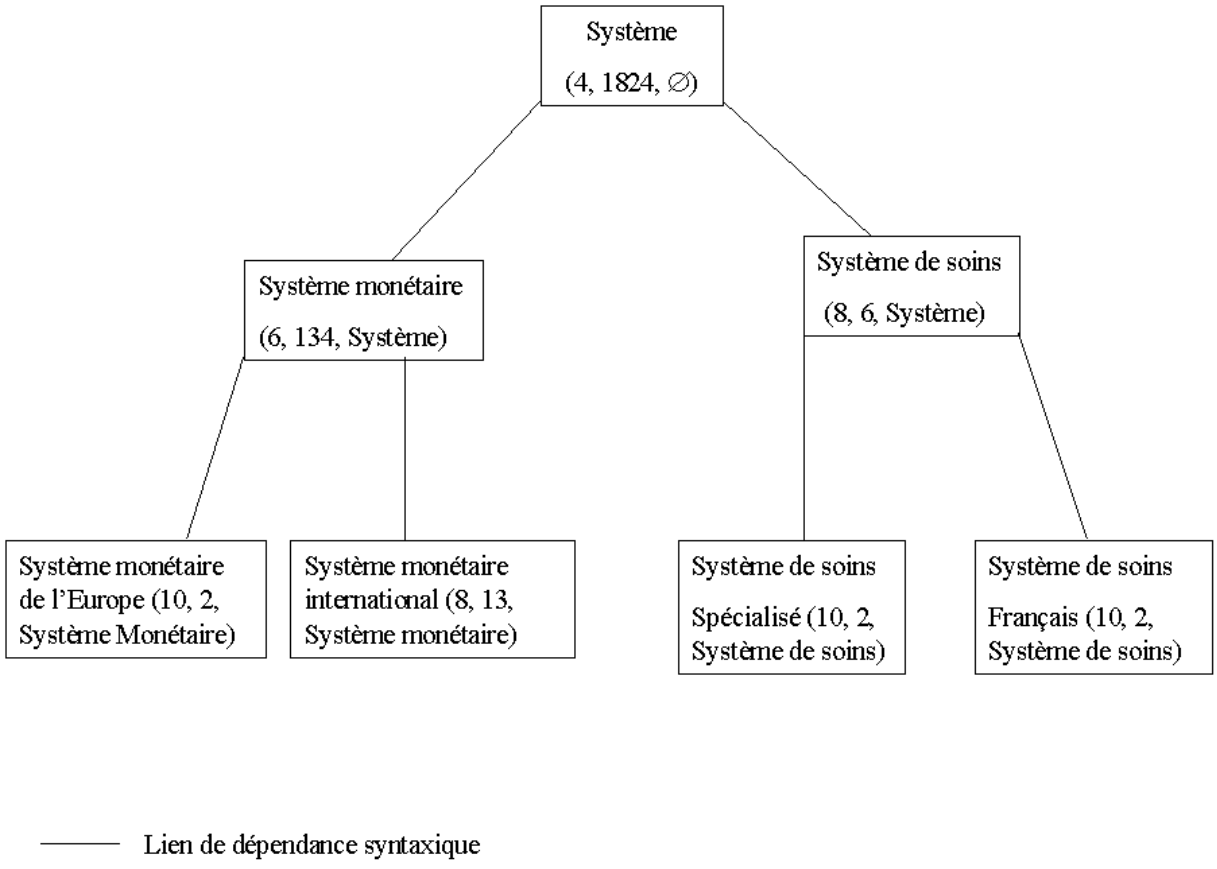

FIG. 9.2 – *Structuration des dépendances syntaxiques des syntagmes nominaux*

9.7 Héritage des règles d'association

Dans la base de connaissance et semblablement aux liens de dépendances, les règles d'association sont représentées avec des liens d'association. Dans la Figure 9.3, quelques règles d'association relatives aux SNs de la figure 9.2 sont illustrées. Certains SNs dont la fréquence est faible ne possèdent pas de règles d'association. C'est le cas dans l'exemple des SNs de taille supérieure à deux constituants[7]. Ces derniers héritent alors les règles

7. système monétaire de l'Europe, système monétaire international, système de soins spécialisés et système de soins français

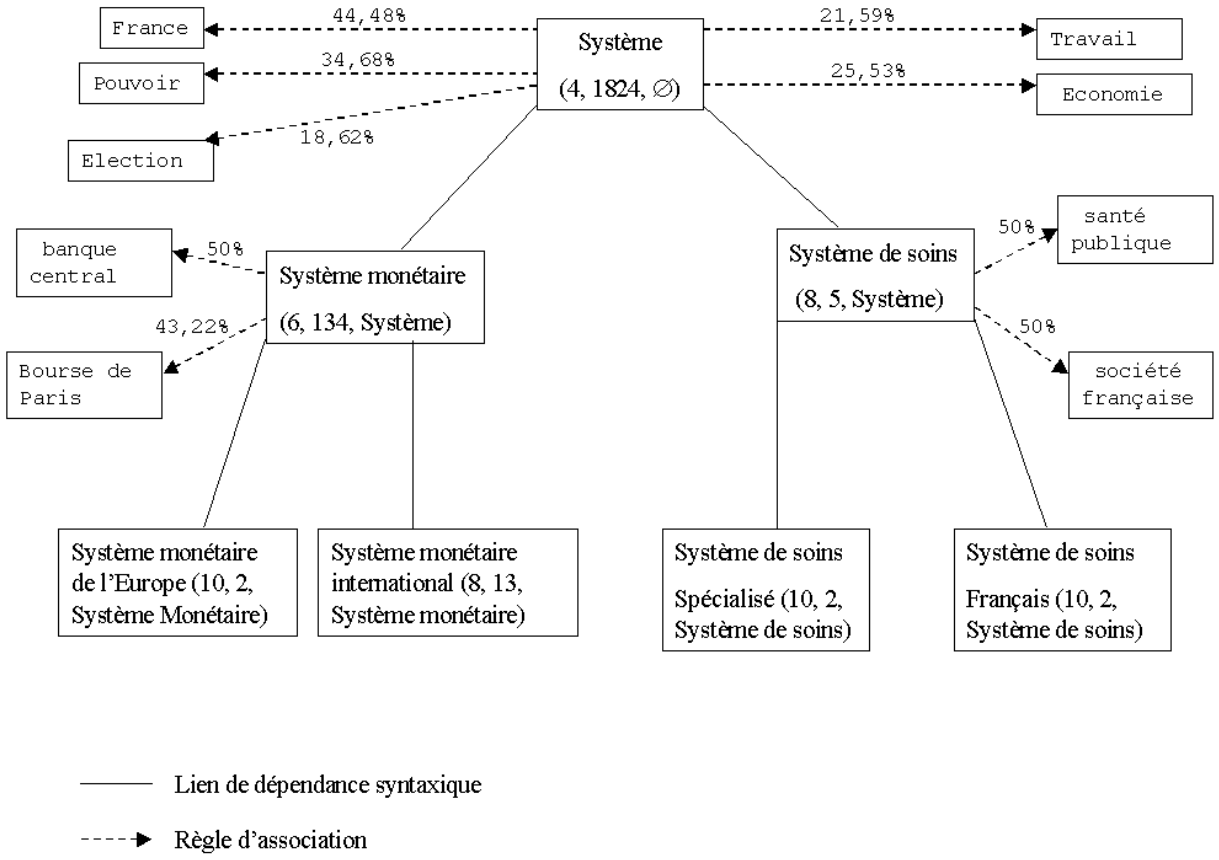

FIG. 9.3 – *Strcuturation des syntagmes nominaux en dépendances syntaxiques et règles d'association*

d'association des têtes desquelles ils sont les expansions selon la règle suivante:

> Si un syntagme nominal SN_1 est la tête du SN_2 et que SN_2 ne possède pas de règles d'association alors SN_2 hérite les règles d'association de SN_1 et les valeurs de confiance des règles d'association héritées diminuent selon le coût de la transformation effectuée.

La valeur de la confiance de la règle d'association héritée correspond à la valeur de la confiance de la règle d'association originale divisée par le coût de la transformation nécessaire pour dériver SN_1 en SN_2.

Si on considère qu'il existe une règle d'association $SN_1 \rightarrow_c SN_3$ dont la valeur de confiance est c et que SN_2 dont la tête est SN_1 ne possède pas de règles d'association

alors ce dernier hérite la règle d'association de sa tête dont la valeur de confiance c' est calculée comme suit:

$$c' = \frac{c}{T(SN_2 \rightarrow SN_1)}$$

SN_2 sera alors lié à SN_3 par la règle d'association $SN_2 \Rightarrow_{c'} SN_3$.

La sémantique de l'héritage des règles d'association est que SN_1 exprime un thème particulier et que SN_2 exprime une partie (plus précise) de ce thème. Si SN_1 est lié à SN_3 par une règle d'association alors SN_3 définit l'environnement de l'utilisation de SN_1 dans le corpus par conséquent définit aussi l'environnement de l'utilisation de SN_2 dans le corpus.

Si on reprend l'exemple précédent, le SN *système de soins spécialisés* qui a comme tête le SN *système de soins* ne possède pas de règles d'association le liant à d'autres SNs à cause de sa faible fréquence dans le corpus. Il va alors hériter les deux règles d'association de sa tête. La valeur de confiance des règles d'association héritées est divisée par la différente de quantité d'information entre *système de soins spécialisés* et *système de soins* (Figure 9.4). La différence de la quantité est le coût de la dérivation de *système de soins spécialisés* en *système de soins* calculée comme suit:

$T($ système de soins spécialisés $\rightarrow$ système de soins$) = Qinf($spécialisés$) = \beta$.

En considérant la valeur de β =2, la valeur des règles d'association héritées est alors:

$c' = \frac{50\%}{2} = 25\%$

système de soins spécialisés est alors associé à deux SNs selon les deux règles d'association suivantes:

système de soins spécialisés $\Rightarrow_{25\%}$ santé publique

système de soins spécialisés $\Rightarrow_{25\%}$ société française

9.8 Conclusion

La structuration des syntagmes nominaux extraits d'une collection textuelle se base à la fois sur les dépendances syntaxiques entre les différentes unités textuelles et les relations d'association entre ces unités. La mesure de la quantité d'information permet d'exprimer le pouvoir évocateur d'un syntagme nominal et de comparer des syntagmes nominaux entre eux. Les syntagmes nominaux qui ne possèdent pas des associations avec

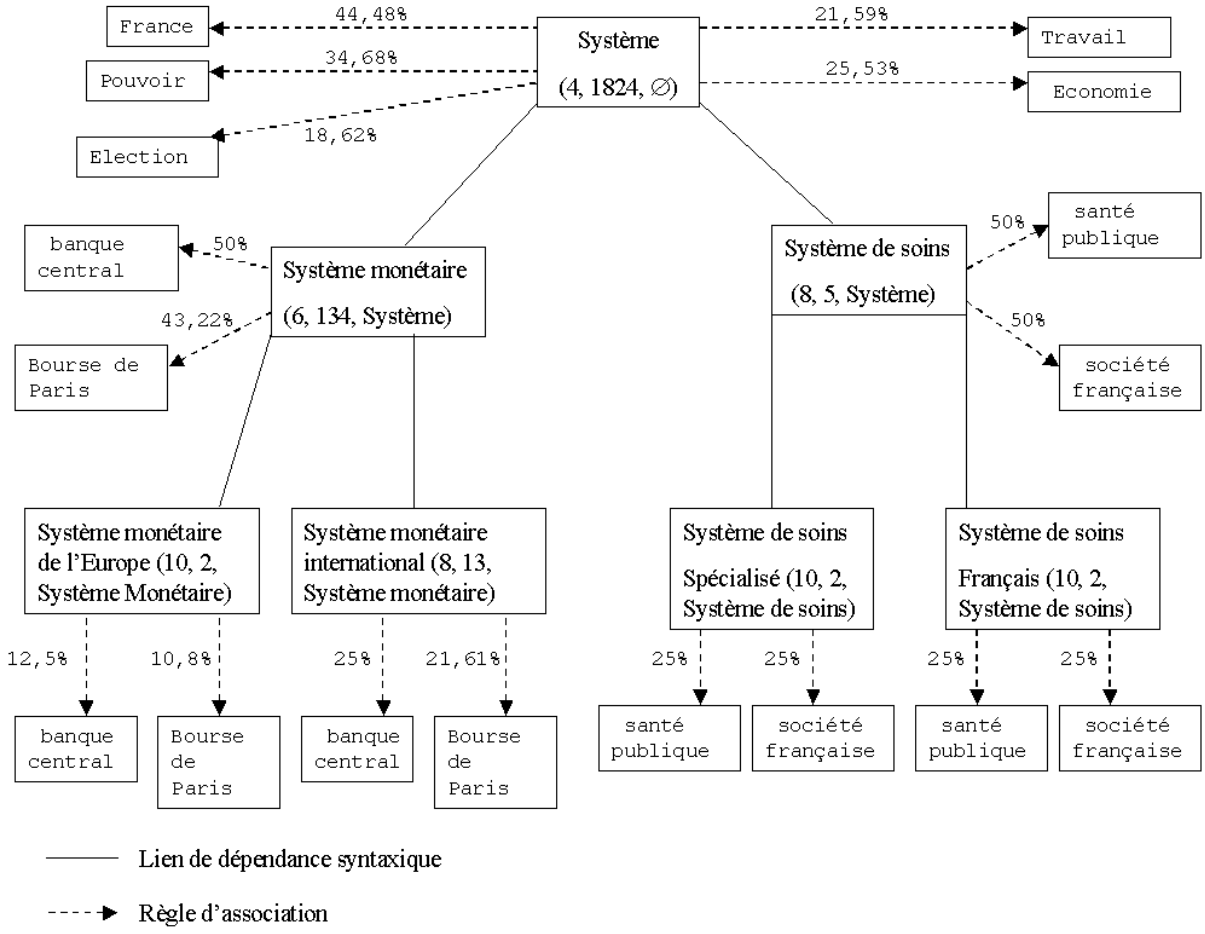

FIG. 9.4 – *Exemple d'héritage de règles d'association*

d'autres syntagmes dans la base de connaissances héritent les associations des unités formant leurs têtes afin de définir l'environnement de leur utilisation dans le corpus.

Cette structuration permet de refléter le contenu d'une collection textuelle. En effet, les éléments constituant cette structure sont des unités textuelles qui sont représentatives du contenu de la collection et chaque unité est accompagnée par ces liens de dépendance ainsi que l'environnement de son utilisation dans la collection. Il est alors intéressant d'exploiter cette structuration par l'utilisateur lors d'une interrogation afin d'élargir ou de focaliser sa stratégie de recherche. Les liens de dépendance et d'association sont alors utilisés pour suggérer à l'utilisateur d'autres termes pour préciser sa recherche.

Si nous reprenons l'exemple de la requête *La Guerre civile en Somalie* présentée dans les sections 6.5 et 8.3.4, en considérant les mêmes paramètres de la section 9.5 et en consultant la base de connaissances avec le syntagme nominal *guerre civile*, les SNs relatifs à *guerre civile* sont:

- guerre civile SUBC ADJQ [6, 62, guerre SUBC]
- reprise de la guerre civile SUBC ARTC ARTD SUBC ADJQ [10, 5, guerre civile SUBC ADJQ]
- origine directe de la guerre civile SUBC ADJQ ARTC ARTD SUBC ADJQ [12, 2, guerre civile SUBC ADJQ]
- raison de la guerre civil SUBC ARTC ARTD SUBC ADJQ [10, 2, guerre civile SUBC ADJQ]
- protagoniste de la guerre civil SUBC ARTC ARTD SUBC ADJQ [10, 2, guerre civile SUBC ADJQ]
- responsable de la guerre civile ADJQ ARTC ARTD SUBC ADJQ [8, 3, guerre civile SUBC ADJQ]

Les syntagmes *raison de la guerre civile* et *origine directe de la guerre civile* qui expriment les raisons de la poursuite des combats en Somalie sont alors ajoutés à la requête ainsi que *protagoniste de la guerre civile* et *responsable de la guerre civile* qui expriment les clans et les chefs de guerre.

Les SNs relatifs au terme *guérilla* et qui sont ajoués à la requête sont:

- mouvement de guérilla SUBC ARTC SUBC [8, 4, guérilla SUBC]
- chef de la guérilla SUBC ARTC ARTD SUBC [8, 2, guérilla SUBC]
- dirigeant de la guérilla ADJQ ARTC ARTD SUBC [6, 4, guérilla SUBC]

Le syntagme *mohamed farah* est une composante du syntagme *mohamed farah aïdid* comme indiqué dans la Figure 9.5 et il est lié par une relation d'association à *faction de l'alliance nationale somalienne*.

Une interrogation de la base de connaissances avec le terme *guerre* permet de trouver le syntagme *chef du guerre somalien* qui est lié avec une relation d'association à *faction somalienne* et *conférence de la réconciliation nationale somalienne*.

La nouvelle requête est alors la suivante;

Exemple 12 *Domaine : International*

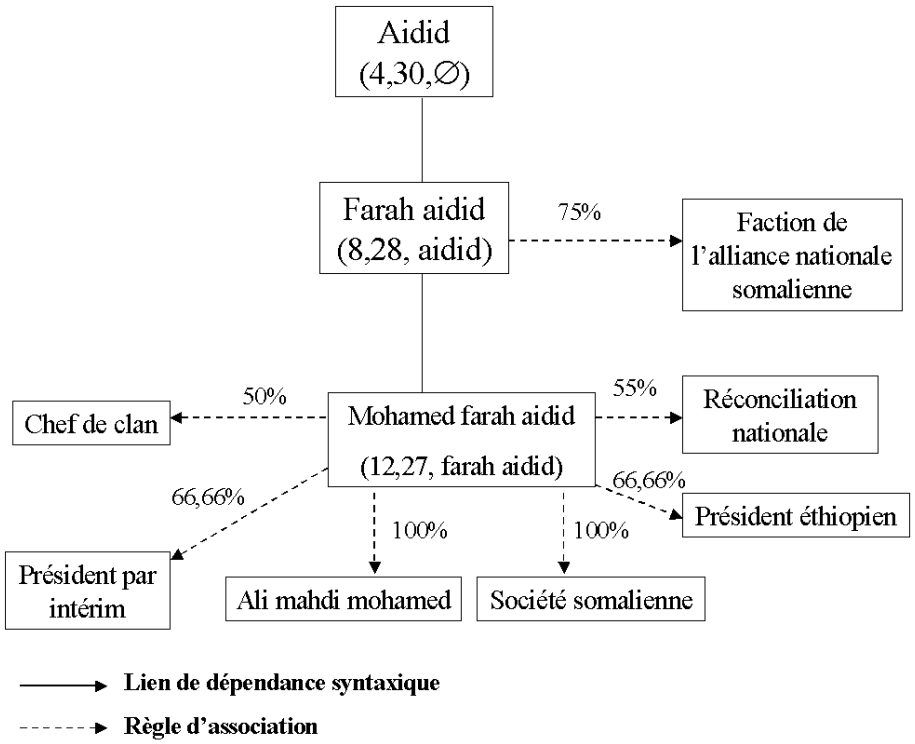

FIG. 9.5 – *Strcuturation des syntagmes nominaux relatives à mohamed aidid en dépendances syntaxiques et règles d'association*

Sujet : La Guerre civile en Somalie

Question : Quelles sont les raisons de la poursuite des combats en Somalie ? Quel rôle l'ONU peut jouer en Somalie ?

