EXPLOITATION DE CONNAISSANCES SÉMANTIQUES EXTERNES DANS LES REPRÉSENTATIONS VECTORIELLES EN RECHERCHE DOCUMENTAIRE

THÈSE N^{O} 3654 (2006)

PRÉSENTÉE LE 14 DÉCEMBRE 2006

À LA FACULTÉ INFORMATIQUE ET COMMUNICATIONS

Laboratoire d'intelligence artificielle

SECTION D'INFORMATIQUE

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

Florian SEYDOUX

ingénieur informaticien diplômé EPF

de nationalité suisse et originaire de Hermance (GE)

acceptée sur proposition du jury:

Prof. K. Aberer, président du jury

Dr M. Rajman, Dr J.-C. Chappelier, directeurs de thèse

Prof. E. Gaussier, rapporteur

Prof. E. Sanchez, rapporteur

Prof. J. Savoy, rapporteur

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Lausanne, EPFL

2006

Remerciements

Bien qu'une thèse soit généralement considérée comme un travail personnel, peu pourraient aboutir sans le concours bienveillant de nombre de personnes.

Cette thèse ne fait pas exception, loin s'en faut.

Mes remerciements vont en premier lieu à Jean-Cédric Chappelier, qui ne s'est pas contenté de m'encadrer durant ce travail, mais en a de plus insufflé l'idée ; il a non seulement su me tirer de mes errements mais m'a également motivé et soutenu tout au long de ces années, en sachant se montrer humain et compréhensif dans les moments difficiles.

Mes remerciements vont également à Martin Rajman, qui a veillé, avec le recul nécessaire, à la réussite de ce travail, et m'a permis de découvrir l'univers TALN au cours d'années passionnantes et riches d'enseignements.

Je remercie également :

Les Professeurs Karl Aberer, Jacques Savoy, Eric Gaussier et Eduardo Sanchez, qui ont accepté de me faire l'honneur de participer à mon jury de thèse.

Mes collègues et toute l'équipe du LIA, en particulier Alex Trutnev, Antoine Rozenknop et Romaric Besançon, qui m'ont montré la voie à suivre, ainsi que Marie Decrausat, pour sa disponibilité et son efficacité de tous les instants.

Mes parents, dont le soutien notamment logistique fut plus déterminant qu'il n'y paraît. Ma sœur, qui a impitoyablement traqué les innombrables fautes d'orthographe semées tout au long de ce document (je reste néanmoins persuadé d'avoir réussi à en faire subsiter quelques-unes).

Mes amis, en particulier Marc Martin et Florian Steffen, pour ces moments de détente, de dérision, et juste ce qu'il faut de sport.

Ce n'est malheureusement qu'à titre posthume que je peux témoigner de mon infinie gratitude envers celle qui m'aura supporté (dans tous les sens du terme) ces années durant. À toi, Régine, je dédie ce travail. C'est la moindre des choses, si l'on songe à tous ces moments dont il nous aura privé, tellement précieux et hélas si rares.

Version abrégée

Résumé en français

Les travaux présentés dans ce mémoire de thèse traitent d'un certain nombre de problématiques rencontrées en recherche documentaire (RD), tâche que l'on peut résumer comme consistant à identifier, dans une collection de documents (au sens large), celui ou ceux porteurs d'une information recherchée, i.e. pertinents par rapport à une requête exprimée par un utilisateur. Dans le cas de documents de nature textuelle, auxquels nous nous sommes limités dans le cadre de cette thèse, une part importante de la difficulté réside dans l'ambiguïté inhérente aux langues humaines. L'interaction avec l'utilisateur est également abordée dans notre travail, par l'étude d'un outil d'accès en langage naturel à une base de données. Finalement, quelques techniques permettant la visualisation des bases documentaires de grande taille sont présentées.

Dans ce mémoire, nous décrivons tout d'abord les principaux modèles de RD, en mettant en évidence les relations qui existent avec les techniques manuelles de RD et de recherche en document, développées au cours des siècles. Nous présentons notamment le principe de l'indexation des documents, permettant de représenter ces derniers dans un espace multidimensionnel, et l'utilisation de cette représentation par le modèle vectoriel. Après avoir passé en revue les principales améliorations apportées ces dernières années aux systèmes de recherches vectoriels, tant sur le plan des pré-traitements des collections, du mécanisme d'indexation, et des mesures de similarité entre documents, nous détaillons les cas récents d'utilisation de ressources sémantiques additionnelles (dictionnaires, thésaurus, réseaux sémantiques, ontologies) rapportées dans la littérature scientifique, en particulier dans une optique d'indexation.

Nous présentons ensuite plus en détail le principe d'indexation sémantique de documents textuels à partir de thésaurus, consistant à intégrer dans l'espace de représentation des documents une partie au moins du contenu informationnel de ressources sémantiques hiérarchisées. Nous proposons un cadre général permettant de décrire et positionner différentes techniques envisageables pour réaliser l'indexation sémantique, en adaptant si possible la richesse des descriptions issues des ressources sémantiques aux données à représenter. Nous utilisons ce cadre pour dégager trois familles de critères utilisables pour l'indexation sémantique, chacune ayant ses particularités propres. Pour chacune de ces familles, nous donnons les algorithmes permettant la mise en œuvre des critères.

Les deux premières familles permettent de considérer plusieurs critères déjà connus de sélection de termes. Nous montrons en outre que bon nombre de ces critères ne sont en fait que peu efficaces pour la tâche considérée. La troisième famille nous permet d'introduire un critère totalement nouveau, le critère de *coupe de redondance minimale* (*CRM*), construit sur la base de la théorie de l'information, et permettant d'obtenir des termes d'index ayant une probabilité d'occurrence dans la collection de documents la plus équilibrée possible. Finalement, nous traitons le cas d'index sémantiques indépendants des données (déterminés statiquement), avec paramétrisation du degré de généralité des index.

Une partie des critères proposés pour l'indexation sémantique fait l'objet d'une évaluation empirique, évaluation qui est présentée à la suite. Pour juger de la pertinence de ces critères, nous avons utilisé un système vectoriel largement répandu (le système *Smart*) et avons mesuré les performances de RD obtenues sur un certain nombre de collections de références, indexées sur la base de ces critères, en prenant en compte la relation sémantique fortement structurante d'hyper/hyponymie « est-un » issue de deux ressources sémantiques différentes. En confrontant les résultats obtenus, et en les comparant aux performances d'une indexation traditionnelle (utilisant les lemmes des mots des documents comme espace de représentation), nous pouvons conclure d'une part à la pertinence des indexations sémantiques en RD, et d'autre part à la qualité indéniable de notre critère *CRM*.

En matière d'interaction homme-machine, nous présentons un schéma général permettant de construire de manière relativement rapide et systématique des systèmes à initiative mixte, laissant à l'utilisateur humain une large latitude dans la conduite du dialogue. Ce schéma est à la fois utilisable dans des applications typiques de recherche d'information dans une base de données (la base est cachée à l'utilisateur, mais celui-ci sait exactement quelle information il désire) et dans des applications de conseils, pour lesquelles l'utilisateur n'a pas nécessairement d'idée précise sur ce qu'il désire, et attend de la part du système non seulement qu'il l'aide à préciser ses souhaits, mais également un ensemble de propositions comme résultat final. Nous mettons en particulier l'accent sur les techniques permettant d'obtenir un système robuste, capable de pallier dans une large mesure les erreurs de reconnaissance vocale.

En matière de visualisation de grandes collections de données textuelles, nous présentons une application de l'analyse des correspondances (permettant de mettre en évidence des similitudes ou des oppositions entre différents groupes, construits sur la base des traits additionnels) au cas de données issues de bases de brevets. Il est ainsi possible de déterminer, pour divers groupes (pays, sociétés, etc), les éléments spécifiques communs à certains de ces groupes (similitudes), ou au contraire les opposant (différences). Nous proposons par ailleurs une méthode (basée sur le principe de réplication *bootsrap*) permettant de déterminer un intervalle de confiance pour les positionnements relatifs des différents groupes, et ainsi de juger immédiatement de la fiabilité des similitudes ou oppositions visuellement apparentes. Ces outils sont utilisés dans le cadre d'une méthodologie d'analyse de bases de brevets, permettant de réaliser des comparaisons multi-critères de l'activité « d'innovation » de différents pays, de différents secteurs d'activité ou encore de grandes companies. Ils présentent également un intéret pour l'identification de concurrents dans un secteur donné, ou l'étude des interactions pouvant exister entre différents domaines d'activité technologique ou différents pôles d'innovation à l'intérieur de ces domaines.

Mots clefs

Recherche Documentaire, Coupe de Redondance Minimale, Indexation Sémantique, WordNet, EDR, Dialogue Homme-Machine, Analyse Factorielle, Analyse de Données

English Summary

The work presented in this thesis deals with several problems met in information retrieval (IR), task which one can summarise as identifying, in a collection of "documents", a subset of documents carrying a sought information, *i.e.*. relevant for a request expressed by a user.

In the case of textual documents, to which we limited ourselves within the framework of this thesis, a significant part of the difficulty lies in ambiguity inherent to human languages. The interaction with the user is also approached in our work, by studying a tool enabling a natural language access to a database. Finally, some techniques which permit the visualisation of large collections of documents are also presented.

In this document we first of all describe the principal models of IR by highlighting the relations which exist with some manual technics of IR and document retrieval, developed during the past centuries. We present the principle of document indexing, allowing us to represent documents in a multidimensional space, and the use of this representation by a vectorial model. After having reviewed the principal improvements made these last years with vectorial research systems, including the preprocessings of collections, the indexing mechanism and measurements of similarities between documents, we detail some recent usecases of additional semantic resources (semantic dictionaries, thesaurus, networks, ontologies) reported in scientific literature for the indexing task.

We then present more in detail the semantic indexing principle of textual documents by using a thesaurus, consisting in integrating in the document's representation space at least part of the informational contents of hierarchical semantic resources. We propose a general framework allowing us to describe and position various possible techniques to carry out the semantic indexing by adapting, if possible, the specificity of the descriptions resulting from the semantic resources to the data to be represented. We use this framework to describe three families of criteria usable for semantic indexing, each one having its own characteristics. For each of these families, we give the specific algorithms allowing the computation of the criteria. The first two families allow us to consider several criteria already known in feature selection. Moreover we show that, unfortunately, many of these criteria are in fact not very effective for the considered task. The third family allows us to introduce a completely new criterion, the *Minimum Redundancy Cut* criterion (MRC), built on the basis of the information theory and allowing us to obtain index terms having a probability of occurrence in the collection of documents as well balanced as possible. Finally, we treat the case of semantic index independent of the data (statically choosen), allowing a parameterisation of the level of generality of the index terms.

Some of the criteria suggested for semantic indexing has been empirically evaluated. To judge their relevance, we used a well known vectorial system (the Smart IR system) and measured the performances of IR obtained with various reference collections. Those collections was indexed on the basis of the studied criterion, by taking into account the strongly structuring semantic relation of hyper/hyponymy ("is-a" relation), given by two different semantic resources. By comparing results obtained with the performances of a traditional indexing (using the lemmas of the words as representation space), we can show on one hand the relevance of the semantic indexings (in RD) and on the other hand the quality of the proposed criterion (MRC).

Concerning man-machine interaction, we present a general outline allowing to build in a relatively fast and systematic way systems with mixed initiative, giving the human user a large (and natural) latitude in the control of the dialogue. This outline is usable in typical database research-task applications (where the database is hidden to the user, but the latter knows exactly which information they wish to find) as well as advice-task applications, for which the users does not necessarily have a precise idea of their needs, and uses the system not only for specifing their wishes, but also a set of propositions as a final result. We particularly stress the techniques allowing us to obtain a robust system, able to deal with speech recognizer failures.

Concerning the visualisation of large textual data collections, we present an application of the correspondences analysis (allowing to highlight similarities and oppositions for various groups of entity, built on the basis of additional features present in the DB) to the case of patents data. In addition, we propose a method (based on the bootstrap replication principle) allowing us to determine a confidence interval for relative positionings of various groups, thus permit to immediately judge the reliability of the visually apparent similarities or oppositions.

Keywords

Information Retrieval, Minimum Redundancy Cut, Semantic Indexing, WordNet, EDR, Human-Machine Dialogue, Factorial Analysis, Data Analysis

Table des matières

Notations mathématiques

Les notations utilisées dans ce document sont le plus souvent définies (ou du moins explicitées) lorsqu'elles sont introduites pour la première fois ; néanmoins, afin de faciliter une lecture non linéaire, elles sont également regroupées et définies de manière formelle ci-après.

NOTATION :

$x \triangleq \ldots$: valeur de x, par *définition*.

$x \widehat{=} \ldots = \widehat{x}$: valeur de x, par *estimation* ; $\widehat{x}$ est une estimation de x.

$A \setminus x$: ensemble A privé de l'élément x (où des éléments de x, si ce dernier est un ensemble).