Compléments : Les documents pertinents devront mettre en lumière les oppositions de clan et de chefs dans la poursuite des combats et ne devront pas ignorer que le comportement de l'ONU en Somalie n'obéit pas à une cohérence marquée.

Concepts: Etats-Unis, France, Négociations, Désarmement, Aide humanitaire, Guerilla.

discussion conférence réunion administration communauté nation représentant position conflit territoire paix solution

.SN

guerre_civile

combat_en_somalie

aide_humanitaire

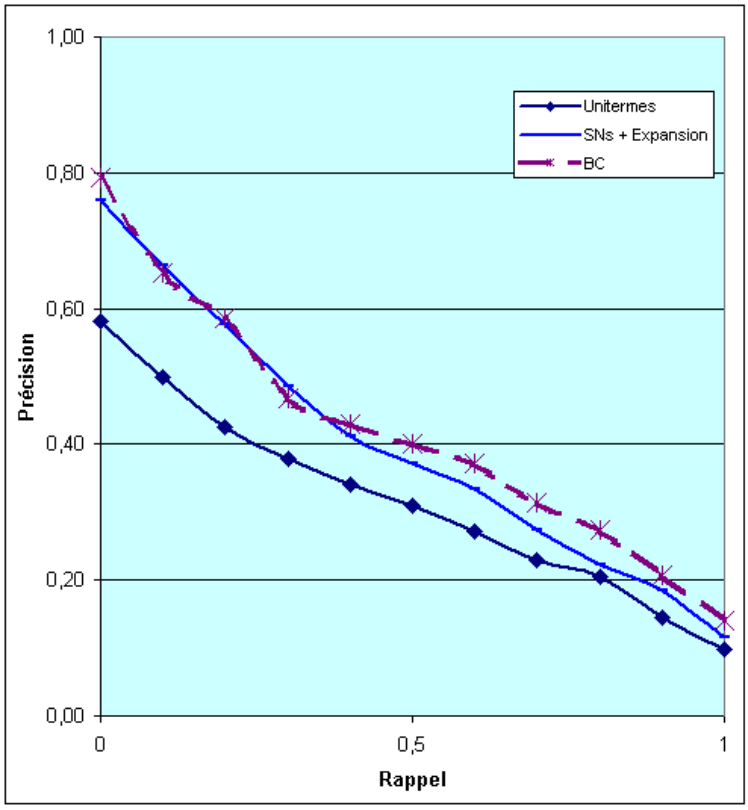

FIG. 9.6 – *Courbes de Rappel-Précision de la collection OFIL en utilisant la base de connaissances*

sanglant_affrontement
volonté_collectif
rapprochement_régional
propre_indépendance
communauté_mondiale
force_militaire
population_somalienne
mohamed_farah
alliance_nationale_somalienne
raison de la guerre civile
origine directe de la guerre civile
protagoniste de la guerre civile

responsable de la guerre civile
mouvement de guérilla
chef de la guérilla
dirigeant de la guérilla
mohamed farah aïdid
faction de l'alliance nationale somalienne
ali mahdi mohamed
société somalienne
président ethiopien
réconciliation nationale
chef de clan
président par intérim
chef du guerre somalien
faction somalienne
conférence de la réconciliation nationale somalienne

Les performances obtenues en appliquant cette stratégie, désignée par BC (Base de Connaissances) dans les Figures 9.6 et 9.7, sur les requêtes de la collection OFIL montrent une augmentation nette des taux de rappel et de précision par rapport aux performances des stratégies précédentes comme le montre la Figure 9.6. En effet, la précision moyenne en 11 points de rappel est de $42,28\%$ avec une augmentation de $10,64\%$ par rapport à la stratégie des unitermes. La précision moyenne de la septième requête d'OFIL (La Guerre civile en Somalie) est passée à $37,99\%$ contre seulement $32,61\%$ dans le cas des unitermes. L'application de cette stratégie sur les requêtes de la collection INIST montre aussi une augmentation des taux de rappel et de précision par rapport aux performances des stratégies précédentes. Cette augmentation est moins importante que dans le cas de la collection OFIL. En effet, comme le montre la figure 9.7, la précision moyenne en 11 points de rappel en utilisant la base de connaissances est de $26,66\%$ avec une augmentation de $4,58\%$ par rapport à la stratégie des unitermes.

Cependant, comme nous l'avons souligné dans la section 8.4, une indexation avec le modèle vectoriel n'est pas adaptée à une indexation avec des syntagmes nominaux. En effet, le modèle vectoriel présente deux inconvénients majeurs qui sont le principe d'indépendance des termes d'indexation et la pondération de ces derniers que nous détaillons dans le chapitre suivant où nous proposons cette structuration pour une exploitation des connaissances dans le cadre d'une indexation avec des syntagmes nominaux.

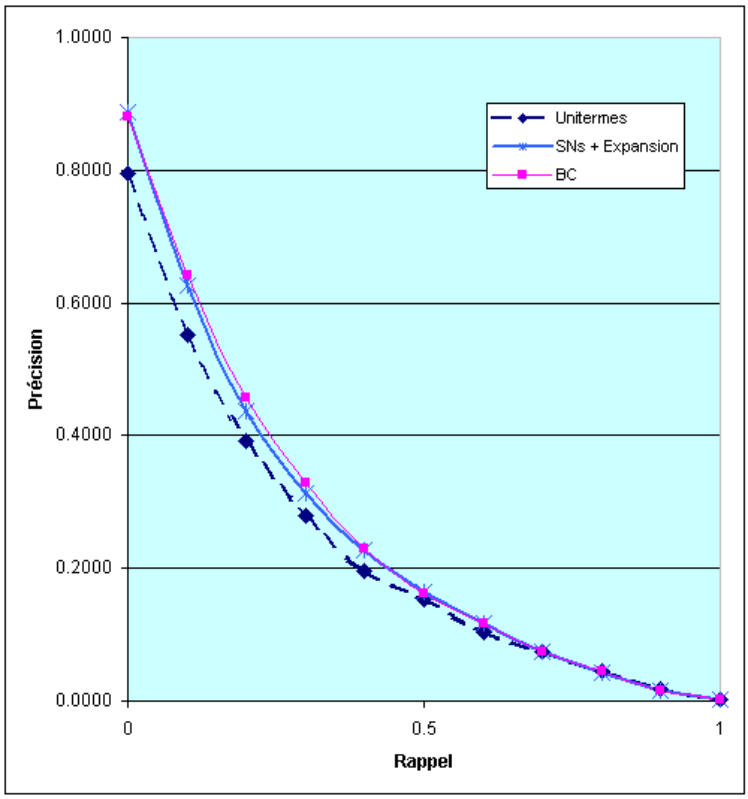

FIG. 9.7 – *Courbes de Rappel-Précision de la collection INIST en utilisant la base de connaissances*

Chapitre 10

Un modèle d'indexation relationnelle basé sur les syntagmes

Il n'y a point de génie sans un grain de folie.
ARISTOTE

Parler d'indexation sous-entend la définition d'un langage d'indexation, (permettant la représentation des concepts des documents du corpus) et un processus d'indexation, (permettant l'extraction, à partir des documents du corpus, de termes d'indexation, c'est à dire leur représentation conformément au langage d'indexation). Nous remarquons que la plupart des travaux sur l'indexation par des termes complexes sont appliqués à la langue anglaise. En effet, les travaux sur d'autres langues sont récents ce qui peut s'expliquer par la disponibilité récente de corpus, notamment de corpus français, et d'analyseurs linguistiques. Le relatif retard des francophones en la matière pouvant être considéré comme un avantage s'il permet d'éviter les écueils rencontrés par les auteurs anglo-saxons principalement au niveau de la prise en compte de la dépendance entre termes d'indexation et de la pondération de ces derniers:

– Dépendance entre termes d'indexation

  Les travaux actuels (section 7.4.2) se basent sur l'hypothèse d'une indépendance entre les termes d'indexations extraits. Cette indépendance est, par exemple, cruciale pour le bon fonctionnement du modèle vectoriel de RI. En effet, chaque terme d'indexation dans le modèle vectoriel constitue une dimension de l'espace vectoriel. Cette hypothèse ne considère pas d'éventuelles relations entre termes d'indexation. Elle est alors dans la plupart des cas une hypothèse abusive. Elle peut se justifier à la limite dans le cas de termes simples fortement tronqués, c'est à dire réduits à leur racine. Par contre, l'usage de termes composés comme termes d'indexation renforce

leur dépendance. En effet, la composition de terme produit des termes plus spécifiques. De plus, les variations et les constructions elliptiques, lorsqu'elles ne sont pas explicitement traitées, provoquent également une dépendance entre des termes et sont en fait des synonymes contextuels. La dépendance entre les termes d'indexation complexes doit alors être explicitement étudiée et intégrée dans un modèle d'indexation.

– Pondération des termes d'indexation

  Une indexation avec des syntagmes nominaux doit proposer un schéma de pondération adapté. Ce dernier doit tenir compte des caractéristiques des syntagmes nominaux principalement leurs faibles occurrences dans les documents par rapport aux unitermes. Les schémas de pondérations basées sur la fréquence des unités textuelles ne sont pas adaptés dans le contexte d'indexation avec des syntagmes nominaux.

Il est donc clair que notre objectif est de s'éloigner du signal pour se rapprocher du sens. Notre modèle doit alors manipuler des index structurés se situant le plus près possible de la *signification*. Notre objectif est de proposer un modèle qui puisse être réalisable avec la technologie actuelle. La contrainte du temps d'indexation est une contrainte forte dès lors que l'on désire indexer une masse d'information en rapport avec la production électronique actuelle de documents. La contrainte du temps d'interrogation est une contrainte encore plus forte pour la réalisation d'un SRI interactif.

Nous avons choisi de travailler dans le contexte d'une indexation non contrôlée. En effet, une indexation contrôlée est le résultat d'un long travail dont la partie la plus importante est souvent de nature terminologique. C'est une tâche évidemment coûteuse. De plus, l'explosion constante de la quantité d'information rend le maintient et la mise à jour des vocabulaires contrôlés (souvent sous la forme de thésaurus) difficile. Nous avons donc opté pour le choix d'une indexation non contrôlé et choisi de ne pas faire appel à des connaissances sémantiques prédéfinies pour être indépendant du domaine couvert par une collection.

10.1 Modèle d'indexation relationnelle syntagmatique

Nous proposons un nouveau paradigme d'indexation relatif à la notion de syntagme. Ce paradigme suppose la traduction de syntagmes extraits des corpus à indexer en terme d'index syntagmatiques possédant une sémantique ensembliste. Cette transformation nous permet de bien séparer les problèmes liés à la définition d'un modèle de recherche d'information, de celui du repérage automatique des termes d'indexation. Nous désirons donc

volontairement nous détacher du texte vu comme un signal, pour construire des index plus conceptuels.

10.1.1 Terme d'indexation syntagmatique (TIS)

Ce que nous désignons par terme d'indexation syntagmatique (TIS) est un terme du langage d'indexation obtenu par une analyse des syntagmes. Autrement dit, un terme d'indexation syntagmatique est la représentation d'au moins un syntagme du corpus. Pour une représentation des TIS en adéquation avec les besoins de la RI, nous proposons un formalisme qui met en relations les éléments lexicaux et identifie leurs positions sur l'axe syntagmatique. En effet, l'utilisation directe de l'arbre de dérivation syntaxique ne semble pas donner de résultats satisfaisants comme le montre [SOK94]. Une représentation en terme de dépendances syntaxiques (section 9.2), adoptée par la majorité des travaux en indexation par des termes complexes, est plus adéquate à notre besoin. Cette représentation décrit un TIS en terme de relations de dépendances exprimant des relations naturelles entre ses composantes. Notre modèle se base alors sur la syntaxe qui associe à chaque SN une description structurale fondée sur la notion de dépendance *tête expansion*. Un terme d'indexation syntagmatique I est alors une structure :

$$I = [T\ R_1\ [A_1]\ R_2[\ A_2]\ ...R_n\ [A_n]]$$

où T désigne la tête du terme, $A_1, A_2, ..., A_n$ désignent les expansions, et $R_1, R_2..., R_n$ qualifie la relation entre la tête et les expansions. Cette qualification de la relation est optionnelle. La tête est un atome syntagmatique. Par *atome syntagmatique* nous entendons une succession de mots possédant une signification propre et étant sujet à aucune variation autre que flexionnelle. Ces termes peuvent être des mots composés comme *hot dog* ou *pomme de terre*. Les expansions peuvent être soit un atome, soit à leur tour des termes d'indexation syntagmatiques. L'ordre entre les expansions n'est pas significatif, par contre, tous les expansions doivent être différentes. Par exemple le syntagme *séparation de la République fédérale tchèque*, peut être traduit par le TIS :

[séparation R_1 [république R_2 [fédérale] R_3 [tchèque]]]

Où la tête est *séparation* et l'expansion est *république fédérale tchèque* qui lui même est un syntagme traduit par la tête *république* et les deux expansions *fédérale* et *tchèque*.

Notre but n'est pas de construire un index à partir de tous les syntagmes mais de prendre en compte la variation dans l'expression de ces syntagmes. Cette variation permet d'exprimer un index unique sous des différentes formes syntaxiques. Ce point est directement lié aux travaux de Jacquemin [Jac97] sur la complexité morphosyntaxique terminologique où, dans le cadre d'une indexation contrôlée, la variation est utilisée pour

cerner les expansions possibles des termes vers leurs variantes. A l'inverse, nous exploitons le phénomène de la variation dans le contexte d'une indexation libre où le but est de prendre en compte la variation dans l'expression de ces syntagmes. En terme de dépendance, une variation d'un syntagme est un syntagme ayant la même tête et les mêmes expansions. Si nous reprenons les exemples de *république tchèque* et *république fédérale tchèque*, ces deux exemples sont deux variantes qui se réduisent à une constante du type $[[\text{ république }]R_1[\text{ tchèque }]]$ où *république* est la tête et *tchèque* est une expansion qui peut être imbriquée. La structuration présentée dans le chapitre 9 permet de tenir compte de cette variation.

D'autres types de variations tels que les variations flexionnelles et syntaxiques doivent aussi être prises en compte. La variation flexionnelle permet d'identifier pour chaque terme, les formes singuliers/pluriels des noms, et les formes infinitives, participes passés et gérondives des noms/verbes. Ces variations peuvent être résolues en détectant des phénomènes linguistiques particuliers.

10.1.2 Phénomènes linguistiques pour la RI

Nous examinons dans cette section les cas de variation qui nous semble prioritaire de traiter dans le cadre de la recherche d'information. Dans [Jac97] on trouve une étude détaillée sur la plupart des phénomènes linguistiques.

Le figement :

Lorsque le sens d'un groupe de mots ne peut pas être déduit du sens des mots qui le composent, nous parlons alors d'expression figée. Si on se réfère à Saussure [Sau72], le figement, appelé agglutination, est le fait de

> deux ou de plusieurs termes originairement distincts, mais qui se rencontrant fréquemment en syntaxe, au sein d'une phrase, se soudent en une unité absolue et difficilement analysable.

En linguistique, les mots figés sont désignés sous le terme de *locutions* et particulièrement de *noms composés* dans le cadre des locutions nominales. Les noms composés ont les caractéristiques suivantes [Gro96]:

- L'interdiction de paraphrasage synonymique:
  On ne peut pas substituer un mot d'une suite figée par son synonyme. On ne peut pas remplacer par exemple noir par sombre dans *caisse noire*.
- L'impossibilité d'insertion:
  On peut introduire des expansions, par exemple adjectivaux, dans les groupes nominaux mais ce n'est pas le cas quand il s'agit d'un figement. On peut par

exemple qualifier *pomme de terre* en ajoutant un adjectif avant ou après : *une bonne pomme de terre, une pomme de terre cuite* ; mais on ne peut pas dire une *pomme bonne de terre*.

– L'absence de libre actualisation des éléments composants:

  Les noms composés ont une détermination globale et les composants ne peuvent pas avoir leur propre actualisation.

– Un nom composé est une non-prédiction:

  Les noms composés sont préconstruits, tout comme les mots simples, et font parti d'un vocabulaire au contraire des groupes nominaux qui sont construits suivant les règles de la grammaire.

Sauf présence d'un trait d'union, il est très difficile à un programme informatique de décider si un groupe nominal rencontré dans un texte est un nom composé ou non d'autant plus que la typologie des noms composés est la même que les groupes nominaux. Les travaux informatiques pour repérer les suites figées dans un texte ont adopté une double démarche [Gro96]. La première démarche consiste à chercher les associations entre les mots voisins à partir de schémas de figement potentiels. La deuxième démarche se base sur la structure de surface en cernant les paramètres linguistiques qui caractérisent les mots figés tel que l'apostrophe ou le trait d'union.

Pour le repérage des noms composés, nous nous basons sur les règles d'association entre les composants d'un syntagme nominal. Si les syntagmes nominaux sont liés par des règles d'association qui ont toutes une mesure de confiance proche de 100%, ils sont alors considérés des noms composés dans la collection utilisée. Nous incluons alors dans la catégories des noms composées les syntagmes nominaux qui ne sont pas figés mais dont le comportement dans la collection est identique à celui des noms composés. L'absence de variation d'un syntagme peut être considérée comme un signe de stabilisation d'un concept et le syntagme est considéré alors figé. Les composants des syntagmes figés ne peuvent pas être séparés. Le syntagme dans sa globalité est alors un atome et peut jouer le rôle de tête ou d'expansion dans un TIS. Les noms propres sont aussi considérés figés.

Par exemple, les syntagmes *metteur en scène* et *margaret thatcher* n'ont pas de variation dans la collection et les valeurs de confiance des règles d'association entre les composants sont proches de 100%:

– metteur $\Rightarrow_{96,25\%}$ scène et scène $\Rightarrow_{92,66\%}$ metteur

– margaret $\Rightarrow_{98,36\%}$ thatcher et thatcher $\Rightarrow_{97,48\%}$ margaret

Ces syntagmes sont alors des atomes syntagmatiques. Ils sont représentés sous la forme des TIS suivants:

[[metteur en scène] R_1 [britannique]]
[[ministre] R_1 [premier] R_2 [margaret thatcher]]

L'anaphore elliptique :

Notre étude est partiellement motivée par la justesse du décomptage des syntagmes candidats pour l'indexation. Il est donc important de tenir compte des anaphores elliptiques.

L'anaphore est traditionnellement définie comme étant toute reprise d'un élément antérieur dans un texte, dans la même phrase ou dans une phrase précédente, par un souci d'économie langagière. Cet élément antérieur est également appelé antécédent. On dit alors qu'une expression est anaphorique si son interprétation référentielle dépend de cet antécédent.

Par exemple dans les phrases *La république fédérale tchèque est .... Cette république possède ...*, la seconde apparition de *république* est une référence anaphorique au syntagme initial *république fédérale tchèque*. Une anaphore peut ainsi permettre une forme d'abréviation qui permet de désigner une entité sans la nommer explicitement ni la décrire.

On distingue plusieurs types d'anaphore d'un syntagme nominal . En effet, un syntagme nominal peut être anaphorisé par :

- un autre syntagme nominal comprenant un nom identique à celui de son antécédent mais marqué par un autre déterminant (la république fédérale tchèque - Cette république), et éventuellement des expansions du nom différentes (Un soldat du régiment royal irlandais.... Le jeune soldat était rentré ...).

  C'est le cas de l'anaphore par reprise partielle.

- un pronom personnel (le montant du SMIC reste inchangé en métropole. Il augmente de 3 % dans les départements d'outre-mer), indéfini (l'un - l'autre,...) ou démonstratif (ils ont adopté un chiot Terre-neuve. Celui-ci deviendra très vite un membre à part entière de leur famille).

  C'est la cas de l'anaphore pronominale.

- synonyme, périphrase ou encore hyperonyme de son antécédent (Un joueur de foot a été contrôlé positif à la nandrolone. Ce sportif a consommé ...).

  C'est le cas de l'anaphore par lien sémantique.

Pour résoudre tous les problèmes d'anaphore dans un texte, il serait nécessaire d'analyser le contexte et d'effectuer une analyse en profondeur du texte. Or, notre étude ne tient pas en compte de ce type d'analyse vue la lourdeur qu'elle présente pour un SRI, nous ne pensons pas être en mesure, à court terme, de traiter tous les cas d'anaphore. Nous éliminons les cas de l'anaphore par lien sémantique qui demandent des connaissances paradigmatiques. Nous ne nous intéressons alors qu'à la résolution de l'anaphore pronominale et l'anaphore par reprise partielle.

La coordination :

La coordination permet de factoriser les têtes de deux syntagmes en coordonnant leurs expansions. Elle peut aussi factoriser les expansions et coordonner les têtes. Par exemple, dans la phrase *Il n'est pas vrai que l'unique voie démocratique pour la séparation de la République fédérale tchèque et slovaque ait été le référendum*. Le système extrait les syntagmes suivants :

[voie R_1 [unique]R_2 [démocratique] R_3[séparation R_4 [république R_5 [fédérale] R_6 [tchèque]] R_7 [république R_8 [fédérale][slovaque]]]

10.1.3 Qualificateurs de relations dans les syntagmes

L'utilisation des relations explicites au cours de l'indexation a été mise en évidence dans plusieurs travaux de recherche d'information dont la plupart restent théoriques [Oun98, Far80a]. En effet, l'identification automatique correcte de relations sémantiques entre unités textuelles est très difficile. Elle fait intervenir des connaissances générales et des connaissances sur le domaine. C'est pourquoi plusieurs travaux se sont orientés vers les relations syntaxiques pour substituer aux relations sémantiques [Hea92, Mor99].

Notre modèle se concentre sur les relations qui existent entre les éléments d'un syntagme. Ce sont des relations dites locales qui expriment la nature de l'attraction entre la tête du syntagme et ses expansions. Les relations doivent leurs qualifications aux rôles que jouent les expansions dans le syntagme. Dans ce qui suit nous présentons quelques exemples de relations. La qualification de la relation est réalisée après la structuration du syntagme et d'après la nature des termes utilisés.

QUAL : dans une relation de qualification, l'expansion joue un rôle de qualifieur de la tête comme c'est le cas lorsque l'expansion est un adjectif.

[république QUAL [fédérale]]

SPEC : dans le cas où l'expansion est un substantif avec absence de préposition, il s'agit d'une spécification de la tête.

[presse SPEC [papier]]

OBJ, AGT : dans le cas où l'expansion est un substantif avec présence de préposition:

- si l'expansion dénote une action comme *le vote, la pollution*. Cette dénotation peut être détectée par l'existence d'une forme verbale du substantif de tête. Dans ce cas, les qualificatifs de relations exprimant l'objet et l'agent de l'action peuvent être utilisés (AGT, OBJ). Par exemple, dans la phrase *De plus, lors du vote par l'Assemblée fédérale de la séparation de la Fédération, la coalition a obtenu ...*, le syntagme suivant est extrait :

  [vote AGT [assemblée [fédérale]] OBJ [séparation OBJ [Fédération]]]

- si l'expansion ne dénote pas une action alors il s'agit d'une spécification de la tête dans *la rubrique du magazine*.

  [rubrique SPEC [magazine]]

PART : cette relation importante exprime la relation d'holonymie. C'est la relation qui lie un composant à son objet composé.

10.1.4 Caractéristiques d'un TIS

Ce modèle très simple de représentation, va tout de même nous permettre de mettre en place une nouvelle fonction d'indexation. Il nous permet de mettre en évidence des aspects de la recherche d'informations trop rarement étudiés dans les systèmes actuels. Nous avons retenu trois caractéristiques principales que nous considérons essentielles pour une indexation syntagmatique : la pertinence d'un TIS, la qualité d'un TIS et la quantité d'information d'un TIS.

- Pertinence d'un TIS

  Le modèle d'indexation doit permettre de valuer un terme d'indexation pour exprimer l'importance qu'il a par rapport au document tout entier qu'il doit indexer. Cette valeur est le plus souvent calculée à partir de valeur fréquentielle d'occurrences des termes. On désignera cette valeur par Pertinence d'un terme indexation syntagmatique.

  > La pertinence Pert(I,D) d'un TSI I est une valeur exprimant l'importance de ce terme dans le document D.

  Cette définition tient compte des informations linguistiques mais aussi de critères statistiques ce qui reflète l'importance, au sens RI, du syntagme dans un document.

– Qualité d'un TIS

  Comme nous proposons d'utiliser des structures d'indexation complexes extraites automatiquement, cette extraction ne peut pas être considérée comme totalement fiable. En effet, l'analyse de surface que l'on propose ne tient pas compte de certains phénomènes linguistiques et ne résout pas tous les problèmes d'ambiguïté possibles. L'utilisation dans la correspondance d'un TIS doit alors tenir compte d'un facteur de fiabilité :

  > La qualité Qual(I,P,D) d'un TSI I, extrait de la phrase P d'un document D, est définie comme la probabilité que ce terme reflète effectivement l'information initialement contenue dans la phrase de ce document.

  Cette valeur doit être, dans l'absolu, estimée manuellement comme *la proportion d'experts en accord avec la construction de ce terme d'indexation*. La mesure de pertinence d'un TIS doit être comprise comme indépendante de sa qualité. Elle représente sa pertinence en supposant le terme d'excellente qualité.

– Quantité d'information d'un terme d'indexation syntagmatique

  nous avons introduit cette grandeur dans le chapitre 7.

10.1.5 Index syntagmatique

La définition d'un TIS et de ces caractéristiques nous permettent de définir un index syntagmatique (IS) comme un ensemble de termes d'indexation syntagmatiques :

> Un index syntagmatique d'un document D est un ensemble de termes d'indexations syntagmatiques, où chaque terme est muni de sa valeur de pertinence, sa valeur de qualité et sa quantité d'information.

$$IS(D) = \{(I, Pert(I, D), Qual(I, D), Qinf(I))\}$$

Nous allons maintenant détailler comment mettre en place un processus de correspondance, d'abord entre les TIS puis globalement pour l'index syntagmatique.

10.2 Fonction de correspondance entre TIS

Dans notre modèle de RI, une requête est exprimée par un ensemble de termes d'indexations syntagmatiques où la valeur de pertinence n'est pas prise en compte, car nous

considérons que, ce terme ayant été volontairement proposé par l'utilisateur dans sa requête, il ne peut être que pertinent.

> Une requête syntagmatique RS issue d'une requête en langue naturelle R est un ensemble de termes d'indexations syntagmatiques, où chaque terme est muni de sa valeur de qualité relative à la requête et sa quantité d'information.

$$RS = \{(I, Qual(I, R), Qinf(I))\}$$

La correspondance entre document et requête consiste alors à calculer le degré de correspondance entre deux ensembles de TIS. Nous proposons de baser le calcul de la correspondance entre ces deux ensembles, sur une mesure de similarité terme à terme.

Dans notre modélisation, la tête du TIS est une entité prépondérante (sections 7.5.2 et 9.2). Il semble alors important de privilégier une correspondance basée en priorité sur les têtes de ces deux termes. Les expansions, pouvant être eux mêmes des termes, le terme de la requête peut être une expansion d'un terme d'index plus complexe. Par exemple, avec le terme d'indexation suivant:

[séparation R_1 [républiqueR_2 [fédérale] R_3 [tchèque]]].

Il faut pouvoir mesurer une correspondance avec la RS :

[république R_3 [tchèque]]

ou bien avec la RS :

[république R_4 [petite] R_3 [tchèque]].

Pour solutionner ce problème de correspondance, nous proposons d'adopter la vision de Nie dans [Nie90] qui consiste à considérer le processus de calcul de la correspondance comme un processus incertain de déduction. Ce point de vue est une concrétisation de la suggestion de [Rij86] selon laquelle la correspondance entre un document est une requête devrait s'exprimer par la mesure de la force de l'implication entre l'ensemble des propositions du document et celui de la requêtes. Ainsi la correspondance, en terme de logique, est la satisfaction de l'implication $D \rightarrow Q$ où D est un document et Q est une requête. Dans la plupart des cas, cette implication ne peut pas être évaluée à vrai ou faux. C'est pour quoi une probabilité de l'implication P($D \rightarrow Q$) est calculée pour mesurer la force de l'implication.

Soient les exemples suivant :

Exemple 13 $IS(D_1) = \{([$*référendum* R_1 *[démocratique*$]]$,
$Pert([$*référendum* R_1 *[démocratique*$]], D_1)$,
$Qual([$*référendum* R_1*[démocratique*$]], R_1)$,
$Qinf([$*référendum* R_1 *[démocratique*$]]))$,
$([$*république* R_2 *[tchèque*$]]$,
$Pert([$*république* R_2 *[tchèque*$]], D_1)$,
$Qual([$*république* R_2 *[tchèque*$]], R_2)$,
$Qinf([$*république* R_2*[tchèque*$]]))\}$

$RS_1 = \{([$*référendum* R_1*[démocratique*$]]$,
$Qual([$*référendum*$[R_1$*démocratique*$]], R_1)$,
$Qinf([$*référendum* R_1*[démocratique*$]))$,
$([$*république* R_2 *[tchèque*$]]$,
$Qual([$*république* R_2 *[tchèque*$]], R_2)$,
$Qinf([$*république*R_2*[tchèque*$]]))\}$

Exemple 14 $IS(D2) = \{([$*référendum* R_1 *[démocratique*$]]$,
$Pert([$*référendum* R_1 *[démocratique*$]], D_2)$,
$Qual([$*référendum* R_1 *[démocratique*$]], R_1)$,
$Qinf([$*référendum* R_1 *[démocratique*$]]))$,
$([$*république* R_2 *[fédérale]* R_3 *[tchèque* $]]$,
$Pert([$*république* R_2 *[fédérale]* R_3 *[tchèque* $]], D_2)$,
$Qual([$*république* R_2 *[fédérale]* R_3 *[tchèque* $]], (R_2, R_3))$,
$Qinf([$*république* R_2 *[fédérale]* R_3 *[tchèque* $]])$

$RS_1 = \{([$*référendum* R_1*[démocratique*$]]$,
$Qual([$*référendum*$[R_1$*démocratique*$]], R_1)$,
$Qinf([$*référendum* R_1*[démocratique*$]))$,
$([$*république* R_2 *[tchèque*$]]$,
$Qual([$*république* R_2 *[tchèque*$]], R_2)$,
$Qinf([$*république*R_2*[tchèque*$]]))\}$

Le premier exemple est le cas où la requête syntagmatique RS1 est identique à l'index syntagmatique du document D1. Dans ce cas, l'implication $D \rightarrow Q$ est valuée à 1. Par contre, dans le deuxième exemple cette implication n'est pas sûre. En effet, la requête n'est pas exactement identique au document. La mesure de correspondance est alors transformée en une mesure de la certitude de l'implication de la requête par le document, notée $P_K(D \rightarrow Q)$ où K est l'ensemble de connaissances utilisé lors de la transformation. Quand une requête Q ne peut pas être directement dérivée du document D, alors soit le IS(D) doit être élargi soit certains TIS de IS(D) doivent être modifiés. L'ensemble des

connaissances K ne s'attache qu'aux connaissances linguistiques sur les SNs qui s'intéressent aux catégories grammaticales des composantes d'un SN, les règles syntaxiques de l'agencement de ces composantes, les interdictions et les possibilités d'insertion d'un nouvel élément ou de la suppression d'un élément existant du syntagme nominal comme indiqué dans la section 9.4.

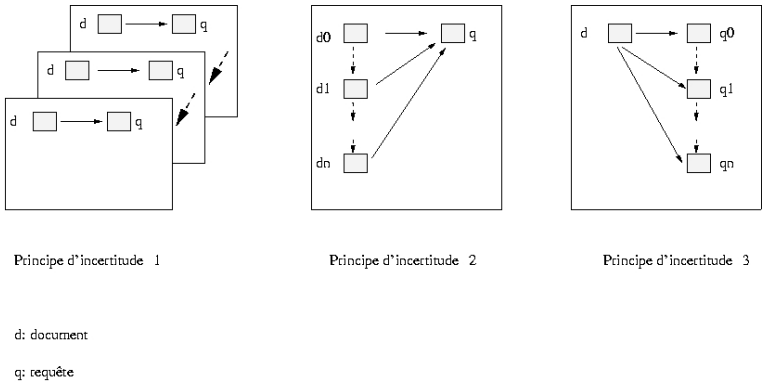

FIG. 10.1 – *Principes d'incertitude*

Nie dans [Nie90] propose dans son modèle logique modale floue général, qui s'étend à tous les types de dérivation, trois principes d'incertitude (Figure 10.1):

– dans le premier principe d'incertitude, les transformations sont effectuées sur le document et sur la requête.

– dans le deuxième principe d'incertitude, les transformations sont effectuées sur le document.

– dans le troisième principe d'incertitude, les transformations sont effectuées sur la requête.

Le troisième principe est écarté car il nécessite un méta-modèle lourd à manipuler [Nie90]. Le premier principe et le deuxième sont dual, nous nous focalisons sur le deuxième principe qui est dans notre cas plus aisé à mettre en oeuvre et dont le principe est le suivant [Nie90]:

> Etant donnés deux ensembles d'information x et y, la mesure d'incertitude de $x \rightarrow y$ relative à un certain ensemble K de connaissances donné est déterminée par l'extension nécessaire de x en x' et la certitude de la vérification de $x' \rightarrow y$.

10.3 Définition de la fonction de correspondance

Pour définir la fonction de correspondance, nous adaptons le formalisme de Nie dans [Nie90] à notre modèle d'indexation:

Soit un ensemble de termes d'indexation IS(D) qui contient les TIS du document D. Il existe un ensemble de connaissances syntaxiques K qui permettent d'obtenir un nouveau TIS à partir d'un TIS existant. La relation de dérivation entre un TIS initial I_1 et le nouveau TIS I_2 est notée $I_1 \rightarrow I_2$. Le coût de cette transformation est proportionnel à la différence de quantité d'information comme indiqué dans la section 9.4.

10.4 Pondération d'une correspondance

Nous définissons un chemin de dérivation comme étant une transformation possible d'un TIS. Dans la structure des syntagmes nominaux présentée dans le chapitre 9, un chemin de dérivation correspond à l'ensemble des liens syntaxiques qui mènent d'un syntagme à un autre syntagme. La pondération du calcul de la correspondance entre deux termes correspond au minimum de la somme des coûts de tous les chemins de dérivations possibles ce qui se traduit par la perte minimale d'information due à la transformation. Le coût d'un chemin de dérivation est défini comme la somme des coûts des dérivations élémentaires.

Nous sommes donc maintenant à même de définir une correspondance entre index et requête syntagmatique sur la base de cette correspondance de termes.

10.5 Fonction de correspondance entre index et requête

La correspondance entre un index et une requête utilise donc la correspondance entre deux TIS. La correspondance peut s'établir entre un TIS de la requête et un TIS du document, mais également entre deux TIS du document. Nous devons par cette situation tenir compte de la dépendance entre les TIS, au sein même d'un document. Globalement, la confrontation d'une requête avec un document se présente comme l'introduction dans un réseau de dépendance de TIS dans l'index, des nouveaux TIS appartenant à la requête. C'est le réseau obtenu dans sa globalité qui doit être évalué.

Nous calculons de la manière suivante, la distance d en terme de coût entre un TIS I_r de la requête et un index syntagmatique $IS(D)$, c'est à dire un ensemble de TIS I_d qui forment l'index du document D comme décrit en 10.1.5:

$$d(I_r, IS(D)) = min_{I_d \in IS(D)}(\frac{|T(SN \rightarrow SN')| * Pert(I_d, D)}{Qual(I_r, R) * Qual(I_d, D)})$$

Cette formule exprime que la distance d'un terme de la requête à l'index du document, se calcule en recherchant le coût de dérivation minimum par rapport à tous les TIS du document, en tenant compte de la qualité des deux termes et de la pertinence du TIS dans son document. Nous supposons bien sur qu'aucun terme n'a pas une qualité nulle. Finalement, la mesure de pertinence P entre un index $IS(D)$ et une requête syntagmatique RS peut se calculer comme une combinaison $\bigodot$ de la distance $d(I_r, IS(D))$ de chaque TIS de la requête :

$$P(IS(D), RS) = \frac{1}{1 + \bigodot_{I_r \in RS} d(I_r, IS(D))}$$

Cette formule assure une mesure de pertinence comprise entre 0 et 1. La combinaison $\bigodot$ est fonction du sens que l'on veut donner à la requête. Lorsque l'on veut privilégier le meilleur terme de la requête, alors on peut choisir la fonction min. De manière plus neutre on peut se contenter de la somme des distances de tous les termes de la requête.

10.6 Conclusion

Nous avons proposé un modèle général décrivant les éléments à prendre en compte lors de l'utilisation des syntagmes comme support des termes d'indexation. L'originalité de notre approche tient dans l'introduction dissociée, au niveau de la modélisation, des notions de pertinence, qualité et quantité d'information d'un terme d'indexation. La plupart des modèles de RI fondent leur calcul de correspondance, au mieux sur une vision probabiliste ou logique, et au pire sur un calcul donné a priori. Nous pensons que les notions présentées dans ce chapitre doivent être le fondement de tout SRI désirant se rapprocher du sens et s'éloigner du signal. Pour cela, nous avons associé à notre langage d'indexation une sémantique qui, bien qu'embryonnaire, nous fournit une justification pour définir la notion de coût d'une correspondance.

Le modèle proposé reste une vision théorique de ce qui doit être réalisé en pratique pour faire fonctionner un SRI à base de syntagmes. Il est en cours d'expérimentations dans notre système IOTA. Ce système est issu de divers travaux de notre équipe notamment [Ker84, CDBK86, Pal90]. L'implantation de la correspondance s'inspire des travaux sur la mise en oeuvre de la projection dans le cadre d'un SRI basé sur les graphes conceptuels décrite dans [Oun98].

Chapitre 11

Conclusion et apport

En toute chose, c'est la fin qui est essentiel.
ARISTOTE

Nous avons présenté dans cette thèse une approche de traitement de données textuelles qui associe la finesse d'analyse d'une approche linguistique à la capacité d'une approche statistique d'absorber de gros corpus. En combinant ces deux approches complémentaires, notre objectif est d'extraire des connaissances à partir du texte que nous voulons être utiles à un système de recherche d'information.

L'approche statistique se base sur la fouille de données textuelles. La technique de règles d'association permet d'extraire des connaissances relatives aux termes du corpus et qui permet de définir leur contexte d'utilisation en associant un terme à un ensemble de termes par des relations d'association.

L'approche linguistique se base sur les syntagmes nominaux que nous considérons comme les entités textuelles les plus susceptibles de représenter l'information contenue dans le texte. Elle explicite les contraintes linguistiques nécessaires à l'extraction des syntagmes nominaux et explicite les rapports syntagmatiques entre les composantes d'un syntagme nominal. Ces relations syntagmatiques sont exploitées pour la structuration des syntagmes nominaux. La mesure de la quantité d'information permet d'évaluer le pouvoir évocateur de chaque syntagme nominal, de filtrer les syntagmes nominaux et de comparer les syntagmes nominaux entre eux.