$D = \{d_1, \ldots, d_n\}$:
un ensemble de documents, où un document d_i est une entité à indexer ; $|D|$ est le nombre d'éléments de cet ensemble ($|D| = n$).

$\mathcal{M} = \{m_1, \ldots, m_n\}$:
un ensemble de *mots*, où un mot m_i est une entité constitutive des données, considérée unitaire pour l'indexation (élément de vocabulaire) ; en pratique, il s'agira souvent de lemmes, voir de racines (*stems*).

$O = \{o_1, \ldots, o_n\}$:
un ensemble d'*occurrences* de mots (quelconques), où o_i est une de ces occurrences.

$\mathcal{C} = \{c_1, \ldots, c_n\}$:
un ensemble de *concepts*, où un concept c_i est un élément abstrait externe aux données, relatif à un ou plusieurs mots (éventuellement de manière indirecte, par le biais d'autres concepts).

$G = [\mathcal{S}, \mathcal{R}]$: un graphe, avec $\mathcal{S} = \{s_1, \ldots, s_n\}$ l'ensemble des sommets (nœud) et $\mathcal{R} = \{r_1, \ldots, r_n\}$ l'ensemble des arrêtes (arcs).

$\Upsilon = \{\Gamma_1, \ldots, \Gamma_n\}$:
un ensemble de *coupes* (Γ_i) dans un graphe (arbre ou DAG) (c.f. définition 3.5.1 [page 34]).

$M = (\Gamma, \theta)$: un modèle probabilisé de coupe, avec θ un vecteur de probabilités associées aux éléments de Γ (c.f. définition 3.5.1 [page 34]).

$s^{\downarrow}$; $s_r^{\downarrow}$: l'ensemble des sommets successeurs (*dominés*) de s ; respectivement l'ensemble des sommets successeurs de s via un arc de l'ensemble r.

$s^{\uparrow}$; $s_r^{\uparrow}$: l'ensemble des sommets prédécesseurs (*dominants*) de s ; respectivement l'ensemble des sommets prédécesseurs de s via une arrête de l'ensemble r.

$s^{\Downarrow}$; $s_r^{\Downarrow}$: l'ensemble des sommets résultant de la fermeture transitive de $s^{\downarrow}$; respectivement, l'ensemble des sommets successeurs de s via une arrête de l'ensemble r.

$s^{\Uparrow}$; $s_r^{\Uparrow}$: l'ensemble des sommets résultant de la fermeture transitive de $s^{\uparrow}$.

$\text{rang}(n, E, \mu_r) \in [0, |E| - 1]$:
le nombre de successeurs du minorant de l'ensemble E à parcourir pour atteindre l'élément $n \in E$, où E est un ensemble totalement ordonné par la relation d'ordre r sur la mesure μ (appliquée aux éléments de E). $\text{rang}(n, E, \mu_r) = 0 \Leftrightarrow n$ est le minorant de E muni de la relation d'ordre total μ_r.

$\mathcal{F}(\mathcal{A}, \mathcal{B})$: l'ensemble des applications de type $f : \mathcal{A} \to \mathcal{B}$, faisant correspondre à tout élément de $\mathcal{A}$ un et un seul élément de $\mathcal{B}$ (injection).

$\chi_{a,b}(x)$: fonction caractéristique d'un domaine ou d'un ensemble de valeur :

$$\chi_{[a,b]}(x) = \begin{cases} 1 & \text{si } x \in [a, b], \\ 0 & \text{sinon.} \end{cases}$$

$\delta_{x,y}; \delta(x, y)$: fonction (symbole de Kronecker) testant l'égalité de deux arguments :

$$\delta_{x,y} = \delta(x, y) = \begin{cases} 1 & \text{si } x = y, \\ 0 & \text{sinon.} \end{cases}$$

La seconde forme est utilisée en présence d'arguments longs.

Chapitre 1

Introduction

À mon épouse Régine, incarnation du courage et de la candeur (09/11/1971 – 01/11/2005)

« *Mets-moi comme un sceau sur ton cœur, Comme un sceau sur ton bras ; [...]* »

Cantique des Cantiques (8.6)

1.1 Problématique

Consacrée à la fin du 20e siècle, l'hégémonie de l'ordinateur comme outil central pour la gestion de « l'information », quelqu'en soit sa nature ou presque (audio, vidéo, ... et naturellement textuelle) n'est plus à démontrer. La disponibilité croissante des données, grâce aux technologies réseaux et notamment *internet*, résulte en d'immenses collections de *documents* accessibles aussi bien au spécialiste qu'au simple profane.

Mais l'accessibilité de ces documents n'est en elle-même pas suffisante pour en permettre l'exploitation. Elle en devient même un obstacle : en effet, pour qu'un document puisse être « exploité » (pour que l'information qu'il contient puisse être portée à la connaissance de celui qui la cherche), encore faut-il qu'il soit *connu*. Or la profusion des documents accessibles conduit à noyer dans un fond documentaire (non structuré) ceux potentiellement intéressants ; le cas d'*internet* est révélateur : une étude menée en 2000 par la compagnie *BrightPlanet* concluait qu'environ 1/500e de l'information disponible sur internet était indexée par les moteurs traditionnels de recherche [Sullivan, 2000].

Les techniques initialement mises en œuvre pour permettre la recherche d'information dans les premières bases documentaires informatisées se résumaient le plus souvent à la simple traduction des procédés utilisés dans les bibliothèques, à savoir essentiellement la structuration des documents de manière hiérarchique et/ou thématique, ainsi que l'adjonction aux documents de descripteurs en nombre réduit, permettant une recherche de type « base de données ». Mais ces procédés, coûteux à réaliser de manière manuelle, ne sont plus envisageables à l'avenir, notamment en raison de la masse considérable de données à traiter, et de leur forte volatilité.

Des techniques automatiques de structuration, de classification et d'indexation ont donc été mises au point, et constituent les éléments clés pour la création de systèmes d'information effectivement utilisables à grande échelle.

Dans ce cadre, l'arrivée du modèle vectoriel marque le passage d'une recherche de type « base de données » à une recherche par « similarité documentaire », permettant de retrouver, en partant d'un document donné, ceux qui, dans la collection, lui sont le plus « similaire ». Cette manière de procéder offre de nombreux avantages, détaillés en section 2.2.2 [page 8], notamment celui

de permettre d'utiliser les éventuels documents pertinents obtenus au cours d'une première phase de recherche.

Néanmoins, les techniques utilisées jusqu'à présent – en particulier pour l'indexation des documents et les calculs de similarités – ne permettent pas d'exploiter tout le potentiel de ce modèle. Les limitations sont particulièrement évidentes dans le cas de documents de type audio ou vidéo, pour lesquels ces techniques ne sont simplement pas applicables, l'information présente dans les documents échappant en effet totalement au modèle. Il est dès lors nécessaire, pour pouvoir traiter des documents de cette nature, de munir le modèle d'une représentation adéquate des données contenues dans les documents, et de disposer de moyens permettant de calculer automatiquement ces représentations. Ces limitations se retrouvent également (quoique dans une moindre mesure) en présence de documents textuels, du fait de leur pouvoir expressif limité, mal adapté à la complexité de l'information contenue dans les documents (notamment, ces techniques traitent essentiellement le *signifiant*, mais très peu le *signifié*).

On constate de plus dans les campagnes d'évaluation de systèmes de recherche documentaire (*TREC*, NTCIR, etc.) que la plupart des systèmes ont atteint un plateau de performances, alors que la marge d'amélioration est encore grande (selon les mesures utilisées pour l'évaluation) ; ceci qui tend à indiquer que les optimisations (de nature essentiellement statistiques) du modèle ont atteint leurs limites. Clairement, les systèmes vectoriels tels que conçus jusqu'à présent ne prennent pas suffisamment d'informations en compte, tant pour la représentation des documents que pour l'évaluation de leurs similarités.

Parallèlement, on peut constater une disponibilité croissante de ressources électroniques de qualité, de type dictionnaire, thésaurus, ou réseau sémantique. Un certain nombre d'entre elles sont mêmes devenues incontournables ; c'est souvent le cas dans des domaines spécialisés – par exemple les thésaurus *MeSH* et *UMLS* pour le domaine médical – mais des ressources généralistes connaissent également un franc succès, à l'image de l'incontournable projet *WordNet*. Ce succès s'explique en premier lieu par le besoin de référentiel commun, le moins ambigu possible (vocabulaire contrôlé, définitions unanimement admises), mais également par l'accessibilité (incluant le coût et l'utilisabilité) de ressources de qualité ; ce phénomène s'amplifie et se nourrit de lui-même, grâce à l'émergence de nouvelles ressources dans des domaines de plus en plus divers (génomique, droit, machines-outils, etc.). Néanmoins, l'utilisation de ces ressources reste le plus souvent à un niveau proche de l'utilisateur, et leur intégration dans les couches profondes des outils de traitement automatique de la langue, notamment en recherche documentaire, reste encore marginale.

L'objectif principal des travaux présentés dans ce mémoire vise donc l'intégration, dans la représentation des documents d'un modèle vectoriel, d'une partie au moins de l'information contenue dans des ressources sémantiques hiérachisées. Pour réaliser cette intégration, nous avons choisi de remplacer la représentation usuelle des documents au moyen d'un espace de « mots » (qu'il s'agisse de graphies, racines ou lemmes) par une représentation dans un espace incluant les éléments conceptuels définis dans ces ressources (en particulier la hiérarchisation).

Une telle modification de l'espace de représentation, bien que paraissant simple dans son principe, n'est pas sans poser un certain nombre de difficultés. Outre les problèmes usuels de mise en correspondance des informations documents-ressources, sur lesquels nous ne nous attardons pas dans ce travail (ils représentent à eux seuls un vaste champ d'étude : la problématique de la désambigüisation sémantique), se pose notamment la question de la quantité et de la qualité de l'information additionnelle à conserver pour la représentation.

Nous avons envisagé différentes possiblités pour répondre à ces questions et avons en particulier étudié l'utilisation d'un critère original issu de la théorie de l'information : la *Coupe de Redondance Minimale* (*CRM*). Nous avons évalué l'intéret de ce critère appliqué au modèle vectoriel dans le cadre d'une tâche de recherche documentaire *ad hoc*, en comparant l'utilisation de deux ressources sémantiques différentes : la ressource à large couverture *WordNet* d'une

part, et d'autre part le dictionnaire *EDR*, constitué à la fois d'un réseau sémantique à large couverture et d'un réseau spécialisé. Les résultats que nous avons obtenus sont prometteurs et montrent clairement que l'intégration de ressources de ce type pour la représentation des documents est une piste pour améliorer les performances des outils de recherche documentaire. Par ailleurs, le critère proposé pour réaliser cette intégration offre une souplesse intéressante pour l'intégration de ressources multiples.

Un second objectif de nos travaux concerne deux domaines connexes à la recherche documentaire et d'une grande importance pour la mise en œuvre effective d'un système : l'interaction homme-machine dans le cadre d'une application multimodale de recherche interactive d'information, et diverses techniques d'analyse et de représentation visuelle de grandes bases de documents, s'appuyant à la fois sur le contenu textuel des documents et sur un ensemble fixe de descripteurs.

En effet, en plus du problème de l'identification d'un document pertinent dans une collection de grande taille, la qualité d'un processus de recherche dépend dans une large mesure de la manière dont s'effectuent les interactions entre l'utilisateur et le système, tant durant la phase de recherche proprement dite que lors de la transmission des résultats à l'utilisateur. Ainsi, Calvin N. Mooers (qui introduisait en 1948 le terme de « recherche d'information » – *information retrieval*) est à l'origine de l'énoncé suivant, illustrant parfaitement le défi auquel doit répondre un système de recherche documentaire dans son ensemble : « *La recherche d'information est arrêtée dès l'instant où il est plus pénible à l'usager de rechercher cette information que de s'en priver* »[Mooers, 1960].

1.2 Contexte du travail

Les travaux présentés dans la première partie de ce mémoire, portant sur l'indexation sémantique, ont été (partiellement) financés par le Fond National Suisse pour la Recherche Scientifique,[1] et constituent donc un projet de recherche pure, réalisé de manière autonome.

Les travaux de la seconde partie ont, pour leur part, été réalisés dans un contexte totalement différent. La problématique de l'interaction homme-machine a été abordée dans le cadre d'un projet CTI/KTI (Commission Suisse pour la Technique et l'Innovation),[2] en collaboration avec l'institut de recherche IDIAP (Martigny, VS) et un certain nombre de partenaires industriels[3] ; contexte induisant naturellement un certain nombre de contraintes, liées à la fois à l'aspect collaboratif du projet et aux attentes des partenaires privés.

Les travaux effectués sur la représentation des bases de grandes tailles ont quant à eux été réalisés dans le cadre d'un projet Européen[4] visant à définir un certain nombre d'indicateurs statistiques permettant de mesurer l'innovation en Europe (par le biais des brevets qui y sont déposés), et réalisés en collaboration avec de nombreux partenaires académiques, institutionnels et industriels, répartis à travers l'Europe.