On discute ci-après les contributions de ce travail et un certain nombre de perspectives futures.

11.1 Contribution de cette recherche

La thèse présente trois principales contributions dans le domaine de la recherche d'information. Premièrement, elle démontre que la combinaison d'une approche statistique et d'une approche linguistique permet d'affiner les connaissances extraites et d'augmenter les performances d'un SRI. Elle démontre que des méthodes statistiques et linguistiques simples et faciles à implémenter peuvent être d'un grand intérêt pour le repérage et l'organisation de connaissances existantes dans les textes. En effet, elle se base sur l'analyse statistique de la distribution des termes et l'analyse linguistique superficielle du texte qui élimine la détermination de la structure linguistique profonde. Cette approche possède de nombreux avantages par rapport à d'autres méthodes traditionnelles. En effet, elle élimine le besoin de disposer de connaissances sémantiques ou paradigmatiques externes au SRI et se focalise sur l'extraction de connaissances à partir du texte. Elle permet d'appliquer des techniques qui respecte deux contraintes fortes que nous nous sommes fixées : la contrainte de temps de traitement et la contrainte de la qualité et de l'utilité des connaissances extraites essentielles pour la conception du SRI souhaité. Nous avons démontré que notre approche permet de traiter des grandes collections de texte avec des temps de traitement raisonnables et que la qualité de ces connaissances est validée qualitativement et quantitativement.

Deuxièmement, nous introduisons la mesure de la quantité d'information comme un des critères pour mesurer l'importance des termes ainsi que leur pouvoir de représenter le contenu textuel des documents. Ce critère, qui vient se joindre au critère statistique, mesure le pouvoir évocateur d'un terme et permet de comparer les termes entre eux. Ce critère permet la prise en compte des termes complexes dans le contexte de la recherche d'information. En effet, dans l'objectif d'aller vers plus de précision dans les résultats des SRIs, l'intégration des termes complexes est une condition nécessaire. Ces termes complexes, que nous avons identifiés comme étant des syntagmes nominaux, doivent être considérés dans les phases d'indexation, de correspondance et d'interrogation d'un SRI. Pour cela, les mesures statistiques ne permettent pas à elles seules de prendre en compte les caractéristiques de syntagmes nominaux. D'où l'introduction de la mesure de la quantité d'information que ce soit pour la structuration des syntagmes nominaux extraits que pour l'indexation des données textuelles.

La thèse présente aussi des contributions secondaires. Nous avons montré que l'organisation des termes sous forme d'une tête et des expansions permet d'avoir une vue générale du contenu du corpus. Cette organisation qui tient aussi en compte les règles d'association entre les entrées de la structure ainsi qu'un processus d'héritage des règles d'association permet de normaliser les connaissances extraites et de les organiser selon une structure facilement exploitable dans le cadre de la recherche d'information que ce soit à l'interrogation qu'à l'indexation ou la correspondance.

11.2 Perspectives

Dans notre travail, nous avons défini et expérimenté les éléments essentiels pour une acquisition de connaissances à partir des textes et leur intégration dans un SRI. A partir de cette base, plusieurs perspectives sont envisageables tant sur le plan théorique que sur le plan pratique. La première perspective offerte par notre travail se situe au niveau du processus d'indexation et de correspondance. Comme nous l'avons indiqué dans la section 10.6, le modèle proposé reste une vision théorique. Un système de recherche d'indexation est en cours d'expérimentations dans notre système IOTA et qui intègre des travaux de notre équipe que ce soit au niveau d'une indexation avec des termes complexes ou bien au niveau d'une correspondance basée sur les graphes conceptuels.

Le travail présenté dans cette thèse peut être complété et étendu pour différentes applications, certaines ont été déjà mentionnées dans les chapitres, mais que nous résumons ici.

Applications aux relations de thésaurus :

Un thésaurus au sens linguistique peut être vu comme un dictionnaire de concepts d'un domaine de connaissances spécifiques décrits par des termes, adapté à la recherche documentaire. Une entrée du thésaurus peut être un concept décrit par un terme ou un concept décrit par des termes équivalents ou termes quasi-synonymes. Notre méthode peut apporter un plus à cette construction en enrichissant le thésaurus de manière incrémental en complétant la description des concepts avec d'autres nouvelles formes syntaxiques.

Applications aux corpus multilingues :

L'application de notre approche dans le cadre de la recherche d'information multilingue nous semble pertinente. Cette application ne consiste pas à traduire les requêtes (Cross Language Information Retrieval) ou à étendre les étendre en utilisant un thesaurus multilingue mais à interroger un corpus multilingue pour rechercher des documents écrits dans des langues différentes à l'aide d'une unique requête [RCP99]. Cette perspective consiste à émerger dans plusieurs langues des syntagmes nominaux afin que la recherche multilingue se fasse au niveau des termes d'indexation et non plus par traduction des requêtes.

Applications à la veille stratégique sur Internet :

La veille stratégique consiste à exploiter au maximum le potentiel d'information d'Internet dans le but de répondre rapidement aux demandes d'information stratégiques. Les services de veille sont essentiels pour toutes les entreprises soucieuses

de bien connaître leur environnement commercial (concurrents, législation, innovation technologique, symposium, etc.). Pour maîtriser les flux d'information devant conduire à des prises de décisions stratégiques, les veilleurs doivent disposer d'outils efficaces, évolutifs, prenant en compte les besoins informationnels des entreprises et adaptés aux usages réels. Dans ce contexte, les entreprises ont de plus en plus recours à des entrepôts de documents [1] [MCD+01] et des applications de traitement automatique de l'information qui permettent notamment de rechercher et de filtrer l'information pertinente, de la router vers les bons destinataires, de la traduire, de détecter les "signaux faibles" dans les grandes masses d'information circulant sur l'Internet, etc. L'utilisation de notre formalisme dans ce contexte pourrait être intéressant dans le sens où il permet d'extraire des connaissances qui pourraient apporter une aide aux utilisateurs afin de converger rapidement vers une information pertinente et de disposer de la terminologie évolutive de leurs domaines.

1. en anglais *document warehouse*

Annexe A

Les relations de Farradane

D'après Farradane [Far80a, Far80b], la connaissance sur un concept commence par associer le concept à des mots et ensuite à un ou plusieurs concepts. En une deuxième étape, ce concept occurre, dans certaines situations, dans la présence d'un autre concept (concurremment avec). Dans certain cas, cette occurrence n'est pas vérifiée alors elle est occasionnelle (association temporaire). L'étape suivante est la reconnaissance des caractéristiques communes entre concepts et la distinction des concepts qui n'ont pas des caractéristiques communes.

Farradane définit un concept comme étant chaque unité concrète ou une idée abstraite, de n'importe quel niveau de complexité. Un concept est dénoté par sa désignation qui est une association de symboles. Une fois définit, ce mot peut être utilisé tel qu'il est. C'est donc une entité discrète qui correspond à une image mentale : toute pensée ou idée au moyen de laquelle l'esprit appréhende les choses ou parvient à les reconnaître. Une relation entre deux concepts existe si une implication entre eux existe dans l'esprit. L'esprit a deux mécanismes pour interconnecter les concepts qui sont l'association et la discrimination. Pour un besoin d'indexation relationnelle, Farradane identifie les concepts comme étant des substantifs ou des verbes. Les autres catégories, tel que les adjectifs ou les adverbes, ne peuvent pas être utilisées qu'avec un substantif.

- coopération ou accord (Concurrence) : ou encore juxtaposition (mentale) d'un objet avec un autre. Elle exprime aussi la relation de la forme bibliographique : exemple chimie β encyclopédie exprimée linguistiquement sous la forme : l'encyclopédie de la chimie. Elle exprime aussi la durée et les actions futures.

- équivalence.

  Elle exprime un certain degré de l'équivalence jusqu'à l'équivalence complète (les cas de synonymie). Elle peut être utilisée pour l'introduction des noms propres. Elle

exprime aussi l'idée qu'un objet "peut être considéré comme" ou "utilisé comme ", "autre chose", exemple sodium /= solvent.

– clarté ou netteté (distinctness).

  Elle exprime la relation d'imitation ou substitution; exemple recherche d'information /) modèle, mathématique.

– auto-activité (self-activity) ou activité propre.

  Elle exprime les actions exemple homme /* marche.

– dimensionnelle.

  Elle exprime une position dans l'espace ou le temps, des états temporaire et certaines propriétés temporaires. L'espace concerne seulement les positions relatives exemple : au-dessus /+ table, et les positions actuelles exemple industriel + angleterre.

– action.

  elle est utilisée pour chaque objet agissant ou affectant un autre objet exemple eau /- purifier.

– association.

  Elle exprime différentes formes d'association qui ne sont pas spécifiées tel que prison /; disgrâce, ou la relation "un agent de" ou "un outil de ". Elle peut être utilisée aussi pour des propriétés abstraites tel que image/; beauté ou bien pour des relations indirectes ou non calculées (imposées par l'esprit humain) tel que machine /; efficacité. Elle peut exprimer aussi des actions passées.

– appartenance.

  Elle exprime la relation d'appartenance tel que table /(pied, ou l'organe d'un corps, ou un ingrédient intrinsèque exemple thé /(caféine. Elle exprime aussi la relation générique : générique /(spécifique. Elle exprime aussi propriétés physiques intrinsèques d'un objet exemple : métal /(densité.

– dépendance fonctionnelle.

  Elle exprime le fait qu'un objet cause ou produit quelque chose exemple auteur/ : livre. Elle exprime "cause et effet" spécialement appliquée dans l'indexation des réactions chimiques.

	Mécanisme Associatif		
Conceptualisation	Conscience	Association Temporaire	Association Fixe
Simultanée	Juxtaposition	Activité propre	Association
Concepts non-distincts	Équivalence	Dimensionnelle	Appartenance
Concepts distincts	Substitution	Action	Dépendance fonctionnelle

TAB. A.1 – *Les catégories des relations de Farradane*

Annexe B

Mesures de similarité

L'évaluation de la similarité est utilisée pour une large de variété de traitements [RL98]:

- l'identification de structures cachées et pour la prédiction en analyse de données textuelles (ADT). On utilise généralement des profils lexicaux et la distance du "chi-deux χ^2".

- l'évaluation des similarités entre documents et requêtes en recherche documentaire (RD). On utilise généralement des distributions de mots-clés ou de vecteurs contextuels de co-occurrences et des similarités dérivées de mesures à base de cosinus.

- produire des représentations synthétiques de vaste collections de documents en fouille de données textuelles. On utilise généralement des distributions de mots-clés ou de vecteurs contextuels de co-occurrences et des similarités issues de la théorie de l'information tel que *la distance de Kullback-leiber à base d'entropie relative.*

Pour le calcul de similarités, les textes sont représentés en unités linguistiques (mots). Ils sont d'abord décomposés en unités lexicales. Des traitements additionnels tel que l'étiquetage morpho-syntaxique, la lemmatisation et la détection automatique des séquences répétitives permettent d'intégrer des connaissances linguistiques plus sophistiquées.

Un document est représenté sous la forme d'un tuple de valeurs D.

B.0.1 Distance de *chi-deux* χ^2 *[SHP95]*

Chaque document est représenté par un profil lexical : un tuple D_i contient les fréquences des unités textuelles dans le document ($D_i = (f_{i.j})_j$ où $f_{i.j}$ est la fréquence de la

j^{me} unité dans le document D_i). Formellement, la distance de chi-deux χ^2 est la suivante:

$$d_{\chi^2}(D_i, D_{i'})^2 = \sum_j \frac{1}{f_{.j}}[\frac{f_{i,j}}{f_{i.}} - \frac{f_{i',j}}{f_{i'.}}]^2$$

où $f_{i.} = \sum_j f_{i,j}$ et $f_{.j} = \sum_i f_{i,j}$.

Une pondération peut être utilisée dans la représentation des documents : $D_i = (w_{i,j})$ avec $w_{i,j} = \frac{p_{i,j}}{\sqrt{f_{.j}}}$ où $p_{i,j}$ est la fréquence relative de la j^{me} unité dans le document $D_i(p_{i,j} = \frac{f_{i,j}}{f_{i.}})$ et $d_{\chi^2}(D_i, D_{i'}) = |D_i - D'_i|$.

Les propriétés de la distance de chi-deux sont les suivantes :

- deux textes ayant le même profil lexical pourront être indifféremment considérés comme une seule entité ou deux entité distinctes sans que cela n'affecte en rien les autres distances. La distance de chi-deux vérifie l'équivalence distributionnelle.

- c'est une mesure de proximité sensible aux "différences hors intersection". Cette dernière joue un rôle important dans le calcul de la valeur de dissimilarité par conséquence la distance de chi-deux est peu adaptée aux situations où les tailles des entités textuelles comparées sont fortement différentes.

B.0.2 Similarité à base de cosinus

Les métriques à base de cosinus sont utilisées dans le cadre du modèle vectoriel classique. Différentes variations de cette approche sont utilisées dans le système SMART [Sal71] tel que l'atn et l'atc.

Un document D_i est représenté comme suit : $D_i = (w_{i,k})_k$ avec

$$w_{i,k} = 0.5(\frac{1 + p_{i,k}}{max_l(p_{i,l})}).log(\frac{N}{n_k})$$

si $p_{i,k} \neq 0$ sinon $w_{i,k} = 0$.

$w_{i,k}$ est le poids du terme T_k dans le document D_i, $p_{i,k}$ est la fréquence relative de T_k dans D_i, N est le nombre total de documents dans la base documentaire et n_k le nombre de documents contenant le terme T_k. On a donc:

- dissimilarité SMART atn : $atn(D_i, D_j) = D_i \bullet D_j$ où $\bullet$ représente le produit scalaire.

- dissimilarité SMART atc : $atc(D_i, D_j) = cos(D_i, D_j)$.

la dissimilarité atn n'est sensible qu'aux parties communes des entités textuelles comparées c'est à dire les parties partagées par les profils lexicaux. La dissimilarité atn est sensible au nombre de mots communs entre les documents comparés alors que la dissimilarité atc est sensible à la "proportion" de mots communs entre les documents comparés. La similarité à base de cosinus est adapté à la recherche documentaire.

B.0.3 La distance de Kullback-leibler ou la mesure d'entropie relative

Pour l'exploration de grandes collections de document (fouille de données textuelles), les distributions de co-occurrences de mots-clés sont souvent utilisées [FD95]. Une mesure est nécessaire pour quantifier le degré d'intérêt d'une distribution observée par rapport à un modèle donné. Une mesure souvent utilisée est la mesure de l'entropie relative ou distance de Kullback-leibler qui quantifie le degré de "surprise" associé à l'observation d'une distribution p alors qu'une distribution q était attendue:

$$KL_0(D_i, D_j) = \sum_{k|p_{i,k}.p_{j,k}\neq 0} p_{i,k}.log(\frac{p_{i,k}}{p_{j,k}})$$

.

Une version symétrisée de la distance de Kullback-leibler pour faciliter la comparaison avec les autres mesures de similarité est la suivante:

$$KL(D_i, D_j) = \sum_{k|p_{i,k}.p_{j,k}\neq 0} ((p_{i,k} - p_{j,k}).(log(p_{i,k}) - log(p_{j,k})))$$

. La distance KL peut être utilisée pour comparer des représentations de tailles sensiblement différentes.

B.0.4 Coefficient de cohérence

Salton a proposé un procédé d'extraction limité aux expressions de deux mots et se base sur la cooccurrence des termes utilisant un coefficient de cohérence (coh) qui représente la proportion des cas de cooccurrence de deux termes T_i et T_j selon la formule suivante:

$$coh = a * \frac{pf_{ij}^2}{f_i f_j}$$

où

- pf_{ij} est le nombre de fois où les termes T_i et T_j cooccurrent.

– f_i et f_j représentent les fréquences respectives de T_i et de T_j

– a est une constante liée à la taille du corpus.

Les mesures de similarités présentés sont toutes définies en termes de fréquences relatives et non pas en termes de fréquences absolues. Les mesures de chi-deux et de SMART sont associées à une pondération des pondération des dimensions de l'espace de représentation (par le biais du facteur $log\frac{N}{n_k}$ pour les dissimilarités en RI et le facteur $\frac{1}{\sqrt{f_{.j}}}$ pour la distance de chi-deux en analyse de données textuelles.). Ces. pondérations sont des procédures de normalisation dont l'objectif est d'intégrer dans l'évaluation des similarités, la notion de "pouvoir de discrimination" sélectivement associé avec les différentes dimensions de l'espace de représentation.

Annexe C

Algorithme APRIORI [AS94]

L'algorithme APRIORI est un algorithme itératif de recherche des itemsets fréquents par niveau : durant le k^{me} itération, un ensemble d'itemsets candidats de taille k est généré et un balayage de contexte est réalisé afin de supprimer les candidats infréquents. L'ensemble des k-items fréquents ainsi généré est utilisé lors de l'itération $k+1$ suivante pour générer les candidats de taille $k+1$.

Soient T l'ensemble des transactions, n-itemsets est l'ensemble des itemsets de taille n noté Ln et Ck l'ensemble des candidats k-itemsets (potentiellement les ensembles les plus larges). Lk est l'ensemble des candidats avec un support supérieur au support minimum. L'algorithme APRIORI est le suivant:

L1 = { 1-itemsets fréquents } ; // c'est l'ensemble des itemsets de taille 1
for (k=2 ; Lk-1 $\neq$ Ø ; k++) do
begin
Ck = apriori-gen(Lk-1) ; // génération des candidats
pour toutes les transactions t $\in$ T do
begin Ct = subset(Ck, t) ; // candidats contenus dans t
pour tous les candidats c $\in$ Ct do
c.count ++ ; // calcul du support des nouveaux candidats end
Lk = { c$\in$ Ck | c.count $\geq$ suppot minimum} ;
end
retourne $\bigcup_k$ Lk ;

La fonction apriori-gen prend Lk-1 comme argument et retourne un ensemble composé de tous les ensembles de tailles k-itemsets possibles.

apriori-gen() {
// génération des candidats
insert into Ck
select p.item1, p.item2, ..., p.itemk-1, q.itemk-1

from Lk-1 p, Lk-1 q
where p.item1 = q.item1, ..., p.itemk-2 = q.itemk-2, p.itemk-1 ¡q.itemk-1
// élimination des c $\in CK$ si c n'appartient pas à Lk-1
forall itemsets c $\in$ Ck do
forall (k-1)-itemsets s $\in$ c do
if (s n'appartient pas à Lk-1) then
delete c from Ck;
Soit l'ensemble de données suivant:

TID	A	B	C	D	E
T1	1	1	1	0	0
T2	1	1	1	1	1
T3	1	0	1	1	0
T4	1	0	1	1	1
T5	1	1	1	1	0

TAB. C.1 – *La base des transactions*

Pour l'exemple, le support minimum est fixé à 40%. L'application de l'algorithme APRIORI nous donne les résultats suivants à chaque passage:
Passage 1:

Itemset X	support(X)
A	100%
B	60%
C	100%
D	80%
E	40%

TAB. C.2 – *L1*

Passage 2:
L'ensemble $\{$B, E$\}$ a été éliminé parce que son support 20% est inférieur au support minimum.
Passage 3:
L'ensemble $\{$A, B, E$\}$ a été éliminé parce que son support est inférieur au support minimum.
Passage 4:
L'union de L1, L2, L3 et L4 forme l'ensemble des items fréquents.