---

[1] Projet *EXKNOWTIC*, « *Intégration de sources de connaissances pour l'amélioration des modèles à base de sémantique distributionnelle*, Fond National Suisse n° 2100-066901 & 200020-103529, EPFL.

[2] Projet *InfoVox*, « *Interactive Voice Servers for Advanced Computer Telephony Application* », CTI n° 4247.1, EPFL & IDIAP.

[3] Swisscom SA, VoxCom SA, Omédia SA.

[4] Projet *Sting*, « *Evaluation of Scientific & Technological Innovation and Progress in Europe through Patents* », EU IST99-20847, Computer Technology Institute (Patras).

Chapitre 2

Recherche Documentaire

Résumé

La recherche documentaire, tâche que l'on peut résumer comme consistant à identifier, dans une collection de documents (au sens large), celui ou ceux porteurs d'une information désirée, est une problématique ancienne, apparue pratiquement en même temps que l'écriture. Au fil du temps s'est ajouté à cette problématique celle de la recherche en document (consistant à identifier dans un document volumineux le ou les segments porteurs de l'information recherchée). Les principes alors mis en œuvre pour résoudre ces deux problématiques sont, dans leurs grandes lignes, encore utilisés à ce jour. Néanmoins, tant la généralisation du support numérique que le volume considérable des collections actuelles ont rendu indispensable une automatisation de ces techniques, dans le cadre de vastes systèmes de recherche informatisés. Cette transposition est cependant loin d'être évidente, l'efficacité des méthodes de recherche offertes étant directement liée au degré de « compréhension », par le processus automatique, de l'information contenue dans les documents traités. Dans le cas de documents de nature textuelle (auxquels nous nous limiterons dans le cadre de cette thèse), une part importante de la difficulté réside dans l'ambiguïté inhérente aux langages humains.

Ce chapitre débute par une synthèse des principaux moyens utilisés par le passé pour répondre aux problèmes posés par la recherche documentaire et la recherche en document, et met en évidence les liens existant entre ces techniques et les approches automatisées actuelles. S'en suit une présentation des principaux modèles informatiques de recherche documentaire, notamment le modèle vectoriel, utilisé dans nos travaux. La suite du chapitre présente un état de l'art, centré sur le modèle vectoriel, des améliorations et pistes ayant fait l'objet de publications dans la littérature scientifique de ces dernières années. Sont regroupées d'une part les améliorations (essentiellement statistiques) ne nécessitant par de connaissances « additionnelles » (externes) aux documents (critère de sélection des termes d'index, indexation sémantique latente, pondération des éléments d'un profil d'indexation, etc.), et d'autre part les améliorations faisant usage d'informations externes (par exemple de dictionnaires). La technique d'indexation sémantique au moyen d'un thésaurus (réseau sémantique) est notamment abordée, ainsi que les résultats obtenus par divers chercheurs dans leurs mises en œuvre de cette technique.

2.1 Introduction

La *Recherche Documentaire* s'inscrit dans le cadre plus vaste de la *Recherche d'Information* ; elle s'applique lorsque l'on est en présence d'une collection de *documents*, et consiste globalement à aider l'utilisateur à identifier, dans une collection, un sous-ensemble de documents susceptibles de l'intéresser. Elle n'est le plus souvent qu'une réponse partielle au besoin d'information de l'utilisateur, la tâche de recherche de l'information proprement dite étant du ressort de ce dernier (ou d'un autre système, comme c'est le cas des outils de réponse à des questions).

Historique

La problématique de retrouver un « document » parmi un ensemble s'est posée pratiquement dès les débuts de l'écriture.[1]La méthode simple (simpliste ?) consistant à parcourir l'ensemble des « documents » à la recherche de celui ou ceux désirés s'avérant fastidieuse (et potentiellement risquée pour les documents) lorsque la taille ou le nombre des documents est conséquent, des techniques plus efficaces ont été employées : rangement en fonction d'une classification (le plus souvent relative à la nature du document), étiquettage du contenu, etc. On relevera que la plupart de ces techniques apportaient une réponse similaire à celles (encore utilisées) en recherche de données, à savoir le recours à un ensemble de « descripteurs » (informations bibliographiques le plus souvent, telles que auteur ou nature du document) associés à une « localisation » (ou tout du moins une identification) du document correspondant, soit par le biais d'un référencement (cas des colophons de Mésopotamie), soit en étant physiquement relié au document (cas des inscriptions murales et étiquettes des rouleaux de la bibliothèque d'Edfou, en Égypte) ; c'est-à-dire un *index*.

Au Moyen Âge, afin de faciliter les études théologiques, on constitua des index des mots importants de la Bible et des Évangiles. On assiste donc à l'émergence de deux types d'index (mais remplissant un rôle similaire) : ceux permettant de structurer la connaissance, essentiellement dans un but de classement des documents (recherche *documentaire*), et ceux liés à un document volumineux, facilitant la localisation de passages spécifiques (recherche *en document*).

Avec l'apparition de l'informatique, l'automatisation des processus de recherche documentaire s'est rapidement imposée ; pour exprimer son besoin d'information, l'utilisateur soumet une *requête* au système, à laquelle celui-ci répond par la mise à disposition des documents qui y correspondent (ou du moins identifier comme tels par le système). Utilisée pour décrire les documents (potentiellement) recherchés par l'utilisateur, la requête peut revêtir plusieurs formes, selon le modèle mise en œuvre par le système.

Le rôle de l'index change avec la notion de *requête* ; il s'agit désormais de permettre, tant pour les documents que pour les requêtes, de fournir des descriptions de leurs contenu, qui soit comparables entre elles par le système.

Le cas particulier de la recherche de documents *textuels*, pour lesquels l'information n'est pas (ou très peu) structurée – du moins vu d'un système informatique – et fréquemment ambigüe, s'accorde mal avec les techniques d'appariement exact (entre requête et index), initialement développées pour la recherche en base de données. Il faudra néanmoins attendre la fin des années soixante pour que soit proposée une alternative au modèle *booléen*[2] : le modèle *vectoriel* [Salton, 1968][3] permettant des appariements partiels, grâce auxquels il n'est plus nécessaire de préciser de manière exhaustive les « caractéristiques » des documents recherchés.

[1] Communément admis aux environs de 4000 av. J.-C., en Égypte et en Mésopotamie.

[2] Mettant essentiellemnt en œuvre des principes de recherche en base de données.

[3] Mais aussi Van Rijsbergen (modèle probabiliste), Sparck-Jones (idf), etc.

Avec la multiplication des données informatisées, l'indexation des documents s'est par ailleurs largement automatisée [Luhn, 1957] et l'amélioration des méthodes automatiques d'indexation porte en elle une large part des gains de performances observés ces dernières années.

Un portrait succinct (mais complet) de l'évolution à travers l'Histoire des techniques d'indexation et de recherche documentaire est brossé dans de Loupy [2000].

2.2 Principaux modèles

Les modèles de recherche documentaire les plus courants sont brièvement présentés ici, sous leur forme standard ; quelques-unes des variantes et améliorations apportées à ces modèles sont détaillées dans les sections suivantes (2.4 [page 10] et 2.5 [page 17]). Dans leur version standard, ces modèles mettent habituellement en œuvre le principe du « *sac de mots* », pour lequel seul compte les occurences des mots dans les documents et les requêtes, l'ordre de ces mots n'étant pas considéré.

2.2.1 Modèle *booléen*

Le premier modèle utilisé en recherche documentaire informatisée est le modèle *booléen*.

Mortimer Taube propose, en 1953, le principe d'indexation par *unitermes* (mots clefs), accompagné d'un langage de classifiation à structure combinatoire. L'index des documents se présente sous la forme d'un *fichier inversé*,[4] associant à chaque *uniterme* du jeu d'indexation l'ensemble des documents le contenant.

La requête de l'utilisateur est constituée en combinant des unitermes entre eux, au moyen des opérateurs booléens NON, ET et OU. Le système considérera comme *pertinents* (*i.e.* répondant au besoin d'information de l'utilisateur tel qu'exprimé par la requête) l'ensemble des documents satisfaisant l'expression logique constituée par la requête.

Le modèle booléeen est un modèle à *appariement exact*, avec une valuation binaire de la pertinence ; la réponse à une requête est donc un sous-ensemble de la collection de documents. Outre son efficacité, permettant un temps de réponse très faible, pratiquement indépendant de la taille de la collection considérée, la simplicité du modèle et sa prédictibilité en ont fait son succès (écriture de requêtes structurées au moyen d'une grammaire très simple, bien que parfois contre-intuitive).

Néanmoins, ce modèle implique, pour donner satisfaction, que l'utilisateur connaisse bien à la fois le jeux d'indexation et les documents qu'il cherche. En effet, l'impossibilité de retourner des documents ne répondant que partiellement à la requête (ainsi que la grande subjectivité dans le choix de termes à employer pour indexer un document,[5] conduit souvent à une recherche ne donnant soit aucun résultat (requête trop contrainte ou emploi de terme hors-lexique) soit beaucoup trop de documents (requête insuffisamment contrainte). Par ailleurs, le modèle ne permettant pas de déterminer un ordre sur les documents jugés pertinents, l'utilisateur est obligé d'examiner l'ensemble des réponses pour trouver les documents qu'il cherche.

Outre les améliorations portant sur la constitution des index[6] (voir 2.4 [page 10] et 2.5 [page 17]), les principales extensions du modèle portent sur le langage de formulation des requêtes : opérateurs supplémentaires, tel ADJ, permettant d'imposer un ordre sur l'occurrence de deux

[4] Un *fichier inversé* ou *index inversé* est un mécanisme permettant d'indexer un texte ou une collection par les mots contenus (qui représente l'entrée de l'index), à la manière d'un index présenté à la fin d'un ouvrage.

[5] Qui transparait dans le faible degré d'accord entre deux humains réalisant cette tâche.

[6] Sur ce point, le modèle booléen ne se distingue pas des autres modèles impliquant la constitution explicite d'un index, tel le modèle vectoriel, présenté ci-après.

termes[7] ; caractères *jockers*, tels « ? » et « * » permettant de substituer n'importe quel caractère respectivement chaîne de caractères ; interrogation par champs logique (ex : « recherche » ET « information » DANS titre), etc). Pour une description plus complète du modèles et de ces extensions, consulter par exemple Salton *et al.* [1982], ou encore Gaussier et Stéfanini [2003] (pp. 33–40).

2.2.2 Modèle *vectoriel*

En 1968, Salton propose un modèle, appelé *modèle vectoriel* [Salton, 1968], palliant nombre des défauts du modèle booléen. Basé sur une intuition géométrique, ce modèle propose de représenter documents et requêtes par des vecteurs, dans un espace dont les dimensions correspondent aux termes de l'index.

Concrètement, il s'agit d'une part de remplacer l'index sous forme de fichier inversé du modèle booléen par une *indexation pondérée* de chaque document (on recense, pour chaque document, l'ensemble des termes d'index qu'il contient, et on associe un *poids* à chacun d'eux), et d'autre part de définir une mesure permettant d'évaluer la *similarité* de deux vecteurs de l'espace.

Dans un espace d'indexation $\mathcal{T} = \{t_1, \ldots, t_n\}$, où les t [i] sont les termes d'index, un document (ou une requête) d_j sera représenté par le vecteur (appelé *vecteur-profile*) :

$$\overrightarrow{d_j} = [w_{j1}, \ldots, w_{jk}, \ldots, w_{jn}]^{\mathrm{T}}$$

avec w_{jk} la pondération (*i.e.* le « poids ») du terme t_k dans le document d_j.

La mesure de la similarité entre deux vecteurs-profiles ($\mathrm{sim}(\overrightarrow{d_i}, \overrightarrow{d_j})$) permet, pour un document d_j donné, d'ordonner l'ensemble des documents de la collection en fonction de leur « similarité » avec d_j.

En représentant de manière identique documents et requêtes, ce modèle permet de mesurer la similarité entre une requête et chacun des documents de la collection. L'utilisateur peut ainsi formuler ses requêtes en language naturel, voire même utiliser un document ; la plupart des fonctions de similarité utilisées permettant un appariement partiel, le système pourra nécessairement proposer des documents en réponse à ces requêtes, qui plus est, en les ordonnant selon la vraissemblance de leur « pertinence » (*i.e.* les valeurs de similarités). C'est également là que réside une des faiblesses du modèle : d'une part le système fournit des « réponses » même à des requêtes n'ayant objectivement aucun rapport avec les documents de la collection, et d'autre part il est loin d'être évident de déterminer combien de documents doivent être considérés comme pertinents pour une requête donnée (*i.e.* après combien de documents faut-il stopper l'extraction, pour une requête donnée). Bien qu'intuitivement la « dispersion » des valeurs de similarité sur les documents de la collection devrait pouvoir être utilisée pour surmonter ces problèmes, cela est en pratique moins évident qu'il n'y parait, en raison de la taille des index et des collections de documents (effet « *curse of dimensionality* », cf. 3.3 [page 29]).