Itemset X	spport(X)
A,B	60%
A,C	100%
A,D	80%
A,E	40%
B,C	60%
B,D	40%
C,D	80%
C,E	40%
D,E	40%

TAB. C.3 – *L2*

Itemset X	spport(X)
A,B,C	60%
A,B,D	40%
A,C,D	80%
A,C,E	40%
A,D,E	40%
B,C,D	40%
C,D,E	40%

TAB. C.4 – *L3*

Itemset X	spport(X)
A,B,C,D	40%
A,C,D,E	40%

TAB. C.5 – *L4*

Annexe D

Catégories grammaticales

Les catégories grammaticale utilisées sont les suivantes:

- SUBC : Substantif commun
- SUBP : Substantif propre
- ADJQ : Adjectif qualificatif
- ADJC : Adjectif numeral cardinal
- ADJO : Adjectif numeral ordinal
- ADJP : Adjectif possessif
- ADJR : Adjectif relatif
- ADJD : Adjectif démonstratif
- ADJI : Adjectif indéfini
- ADJE : Adjectif interrogatif/exclamatif
- ARTD : Article défini
- ARTI : Article indéfini
- ARTC : Article contracte
- PRPS : Pronom personnel sujet
- PRPV : Pronom personnel préverbal

- PRPC : Pronom personnel complément
- PADV : Pronom personnel adverbial
- PRPO : Pronom possessif
- PRDM : Pronom démonstratif
- PRRL : Pronom relatif
- PRIN : Pronom indéfini
- PREP : Préposition
- CONJ : Conjonction
- ADVB : Adverbe
- XECJ : Auxiliaire être conjugue
- XEIF : Auxiliaire être a l'infinitif
- XEPA : Auxiliaire être au participe passe
- XEPR : Auxiliaire être au participe présent
- XACJ : Auxiliaire avoir conjugué
- XAIF : Auxiliaire avoir a l'infinitif
- XAPA : Auxiliaire avoir au participe passe
- XAPR : Auxiliaire avoir au participe présent
- VBCJ : Verbe conjugué
- VBIF : Verbe a l'infinitif
- VBPA : Verbe au participe passé
- VBPR : Verbe au participe présent
- ADVN : Adverbe de négation "ne"
- ADVP : Adverbe de négation "pas"
- LOCP : Locution prépositive

- LOCC : Locution conjonctive
- INTJ : Interjection
- NOMB : Nombre
- PNTF : Ponctuation forte
- PARE : Parenthèses
- OPER : Opérateur
- VIRG : virgule
- DPNT : Deux-points
- ABRV : Abréviation et sigle
- EFCP : Elément de Forme Composée

Annexe E

Exemples de patrons syntaxiques

E.1 Syntagme nominal de longueur 2

Les syntagmes nominaux de longueur 2 sont formés de deux unités lexicales pleines séparées par un blanc.

– SUBC SUBC:

  Ce patron correspond à une succession de substantifs comme *demi antenne, dimension zéro, début mars*.

– SUBC ADJQ :

  Ce patron syntaxique est très fréquents dans les corpus français. Il correspond à un Substantif suivit d'un adjectif tel que *défaillance cardiocirculatoire, résidu médicamental*, etc .

– ADJQ SUBC:

  Ce patron correspond à un adjectif suivit d'un substantif tel que *simple signification, double titre, véritable monopole, large rectoplastie*, etc.

– SUBP SUBP:

  Ce patron correspond à une succession de deux substantifs propres qui correspondent à des noms de personnes ou des entités nommés tel que *jrgen habermas, andré siegfried, philippe morris, general motors*, etc.

– SUBC VBPA :

  Ce patrons correspond à un Substantif suivit d'un verbe au participe passé tel que *répression accrue, tableau recouvert*, etc.

E.2 Syntagme nominal de longueur 3

Les patrons syntaxiques de taille 3 sont construits avec trois unités lexicales pleine ou deux unités lexicales pleines et une préposition.

– SUBC PREP SUBC:

  Ce patron correspond à une succession de deux substantifs avec une préposition entre les deux tel que *constatation de métastase, contrôle par thermométrie*, etc.

– SUBC ADJQ SUBC:

  Ce patron correspond à une succession de deux substantifs avec un adjectif entre les deux tel que *missile français mistral, président français mittérand, voiture japonaise tout-terrain,* etc.

– SUBP SUBP SUBP:

  Ce patron correspond à une succession de trois substantifs propres tel que *josé patrocinio gonzales, Valéry Giscard d'Estaing, japan economic news, charles robert darwin*, etc.

– SUBC SUBC ADJQ:

  Ce patron correspond à une succession de deux substantifs suivies d'un adjectif tel que *agénésie péricardique unilatéral*, etc.

– SUBC SUBC SUBC:

  Ce patron correspond à une succession de trois substantifs tel que *rapport air combustible, commission ad hoc, rapport sine die, compétition off shore, assurance multirisque habitation*, etc.

Annexe F

Règles d'association des SNs relatifs au terme *système* dans la collection OFIL

Dans cet annexe, nous illustrons des exemples de règles d'association relatives aux SNs contenant le terme *système* dont les fréquences documentaires sont supérieures à 1. Chaque ligne constitue une règles d'association où sont présentés dans l'ordre : la confiance de la règle, le support en fréquence documentaire, la prémisse de la règle et enfin la conclusion de la règles.

- 100% 6 monnaie_britannique système_monétaire
- 100% 3 système_multilatéral économie_mondial
- 100% 3 système_hospitalier santé_public
- 100% 3 système_clérical eugen_drewermann
- 100% 3 système_audiovisuel edouard_balladur
- 100% 3 système_audiovisuel conseil_supérieur
- 100% 3 commission_éducation système_éducatif
- 100% 2 élan_populaire système_français
- 100% 2 zone_intertropical système_lymphatique
- 100% 2 yasushi_mieno système_financier
- 100% 2 violation_systématique nation_uni
- 100% 2 tuméfaction_ganglionnaire système_lymphatique

- 100% 2 transport_unique système_informatique
- 100% 2 transition_révolutionnaire système_rentier_bureaucratique
- 100% 2 transition_révolutionnaire système_rentier
- 100% 2 titre_de_transport système_offrant
- 100% 2 titre_de_transport système_informatique
- 100% 2 thimothée_malendoma censure_systématique
- 100% 2 système_rentier_bureaucratique transition_révolutionnaire
- 100% 2 système_rentier_bureaucratique système_rentier
- 100% 2 système_rentier_bureaucratique garde_socialiste
- 100% 2 système_rentier_bureaucratique dernier_congrès
- 100% 2 système_rentier_bureaucratique conseil_exécutif
- 100% 2 système_rentier transition_révolutionnaire
- 100% 2 système_rentier système_rentier_bureaucratique
- 100% 2 système_rentier garde_socialiste
- 100% 2 système_rentier dernier_congrès
- 100% 2 système_rentier conseil_exécutif
- 100% 2 système_pénal procédure_pénal
- 100% 2 système_politique_français vie_politique
- 100% 2 système_politique_français système_politique
- 100% 2 système_mutualisé partenaire_social
- 100% 2 système_mutualisé mme_martine_aubry
- 100% 2 système_mutualisé compensation_salarial
- 100% 2 système_multimédia éducation_nationale
- 100% 2 système_maffieux campagne_électoral

- 100% 2 système_lymphatique zone_intertropical
- 100% 2 système_lymphatique vaccin_expérimental
- 100% 2 système_lymphatique tuméfaction_ganglionnaire
- 100% 2 système_lymphatique processus_cancéreux
- 100% 2 système_lymphatique principal_danger
- 100% 2 système_lymphatique pays_industrialisés
- 100% 2 système_lymphatique pays_concernés
- 100% 2 système_lymphatique pays_africain
- 100% 2 système_lymphatique mononucléose_infectieux
- 100% 2 système_lymphatique infection_viral
- 100% 2 système_lymphatique immunité_durable
- 100% 2 système_lymphatique cancer_primitif
- 100% 2 système_lymphatique anomalie_sanguin
- 100% 2 système_lymphatique angine_sévère
- 100% 2 système_lymphatique affection_fébrile
- 100% 2 système_en_place vie_économique
- 100% 2 système_de_retraite sécurité_social
- 100% 2 système_de_retraite remise_en_cause
- 100% 2 système_de_financement collectivité_local
- 100% 2 système_allemand modèle_européen
- 100% 2 stock_commencé système_financier
- 100% 2 solvabilité_supérieur système_financier
- 100% 2 réduction_commercial système_informatique
- 100% 2 récession_antérieur système_financier

- 100% 2 renseignement_complémentaire système_informatique
- 100% 2 rapatriement_systématique secrétaire_général
- 100% 2 rapatriement_systématique réfugié_haïtiens
- 100% 2 rapatriement_systématique président_en_exil
- 100% 2 rapatriement_systématique nation_uni
- 100% 2 rapatriement_systématique information_transmises
- 100% 2 rapatriement_systématique etats_américain
- 100% 2 rapatriement_systématique dante_caputo
- 100% 2 rapatriement_systématique autorités_américain
- 100% 2 puissant_système système_informatique
- 100% 2 processus_cancéreux système_lymphatique
- 100% 2 performance_scolaire système_éducatif
- 100% 2 offre_de_soins système_hospitalier
- 100% 2 nébuleuse_urbain système_financier
- 100% 2 numéro_appelé système_informatique
- 100% 2 mutation_en_profondeur système_éducatif
- 100% 2 minutage_précis système_informatique
- 100% 2 marc_oraison système_clérical
- 100% 2 inadéquation_entre_formation système_éducatif
- 100% 2 immunité_durable système_lymphatique
- 100% 2 hétérogénéité_croissant système_universitaire
- 100% 2 global_positionning positionning_system
- 100% 2 garde_socialiste système_rentier_bureaucratique
- 100% 2 garde_socialiste système_rentier

- 100% 2 français_langue_étranger système_éducatif
- 100% 2 français_langue système_éducatif
- 100% 2 filière_court système_éducatif
- 100% 2 filière_confondues système_éducatif
- 100% 2 femme_médecin viol_systématique
- 100% 2 exploitation_systématique milliard_de_dollar
- 100% 2 détention_systématique procédure_pénal
- 100% 2 détention_systématique mises_en_examen
- 100% 2 détention_systématique mise_en_détention
- 100% 2 détention_systématique magistrat_instructeur
- 100% 2 détention_systématique avocat_pénalistes
- 100% 2 détention_systématique association_de_magistrat
- 100% 2 désistement_systématique élection_législatif
- 100% 2 désistement_systématique secrétaire_général
- 100% 2 désistement_systématique françois_bayrou
- 100% 2 désistement_systématique assemblée_nationale
- 100% 2 désistement_systématique alain_juppé
- 100% 2 défaillance_majeur système_bancaire
- 100% 2 correction_nécessaire système_économique
- 100% 2 contrat_de_facilities système_informatique
- 100% 2 condition_international système_politique
- 100% 2 collège_unique système_éducatif
- 100% 2 clergé_allemand système_clérical
- 100% 2 city_banks système_financier

- 100% 2 circonspect_gouverneur système_financier
- 100% 2 centre_éducatif système_éducatif
- 100% 2 censure_systématique thimothée_malendoma
- 100% 2 cellule_souche système_immunitaire
- 100% 2 caractère_académique système_éducatif
- 100% 2 caractère_académique système_scolaire
- 100% 2 cancer_primitif système_lymphatique
- 100% 2 avocat_pénalistes détention_systématique
- 100% 2 antoine_prost système_éducatif
- 100% 2 anomalie_sanguin système_lymphatique
- 100% 2 angine_sévère système_lymphatique
- 100% 2 affection_fébrile système_lymphatique
- 75%3 écoute_administratif système_informatique
- 75%3 scolarité_obligatoire système_éducatif
- 75%3 scolarité_obligatoire système_scolaire
- 75%3 formation_technologique système_éducatif
- 75%3 formation_technologique système_scolaire
- 75%3 facilities_management système_informatique
- 66.66% 2 vaccin_expérimental système_lymphatique
- 66.66% 2 trentaine_de_ville système_éducatif
- 66.66% 2 système_universitaire hétérogénéité_croissant
- 66.66% 2 système_universitaire enseignement_supérieur
- 66.66% 2 système_productif cohésion_social
- 66.66% 2 système_multilatéral pays_industrialisés

- 66.66% 2 système_multilatéral institution_international
- 66.66% 2 système_multilatéral comité_monétaire
- 66.66% 2 système_multilatéral bloc_commercial
- 66.66% 2 système_hospitalier service_public
- 66.66% 2 système_hospitalier secteur_public
- 66.66% 2 système_hospitalier personne_âgé
- 66.66% 2 système_hospitalier offre_de_soins
- 66.66% 2 système_hospitalier bruno_durieux
- 66.66% 2 système_global dépenses_public
- 66.66% 2 système_fédéral maison_blanc
- 66.66% 2 système_fédéral déficit_budgétaire
- 66.66% 2 système_clérical marc_oraison
- 66.66% 2 système_clérical littérature_français
- 66.66% 2 système_clérical francis_jammes
- 66.66% 2 système_clérical clergé_allemand
- 66.66% 2 système_clérical abondant_littérature
- 66.66% 2 système_audiovisuel nicolas_sarkozy
- 66.66% 2 système_audiovisuel michel_péricard
- 66.66% 2 système_audiovisuel chaînes_public
- 66.66% 2 système_audiovisuel chaînes_privé
- 66.66% 2 système_audiovisuel ancien_ministre
- 66.66% 2 syndicat_ouvrier système_économique
- 66.66% 2 spéculation_boursier système_financier
- 66.66% 2 situation_pathologique système_nerveux

- 66.66% 2 schéma_régional système_éducatif
- 66.66% 2 pratique_systématique tadeusz_mazowiecki
- 66.66% 2 politique_de_prix système_informatique
- 66.66% 2 méthode_pédagogique système_éducatif
- 66.66% 2 méthode_pédagogique système_scolaire
- 66.66% 2 mononucléose_infectieux système_lymphatique
- 66.66% 2 moelle_osseux système_immunitaire
- 66.66% 2 matière_scolaire système_éducatif
- 66.66% 2 langue_local système_scolaire
- 66.66% 2 inflation_faible système_financier
- 66.66% 2 infection_viral système_lymphatique
- 66.66% 2 idéologie_national système_politique
- 66.66% 2 génération_sacrifié système_éducatif
- 66.66% 2 futur_organisation système_original
- 66.66% 2 dépistage_systématique dépistage_obligatoire
- 66.66% 2 disparité_géographique système_éducatif
- 66.66% 2 caractère_systématique échéance_électoral
- 66.66% 2 caractère_systématique conseil_supérieur
- 66.66% 2 billeteries_automatique système_informatique
- 66.66% 2 agent_infectieux système_nerveux
- 60% 3 yves_gilleron système_informatique
- 60% 3 système_démocratique front_national
- 60% 3 formation_général système_éducatif
- 60% 3 eugen_drewermann système_clérical

- 60% 3 alain_savary système_éducatif
- 57.14% 4 institution_scolaire système_éducatif
- 57.14% 4 filière_technologique système_éducatif
- 50% 3 paul_barril système_informatique
- 50% 3 nom_de_code système_informatique
- 50% 2 zhu_rongji système_politique
- 50% 2 véritable_partenaire système_éducatif
- 50% 2 véritable_partenaire système_scolaire
- 50% 2 thérapie_génique système_immunitaire
- 50% 2 système_socialiste campagne_électoral
- 50% 2 système_national ministère_français
- 50% 2 système_national action_humanitaire
- 50% 2 système_médiatique vie_privé
- 50% 2 système_médiatique société_démocratique
- 50% 2 système_médiatique réflexion_collectif
- 50% 2 système_de_soins société_français
- 50% 2 système_de_soins santé_public
- 50% 2 suffrage_direct système_électoral
- 50% 2 société_traditionnel système_éducatif
- 50% 2 situation_global système_financier
- 50% 2 président_en_exil rapatriement_systématique
- 50% 2 principal_société système_informatique
- 50% 2 politique_tarifaire système_informatique
- 50% 2 pays_capitaliste système_communiste

- 50% 2 note_manuscrit système_informatique
- 50% 2 modèle_européen système_allemand
- 50% 2 francis_jammes système_clérical
- 50% 2 filière_scientifique système_éducatif
- 50% 2 filière_scientifique système_scolaire
- 50% 2 faux_passeport système_mafieux
- 50% 2 décentralisation_engagé système_éducatif
- 50% 2 drame_humain viol_systématique
- 50% 2 cursus_scolaire système_scolaire
- 50% 2 cultures_local système_scolaire
- 50% 2 compétences_acquises système_éducatif
- 50% 2 campagne_tranquille système_mafieux
- 50% 2 bloc_commercial système_multilatéral
- 50% 2 besoins_local système_éducatif
- 50% 2 avocat_de_pari système_judiciaire
- 50% 2 association_de_magistrat détention_systématique
- 45.45% 5 éducation_prioritaire système_éducatif
- 44.44% 4 système_majoritaire scrutin_majoritaire
- 44.44% 4 système_majoritaire assemblée_nationale
- 42.85% 3 système_mafieux élection_législatif
- 42.85% 3 système_mafieux pouvoir_socialiste
- 42.85% 3 système_mafieux front_national
- 42.85% 3 système_mafieux charles_pasqua
- 42.85% 3 système_mafieux ancien_ministre

- 42.85% 3 filière_professionnel système_éducatif
- 42.85% 3 filière_professionnel système_scolaire
- 40% 2 système_social sécurité_social
- 40% 2 système_immunitaire thérapie_génique
- 40% 2 système_immunitaire système_nerveux
- 40% 2 système_immunitaire obstacle_majeur
- 40% 2 système_immunitaire moelle_osseux
- 40% 2 système_immunitaire mise_en_oeuvre
- 40% 2 système_immunitaire cellule_souche
- 40% 2 système_démocratique élection_législatif
- 40% 2 système_démocratique assemblée_nationale
- 40% 2 système_de_contrôle opinion_public
- 40% 2 système_de_contrôle croissance_économique
- 40% 2 système_de_contrôle corps_social
- 40% 2 réflexion_collectif système_médiatique
- 40% 2 reste_limité système_financier
- 40% 2 principal_danger système_lymphatique
- 40% 2 politique_communautaire système_européen
- 40% 2 pertes_enregistrées système_informatique
- 40% 2 nouveau_identité système_proportionnel
- 40% 2 mouvement_référendaire système_politique
- 40% 2 historien_américain système_politique
- 40% 2 formation_linguistique système_éducatif
- 40% 2 dépistage_obligatoire dépistage_systématique

- 40% 2 croissance_exceptionnel système_financier
- 40% 2 compensation_salarial système_mutualisé
- 40% 2 chung_ju système_financier
- 40% 2 cellule_nerveux système_nerveux
- 40% 2 alain_savary système_scolaire
- 38.46% 5 corps_enseignant système_éducatif
- 37.5% 3 système_nerveux biologie_moléculaire
- 37.5% 3 système_judiciaire procédure_pénal
- 37.5% 3 système_judiciaire ministre_délégué
- 37.5% 3 femme_violées viol_systématique
- 37.5% 3 enseignement_secondaire système_éducatif
- 37.5% 3 commission_bancaire système_bancaire
- 37.5% 3 banque_américain système_bancaire
- 36.36% 4 notation_financier système_bancaire
- 33.33% 3 système_majoritaire vie_politique
- 33.33% 3 système_majoritaire système_électoral
- 33.33% 3 système_majoritaire mario_segni
- 33.33% 3 système_majoritaire front_national
- 33.33% 3 système_communiste protection_social
- 33.33% 3 soldat_serbe viol_systématique
- 33.33% 3 groupement_interministériel système_informatique
- 33.33% 3 colonel_jean système_informatique
- 33.33% 3 biologie_moléculaire système_nerveux
- 33.33% 2 système_soviétique élection_législatif