On relèvera également que les dimensions de l'espace vectoriel (engendré par les termes d'index), sont généralement considérées comme orthogonales entre elles, induisant ainsi une indépendance de chacun des termes de l'index (du moins en regard des principales mesures de similarité). Cette hypothèse simplificatrice n'est pas sans conséquences : des termes sémantiquement proches (par exemple « hôpital » et « médecin ») seront considérés comme n'ayant aucun rapport entre eux, exactement de la même manière que des termes n'ayant effectivement aucun rapport (tels « médecin » et « éplucher ») ; remarquons que ce défaut est également présent dans le cadre du modèle booléen (mais ce dernier permet cependant d'imposer que les documents

[7] Et donc se départir quelque peu du principe « sac de mots ».

extraits possèdent à la fois deux (ou plus) termes donnés (« hôpital » *ET* « médecin »), ce qui n'est pas possible avec le modèle vectoriel).

L'une des implémentations les plus connues du modèle vectoriel est certainement le système *Smart*[8] [Salton, 1971, Salton et McGrill, 1983], du moins en ce qui concerne les systèmes expérimentaux.

2.3 Problématiques

Un système de recherche documentaire doit à la fois faire face aux problématiques usuelles des bases de données (en particulier les bases de grandes tailles) et à celles induites par la forme que revêtent les informations contenues dans les documents, à savoir des contenus textuels en langage naturel. Ces dernières représentent cependant le point le plus crucial pour un système de recherche documentaire.[9]

Dans leurs mémoires de thèse respectifs, [de Loupy, 2000, pg. 19–21] et [Besançon, 2001, pg. 26,27] recensent les difficultés suivantes :

Variations de graphie : il arrive fréquemment qu'un même mot admette plusieurs graphies. Outre les variations de casse (impliquant souvent des pertes d'accentuation), on relèvera comme source fréquente les erreurs de frappes (inversion de lettres, etc), et finalement les mots ayant réellement plusieurs graphies, tels que « clé » et « clef », ou certains termes provenant de langues différentes ; De Louppy relève ainsi le cas de « Khadaffi », admettant au moins une dizaine de graphies différentes). Des techniques de corrections lexicale (distance de Levenstein, appariements phonétiques, etc) permettent souvent d'unifier les variations de graphie ; relevons néanmoins que ces variations sont parfois porteuses d'information (par exemple les graphies « Histoire » et « histoire », la première désignant usuellement la discipline ayant comme objet l'étude du passé, et la seconde une anecdote).

Variations morphologiques : qui peuvent être de plusieurs ordres ; on relèvera notamment les variations flexionnelles (typiquement passage au pluriel : « oeuf » et « oeux » ; dépendance au genre : « créateur », « créatrice » ; conjugaison : « voir », « vu » et « verrai », etc.), ainsi que les variations dérivationnelles (passage, pour une même racine, d'une catégorie morphosyntaxique à une autre : « peuple », « peupler », « peuplement »). On fait en général l'hypothèse que les modifications de sens liées aux variations morphologiques sont suffisamment faibles pour pouvoir être ignorées ; le recours à des techniques même peu élaborées de désuffixation (*stemming*), voire des techniques plus évoluées de racinisation (*lemmatisation*) donnent en moyenne de bons résulats. Cependant, il existe malheureusement des cas où cette hypothèse ne se vérifie pas ; ainsi, considérer « eau » en lieu et place de « eaux » dans une expression de type « elle a perdu les eaux » est une source d'erreur quasi certaine.

Mots composés : les mots composés, collocations et expressions idiomatiques sont autant de constructions pour lesquelles le principe du « sac de mots » se révèle insuffisant. Repérer de telles unités et les agréger en un seul uniterme permet d'éviter des confusions (« pieds de biche », « machine à remonter le temps »), mais n'est pas une tâche simple à réaliser (par exemple, l'expression « se casser la pipe » peut être formulée de multiple manières, mais « il a cassé ma pipe » n'en est pas une).

[8] Disponible à l'adresse ftp://ftp.cs.cornell.edu/pub/smart/

[9] En particulier du fait que les difficultés de type « base de données » sont bien connues par ailleurs, et que des méthodes relativement efficaces pour les surmonter existent déjà.

Polysémie et homographie (homonymie) : bien que la problématique soit globalement la même (considérer des unitermes différents pour des mots ayant une même graphie), on distingue généralement les ambiüités *de nature lexicale* (homographie : mots *distincts* ayant la même orthographe,[10] comme « tour » (1. nom féminin : bâtiment construit en hauteur ; 2. nom masculin : limite circulaire ; 3. nom masculin : machine-outils) ou « avocat » (1. juriste, vient du latin *advocatus* ; 2. fruit, vient du Nahualt[11]) des ambiguïtés *de sens* (polysémie : mot unique ayant des sens différents), comme « pieds » d'un être humain et « pieds » de table. La raison est que les ambüités lexicales sont beaucoup plus simples à traiter que les polysémies (dans bien des cas, un simple étiquetage morphosyntaxique permet de distinguer les homographes ; pour les autres, les sens des mots étant très éloignées, il en va généralement de même de leurs contextes, et une désambiguisation sémantique ou des techniques de sémantique distributionnelle [Besançon *et al.*, 2001] ont de bonnes chances de succès). On relevera encore que dans le cas du modèle vectoriel, on pourrait s'attendre à ce que le contexte des mots polysémiques mène, par renforcement mutuel, à la sélection automatique du sens correct ; la pratique montre cependant qu'il n'en est pas vraiment le cas (chapitre 4 [page 55]) – requêtes trop courtes, modèle du sac de mots peut adapté, etc.

Termes sémantiquement liés : il s'agit de la problématique inverse de celle de la polysémie et de l'homographie : il faudrait considérer comme uniterme unique des mots de graphie totalement différentes, mais fortement liés sur le plan sémantique. Le principal lien sémantique à considérer est évidemment la synonymie (« mirroir » et « glace », « voiture » et « automobile »), mais d'autres liens tels que l'hyponymie (« véhicule », « voiture », « camion »), l'antonymie (en particulier les antonymes duals, comme « lune » et « soleil »), ou encore la méronymie (partie de, comme « toit » et « maison ») devraient idéalement être pris en compte. Le but est donc de diminuer, dans les modèles, l'indépendance entre termes d'index sémantiquement liés. Outre la grande diversité des liens possibles et leur « intensité », la tâche est rendue d'autant plus ardue par la polysémie fréquente des termes considérés.

2.4 Améliorations au moyen d'informations latentes

On présente ici brièvement quelques unes des techniques d'amélioration des modèles de recherche documentaire uniquement basées sur des traitements algorithmiques, principalement fondés sur des propriétés statistiques.

Dans pratiquement tous les cas, il s'agira de répondre à une ou plusieurs des problématiques relevées dans la section précédente. On traitera en premier lieu des techniques portant sur les termes d'index, en général appliquables à tous les modèles de recherche, puis quelques techniques spécifiques au modèle vectoriel.

2.4.1 Action sur les termes d'indexation

Pendant longtemps, les index ont été construits de manière manuelle, et ceci tant au niveau de la sélection des termes constituant le jeu d'index que l'indexation des documents elle-même. Cependant, avec l'explosion récente des bases documentaires (en nombre et en taille), une telle indexation n'est plus envisageable[12].

[10] C'est-à-dire des homonymes homographes.

[11] Langue aztèque.

[12] Du moins en ce qui concerne l'indexation ; la sélection manuelle de termes d'index à toujours cours, et à même tendance à se généraliser (lexiques spécifiques, réseau sémantiques, etc.), comme on peut le voir au chapitre 4 [page 55].

À côté de thésaurus « manuels » de plus en plus imposants (à l'image de la « classification universelle » construite en 1876 par Dewey, comportant 10 classes de bases et initialement 1000 sous-classes, portées à plus de 63'000 dans la 16^{e} édition de 1958), des techniques *automatiques* de création d'index (non structurés) ont été mises au point, en particulier sur la base des travaux de Hans Peter Luhn [Luhn, 1957, 1958, 1959, Schultz, 1968]. Mais avec une telle automatisation du processus d'indexation, les ambiguïtés liées à la langue (précédemment évoquées) deviennent problématiques, et la question de la pertinence des entités retenues se pose par ailleurs avec plus d'acuité encore.

La majorité des techniques actuelles reposent sur des termes simples (« mots », incluant éventuellement des mots composés), qui constitue l'approche *sac de mots*. Divers expériences recourant à des unités plus complexes ou plus larges ont néanmoins été menées ; entre autres, l'indexation par des multitermes [Jacquemin, 1998, 2001, Jacquemin *et al.*, 1997], par des groupes syntagmatiques [Chevallet et Haddad, 2001], par des profils de co-occurrences [Besançon *et al.*, 2001], ou encore par des phrases ou des énoncés [Fagan, 1987*a*,*b*, Schütze *et al.*, 1995].

2.4.1.1 Sélection des termes d'indexation

En recherche documentaire classique, les termes de l'index sont le plus souvent sélectionnés en fonction de la collection de documents à représenter (indexation libre ou non contrôlée).

Il paraît évident que des termes très fréquents, porteurs de peu d'information (tels que « un », « les », « alors », « ainsi », etc.) ne sont en général pas de bons candidats pour l'indexation[13] : d'une part un appariement (entre requête et document) réalisé sur la base de termes de cette nature n'a que peu de chance d'être effectivement pertinent, et d'autre part la plupart des documents contenant ces termes dans des proportions a peu près similaires (en fonction de la taille du document), il y a de fortes chances pour que la mesure de similarité entre requête et document soit au mieux augmentée d'un facteur à peu près constant (*i.e.* sans effet aucun), au pire rendue à peu près identique pour tous les documents de la collection. Dans cette optique, on cherchera à retenir les termes en fonction de leur *valeur de discrimination* [Salton *et al.*, 1975], le but étant d'avoir une indexation permettant de maximiser la « dissemblance moyenne » entre documents.

Ce principe de réduction (de la dimension) de l'espace d'indexation en fonction d'un critère de discrimination fait partie d'un domaine de recherche propre (*feature selection*). Les techniques de sélection des termes sur la base d'informations extraites de la collection utilisent généralement une ou plusieurs des mesures suivantes :

- tf_{ij} la *fréquence d'occurrence* absolue du terme t_i dans le document d_j (*term frequency*), *i.e.* le nombre de fois que le terme t_i apparaît dans le document d_j ;
- tf_i la fréquence du terme t_i dans la collection, $\mathsf{tf}_i = \sum_{j \in d_i} \mathsf{tf}_{ij}$;
- df_i la *fréquence documentaire* du terme t_i dans la collection, soit le nombre de documents dans lesquels le terme t_i apparaît.

On trouve parmi ces techniques le *filtrage fréquentiel*, une des forme de sélection les plus simple qui soit : il consiste à ignorer les termes sur-représentés et sous-représentés, sur la base de leur fréquence d'occurrence absolue ; un tel filtrage trouve sa justification dans la *loi de Zipf*[14] [Baayen, 2001, pg. 13–24 notamment]. Salton *et al.* [1975] ont montré qu'une indexation ne

[13] On peut néanmoins trouver des cas où ces mots habituellement vide sont cependant porteurs de sens ; par ailleurs, Roeck *et al.* [2005] démontrent la non-homogénéité de leurs occurrences dans les documents d'une collection, accréditant ainsi le fait que ces termes peuvent effectivement apporter une certaine information.

[14] Loi du nom de son auteur, George Kingsley Zipf (1902-1950), généralisée par la suite par Benoit Mandelbrot (1950), et stipulant que pour un corpus donné, la fréquence d'occurrence d'un terme t_i est liée à son rang n dans l'ordre des termes classés par fréquences par une loi du genre : $\mathsf{tf}_i = {}^{K}/_{n}$, où K est une constante.

retenant que les termes dont la fréquence documentaire se situe entre 1/10 et 1/100 de l'ensemble des documents génère le plus souvent un index ayant un pouvoir de discrimination satisfaisant.

D'autres méthodes de sélection ont également été envisagées, en particulier dans le domaine de la classification automatique,[15] lorsque des données de référence sont disponibles [Sebastiani, 2002].

Pour la plupart, ces méthodes utilisent comme critère de sélection des fonctions tirées de la théorie de l'information : χ^2 (Chi-carré), gain d'information (GI), information mutuelle (IM) qui correspondent à l'intuition que les termes les plus discriminants sont ceux qui sont distribués le plus différemment dans les ensembles d'exemples positifs ou négatifs d'une classe. Les expressions mathématiques de ces fonctions sont données dans la table 2.1, où les t_k sont les termes, les c_i les classes, N est le nombre de documents dans l'ensemble d'entraînement, et les probabilités $p(t_k, c_i)$ (resp. $p(\overline{t_k}, c_i)$) sont estimées en comptant le nombre d'éléments de la classe c_i contenant (resp. ne contenant pas) le terme t_k dans l'ensemble d'entraînement. Toutes ces méthodes demandent donc d'avoir des données de référence pour estimer les probabilités d'avoir ou de ne pas avoir un terme sachant une classe. Ces fonctions sont données pour une classe et un terme : la mesure prise en compte pour la sélection d'un terme t_k dans une collection comptant m classes est, pour une fonction F, $\sum_{i=1}^{m} p(c_i)F(t_k, c_i)$.

Fonction	Forme mathématique
$\chi^2(t_k, c_i)$	$\dfrac{N \times \left(p(t_k, c_i)\ p(\overline{t_k}, \overline{c_i}) - p(t_k, \overline{c_i})\ p(\overline{t_k}, c_i)\right)^2}{p(t_k)\ p(\overline{t_k})\ p(c_i)\ p(\overline{c_i})}$
$GI(t_k, c_i)$	$\displaystyle\sum_{c \in \{c_i, \overline{c_i}\}} \sum_{t \in \{t_k, \overline{t_k}\}} p(t, c) \cdot \log \frac{p(t, c)}{p(t) \cdot p(c)}$
$IM(t_k, c_i)$	$\log \dfrac{p(t_k, c_i)}{p(c_i)\ p(t_k)}$

Table 2.1*:* Diverses fonctions pour la sélection des termes d'indexation

Yang et Pedersen [1997] présentent une comparaison de ces différentes fonctions, et montrent que les fonctions GI et χ^2 ont des performances comparables à la sélection par la fréquence en documents, l'information mutuelle ne donnant, pour sa part, pas de très bons résultats. Ils montrent également qu'il existe une corrélation entre les valeurs de ces fonctions.

D'autres mesures ont également été utilisées ; par exemple, Ng *et al.* [1997] montrent que de meilleurs résultats peuvent être obtenus en utilisant à la place du χ^2 un coefficient de corrélation égal à la racine carrée signée du χ^2 (le carré du χ a en effet la propriété indésirable de considérer de la même façon les facteurs qui indiquent une corrélation soit positive soit négative entre un terme et une classe).

[15] Parce que les algorithmes de classification sont souvent moins adaptés à des espaces de représentation de très grande taille, et qu'une sélection «agressive» des termes est souvent nécessaire.

2.4.1.2 Indexation sémantique latente (LSI/PLSI)

Certaines autres techniques de sélection ont été envisagées, ayant comme propriété d'indexer les documents au moyen de termes « artificiels », absents des données d'origines mais construits à partir d'elles (variables cachées).

C'est le cas du modèle d'*indexation sémantique latente* LSI, et de leur extension probabiliste PLSI [Hofmann, 1999].

Le modèle LSI est une variante du modèle vectoriel standard, visant à prendre en compte de potentiels structures sémantiques implicites (*i.e.* latentes) des unités linguistiques, représentées par leurs dépendances cachées [Deerwester *et al.*, 1990, Landauer *et al.*, 1998], en utilisant une technique de réduction de dimensionnalité usuelle en statistique multidimensionnelle.

Pratiquement, les techniques LSI utilisent une matrice (*documents* $\times$ *termes*) similaire à la matrice de vecteurs-profiles du modèle vectoriel, dans laquelle chaque élément w_{ij} est une pondération, fonction du nombre d'occurrences du terme t_j dans le document d_i. Une décomposition en valeurs singulières (SVD) de cette matrice est effectuée et seuls les k premiers vecteurs propres sont pris en compte (k prend typiquement une valeur entre 100 et 300), constituant les termes d'indexation.

Ce modèle permet donc de représenter les documents dans un espace réduit de dimension k. Il est toutefois important de noter que chacune des dimensions de l'espace de représentation final $\mathbb{R}^k$ correspond à une « combinaison linéaire des unités linguistiques ». L'espace de représentation n'a donc pas pour support un ensemble de termes directement présents dans les documents, ce qui rend les dimensions relativement difficiles à interpréter directement.

Ce modèle a été appliqué à de nombreuses tâches, comme la recherche documentaire, le filtrage, le routage d'information ou la recherche documentaire multilingue [Dumais, 1994, Dumais *et al.*, 1996, Foltz et Dumais, 1992, Schütze *et al.*, 1995].

Papadimitriou *et al.* [1998] proposent d'intégrer dans le modèle LSI une phase stochastique de réduction de dimension pour faire une première approximation réduite de la matrice *documents* $\times$ *termes* avant de procéder à la décomposition en valeurs singulières, cette dernière étant coûteuse pour une matrice de grande taille. La réduction stochastique repose sur le lemme de Johnson-Lindenstrauss, qui affirme que si les points d'un espace vectoriel sont projetés dans un sous-espace aléatoire de dimension assez grande, alors les distances entre les points sont approximativement conservées [Johnson et Lindenstrauss, 1984]. Les auteurs de l'article montrent qu'en appliquant une réduction stochastique de l'espace avant la SVD, ils gardent presque autant d'information qu'en faisant la SVD directement, pour un temps de calcul moindre. D'autres extensions visant à alléger la phase de SVD ont vu le jour, telles que l'indexation « aléatoire » (*random indexing*) [Sahlgren, 2005]), l'indexation sémantique généralisée [Matveeva *et al.*, 2005] ou encore le filtrage polynomial [Kokiopoulou et Saad, 2004].

2.4.2 Pondération des termes dans le modèle vectoriel

La pondération accordée à la contribution d'un terme d'index t_j dans la représentation d'un document d a fait l'objet de nombreuses études [Lee, 1995, Salton et Buckley, 1988, Singhal *et al.*, 1995], et prend en général en compte des facteurs de pondération locale, de pondération globale et de normalisation en fonction de la taille du document.

Pondération locale

La pondération locale prend en compte les informations locales du terme qui ne dépendent que du document. Elle correspond en général à une fonction de la fréquence d'occurrence du terme

dans le document (notée tf pour *term frequency*), c'est-à-dire le nombre de fois où le terme est utilisé dans le document. Les fonctions les plus utilisées sont les suivantes :

- facteur tf : il correspond à la fréquence d'occurrence du terme dans le document ;
- facteur binaire : il vaut 1 si le terme est présent, 0 s'il ne l'est pas. Ce facteur est utilisé pour des représentations de type ensembliste (prenant en compte de mesures de similarité de type Dice ou Jaccard. Il donne aussi souvent une base de comparaison par rapport aux autres pondérations locales proposées ;
- facteur logarithmique : ce facteur est une fonction logarithmique de la fréquence du terme dans le document, valant :

  $$1 + \log(\mathsf{tf})$$

  Cette fonction, initialement proposée par Buckley *et al.* [1992], est motivée par le fait qu'un document qui contient un grand nombre de fois un terme de la requête n'est pas forcément plus pertinent qu'un document qui contient un petit nombre de fois plusieurs termes de la requête. On souhaite donc qu'en général, un plus grand nombre d'occurrences d'un terme ne soit pas dominant par rapport à un plus petit nombre d'occurrences de plusieurs termes ;
- facteur augmenté (*augmented* tf *factor*) [Salton et Buckley, 1988] ce facteur, comme le facteur logarithmique, réduit les différences entre valeurs pour les différents poids accordés aux termes du document, en accordant une valeur minimale aux termes présents dans le document (par exemple 0.5), et en accordant aux termes présents plusieurs fois un poids ne dépassant pas une certaine valeur maximale (par exemple 1.0).

Pondération globale

La pondération globale prend en compte des informations concernant le terme et dépendant de la totalité de la collection. Une pondération prenant en compte l'importance de l'unité linguistique dans la collection améliore les performances dans le cadre de la RD. Un poids plus important doit être donné aux unités linguistiques qui apparaissent moins fréquemment dans la collection : les unités linguistiques qui sont utilisées dans de nombreux documents sont moins utiles pour la discrimination que celles qui apparaissent dans peu de documents. On introduit donc un facteur de pondération globale qui dépend de façon inverse de la fréquence en documents, comme par exemple le facteur idf (pour *inverted document frequency factor*) [Salton *et al.*, 1975] valant, pour une collection de documents D :

$$\mathsf{idf} = \log(\frac{|D|}{\mathsf{df}})$$

où df est la fréquence en documents du terme considéré.

Normalisation

Les pondérations locale et globale sont une bonne approximation de l'importance d'un terme dans un document, mais ne prennent pas en compte un aspect important du document : sa longueur. Les différences de longueur des documents peuvent être dues à plusieurs raisons :

- Verbosité (plus grandes fréquences) : les documents les plus longs auront tendance à utiliser les mêmes mots de façon répétée. En conséquence, les facteurs de fréquences en documents seront plus élevés, et les similarités avec la requête seront également plus grandes pour les documents plus longs.
- Variations thématiques (plus de termes) : les documents les plus longs ont également tendance à parler de plus de choses et donc à utiliser plus de termes distincts, ce qui accroît le nombre de correspondances entre les termes de la requête et les termes d'un long document, et qui augmente donc les similarités avec les documents longs.

Il est donc nécessaire de compenser les différences de longueur entre documents dans la collection.

Plusieurs techniques de normalisation ont été proposées :

- normalisation par le cosinus : la normalisation en fonction de la longueur du document peut être effectuée directement dans le calcul de la similarité, en utilisant une mesure de similarité indépendante de la norme du document, comme l'est par exemple le cosinus.
- max-tf : une autre normalisation proposée est la normalisation de chaque facteur tf par le tf maximum dans le document. Ce type de normalisation répond surtout à la première raison évoquée (*plus grandes fréquences*), et est insuffisante lorsqu'il n'y a pas de stratégie pour résoudre la seconde raison (*plus de termes*).
- normalisation à pivot (*pivoted normalisation*) : une autre normalisation est proposée par Singhal [1997], et repose sur l'idée que les schémas de normalisation qui permettent au système de retourner des documents d'une certaine longueur avec une probabilité proportionnelle à la probabilité qu'un document de cette longueur soit pertinent aura de meilleures performances. À partir de cette idée, Singhal propose les schémas de normalisation à pivot :

  $$1 + \frac{\text{slope}}{(1 - \text{slope}) \times \text{pivot}} \times (\text{ancienne normalisation})$$

  et de normalisation unique à pivot :

  $$(1 - \text{slope}) \times \text{pivot} + \text{slope} \times (\text{nb de termes uniques}),$$

  où *pivot* est la longueur de document pour laquelle la probabilité qu'un document soit retourné sachant sa longueur est égale à la probabilité qu'un document soit pertinent sachant sa longueur (ces probabilités sont estimées après une première exécution du programme de recherche en utilisant une autre mesure de normalisation, comme le cosinus, par exemple), et *slope* est un paramètre du modèle.

2.4.3 Mesures de similarité dans le modèle vectoriel

La mesure de similarité entre documents, établie sur la base de leur représentation dans l'espace vectoriel, a elle aussi fait l'objet de nombreuses études.

Nous présentons ici brièvement les mesures les plus couramment utilisées ; pour une revue plus complète, se référer notamment à Besançon [2001].

2.4.3.1 Mesures ensemblistes

Certaines mesures utilisent seulement l'information de la présence ou de l'absence d'un terme dans un document (correspondant au facteur binaire de pondération locale). Les mesures les plus utilisées dans ce cadre sont les coefficient de Dice et de Jaccard. Dans les définitions de ces mesures, pour un document $d = (w_1, \ldots, w_{|T|})$, nous notons $\{d\}$ l'ensemble des unités linguistiques qu'il contient (*i.e.* $\{d\} = \{u_i \in T | w_i \neq 0\}$).

Coefficient de Dice :

$$\delta_{\text{dice}}(d, d') \;=\; 2 \times \frac{|\{d\} \cap \{d'\}|}{|\{d\}| + |\{d'\}|}$$

Coefficient de Jaccard :

$$\delta_{\text{jaccard}}(d, d') \;=\; \frac{|\{d\} \cap \{d'\}|}{|\{d\} \cup \{d'\}|}$$

Différence symétrique normalisée :

$$\delta_{\triangle}(d, d') \;=\; \frac{|d \triangle d'|}{|\{d\}| + |\{d'\}|} \;=\; \frac{|\{d\} \cup \{d'\} - \{d\} \cap \{d'\}|}{|\{d\}| + |\{d'\}|}$$

2.4.3.2 Mesures géométriques

Mesure du cosinus :
La mesure de similarité la plus utilisée est le cosinus de l'angle entre les vecteurs représentant les documents :

$$\delta_{\cos}(d, d') \;=\; \frac{d \cdot d'}{||d||\;||d'||} \;=\; \frac{\sum_{j=1}^{|T|} w_j \; w'_j}{\sqrt{\sum_{j=1}^{|T|} w_j^2 \; \sum_{j=1}^{|T|} w_j'^2}}$$

Cette mesure est indépendante de la norme des vecteurs représentant les documents, ce qui fournit une forme de normalisation par la longueur des documents, et peut éviter d'augmenter les valeurs des similarités entre documents plus longs. D'autre part, cette mesure ne repose que sur les termes contenus dans les deux documents (*i.e.* l'intersection des documents) : l'amplitude des différences hors intersection n'est pas prise en compte.