- 33.33% 2 système_soviétique élection_anticipé
- 33.33% 2 système_soviétique union_monétaire
- 33.33% 2 système_soviétique politique_étranger
- 33.33% 2 système_soviétique politique_économique
- 33.33% 2 système_soviétique intervention_télévisé
- 33.33% 2 système_soviétique guerre_civil
- 33.33% 2 système_soviétique boris_eltsine
- 33.33% 2 région_par_région système_éducatif
- 33.33% 2 réfugié_haïtiens rapatriement_systématique
- 33.33% 2 péril_vert système_politique
- 33.33% 2 province_japonais système_financier
- 33.33% 2 priorité_nationale système_éducatif
- 33.33% 2 nouveau_établissements système_éducatif
- 33.33% 2 information_transmises rapatriement_systématique
- 31.25% 5 christian_prouteau système_informatique
- 30.76% 4 services_informatique système_informatique
- 30.76% 4 langue_vivant système_éducatif
- 30% 6 enseignement_privé système_éducatif
- 30% 3 établissements_privé système_éducatif
- 30% 3 établissements_financier système_bancaire
- 30% 3 diplôme_universitaire système_éducatif
- 28.57% 30 système_monétaire politique_monétaire
- 28.57% 2 établissements_français système_français
- 28.57% 2 établissement_bancaire système_bancaire

- 28.57% 2 système_mafieux élection_présidentiel
- 28.57% 2 système_mafieux véritable_système
- 28.57% 2 système_mafieux lionel_jospin
- 28.57% 2 système_mafieux grand_dossier
- 28.57% 2 système_mafieux faux_passeport
- 28.57% 2 système_mafieux campagne_tranquille
- 28.57% 2 système_mafieux assemblée_nationale
- 28.57% 2 système_mafieux ancien_président
- 28.57% 2 système_mafieux ancien_collaborateur
- 28.57% 2 science_économique système_économique
- 28.57% 2 rénovation_pédagogique système_éducatif
- 28.57% 2 patronat_français système_éducatif
- 28.57% 2 patronat_français système_scolaire
- 28.57% 2 michel_laval système_informatique
- 28.57% 2 mario_segni système_proportionnel
- 28.57% 2 mario_segni système_politique
- 28.57% 2 intégration_économique système_économique
- 28.57% 2 institution_scolaire système_scolaire
- 28.57% 2 insertion_social système_français
- 28.57% 2 filière_technologique système_scolaire
- 28.57% 2 exclusion_social système_scolaire
- 28.57% 2 conseil_mondial système_électronique
- 28.57% 2 concours_de_recrutement système_éducatif
- 28.57% 2 bac_professionnel système_éducatif

- 28.57% 2 bac_professionnel système_scolaire
- 27.27% 3 éducation_prioritaire système_scolaire
- 27.27% 3 place_financier système_bancaire
- 25% 3 performance_économique système_économique
- 25% 3 enseignement_catholique système_éducatif
- 25% 2 technicien_supérieur système_éducatif
- 25% 2 système_nerveux système_immunitaire
- 25% 2 système_nerveux système_cardio
- 25% 2 système_nerveux situation_pathologique
- 25% 2 système_nerveux prise_en_charge
- 25% 2 système_nerveux obstacle_majeur
- 25% 2 système_nerveux espèce_humain
- 25% 2 système_nerveux cellule_nerveux
- 25% 2 système_nerveux agent_infectieux
- 25% 2 république_italien système_majoritaire
- 25% 2 richesse_nationale système_économique
- 25% 2 personne_citées système_informatique
- 25% 2 nouveau_mouvement système_rentier_bureaucratique
- 25% 2 nouveau_mouvement système_rentier
- 25% 2 mises_en_examen détention_systématique
- 25% 2 fort_pression système_financier
- 25% 2 enseignement_secondaire système_scolaire
- 25% 2 croissance_rapide système_politique
- 25% 2 condition_économique système_éducatif

- 25% 2 comité_monétaire système_économique
- 25% 2 comité_monétaire système_multilatéral
- 25% 2 ascension_social système_éducatif
- 25% 2 alain_prost système_électronique
- 25% 2 action_économique système_communiste
- 25% 2 action_diplomatique viol_systématique
- 23.8% 5 lycée_professionnel système_éducatif
- 23.07% 3 enseignement_professionnel système_éducatif
- 22.22% 2 établissements_bancaire système_bancaire
- 22.22% 2 éducation_civique système_scolaire
- 22.22% 2 école_public système_éducatif
- 22.22% 2 université_français système_français
- 22.22% 2 système_majoritaire élection_législatif
- 22.22% 2 système_majoritaire système_politique
- 22.22% 2 système_majoritaire scrutin_législatif
- 22.22% 2 système_majoritaire république_italien
- 22.22% 2 système_majoritaire représentation_nationale
- 22.22% 2 système_majoritaire proche_avenir
- 22.22% 2 système_majoritaire grand_ligne
- 22.22% 2 système_majoritaire giuliano_amato
- 22.22% 2 système_majoritaire georges_vedel
- 22.22% 2 système_communiste économie_mondial
- 22.22% 2 système_communiste union_soviétique
- 22.22% 2 système_communiste société_français

- 22.22% 2 système_communiste pouvoir_public
- 22.22% 2 système_communiste politique_économique
- 22.22% 2 système_communiste politique_monétaire
- 22.22% 2 système_communiste pays_industriel
- 22.22% 2 système_communiste pays_européen
- 22.22% 2 système_communiste pays_capitaliste
- 22.22% 2 système_communiste partenaire_social
- 22.22% 2 système_communiste intervention_public
- 22.22% 2 système_communiste grand_puissance
- 22.22% 2 pays_signataire système_original
- 22.22% 2 paradis_fiscal système_bancaire
- 22.22% 2 opposition_actuel système_européen
- 22.22% 2 indifférence_général système_scolaire
- 22.22% 2 histoire_politique système_politique
- 22.22% 2 heures_hebdomadaire système_français
- 22.22% 2 gouvernement_croate viol_systématique
- 22.22% 2 femme_musulman viol_systématique
- 22.22% 2 déficit_social système_français
- 22.22% 2 cotisation_familial système_éducatif
- 22.22% 2 canne_à_sucre système_politique
- 21.05% 4 enseignement_général système_éducatif
- 20.83% 5 insertion_professionnel système_éducatif
- 20% 7 système_politique élection_législatif
- 20% 21 système_monétaire banque_central

- 20% 2 union_patronal système_éducatif
- 20% 2 principal_collaborateur système_informatique
- 20% 2 obstacle_majeur système_nerveux
- 20% 2 obstacle_majeur système_immunitaire
- 20% 2 intervention_public système_communiste
- 20% 2 dante_caputo rapatriement_systématique
- 19.04% 20 système_monétaire politique_économique
- 18.18% 2 progrès_accompli système_éducatif
- 18.18% 2 mise_en_détention détention_systématique
- 18.18% 2 luigi_scalfaro système_politique
- 18.18% 2 intérieur_même système_éducatif
- 18.18% 2 hôtel_particulier système_informatique
- 18.18% 2 firme_français système_informatique
- 18.18% 2 domaine_social système_européen
- 18.18% 2 consultation_populaire système_politique
- 17.39% 4 système_informatique services_informatique
- 17.14% 18 système_monétaire union_monétaire
- 17.14% 18 système_monétaire taux_directeur
- 16.66% 3 réserve_fédéral système_bancaire
- 16.66% 3 heures_par_semaine système_éducatif
- 16.66% 2 travaux_pratique système_éducatif
- 16.66% 2 société_démocratique système_médiatique
- 16.66% 2 secteur_bancaire système_bancaire
- 16.66% 2 salaire_direct système_éducatif

- 16.66% 2 représentation_nationale système_majoritaire
- 16.66% 2 performance_économique seul_système
- 16.66% 2 jours_dernier système_électronique
- 16.66% 2 institution_international système_multilatéral
- 16.66% 2 georges_vedel système_majoritaire
- 16.66% 2 conseil_exécutif système_rentier_bureaucratique
- 16.66% 2 conseil_exécutif système_rentier
- 16.66% 2 collectivité_public système_économique
- 15.38% 2 véritable_révolution système_éducatif
- 15.38% 2 vaste_mouvement système_politique
- 15.38% 2 stabilité_monétaire système_économique
- 15.38% 2 stabilité_monétaire système_européen
- 15.38% 2 société_de_services système_informatique
- 15.38% 2 recomposition_politique système_politique
- 15.38% 2 prochain_conseil système_électronique
- 15.38% 2 politique_familial système_français
- 15.38% 2 ligue_lombard système_politique
- 15.38% 2 inégalité_social système_scolaire
- 15.38% 2 intérêts_national système_original
- 15.38% 2 espèce_humain système_nerveux
- 15.38% 2 espèce_humain système_cardio
- 15.38% 2 débat_parlementaire système_français
- 15.38% 2 dernier_congrès système_rentier_bureaucratique
- 15.38% 2 dernier_congrès système_rentier

- 15.38% 2 crédit_national système_bancaire
- 15.38% 2 centre_régional système_éducatif
- 15.23% 16 système_monétaire court_terme
- 14.28% 5 système_politique assemblée_nationale
- 14.28% 3 institution_financier système_bancaire
- 14.28% 2 nouveau_proposition système_éducatif
- 14.28% 2 milicien_serbe viol_systématique
- 14.28% 2 michel_péricard système_audiovisuel
- 14.28% 2 corps_social système_de_contrôle
- 14.28% 15 système_monétaire pays_européen
- 14.14% 14 enseignement_supérieur système_éducatif
- 13.63% 3 intérêt_public système_éducatif
- 13.33% 4 formation_continu système_éducatif
- 13.33% 2 école_privé système_éducatif
- 13.33% 2 etats_américain rapatriement_systématique
- 13.33% 2 contexte_économique système_économique
- 13.33% 2 classes_préparatoire système_éducatif
- 13.33% 2 bloc_communiste système_européen
- 13.33% 2 banque_fédéral système_bancaire
- 13.33% 2 armes_chimique système_original
- 13.33% 14 système_monétaire élection_législatif
- 13.33% 14 système_monétaire valéry_giscard
- 13.33% 14 système_monétaire déficit_budgétaire
- 13.04% 3 système_informatique écoute_administratif

- 13.04% 3 système_informatique yves_gilleron
- 12.9% 8 formation_professionnel système_éducatif
- 12.9% 4 nouveau_système système_informatique
- 12.5% 3 pouvoir_socialiste système_mafieux
- 12.5% 3 insertion_professionnel système_scolaire
- 12.5% 2 principal_acteur système_politique
- 12.5% 2 principal_acteur système_européen
- 12.5% 2 chaînes_privé système_audiovisuel
- 12.38% 13 système_monétaire taux_allemand
- 12.38% 13 système_monétaire crise_monétaire
- 12.38% 13 système_monétaire construction_européen
- 12.12% 4 scrutin_majoritaire système_majoritaire
- 12.12% 4 institut_universitaire système_éducatif
- 12%3 épuration_ethnique viol_systématique
- 11.76% 2 tadeusz_mazowiecki viol_systématique
- 11.76% 2 tadeusz_mazowiecki pratique_systématique
- 11.76% 2 politique_budgétaire système_européen
- 11.76% 2 centre_droit système_politique
- 11.76% 2 bruno_durieux système_hospitalier
- 11.53% 3 giuliano_amato système_politique
- 11.48% 4 système_politique société_civil
- 11.48% 4 système_politique secrétaire_général
- 11.48% 12 système_monétaire françois_mitterrand
- 11.48% 12 système_monétaire communauté_européen

- 11.11% 2 économie_allemand système_financier
- 11.11% 2 organisme_international viol_systématique
- 10.81% 4 langue_étranger système_éducatif
- 10.52% 2 enseignement_général système_scolaire
- 10.52% 2 bourses_français système_informatique
- 10.52% 2 banque_français système_bancaire
- 10.52% 2 autorités_américain rapatriement_systématique
- 10.47% 11 système_monétaire économie_français
- 10.47% 11 système_monétaire gouvernement_français
- 10.34% 3 bettino_craxi système_électoral

Annexe G

Structuration des SNs relatifs au terme *système* dans la collection OFIL

Dans cet annexe, nous illustrons la structuration des SNs relatifs au terme *système* dans la collection OFIL où un SN est défini par rapport à sa tête, sa quantité d'information et sa fréquence (section 9.6) :

SN : [Qinf, freq, X]

où :

– Qinf : quantité d'information

– Freq : fréquence

– X : tête

Nous ajoutons pour chaque SN, ainsi qu'à sa tête, le patron syntaxique qui lui correspond:

– système bancaire italien SUBC ADJQ ADJQ [8, 1, système SUBC]

– système économique SUBC ADJQ [6, 8, système SUBC]

– système nerveux central SUBC ADJQ ADJQ [8, 3, système SUBC]

– système téléphonique interne SUBC ADJQ ADJQ [8, 1, système SUBC]

– nouveau système ADJQ SUBC [6, 12, système SUBC]

– système concentrationnaire SUBC ADJQ [6, 1, système SUBC]

- système universitaire SUBC ADJQ [6, 2, système SUBC]
- système asiatique SUBC ADJQ [6, 1, système SUBC]
- nouveau système électoral ADJQ SUBC ADJQ [8, 2, système SUBC]
- système nerveux SUBC ADJQ [6, 2, système SUBC]
- nouveau système fiscal ADJQ SUBC ADJQ [8, 1, système SUBC]
- système social SUBC ADJQ [6, 3, système SUBC]
- système parlementaire SUBC ADJQ [6, 1, système SUBC]
- système national SUBC ADJQ [6, 2, système SUBC]
- système graphique SUBC ADJQ [6, 1, système SUBC]
- système logique SUBC ADJQ [6, 1, système SUBC]
- système de péage SUBC PREP SUBC [8, 1, péage SUBC]
- système de bloc SUBC PREP SUBC [8, 1, bloc SUBC]
- système sonore SUBC ADJQ [6, 1, système SUBC]
- système audiovisuel SUBC ADJQ [6, 3, système SUBC]
- système communiste SUBC ADJQ [6, 3, système SUBC]
- système constitutionnel SUBC ADJQ [6, 2, système SUBC]
- système policier SUBC ADJQ [6, 1, système SUBC]
- système colonial SUBC ADJQ [6, 1, système SUBC]
- système original SUBC ADJQ [6, 1, système SUBC]
- système protecteur SUBC ADJQ [6, 1, système SUBC]
- système archaïque SUBC ADJQ [6, 1, système SUBC]
- système de soins spécialisé SUBC PREP SUBC ADJQ [10, 2, système de soins SUBC PREP SUBC]
- système judiciaire SUBC ADJQ [6, 3, système SUBC]

– système économique monétaire SUBC ADJQ ADJQ [8, 1, système SUBC]
– système carcéral SUBC ADJQ [6, 2, système SUBC]
– système électronique SUBC ADJQ [6, 4, système SUBC]
– système scolaire classique SUBC ADJQ ADJQ [8, 1, système SUBC]
– système utilitariste SUBC ADJQ [6, 1, système SUBC]
– système électoral proportionnel SUBC ADJQ ADJQ [8, 1, système SUBC]
– mémoire sur le système informatique SUBC PREP ARTD SUBC ADJQ [10, 3, système informatique SUBC ADJQ]
– système majoritaire simple SUBC ADJQ ADJQ [8, 1, système SUBC]
– système multilatéral SUBC ADJQ [6, 1, système SUBC]
– système contractuel SUBC ADJQ [6, 1, système SUBC]
– meilleur système ADJQ SUBC [6, 1, système SUBC]
– système de retardement SUBC PREP SUBC [8, 1, retardement SUBC]
– système féodal SUBC ADJQ [6, 1, système SUBC]
– système monétaire international SUBC ADJQ ADJQ [8, 13, système SUBC]
– pivot de système SUBC PREP SUBC [8, 1, système SUBC]
– système vidéo SUBC ADJQ [6, 3, système SUBC]
– système international SUBC ADJQ [6, 1, système SUBC]
– système odieux SUBC ADJQ [6, 1, système SUBC]
– système immunitaire SUBC ADJQ [6, 2, système SUBC]
– système européen SUBC ADJQ [6, 5, système SUBC]
– système de levier SUBC PREP SUBC [8, 1, levier SUBC]
– système idéal SUBC ADJQ [6, 1, système SUBC]
– système monétaire commun SUBC ADJQ ADJQ [8, 1, système SUBC]

- système allemand SUBC ADJQ [6, 2, système SUBC]
- système hypocrite SUBC ADJQ [6, 1, système SUBC]
- système de parrainage SUBC PREP SUBC [8, 1, parrainage SUBC]
- système de soins français SUBC PREP SUBC ADJQ [10, 2, système de soins SUBC PREP SUBC]
- système de faillite SUBC PREP SUBC [8, 1, faillite SUBC]
- système municipal étrange SUBC ADJQ ADJQ [8, 1, système SUBC]
- système de transfert SUBC PREP SUBC [8, 2, transfert SUBC]
- système rigide SUBC ADJQ [6, 1, système SUBC]
- nouveau système informatique ADJQ SUBC ADJQ [8, 1, système SUBC]
- système de contrôle social SUBC PREP SUBC ADJQ [10, 1, système de contrôle SUBC PREP SUBC]
- système fédéral SUBC ADJQ [6, 2, système SUBC]
- système présidentiel SUBC ADJQ [6, 1, système SUBC]
- système bipolaire SUBC ADJQ [6, 1, système SUBC]
- système constitutionnel actuel SUBC ADJQ ADJQ [8, 1, système SUBC]
- futur système tarifaire ADJQ SUBC ADJQ [8, 1, système SUBC]
- système monétaire de l’europe SUBC ADJQ PREP ARTD SUBC [10, 2, système monétaire SUBC ADJQ]
- système fiscal SUBC ADJQ [6, 3, système SUBC]
- ordre dans le système monétaire SUBC PREP ARTD SUBC ADJQ [10, 1, système monétaire SUBC ADJQ]
- système soviétique SUBC ADJQ [6, 3, système SUBC]
- information sur le système SUBC PREP ARTD SUBC [8, 1, système SUBC]
- seul système économique ADJQ SUBC ADJQ [8, 2, système SUBC]