Parmi les autres mesures géométriques, les distances métriques usuelles reposant sur les normes L1 et L2 sont également utilisées.

Distance L1 :

$$\delta_{\mathrm{L1}}(d, d') \;=\; ||d - d'||_{L1} \;=\; \sum_{j=1}^{|T|} (w_j - w'_j)$$

Distance euclidienne (L2) :

$$\delta_{\mathrm{L2}}(d, d') \;=\; ||d - d'|| \;=\; \sqrt{\sum_{j=1}^{|T|} (w_j - w'_j)^2}$$

Si les vecteurs des documents sont normalisés par la norme euclidienne, la distance euclidienne est monotone par rapport à la mesure du cosinus : on a en fait la relation suivante :

$$\frac{\delta_{\mathrm{L2}}(d, d')^2}{2} \;=\; 1 - \delta_{\cos}(d, d')$$

Les deux mesures seront donc dans ce cas équivalentes pour le classement des document retournés.

2.4.3.3 Mesures Distributionnelles

Dans le cas où les vecteurs sont normalisés par la norme L1 (*i.e.* $\sum_j w_j = 1$), et peuvent donc être interprétés comme des distributions de probabilité, des mesures pour la dissimilarité entre distributions de probabilité sont également proposées :

Distance du χ^2 :
La distance du χ^2 est en fait proche de la distance euclidienne, mais avec une pondération ρ_j associée à chacun des termes de la somme :

$$\delta_{\chi^2}(d, d') = \sqrt{\sum_{j=1}^{|T|} \rho_j \left(w_j - w'_j\right)^2},$$

avec $\rho_j = \frac{|D|}{\sum_{d \in D} w_j)}$ l'inverse de la distribution marginale sur la collection de documents. Au contraire du cosinus, cette mesure est particulièrement sensible aux différences hors intersection.

Divergence de Kullback-Leibler :

La divergence de Kullback-Leibler (KL), appelée aussi *entropie relative* Cover et Thomas [1991], est définie, pour deux distributions q et r, par :

$$D(q||r) = \sum_y q(y) \log \frac{q(y)}{r(y)}$$

Divergence de Jensen-Shannon :

La *divergence de Jensen-Shannon* (aussi appelée *Divergence Totale à la Moyenne*), est définie, pour deux distributions q et r, par :

$$JS(q||r) = D\left(q \middle|\middle| \frac{q+r}{2}\right) + D\left(r \middle|\middle| \frac{q+r}{2}\right)$$

La divergence de Jensen-Shannon est préférée à la divergence de Kullback-Leibler lorsqu'on veut avoir une mesure symétrique.

2.5 Améliorations au moyen d'informations additionnelles

En parallèle aux approches n'utilisant comme seule source d'information les données à traiter, plusieurs autres techniques utilisant des connaissances « externes » ont été mises au point.[16]

2.5.1 Expansion de requête

En recherche documentaire « classique » (par opposition aux tâches de type filtrage, détection de spam, etc.), les requêtes sont habituellement de taille nettement plus faible que les documents.[17] Par ailleurs, il est évident que plus une requête et les documents pertinents qui lui correspondent ont de termes en commun et plus ces documents auront de chances d'être retrouvés par un système de recherche documentaire. Ainsi, plus une requête est courte et plus il sera difficile au système de trouver des documents effectivement pertinents pour celle-ci.

L'idée de réaliser une « expansion » des requêtes courtes est donc rapidement apparue ; en 1968, Salton constate que l'utilisation d'un thésaurus (Harris Synonym) permet d'améliorer les

[16] On remarquera au passage que l'utilisation, dans le cadre du modèle LSI, d'une indexation contrôlée, établie une fois pour toute sur la base d'un corpus donné, est en fait assimilable à une technique d'indexation utilisant une information additionnelle, le jeu d'index pré-calculé.

[17] Aussi bien en taille du vocabulaire (*i.e.* mots différents) qu'en nombre d'occurrences (nombre de mots).

perfomances, pour autant que les termes utilisés pour l'enrichissement soient validés manuellement par l'utilisateur ; a contrario, une expansion automatique, utilisant l'ensemble des termes possibles, dégrade ces performances [Salton, 1968].

En plus de l'enrichissement au moyen de termes liés à ceux de la requête fournis par une ressource additionnelle de type dictionnaire de synonymes (voir par ex. [Moldovan et Mihalcea, 2000, Voorhees, 1994]), une technique couramment employée consiste, pour une requête, à utiliser les termes contenus dans les documents pertinents identifiés lors d'une première recherche, soit en utilisant les k premiers documents retournés par le système, soit en demandant à l'utilisateur de sélectionner ces documents – méthode appelée *retour de pertinence* (Allan [1996], Salton et Buckley [1990], ainsi que Harman [1988], qui arrive à des conclusions similaires à celles de Salton [1968]). Certains encore utilisent un critère statistique, construit sur la collection de documents, pour déterminer les associations de termes à prendre en compte [Qiu, 1994, Qiu et Frei, 1993].

2.5.2 Action sur les termes d'indexation

L'index permettant de construire les représentations des documents étant un élément central de la plupart des tâches de traitements documentaire, une large part des travaux réalisés portent naturellement sur l'amélioration de la représentation des documents. En regard des techniques visant à améliorer le processus d'indexation lui-même (en particulier, les techniques de désambiguisation sémantiques), que nous ne traiterons pas ici,[18] la constitution des index (*i.e.* sélection des termes de l'index) peut être sensiblement améliorée en recourant à des informations additionnelles.

2.5.2.1 Pré-traitements linguistiques

Une des première technique mise en œuvre consiste à retirer des termes d'index potentiels les mots « vides », tels que les articles, déterminant, etc. Outre l'utilisation de techniques statistiques (filtrage fréquentiel), cela peut aussi être réalisé de manière plus fine au moyen *d'anti-dictionnaires* (*stop-list*).

2.5.2.2 Indexation sémantique et conceptuelle – Réindexation au moyen de thésaurus

Les indexations *sémantique* et *conceptuelle* associent le principe d'indexation par des termes artificiels absents des documents indexés (voire par exemple LSI 2.4.1.2) à celui d'enrichir les termes d'indexation par des synonymes ou termes liés (« expansion de requête », 2.5.1), mais sur l'ensemble de la collection, et pas uniquement les requêtes (ou les documents « courts »).

Le principe consiste donc à modifier tout ou partie du jeu d'indexation, en utilisant des termes additionnels issu d'un thésaurus ; le type et la provenance de ce thésaurus permet de distinguer entre indexation *sémantique* et indexation *conceptuelle*, selon la typologie proposée par Mihalcea et Moldovan [2000] (bien que l'on puisse relever dans la littérature une certaine confusion quant à l'emploi de l'un ou l'autre de ces termes) : en indexation sémantique, le thésaurus utilisé est une ressource lexicale pré-existante (de type *WordNet*), tandis qu'en indexation conceptuelle, il s'agit le plus souvent d'une ressource construite à partir de la collection de document, de manière statistique [Kang, 2003, Woods, 1997]. Bien que cette seconde catégorie permette de s'affranchir des problèmes d'ambiguïté sémantique entre le vocabulaire de la collection et les termes du thésaurus, elle n'a pas fait l'objet d'une étude détaillée de notre part, notre propos étant principalement centré sur l'utilisation de ressources externes.

[18] Remarquons cependant que l'expansion de requête peut y être assimilée.

Indexation par classes de synonymes
L'indexation par classe de synonymes a été envisagées à plusieurs reprises [Gonzalo *et al.*, 1998*a*, 2000, 1998*b*, Hotho *et al.*, 2003, Whaley, 1999]. En vue de constituer le jeu d'index (espace de représentation des documents), l'idée est ici d'utiliser une ressource sémantique (la plupart du temps *WordNet*) pour substituer les mots de documents non pas par leurs lemmes ou leurs racine comme à l'accoutumée, mais par le représentant de la classe de synonymes auxquels ils appartiennent. Voorhees [1998], qui utilise *WordNet* à des fins de désambiguisation sémantique, tente de mesurer la pertinence de cette désambuisation en appliquant une tâche de recherche documentaire aux données « désambgüiées » (et donc indexée selon les synsets de *WordNet*). Mais les résultats de l'expérience dénotent une baisse des performances par rapport à une indexation traditionnelle des documents, et Voorhees conclut à la faiblesse de processus de désambiguisation sémantique.

À l'inverse, Gonzalo *et al.* [1998*a*,*b*] qui utilisent également les synsets de *WordNet* comme termes d'index pour leur collection, observent quant à eux une augmentation notable des performances ; leurs résultats sont cependant biaisés par les conditions de l'expérience ; ils s'affranchissent en effet du problème de l'ambiguité de l'affectation d'un mot à une classe de synonymes en utilisant comme base documentaire un corpus désambiguisé à la main pour la ressource *WordNet* (*SemCor*). Pour pouvoir conduire une recherche documentaire avec cette base, cette dernière est par ailleurs « transformée » en corpus d'évaluation : elle est segmentée en différents morceaux, et pour chacun d'eux un résumé est produit, jouant le rôle de requête d'évaluation (requête n'admettant qu'un seul document pertinent).

Indexation par hyperonymes (Onto-matching)
Dans le même ordre d'idée, Kiryakov et Simov [1999] proposent (sans conduire d'évaluation, du moins dans l'article en question), pour l'indexation d'un mot, de ne pas se limiter aux seuls représentants des classes de synonymes, mais d'utiliser la totalité des concepts hyperonymes des classes relatives. Ils anticipent cependant le fait qu'une telle utilisation de l'ensemble des hypernoymes ne donnerait pas de bons résultats, et proposent comme piste de limiter le nombre de termes hyperonymes adjoints à chaque terme de base en ne prenant pas en considération les hyperonymes trop abstraits, ou en demandant à l'utilisateur de définir le sous-ensemble d'hyperonymes à conserver.
Mihalcea et Moldovan [2000] rapportent que l'indexation simultanée graphie – classe de synonymes (issus de *WordNet*) permet d'augmenter de 16% le rappel et de 4% la précision, tandis que l'indexation basée uniquement sur les classes d'hyperonymes augmente le rappel de 28%, mais diminue la précision de 9%. Cependant, il faut relever d'une part la mise en œuvre d'un porcessus de WSD construit pour l'occasion, et relativement complexe. D'autre part ils ne se contentent pas de travailler sur des catégories de mot bien précises, mais utilisent (et indexent) la totalité de l'information disponible (à l'exception de la ponctuation).

Chapitre 3

Utilisation de thésaurus en indexation

Résumé

Ce chapitre présente de manière détaillée différentes possibilités permettant de réaliser une *indexation séamntique* de documents textuels au moyen d'un thésurus ou d'un réseau sémantique, c'est-à-dire une indexation qui prenne en compte une partie au moins du contenu informationnel d'un thésaurus. Nous avons pour cela construit un cadre permettant d'exprimer et de positionner ces différentes méthodes les unes par rapport aux autres, en dégageant un certain nombre de familles de solutions, chacune ayant ses particularités (notamment en terme de complexité algorithmique).

Nous focalisons l'essentiel de notre propos sur l'utilisation de thésaurus conditionnellement aux données à indexer, mais donnons également en fin de chapitre quelques pistes pour une utilisation indépendante des données. Trois classes de critères visant à utiliser au mieux la ressource sémantique sont principalement décrites : les critères « locaux », les critères « globaux séparables » et les critères « globaux non-séparables ». Ces derniers, constituant l'aspect novateur de nos travaux, en particulier par le développement du principe de « coupe de redondance minimale » (*CRM*), sont ceux induisant le plus de dépendances entre les différents éléments de l'index.

Pour chacune des trois famille de critères, nous donnons un ou plusieurs exemples concrets de mise en œuvre, ainsi qu'un algorithme qui en permette le calcul, exact lorsque cela est possible, ou approché lorsque la nature du critère ou la topologie du thésaurus l'impose.

Ce chapitre débute par une introduction présentant de manière détaillée le principe de l'*indexation sémantique*, dont nous explicitons les avantages et les difficultés. L'utilisation d'un thésaurus conditionnellement aux données à indexer est ensuite abordée pour chacune des trois classes de critères présentées. Nous envisageons ensuite le problème de la prise en compte dynamique des informations externes dans l'indexation de documents non connus *a priori*. Nous terminons en considérant le cas d'une utilisation indépendante des données, où l'index issu du thésaurus sémantique est déterminé statiquement.

L'évaluation effective des techniques présentées n'est pas abordée ici, mais fait l'objet du chapitre suivant.

3.1 Préambule

3.1.1 Introduction

Comme cela a été présenté au second chapitre, il existe différentes méthodes permettant la représentation de documents textuels dans un « espace sémantique » vectoriel, se différenciant principalement sur la manière dont les termes d'indexation sont construits ou choisis.

Par ailleurs, on constate que des connaissances additionnelles sur les données textuelles sont de plus en plus souvent disponibles (qu'il s'agisse de thésaurus, de taxonomie thématique ou encore d'« ontologies », généralistes ou spécialisées) ; cependant, ces connaissances *externes*[1] ne sont que rarement prises en compte dans les représentations de types vectorielles ; les quelques techniques reportées en section 2.5 [page 17] ne concernent presque qu'exclusivement des outils de laboratoire, et ne représentent qu'une faible proportion de la littérature relative à la recherche documentaire. On notera cependant que la tendance s'inverse au cours des dernières années, probablement en raison de la disponibilité croissante de ressources de qualité (telle *WordNet*) d'une part, et d'autre part de l'intrication toujours plus forte entre la « linguistique informatique » et d'autres domaines – médical, juridique, etc. – pour lesquels existe le plus souvent un jargon spécialisé, déjà répertorié dans des thésaurus, et largement utilisé par les spécialistes du domaine.

Les modèles usuels consistent généralement, pour représenter les documents, à construire un espace vectoriel dans lequel chaque axe est assigné à un mot (éventuellement une combinaison linéaire de mots, dans le cadre de LSI), indépendamment des relations existant *a priori* entre eux.

Il est cependant possible d'intégrer dans un modèle vectoriel une partie au moins de ces connaissances additionnelles, notamment en agissant :

- lors de l'indexation, par exemple en remplaçant tout ou partie des termes d'index par des termes issus des connaissances additionnelles (aggrégation de synonymes, ajout d'antonymes, etc.) ; les méthodes agissant de la sorte sont pour l'essentiel désignées par le terme d'*indexation sémantique* (voir § 2.5.2.2 [page 18]) ;
- lors du calcul de proximité entre documents, en intégrant, dans la mesure de similarité utilisée pour établir ces proximités, une distortion de l'espace vectoriel en fonction de la « distance sémantique » entre les termes associés aux axes vectoriels, cette « distance » étant construite sur la base de ces connaissances additionnelles.

Dans cette optique (intégration de connaissances externes), nos travaux ont principalement portés sur la première de ces possibilités (*indexation sémantique*), qui fait l'objet de ce chapitre.

3.1.2 Organisation du chapitre

Nous avons développé notre propos sur la base de la structure donnée en figure 3.1 [page ci-contre].

Après une présentation du principe de base de l'*indexation sémantique*, nous détaillons un ensemble de techniques plus ou moins élaborées permettant de surmonter la problématique posée par ce type d'indexation. Cette problématique se résume en grande partie à isoler, lors d'une phase précédant l'indexation proprement dite, un ensemble le plus pertinent possible de termes d'index.

Cette sélection des termes peut être réalisée en prenant ou non en compte les données à indexer (en supposant ces données connues *a priori*). Nous avons distingué deux familles de critères de

[1] « Externes » car non présentes dans les documents à traiter, du moins sous une forme explicite et complète.

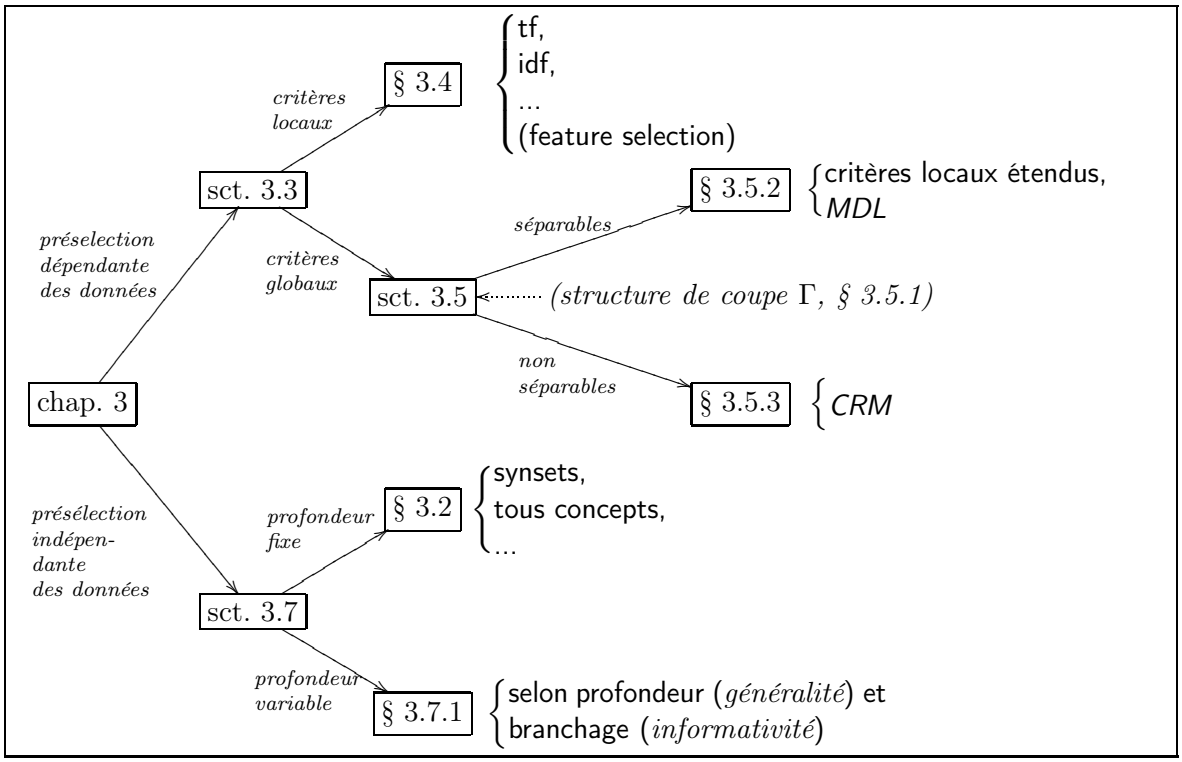


Figure 3.1*:* Organisation du chapitre

décision : les critères « locaux », avec lesquels chaque terme potentiel est évalué indépendamment des autres, et les critères « globaux », pour lesquels la décision est prise non plus terme par terme, mais sur un ensemble déterminé au moyen d'une structure particulière (structure de « coupe »). Cette dernière famille se décompose en deux sous-familles regroupant les critères « séparables », dont l'évaluation globale peut être obtenue par la composition d'évaluations partielles, et les critères « non-séparables », pour lesquelles l'évaluation ne peut se faire que globalement.

Nous proposons également une méthode de sélection utilisable lorsque les données à indexer ne sont pas connues *a priori*, qui se démarque des techniques d'*indexation sémantique* actuellement reportées dans la littérature, consistant à préselectionner dans les ressources sémantiques, les termes à profondeur constante.

3.2 Principes de l'indexation sémantique

L'intégration par le biais des termes d'index d'une structure externe *formelle* dans l'espace de représentation des documents n'est de loin pas triviale ; elle implique de fait une représentation permettant de fusionner des connaissances formelles et des connaissances numériques.

Dans cette optique, le principe de l'*indexation sémantique* consiste à projeter les données à indexer (*i.e.* les « mots » et multi-termes constitutifs de ces données) dans une ressource sémantique, et en extraire un jeu d'indexation mêlant termes initiaux issus des données et termes additionnels (abstraits) provenant de la ressource.

Postulat :

Par la suite, on admettra que les connaissances sémantiques additionnelles utilisées sont issues de ressources externes organisées sous forme de **réseau sémantique** ou de **thésaurus**, tels que définis ci-après.

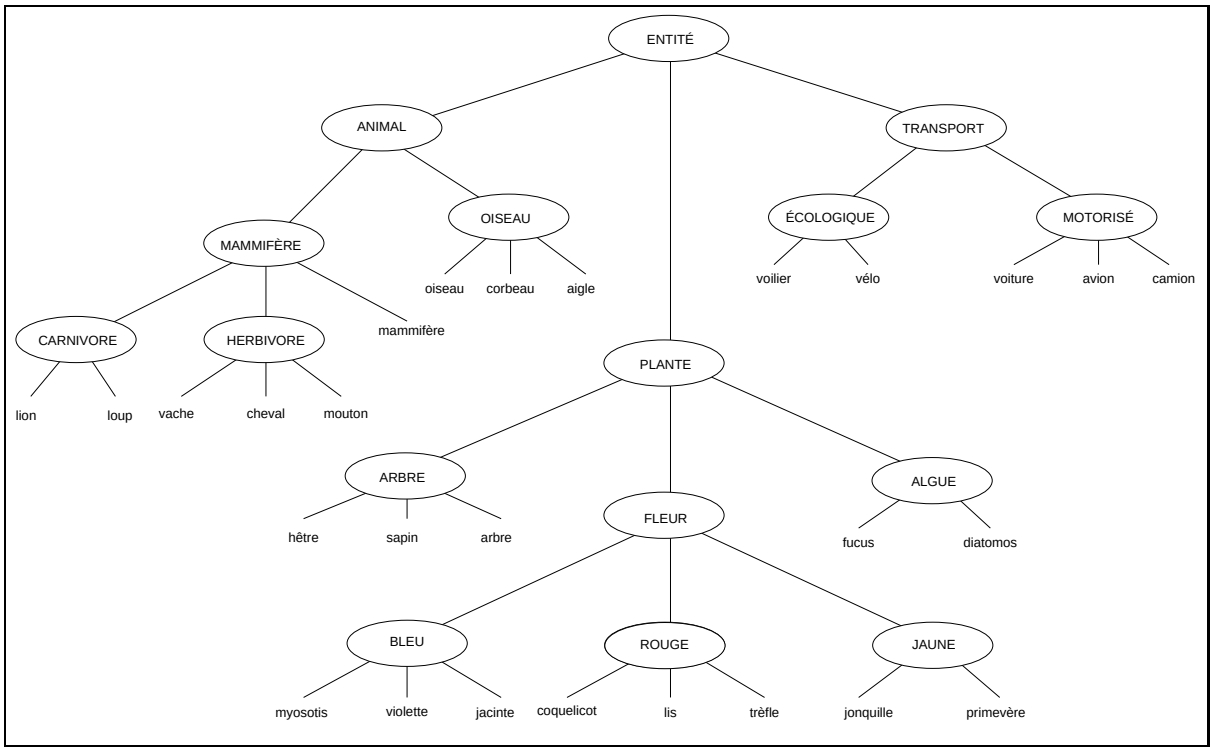


Figure 3.2*:* Exemple (artificiel) de thésaurus organisé selon une relation « est-un » (hyperonymie).

Définitions

- Un **réseau sémantique** est assimilé à un graphe orienté $G = [\mathcal{S}, \mathcal{R}]$ dont les sommets $\mathcal{S}$ sont les entités définies par le réseau, et les arcs $\mathcal{R}$ les relations sémantiques existantes entre ces entités. Le réseau étant lexicalisé, on considère deux catégories de sommets : les sommets $\mathcal{M}$ modélisant le *lexique* du réseau (*i.e.* le signifiant) et les sommets $\mathcal{C}$ modélisant les *concepts* (abstraits) propres au réseau sémantique (*i.e.* le signifié) ; on a donc $\mathcal{S} = \mathcal{M} \cup \mathcal{C}$. Par ailleurs, on considérera autant de catégories d'arcs qu'il y a de types de relations sémantiques définies par le réseau.
- Un **thésaurus** est la restriction d'un réseau sémantique aux relations inclusives de filiation (induisant une hiérarchie) ; pour un même type de relations sémantiques, un thésaurus est modélisé par un graphe orienté *sans cycle* (DAG).

En figure 3.2 est donné le graphe (en l'occurrence un arbre) correspondant à un thésaurus hypothétique, structuré par des relations de type « est-un ». Les « concepts » (abstraits) sont cerclés et notés en haut de casse, à l'inverse des « mots » notés en bas de casse ; nous conserverons cette convention par la suite.

Remarque :

On considérera principalement des réseaux sémantiques de type **thésaurus** (selon notre définiton), *i.e.* des réseaux pour lesquels il est possible d'identifier des sommets initiaux et terminaux, ces derniers correspondant par ailleurs aux « mots » présents dans les données. Par abus de language, on généralisera le jargon des arbres aux DAG : on parlera ainsi de *racines* pour les sommets initiaux, de *feuilles* pour les terminaux, etc.