- système japonais SUBC ADJQ [6, 1, système SUBC]
- système lymphatique SUBC ADJQ [6, 2, système SUBC]
- système public SUBC ADJQ [6, 1, système SUBC]
- système fragile SUBC ADJQ [6, 1, système SUBC]
- système audiovisuel français SUBC ADJQ ADJQ [8, 1, système SUBC]
- système paritaire SUBC ADJQ [6, 1, système SUBC]
- système parallèle SUBC ADJQ [6, 1, système SUBC]
- système inconnu SUBC ADJQ [6, 1, système SUBC]
- système de bouton SUBC PREP SUBC [8, 1, bouton SUBC]
- système ecclésial SUBC ADJQ [6, 1, système SUBC]
- système majoritaire SUBC ADJQ [6, 3, système SUBC]
- système de caméra SUBC PREP SUBC [8, 1, caméra SUBC]
- système de licence SUBC PREP SUBC [8, 1, licence SUBC]
- système de freinage SUBC PREP SUBC [8, 1, freinage SUBC]
- système hospitalier français SUBC ADJQ ADJQ [8, 2, système SUBC]
- système de contrôle SUBC PREP SUBC [8, 3, contrôle SUBC]
- système de prélèvement SUBC PREP SUBC [8, 1, prélèvement SUBC]
- système de calcul SUBC PREP SUBC [8, 1, calcul SUBC]
- système existant SUBC ADJQ [6, 1, système SUBC]
- système de réserves SUBC PREP SUBC [8, 1, réserves SUBC]
- système idéologique SUBC ADJQ [6, 1, système SUBC]
- système autonome SUBC ADJQ [6, 1, système SUBC]
- système socialiste SUBC ADJQ [6, 2, système SUBC]
- esprit de système SUBC PREP SUBC [8, 6, système SUBC]

- ancien système ADJQ SUBC [6, 4, système SUBC]
- système productif SUBC ADJQ [6, 1, système SUBC]
- système de bourses SUBC PREP SUBC [8, 1, bourses SUBC]
- système douanier SUBC ADJQ [6, 1, système SUBC]
- système scolaire SUBC ADJQ [6, 8, système SUBC]
- système mixte SUBC ADJQ [6, 1, système SUBC]
- système solaire SUBC ADJQ [6, 5, système SUBC]
- grand système ADJQ SUBC [6, 2, système SUBC]
- système ferroviaire original SUBC ADJQ ADJQ [8, 1, système SUBC]
- système clérical contraignant SUBC ADJQ ADJQ [8, 1, système SUBC]
- système électoral SUBC ADJQ [6, 1, système SUBC]
- système défensif SUBC ADJQ [6, 2, système SUBC]
- système monétaire européen SUBC ADJQ ADJQ [8, 45, système SUBC]
- système administratif SUBC ADJQ [6, 1, système SUBC]
- système commercial SUBC ADJQ [6, 1, système SUBC]
- système de transports SUBC PREP SUBC [8, 1, transports SUBC]
- système capitaliste SUBC ADJQ [6, 1, système SUBC]
- système informatique SUBC ADJQ [6, 5, système SUBC]
- système proportionnel SUBC ADJQ [6, 4, système SUBC]
- intégration de système SUBC PREP SUBC [8, 1, système SUBC]
- système actuel SUBC ADJQ [6, 10, système SUBC]
- système de santé pour le personne âgé SUBC ARTC SUBC PREP ARTD SUBC ADJQ [14, 1, personne âgé SUBC ADJQ]
- système de gouvernement SUBC PREP SUBC [8, 1, gouvernement SUBC]

- système hydraulique SUBC ADJQ [6, 2, système SUBC]
- système fruste SUBC ADJQ [6, 1, système SUBC]
- système musculaire SUBC ADJQ [6, 1, système SUBC]
- système condamné SUBC ADJQ [6, 1, système SUBC]
- système souple SUBC ADJQ [6, 1, système SUBC]
- système financier SUBC ADJQ [6, 5, système SUBC]
- système alterné SUBC ADJQ [6, 1, système SUBC]
- système autoritaire SUBC ADJQ [6, 1, système SUBC]
- système carcéral égyptien SUBC ADJQ ADJQ [8, 1, système SUBC]
- système de couleur SUBC PREP SUBC [8, 1, couleur SUBC]
- système totalitaire SUBC ADJQ [6, 3, système SUBC]
- système planétaire voisin SUBC ADJQ ADJQ [8, 1, système SUBC]
- système culturel SUBC ADJQ [6, 1, système SUBC]
- système électoral italien SUBC ADJQ ADJQ [8, 1, système SUBC]
- système uninominal SUBC ADJQ [6, 1, système SUBC]
- système de quota SUBC PREP SUBC [8, 1, quota SUBC]
- fameux système binaire ADJQ SUBC ADJQ [8, 1, système SUBC]
- système clanique traditionnel SUBC ADJQ ADJQ [8, 1, système SUBC]
- actuel système digital ADJQ SUBC ADJQ [8, 1, système SUBC]
- système métaphorique SUBC ADJQ [6, 1, système SUBC]
- système de soins SUBC PREP SUBC [8, 6, système SUBC]
- système espagnol SUBC ADJQ [6, 1, système SUBC]
- système dual SUBC ADJQ [6, 1, système SUBC]
- système de notation SUBC PREP SUBC [8, 1, notation SUBC]

- système informatique central SUBC ADJQ ADJQ [8, 1, système SUBC]
- système de pouvoir SUBC PREP SUBC [8, 1, pouvoir SUBC]
- système du retraite par répartition SUBC ARTC SUBC PREP SUBC [12, 1, retraite par répartition SUBC PREP SUBC]
- système imaginaire SUBC ADJQ [6, 1, système SUBC]
- système démocratique SUBC ADJQ [6, 3, système SUBC]
- vieux système ADJQ SUBC [6, 1, système SUBC]
- système incompatible SUBC ADJQ [6, 1, système SUBC]
- système juridique SUBC ADJQ [6, 1, système SUBC]
- système libéral SUBC ADJQ [6, 1, système SUBC]
- système philosophique SUBC ADJQ [6, 1, système SUBC]
- système français SUBC ADJQ [6, 3, système SUBC]
- système transposé SUBC ADJQ [6, 1, système SUBC]
- système belge SUBC ADJQ [6, 1, système SUBC]
- système pénal SUBC ADJQ [6, 1, système SUBC]
- système monétaire SUBC ADJQ [6, 3, système SUBC]
- système nippon SUBC ADJQ [6, 1, système SUBC]
- système bancaire SUBC ADJQ [6, 8, système SUBC]
- système médiatique SUBC ADJQ [6, 2, système SUBC]
- perte du confiance dans le système éducatif SUBC ARTC SUBC PREP ARTD SUBC ADJQ [14, 1, système éducatif SUBC ADJQ]
- propre système informatique ADJQ SUBC ADJQ [8, 1, système SUBC]
- système social généreux SUBC ADJQ ADJQ [8, 1, système SUBC]
- système commun SUBC ADJQ [6, 1, système SUBC]

- système porteur SUBC ADJQ [6, 2, système SUBC]
- système éducatif français SUBC ADJQ ADJQ [8, 2, système SUBC]
- système évolutif SUBC ADJQ [6, 1, système SUBC]
- système multimédia SUBC ADJQ [6, 1, système SUBC]
- système étatique uniforme SUBC ADJQ ADJQ [8, 1, système SUBC]
- système hétérogène SUBC ADJQ [6, 1, système SUBC]
- système de mensualité SUBC PREP SUBC [8, 1, mensualité SUBC]
- système monolithique SUBC ADJQ [6, 1, système SUBC]
- système unique SUBC ADJQ [6, 1, système SUBC]
- système productif allemand SUBC ADJQ ADJQ [8, 1, système SUBC]
- système libre SUBC ADJQ [6, 1, système SUBC]
- système éducatif SUBC ADJQ [6, 26, système SUBC]
- système de crédit SUBC PREP SUBC [8, 1, crédit SUBC]

Bibliographie

[AHKV97] H. Ahonen, O. Heinonen, M. Klemettinen, and A. Verkamo. Applying data mining techniques in text analysis. Technical report, Department of Computer Science, University of Helsinki, 1997.

[Ahl88] T. Ahlswede. Automatic construction of phrasal thesaurus for an information retrieval system from a machine readable dictionnary. In *2ème Conférence de Recherche d'Information Assistée par Ordinateur (RIAO'1998), Cambridge, USA*, pages 597–608, 1988.

[AIS93] R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases. In *Proceedings of the ACM SIGMOD Conference on Management of Data, Washington, USA*, pages 207–216, Mai 1993.

[Alt] http://www.altavista.com.

[Ama] http://amaryllis.inist.fr.

[Ama00] M. Amar. *Les Fondements théoriques de l'indexation une approche linguistique*. ADBS éditions, Paris, 2000.

[APWZ95] R. Agrawal, G. Psaila, E. Wimmers, and M. Zait. Qurying shapes of histories. In *20ème Conference on Very Large Databases, Santiago, Chili*, pages 502–514, Septembre 1995.

[AS94] R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In *20ème Conference on Very Large Databases, Santiago, Chili*, pages 487–499, Septembre 1994.

[AS95] R. Agrawal and R. Srikant. Mining sequential patterns. In *11ème International Conference on Data Engineering (ICDE'1995), Taipei, Taiwan*, pages 3–14, Mars 1995.

[Ass] http://abu.cnam.fr.

[Ass98] H. Assadi. Construction d'ontologies à partir de textes techniques; application aux systèmes documentaires, Octobre 1998. Thèse de doctorat, Université Paris 6.

[ATK97] A. Arampatzis, T. Tsoris, and C.H.A. Koster. Irena: Information retrieval engine based on natural language analysis. In *5ème Conférence de Recherche d'Information Assistée par Ordinateur (RIAO'1997), Montreal, Canada*, pages 159–175, Juin 1997.

[ATKW98] A. T. Arampatzis, T. Tsoris, C. H. A. Koster, and T. P. Van Der Weide. Phrase-based information retrieval. *Information Processing and Management*, 34(6):693–707, 1998.

[AvdWK+00] A. Arampatzis, T.P. van der Weide, C.H.A. Koster, , and P. van Bommel. An evaluation of linguistically-motivated indexing schemes. In *Proceedings of BCS-IRSG 2000 Colloquium on IR Research, Cambridge, England*, pages 34–45, Avrile 2000.

[AWKB00] A. Arampatzis, Th.P. Van Der Weide, C.H.A. Koster, and P. Van Bommel. Linguistically-motivated information retrieval. In *Encyclopedia of Library and Information Science, Marcel Dekker, Inc., New York, Basel*, 2000.

[BC99] D. Bourigault and J. Charlet. Construction d'un index thématique de l'ingénierie des connaissances. In *Actes de la Troisième Conférence sur l'Ingénierie des Connaissances (IC'99), Palaiseau, France*, pages 107–118, Juin 1999.

[BDO95] M. W. Berry, S. T. Dumais, and G. O'Brien. Using linear algebra for intelligent information system. *Revue SIAM*, 37(4):573–595, 1995.

[Ben73] J. P. Benzecri. *L'analyse de données, Tome 1 et Tome 2*. Dunod, 1973.

[Ber97a] H. Berghel. Cyberspace 2000 : dealing with information overload. *Communications of the ACM*, 40(2):19–24, Février 1997.

[Ber97b] C. Berrut. Indexation des données multimédia, utilisation dans le cadre d'un système de recherche d'information, Octobre 1997. Habilitation à diriger des recherches, Université Joseph Fourrier.

[BFM98] P. Bradly, U. Fayyad, and O. Mangasarian. Mathematical programming for data mining: Formulations and challenges. Technical report, Departement of computer sciences, University of Wisconsin, Madison, 1998.

[BHR96] P. Bowden, P. Halstead, and T. Rose. Explicit relation markers. In *Proceedings of the 9th European Knowledge Acquisition Workshop (EKAW'96), Nottingham, United Kingdom*, pages 147–162, Mai 1996.

[Bog94] P. Bogaards. *Le vocabulaire dans l'apprentissage des langues étrangères*. CREDIF/Saint-Cloud/France, Hatier-Didier/Paris (LAL : Langues et apprentissage des langues), 1994.

[Bou92] D. Bourigault. Lexter, un logiciel d'extraction terminologique. In *2ième Colloque international de TermNet, Avignon, France*, Juin 1992.

[Bou93] D. Bourigault. Analyse syntaxique locale pour le repérage de termes complexes dans un texte. *TAL*, 34(2):105–117, 1993.

[Bou97] F. Bourdoncle. Livetopics: Recherche visuelle d'information sur l'internet. *La Documentation Française*, (74):36–38, 1997.

[BRC99] R. Besançon, M. Rajman, and J.C. Chappelier. Textual similarities based on a distributional approach. In *International Workshop on Similarity Search (IWOSS'1999), Firenze, Italie*, pages 180–184, Septembre 1999.

[Bru90] M. F. Bruandet. Construction automatique d'une base de connaissances du domaine dans un système de recherche d'informations, 1990. Habilitation à Diriger des Recherches, Université Joseph Fourier.

[BS96] R. Border and F. Song. Knowledge-based approaches to query expansion in information retrieval. In *Proceedings of the 11th Conference of the canadian Society for Computational Studies of Intelligence, Toronto, Ontario, Canada*, pages 146–158, Mai 1996.

[CB97] J. P. Chevallet and M. F. Bruandet. Impact de l'utilisation de multi termes sur la qualité des réponses d'un système de recherche d'information à indexation automatique. In *Organisation des connaissances en vue de leur intégration dans les systèmes de représentation et de recherche d'information Collection UL3 Lilles, USBN 2-84467-002-4, Lille, France*, pages 223–238, 16–17 octobre 1997.

[CDBK86] Y. Chiaramella, B. Defude, M. F. Bruandet, and D. Kerkouba. Iota: a full test information retrieval system. In *ACM conference on research and development in information retrieval (SIGIR'1986), Pisa, Italie*, pages 207–213, Septembre 1986.

[CDG+98] S. Chakrabarti, B. E. Dom, D. Gibson, R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. Spectral filtering for resource discovery. In *21ème Conference on Research and Development in Information Retrieval (SIGIR'1998), Workshop on Hypertext Analysis, Melbourne, Australie*, pages 558–567, 1998.

[CGH00] J. P. Chevallet, M. Géry, and M. H. Haddad. Campagne de tests amaryllis ii : Expérimentations et résultats. In *Atelier final de la campagne Amaryllis II, Paris*, Avril 2000.

[CH90] W. K. Church and P. Hanks. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29, 1990.

[CH01] J.P. Chevallet and M. H. Haddad. Proposition d'un modèle relationnel d'indexation syntagmatique : mise en oeuvre dans le système iota. In *Congrès Informatique des Organisations et Systèmes d'Information et de Décision (INFORSID'2001), Martigny, Suisse*, pages 465–483, Juin 2001.

[CMS97] R. Cooley, B. Mobasher, and J. Srivastava. Web mining: Information and pattern discovery on the world wide web. In *9ème IEEE International Conference on Tools with Artificial Intelligence (ICTAI'97), Newport Beach, USA*, pages 558–567, Novembre 1997.

[CN97] J. P. Chevallet and J. Nie. Intégration des analyses du français dans la recherche d'information. In *5ème Conférence de Recherche d'Information Assistée par Ordinateur (RIAO'1997), Montreal, Canada*, pages 761–772, Juin 1997.

[CP98] M. De Calmès and G. Pérennou. Bdlex : a lexicon for spoken and written. In *International Conference on Language Resources and Evaluation, Grenade, Espagne*, pages 129–136, Mai 1998.

[Cro90] C.J. Crouch. An approach to the automatic construction of global thesauri. *Information Processing and Management*, 26(5):629–640, 1990.

[CTL91] W. B. Croft, H. R. Turtle, and D. Lewis. The use of phrases and structured queries in information retrieval. In *14ème Conference on Research and Development in Information Retrieval (SIGIR'1991), Illinois, USA*, pages 32–45, Octobre 1991.

[CY92] C. J. Crouch and B. Yang. Experiments in automatic statistical thesaurus construction. In *Annuel International ACM SIGIR conference on research*

and development in information retrieval (SIGIR'92), ACM press, Copenhagen, Danemark, pages 77–88, July 1992.

[Dai94] B. Daille. Approche mixte pour l'extraction automatique de terminologie : statistiques lexicales et filtres linguistiques, 1994. Thèse de doctorat, Université Paris 7.

[Dai02] B. Daille. Découvertes linguistiques en corpus, 2002. Mémoire d'Habilitation à diriger des recherches, Université de Nantes.

[DDK+90] S. Deerwester, S. T. Dumais, Landauer T. K., Furnas G. W., and Harshman R. A. Indexing by latent semantic analysis. *the Society for information science*, 41(6):391–407, 1990.

[Deb82] F. Debili. Analyse syntaxico-semantique fondee sur une acquisition automatique de relations lexicales-semantiques, Janvier 1982. Thèse de doctorat, Université Paris XI.

[Fag87] J. L. Fagan. Experiments in automatic phrase indexing for document retrieval: A comparison of syntactic and non-syntactic methods, Septembre 1987.

[Far80a] J. Farradane. Relational indexing. part i. *Journal of Information science*, 1(5):267–276, 1980.

[Far80b] J. Farradane. Relational indexing. part ii. *Journal of Information science*, 1(6):313–324, 1980.

[FD95] R. Feldman and I. Dagan. Kdt- knowledge discovery in texts. In *1ère International Conference on knowledge discovery (KDD'1995), Montreal, Canada*, pages 112–117, Aout 1995.

[Feu88] J. Feuillet. Introduction à l'analyse mophosyntaxique, 1988.

[FFH+98] R. Feldman, M. Fresko, H. Hirsh, Y. Aumann, O. LiphstatI, Y. Schler, and M. Rajman. Knowledge management: A text mining approach. In *Proc. of the 2nd Int. Conf. on Practical Aspects of Knowledge Management (PAKM98), Basel, Suisse*, pages 9.1–9.10, Octobre 1998.

[FH96] R. Feldman and H. Hirsh. Mining associations in text in the presence of background knowledge. In *2ème International Conference on Knowledge Discovery (KDD-96), Portland, USA*, pages 343–346,, Aout 1996.

[Flu77] C. Fluhr. Algorithmes à apprentissage et traitement automatique des langues, 1977. Thèse de doctorat, Université Paris-sud.

[FPSS98] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. From data mining to knowledge discovery: an overview. *Advances in Knowledge Discovery and Data Mining, MIT Press*, 1:1–36, 1998.

[Gar98] D. Garcia. Analyse automatique de textes pour l'organisation causale des actions, réalisation du système informatique coatis, 1998. Thèse de doctorat, Université Paris 4.

[GC01] M. Géry and J-P. Chevallet. Toward a structured information retrieval system on the web: Automatic structure extraction of web pages. In *International Workshop on Web Dynamics, Londres, UK*, Janvier 2001.

[Gen94] E. Genevay. Ouvrir la grammaire. *Langue Et Parole LEP Loisirs et Pédagogie*, 1994.

[GH99] M. Géry and M. H. Haddad. Knowledge discovery for automatic query expansion on the world wide web. In *International Workshop on the World-Wide Web and Conceptual Modeling, Paris, France*, pages 334–347, Novembre 1999.

[GJSS89] U. Guntzer, G. Juttner, G. Seegmuller, and F. Sarre. Automatic thesaurus construction by machine learning from retrieval sessions. *Information Processing and Management*, 25(3):265–273, 1989.

[Gon99] Y. Gong. Advancing content-based image retrieval by expoliting image color and region features. *Multimedia Systems*, 7(6):449–457, 1999.

[GPW98] G. Gardarin, P. Pucheral, and F. Wu. Bitmap based algorithms for mining association rules. In *Quatorzième Journées Bases de Données Avancées, Hammamet, Tunisie*, pages 157–175, Octobre 1998.

[Gér99] Mathias Géry. SmartWeb : Recherche de Zones de Pertinence sur le World Wide Web. In *Congrès Informatique des Organisations et Systèmes d'Information et de Décision (INFORSID'99)*, pages 133–147, La Garde, France, Juin 1999.

[Gre92] G. Grefenstette. Use of suntactic context to produce term association lists for text retrieval. In *Conference on Research and Development in Information Retrieval (SIGIR'1992), Copenhagen, Denmarke*, pages 89–97, Juin 1992.

[Gre93] G. Grefenstette. Automatic thesaurus generation from raw text using knowledge-poor techniques. In *Making Sense of Words, 9th Annual Conference of the University of Waterloo Centre for the Oxford English Dictionary and Text Research, Cambridge*, Septembre 1993.

[Gro96] G. Gross. *Les expressions figées en français, noms composés et autres locutions*. Ophrys, Gap-Paris, 1996.

[GS99] H. Greisdorf and A. Spink. Regions of relevance: Approaches to measurement for enhanced precision. In *21ème British Computer Society Information Retrieval Sub Group Annual Colloquium on IR Research*, pages 1–33, Glasgow, Scotland, Avril 1999.

[GSEM97] W. Gotthard, R. Seiffert, L. England, and A. Marwick. Text mining. Technical report, Research Division, IBM, Allemagne, 1997.

[Gue84] M. Le Guern. Les descripteurs d'un système documentaire : essai de définition. *Condenser, Suppl*, 1:163–169, 1984.

[GV02] M. Géry and M. H. Haddad D. Vaufreydaz. Web as huge information source for noun phrases: Integration in the information retrieval process. In *The 2002 International Conferencen Information and Knowledge Engineering (IKE), Las Vegas, USA*, Juin 2002.

[Had98] M. H. Haddad. Etude de l'intégration du datamining dans un système de recherche d'information, Juin 1998. Rapport de DEA, Université Joseph Fourrier.

[Had02] M. H. Haddad. Combining text mining and nlp for information retrieval. In *International Conference Conference on Artificial Intelligence (IC-AI), Las Vegas, USA*, Juin 2002.

[Har71] Z. S. Harris. *Structures Mathématiques du Langage*. Dunod, Paris, 1971.

[HBGN+97] B. Habert, S. Bertrand-Gastaldy, A. Nazarenko, F.Dupuis, E. Naulleau, M. Lemieux, and C. Delisle. Recyclage d'analyses syntaxiques automatiques pour le repérage de variantes de termes. In *Actes Coopération franco-québecoise en ingéniérie de la langue, Montréal, Canada*, pages 751–760, 1997.

[HCTH99] D. Hawking, N. Craswell, P. Thistlewaite, and D. Harman. Results and challenges in web search evaluation. In *8ème World Wide Web Conference, Toronto, Canada*, pages 243–252, Toronto, Canada, Mai 1999.

[Hea92] M. A. Hearst. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the Fourteenth International Conference on Computational Linguistics, Nantes, France*, pages 539–545, Juillet 1992.

[HGS+97] D. Hull, G. Grefenstette, B.M. Schulze, E. Gaussier, H. Schutze, and J. Pedersen. Xerox trec-5 site report. routing,filtering, nlp and spanish track. In *E. Voorhees and Donna K. Harman, editor, The Fifth Text REtrieval Conference (TREC-5)*, pages 167–180, 1997.

[HM01a] M. H. Haddad and P. Mulhem. Association rules for symbolic indexing of still images. In *International Conference Conference on Artificial Intelligence (IC-AI), Las Vegas, USA*, pages 469–475, Juin 2001.

[HM01b] M. H. Haddad and P. Mulhem. Utilisation de fouille de données pour l'indexation automatique des images. In *Congrès Informatique des Organisations et Systèmes d'Information et de Décision (INFORSID'2001), Martigny, Suisse*, pages 405–418, Juin 2001.

[HNG98] T. Hamon, A. Nazarenko, and C. Gros. A step towards the detection of semantic variants of terms in technical documents. In *Proceedings of International Conference on omputational Linguistics (COLING'98), Montréal, Canada*, pages 498–504, Aout 1998.

[Hul94] D. Hull. Improving text retrieval for the routing problem using latent semantic indexing. In *Annuel International ACM SIGIR conference on research and development in information retrieval (SIGIR'94), ACM press, Dublin, Irelande*, pages 282–291, Juillet 1994.

[IS99] F. Ibekwe-Sanjuan. L'analyse formelle de corpus terminologiques. In *Septièmes journéesde la Société francophone de classification, Nancy, France*, pages 155–162, Septembre 1999.

[Iss97] F. Issac. Analyse syntaxique et apprentissage des langues, Septembre 1997. Thèse de doctorat, Université Paris Nord.

[Jac97] C. Jacquemin. Variation terminologique: Reconnaissance et acquisition automatiques de termes et de leurs variantes en corpus, 1997. Habilitation à diriger des recherches, Université de Nantes.

[JC94] Y. Jing and W. B. Croft. An association thesaurus for information retrieval. In *4ème Conférence de Recherche d'Information Assistée par Ordinateur (RIAO'1994), New York, U.S.A*, pages 146–160, 1994.

[Jou93] C. Jouis. Contributions à la conceptualisation et à la modélisation des connaissances à partir d'une analyse linguistique de textes. réalisation d'un prototype : le système seek, 1993. Thèse de doctorat, EHESS, Paris.

[Ker84] D. Kerkouba. Une méthode d'indexation automatique des documents fondée sur l'exploitation de leurs propriétés structurelles. application à un corpus technique., 1984. Thèse de doctorat, Institut National Polytechnique de Grenoble.

[KF93] H. Kozima and T. Furugori. Similarity between words computed by spreading activation on an english dictionary. In *Proceedings of the 6th Conference of the European Chapter of the Association for Computational Linguistics, Utrecht, The Netherlands*, pages 232–239, 1993.

[Kho95] C. S. G. Khoo. Automatic identification of causal relations in text and their use for improving precision in information retrieval, 1995. Thèse de doctorat, Syracuse University.

[Kho97] C. S. G. Khoo. The use of relation matching in information retrieval. *Singapore Libraries*, 26(1):3–22, 1997.

[KP98] W. Kraaij and R. Pohlmann. Comparing the effect of syntactic vs. statistical phrase indexing strategies for dutch. In *Proceedings of Second European Conference on Research and Advanced Technology for Digital Libraries (ECDL'1998), Heraklion, Crete*, pages 605–614, Septembre 1998.

[LAS97] B. Lent, R. Agrawal, and R. Srikant. Discovering trends in text databases. In *Proc. of the 3rd Int'l Conference on Knowledge Discovery in Databases and Data Mining (KDD'1997), Newport Beach, California*, pages 227–230, Aout 1997.

[LCBD98] F. LePriol, J. P. Chevallet, M. F. Bruandet, and J. P. Desclés. Intégration d'un système statistique (iota) et d'un système sémantique (seek) dans une chaîne de traitement permettant l'extraction de terminologie. In *Ingénierie des Connaissances (IC'98), pont-à-Mousson, France*, pages 33–40, Mai 1998.

[Lin98] D. Lin. Extracting collocations from text corpora. In *First Workshop on Computational Terminology, Montreal, Canada.*, 1998.

[Lyo70] J. Lyons. *Introduction à la linguistique théorique*. Traduction de F. Dubois-Charlier et D. Robinson, Larousse, Paris, 1970.

[Mar88] J. M. Marandin. À propos de la notion de thème de discours. Éléments d'analyse dans le récit. *Langue française*, 78:67–87, 1988.

[Mas02] F. Masseglia. Algorithmes et applications pour l'extraction de motifs séquentiels dans le domaine de la fouille de données : de l'incrémental au temps réel, Janvier 2002. Thèse de doctorat, Université de Montpellier.

[MBF+90] G. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. J. Miller. Introduction to wordnet: an on-line lexical database. *International Journal of lexicography*, 3(4):235–244, 1990.

[MBSC97] M. Mitra, C. Buckley, A. Singhal, and C. Cardi. In analysis of statistical and syntactic phrases. In *5ème Conférence de Recherche d'Information Assistée par Ordinateur (RIAO'1997), Montreal, Canada*, pages 200–214, Juin 1997.

[Mcc96] J. F. Mccarthy. A trainable approach to coreference resolution for information extraction, Septembre 1996. Thèse de doctorat,Université de Massachusette.

[MCD+01] J. Mothe, C. Chrisment, T. Dkaki, B. Dousseta, and D. Egret. Information mining: use of the document dimensions to analyse interactively a document set. In *Proceedings of BCS-IRSG 2001 Colloquium on IR Research, Darmstadt, Allemagne*, pages 66–77, Avrile 2001.

[Met] http://www.metacrawler.com.

[MM00] B. H. Murray and A. Moore. Sizing the web. Technical report, Cyveillance, Inc, 2000.

[Mor98] E. Morin. Prométhée : un outil d'aide à l'acquisition de relations sémantiques entre termes. In *5e conférence annuelle sur le traitement automatique des langues naturelles, Paris, France*, pages 172–181, Juin 1998.

[Mor99] E. Morin. Extraction de lien sémantique entre termes à partir de corpus de textes techniques, Décembre 1999. Thèse de doctorat, Institut de recherche en informatique de Nantes.

[MPC99] F. Masseglia, P. Poncelet, and R. Clichetti. Analyse du comportement des utilisateurs sur le web. In *Congrès Informatique des Organisations et Systèmes d'Information et de Décision (INFORSID'1999), La Garde*, pages 393–412, Mai 1999.

[MR97] M. Magennis and C. J. V. Rijsbergen. The potential and actual effectiveness of interactive query expansion. In *International Conference on Research and Development in Information Retrieval (SIGIR'97), Philadelphia USA*, pages 324–332, Juillet 1997.

[Nau98] E. Naulleau. Apprentissage et filtrage syntaxico-sémantique de syntagmes nominaux pertinents pour la recherche documentaire, Janvier 1998. Thèse de doctorat, Université Paris 13 - Villetaneuse.

[Nau99] E. Naulleau. Profile-guided terminology extraction. In *Proceedings of the 5th International Congress on Terminology and Knowledge Engineering, Innsbruck, Austria*, pages 222–233, Aout 1999.

[Naz94] A. Nazarenko. Compréhension du langage naturel : le problème de la causalité., Janvier 1994. Thèse de doctorat, Université de Paris-Nord XIII.

[Nie90] J-Y. Nie. Un modèle logique général pour les systèmes de recherche d'information. application au prototype rime, Juillet 1990. Thèse de doctorat, Université Joseph Fourrier.

[Nie97] J.Y. Nie. Using terminological bases in information retrieval. In *5ème Conférence de Recherche d'Information Assistée par Ordinateur (RIAO'1997), Montreal, Canada*, pages 85–97, Juin 1997.

[Oue99] R. Oueslati. Acquisition de connaissances à partir de corpus, Juillet 1999. Thèse de doctorat, Ecole Nationale Supérieure des Arts et Industries de Strasbourg.

[Oun98] I. Ounis. un modèle d'indexation relationnel pour les graphes conceptuels fondé sur une interprétation logique, Février 1998. Thèse de doctorat, Université Joseph Fourrier.

[Pal90] P. Palmer. Etude d'un analyseur de surface de la langue naturelle. application à l'indexation automatique des textes, Septembre 1990. Thèse de doctorat, Université Joseph Fourrier.

[Par96] F. Paradis. Un modèle d'indexation pour les documents textuels structurés, Novembre 1996.

[PCY95] J. S. Park, M. Chen, and P. S. Yu. An effective hash based algorithm for mining association rules. In *Proc. of the 1995 ACM SIGMOD Conference*, pages 175–186, Mai 1995.

[Pri00] F. Le Priol. Extraction et capitalisation automatiques de connaissances à partir de documents textuels. seek-java : identification et interprétation de relations entre concepts, Décembre 2000. Thèse de doctorat, Université Paris 4.

[RB97] M. Rajman and R. Besancon. Text mining: Natural language techniques and text mining applications. In *Proc. of the 7th IFIP 2.6 Working Conference on Database Semantics (DS-7), Chapam and Hall IFIP Proceedings serie,Leysin, Suisse,* Octobre 1997.

[RCP99] C. Roussey, S. Calabretto, and J. M. Pinon. Semantic indexing of multilingual document for information retrieval. In *Actes de la 5ème Conférence Internationale de l'ISAS (Information Systems Analysis and Synthesis), Orlando, USA*, pages 258–265, Août 1999.

[Rij79] C.J. Van. Rijsbergen. *Information Retrieval*. Butterworths, London, 1979.

[Rij86] C. J. Van. Rijsbergen. A new theoretical framework for information retrieval. In *Annuel International ACM SIGIR conference on research and development in information retrieval (SIGIR'86), ACM press, Pisa, Italie*, pages 194–200, Septembre 1986.

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

[RL98] M. Rajman and L. Lebart. Similarités pour des données textuelles. In *4th International Conference on Statistical Analysis of Textual Data (JADT'98), Nice, France,* pages 545–555, Février 1998.

[Rob73] R. H. Robins. *Linguistique générale: une introduction*. Paris. A. Colin, 1973.

[Rug97] G. Ruge. Automatic detection of thesaurus relations for information retrieval applications. In *Christian Freksa, Matthias Jantzen, Rüdiger Valk (eds.): Foundations of Computer Science. Springer, Berlin*, pages 499–506, 1997.

[Sal68] G. Salton. *Automatic information organization and retrieval*. Mc Graw-Hill, New York, 1968.

[Sal71] G. Salton. *The SMART retrieval system : experiments in automatic document processing*. Prentice-Hall series in automatic computation, New Jersey, 1971.

[Sal89] G. Salton. On the use of clustered file organisation in the information search and retrieval. Technical report, departement of computer science, Cornell university, Ithaca, 1989.

[Sau72] F. De Saussure. *Cours de linguistique générale*. Ed. critique de Tullio de Mauro, Paris, Payot., 1972.

[Sav] http://www.savvysearch.com.

[SC94] T. Strzalkowski and J. P. Carballo. Natural language information retrieval: Trec-3 report. In *Donna K. Harman, editor, The Third Text REtrieval Conference (TREC-3)*, pages 39–54, 1994.

[SC95] T. Strzalkowski and J. P. Carballo. Natural language information retrieval: Trec-4 report. In *Donna K. Harman, editor, The Fourth Text REtrieval Conference (TREC-4)*, pages 245–258, 1995.

[SFLG+97] T. Strzalkowski, J. Wang F. Lin, L. Guthrie, J. Leistensnider, J. Wilding, J. Karlgren, T. Straszheim, and J. Carballo. Natural language information retrieval: Trec-5 report. In *E. Voorhees and Donna K. Harman, editor, The Fifth Text REtrieval Conference (TREC-5)*, pages 291–334, 1997.

[SHP95] H. Schutze, D. A. Hull, and J. O. Pedersen. A comparison of classifiers and document representations for the routing problem. In *Annuel International ACM SIGIR conference on research and development in information retrieval (SIGIR'95), ACM press, Seattle, USA*, pages 229–237, 1995.

[Sim00] J. L. Simoni. Acces à l'information à l'aide d'un graphe de termes construit automatiquement (intégration de l'interrogation et de la navigation), Janvier 2000. Thèse de doctorat, Université Paris 7.

[SLPC97] T. Strzalkowski, F. Lin, and J. Perez-Carballo. Natural language information retrieval: Trec-6 report. In *E. M. Voorhees and Donna K. Harman editor, The Sixth Text REtrieval Conference (TREC-6)*, pages 347–366, 1997.

[SM83] G. Salton and McGill. *Introduction to Modern Information Retrieval*. Mc Graw-Hill, New York, 1983.

[Sma93a] F. Smadja. Retrieving collocations from text: Xtract. *Computational Linguistics*, 19(1):143–177, 1993.

[Sma93b] F. Smadja. Xtract: an overview. *Computer and the Humanities*, 26:399–413, 1993.

[Sme89] Alan F. Smeaton. Information retrieval and natural language processing. In *proceedings of a conference jointly sponsored by ASLIB, University of York,*, page 2, Mars 1989.

[SOK94] A. Smeaton, R. O'Donnell, and F. Kelledy. Indexing structures derived from syntax in trec-3: System description. In *Donna K. Harman, editor, The Third Text REtrieval Conference (TREC-3)*, pages 55–67, 1994.

[SR88] F. Smeaton and C. V. Rijsbergen. Experiments on incorporating syntactic processing of user queries into a document retrieval strategy. In *Proceedings of the 11th International Conference on Research and Development in Information Retrieval, ed. Y. Chiaramella, Grenoble, France*, pages 31–51, Juin 1988.

[SSWB00] T. Strzalkowski, G. C. Stein, G. B. Wise, and A. Bagga. Towards the next generation information retrieval. In *6ème Conférence de Recherche d'Information Assistée par Ordinateur (RIAO'2000), Collège de France, Paris, France*, pages 1196–1207, Avril 2000.

[Str93] T. Strzalkowski. Natural language processing in large-scale text retrieval tasks. In *Text REtrieval Conference (TREC-1)*, page 173, 1993.

[SW75] G. Salton and A. Wong. On the role of word and phrases in automatic text analysis. Technical report, departement of computer science, Cornell university, Ithaca, 1975.

[TR98] W. Theilmann and K. Rothermel. Domain experts for information retrieval in the world wide web. In *Lecture Notes in Artificial Intelligence 1435, Springer-Verlag, Berlin, Heidelberg, New York*, pages 216–227, July 1998.

[TRE] http://trec.nist.gov.

[VG01] D. Vaufreydaz and M. Géry. Internet evolution and progress in full automatic french language modelling. In *IEEE Workshop on Automatic Speech Recognition and Understanding, Madonna di Campiglio Trento, Italie*, Décembre 2001.

[Voo93] E. M. Voorhees. Using wordnet to disambiguate word senses for text retrieval. In *Annuel International ACM SIGIR conference on research and development in information retrieval (SIGIR'93), ACM press, pittsburgh, USA*, pages 171–180, July 1993.

[Voo94] E. M. Voorhees. Query expansion using lexical-semantic relations. In *Annuel International ACM SIGIR conference on research and development in information retrieval (SIGIR'94), ACM press, Dublin, Irelande*, pages 63–69, Juillet 1994.

[Voo98] E. Voorhees. Using wordnet for text retrieval. In *C. Fellbaum, (Ed.), WordNet: An Electronic Lexical Database, Cambridge, Massachusetts, USA: The MIT Press*, pages 285–303, 1998.

[Voo99] E. Voorhees. Natural language processing and information retrieval. In *M. T. Pazienza, (Ed.), Information Extraction: Towards Scalable, Adaptable Systems, Springer*, pages 32–48, 1999.

[WBH+00] W. A. Woods, L. A. Bookman, A. Houston, R. J. Kuhns, P. Martin, and S. Green. Linguistic knowledge can improve information retrieval. In *Sixth Annual Applied Natural Language Processing Conference*, pages 262–267, Mai 2000.

[Wu97] F. Wu. Algorithme génétique et data mining, Octobre 1997. Rapport de DEA, Université de versaille.

[XC00] Jinxi Xu and W. Bruce Croft. Improving the effectiveness of information retrieval with local context analysis. *ACM Transactions on Information Systems*, 18(1):79–112, 2000.

[Yah] http://www.yahoo.com.

[YP97] Y. Yang and J. O. Pedersen. A comparative study on feature selection in text categorization. In *Proceedings of the Fourteenth International Conference on Machine Learning (ICML'97)*, pages 412–420, Juillet 1997.

[Zak99] M. J. Zaki. Parallel and distributed association mining: A survey. *IEEE Concurrency*, 7(4):291–298, 1999.

[ZTME96] C. Zhai, X. Tong, N. MilicFrayling, and D. Evans. Evaluation of syntactic phrase indexing - clarit nlp track report. In *E. Voorhees and Donna K. Harman, editor, The Fifth Text REtrieval Conference (TREC-5)*, pages 347–358, 1996.