3.2.1 Exemple d'indexation sémantique

Afin de mettre en évidence l'intérêt apporté par l'*indexation sémantique*, nous allons illustrer son fonctionnement sur un exemple (artificiel), et observer les conséquences d'une telle indexation sur une mesure des similarités entre documents (la similarité « cosinus »)

Admettons que l'on souhaite indexer les quatres « documents » suivants[2] :

d_1	coquelicot arbre vélo coquelicot lion
d_2	lion vache loup mouche
d_3	trèfle voiture sapin loup
d_4	mammifère lion vache lion voiture loup

Une indexation simple, utilisant d'une part les « mots » (lemmes) des documents comme espace de représentation, et d'autre part le dénombrement des occurrences des termes d'index comme mesure de la contribution du terme au document, produira les **profils d'indexation** suivants :

profils mots/document	d_1	d_2	d_3	d_4
arbre	1	0	0	0
coquelicot	2	0	0	0
lion	1	1	0	2
loup	0	1	1	1
mammifère	0	0	0	1
mouche	0	1	0	0
sapin	0	0	1	0
trèfle	0	0	1	0
vache	0	1	0	1
vélo	1	0	0	0
voiture	0	0	1	1

En admettant que l'on dispose, comme ressource sémantique aditionnelle, du thésaurus de la figure 3.2 [page précédente], on peut sur la base de ce thésaurus construire le graphe couvrant les données à indexer[3] augmenté des termes non couverts par la ressource[4] ; on obtient ainsi le graphe donné en figure 3.3 [page suivante].

Dès lors, on peut choisir comme espace de représentation des documents tout ou partie des sommets de ce graphe ; l'extension de l'espace possible de représentation des documents est le principe de base de l'*indexation sémantique*.

Pour notre exemple, considérons deux espaces de représentation distincts : le premier constitué par les concepts directement associés aux mots et uniquement ceux-ci (indexation « synsets », les concepts de l'index étant ceux mis en évidence dans la figure 3.4 [page suivante]), et le second constitué de tous les sommets du graphe de la figure 3.3 [page suivante] (indexation « mots+concepts »).

[2] Pour lesquels on suppose qu'un filtrage sur les catégories morphosyntaxique ainsi qu'une lemmatisation ont préalablement eu lieux.

[3] Et strictement restreint à ces données ; on procédera ainsi en pratique, pour des raisons d'efficacité (les thésaurus utilisés étant très larges et les sommets ainsi ignorés non pertinents).

[4] On a ici fait le choix de les ajouter comme sommets isolés, traités séparément du reste du graphe (en particulier lorsqu'il s'agira de construire des critères de sélection des termes) ; on peut cependant envisager d'ajouter ces sommets comme fils du sommet racine et les traiter comme s'ils avaient fait partie de la ressource. Au final, le résultat est le même ; le premier choix s'avère cependant en pratique moins coûteux en temps de calcul, le second permettant des traitements plus uniformes.

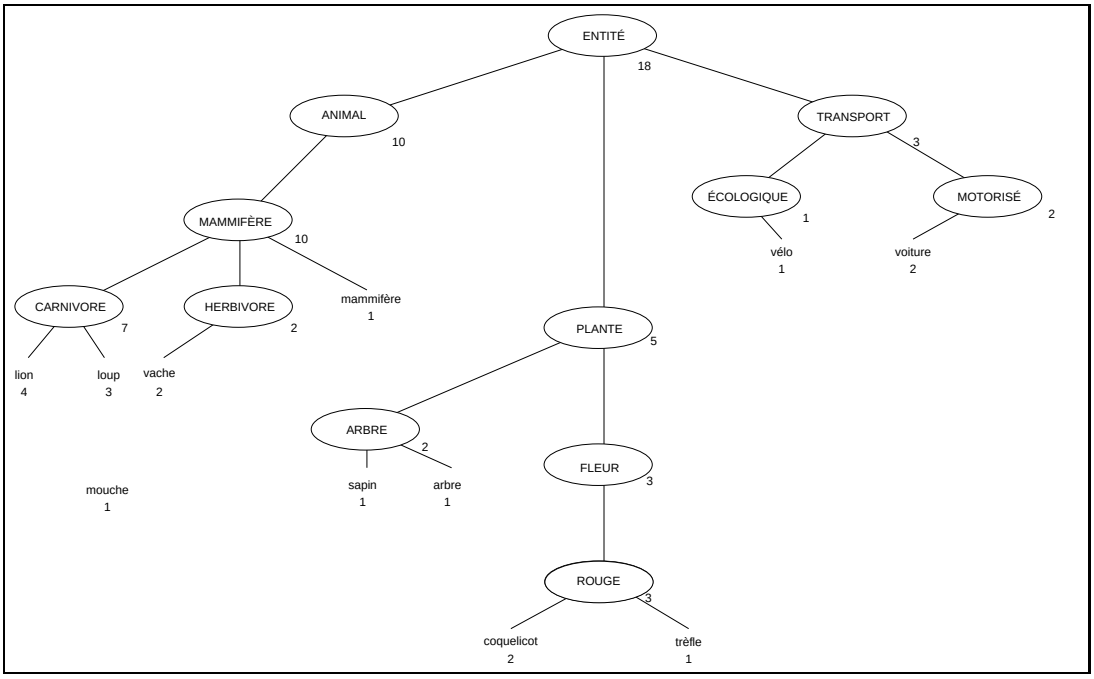


Figure 3.3*:* Thésaurus restreint aux termes présents dans les documents.

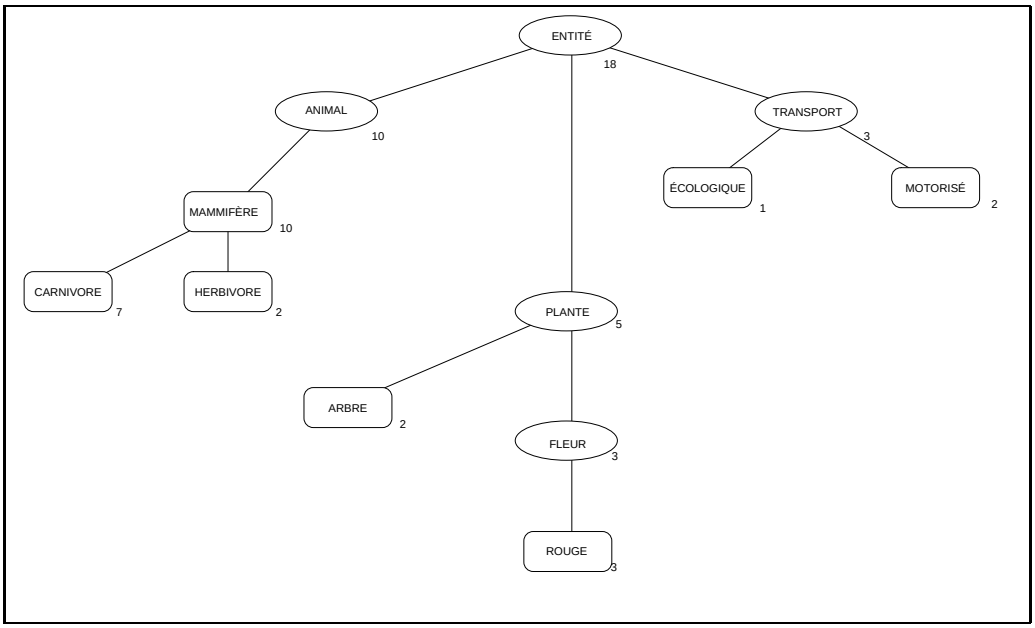


Figure 3.4*:* Thésaurus restreint aux concepts couvrant les termes présents dans les documents (les concepts participant au jeu d'index étant encadrés).

En faisant de plus le choix de comptabiliser les occurrences de *tous* les « descendants »[5] des termes retenus (cas du concept MAMMIFÈRE [6]), les profils des documents issus de l'index « synsets » sont les suivants :

profils synsets	d_1	d_2	d_3	d_4
MAMMIFÈRE	1	3	1	5
ARBRE	1	0	1	0
ÉCOLOGIQUE	1	0	0	0
MOTORISÉ	0	0	1	1
HERBIVORE	0	1	0	1
CARNIVORE	1	2	1	3
ROUGE	2	0	1	0

En appliquant le même principe pour comptabiliser les occurrences des termes d'index, on obtient les profils suivants pour l'indexation « mots+concepts » :

profils mots+concepts	d_1	d_2	d_3	d_4
ENTITÉ	5	3	4	6
ANIMAL	1	3	1	5
PLANTE	3	0	2	0
TRANSPORT	1	0	1	1
FLEUR	2	0	1	0
MAMMIFÈRE	1	3	1	5
ARBRE	1	0	1	0
ÉCOLOGIQUE	1	0	0	0
MOTORISÉ	0	0	1	1
HERBIVORE	0	1	0	1
CARNIVORE	1	2	1	3
ROUGE	2	0	1	0
arbre	1	0	0	0
coquelicot	2	0	0	0
lion	1	1	0	2
loup	0	1	1	1
mammifère	0	0	0	1
mouche	0	1	0	0
sapin	0	0	1	0
trèfle	0	0	1	0
vache	0	1	0	1
vélo	1	0	0	0
voiture	0	0	1	1

L'impact de l'*indexation sémantique* sur la « ressemblance » entre ces documents apparaît clairement dans la table suivante, qui reporte la mesure de similarité cosinus (cf. 2.4.3 [page 15]) entre les documents, établie respectivement sur la base des profils « mots », des profils « synsets » et des profils « mots+concepts ».

[5] On étend, par transitivité, les relations inclusives modélisées ; dans notre exemple, la (seule) relation « est-un ».

[6] MAMMIFÈRE fait partie du jeu d'index, car c'est le concept directement attaché au mot mammifère ; lors du calcul de sa contribution aux différents documents cependant, il sera comptabilisé à la fois pour les occurrences de mammifère et pour celles des mots vache, lion et loup, en raison de sa subordination des concepts auxquels ces mots sont rattachés.

index / Sim-Cos	$S(d_1, d_2)$	$S(d_1, d_3)$	$S(d_1, d_4)$	$S(d_2, d_3)$	$S(d_2, d_4)$	$S(d_3, d_4)$
« mots »	0.19	0.00	0.27	0.25	0.71	0.35
« synsets »	0.47	0.79	0.47	0.60	0.98	0.67
« mots » + « concepts »	0.54	0.12	0.60	0.62	0.96	0.70

L'indexation « synsets » correspond au *remplaçement*, dans le jeu d'index, des mots par les concepts qui leur sont associés dans la ressource, tandis que l'indexation « mots+concepts » correspond à *l'extension* du jeu d'index constitué des mots par les concepts introduits par la ressource.

On constate notamment que l'utilisation de la ressource permet de rendre relativement « équidistants » au document 1 les documents 2 et 4, et de rapprocher les documents 1 et 3 (trop fortement sans doute avec le second jeu d'index), mettant ainsi en évidence une ressemblance qui n'était pas capturée par les profils « mots ». C'est précisément ce type de résultats que l'on souhaite obtenir sur des données réelles, à l'aide des méthodes présentées dans ce chapitre.

3.2.2 Problématique

Choisir de manière *automatique* les termes à utiliser pour l'indexation est délicat. On peut déjà s'en rendre compte avec l'exemple jouet précédent : des termes trop généraux (*i.e.* haut dans la structure hiérarchique) dégraderont les performances du système en associant entre eux des documents peu pertinents, induisant au final une diminution de la précision, tandis que des termes trop spécifiques conserveront une distinction entre mots de sens proches, et ne permettront pas d'associer des documents sémantiquement voisins, induisant au final une diminution du rappel (pour un même nombre de documents extraits).

La problématique principale de l'*indexation sémantique* est donc de choisir correctement, parmi les éléments des ressources sémantiques, les éléments constitutifs du jeu d'index. Quel degré de généralité doivent avoir les termes choisis ? Faut-il remplacer les profils-mots, ou leur ajouter des termes supplémentaires ? Dans ce cas, comment pondérer les profils pour « contrôler » leur influence respective ?

On tentera, dans la suite de ce chapitre, d'apporter des éléments de réponse à ces questions. Les expériences rapportées dans la littérature (voir 2.5.2.2 [page 18]) privilégient principalement deux approches (que nous avons illustrés à l'aide de notre exemple jouet précédent) : l'une consiste à utiliser la totalité des termes possibles (l'ensemble des concepts en sus des mots), et la seconde à n'utiliser que les concepts directement associés aux mots. Entre ces deux extrêmes, nous avons choisi d'explorer une voie médiane, en considérant un sous-ensemble de termes qui ne soient pas nécessairement tous, dans la ressource sémantique, à une même « profondeur » par rapport aux mots, autorisant aussi bien des degrés de généralisation plus élevés qu'une absence de généralisation.

Ces différents cas de figure sont schématisés par l'illustration 3.5 [page suivante] ; l'index (a) représente une indexation traditionelle, par les mots (graphie, lemmes ou racines) ; l'index (b) représente l'expansion du jeu d'index par l'ensemble des termes possibles, il produit donc nécessairement des représentations beaucoup plus grandes ; l'index (c) représente une indexation par les concepts associés aux mots, les représentations sont ainsi moins grandes, des familles de mots ($\approx$ synonymes) étant factorisés par un seul concept ; et finalement l'index (d), qui étend le principe de factorisation à des concepts hyperonymes plus généraux dans certains cas, et conserve une distinction fine dans d'autres. Les index (a), (b) et (c) correspondent respectivement aux indexations « mots », « mots+concepts » et « synsets » de l'exemple de la section précédente ; les index des la dernière famille (d) étant ceux sur lesquels nous avons concentrés nos travaux.

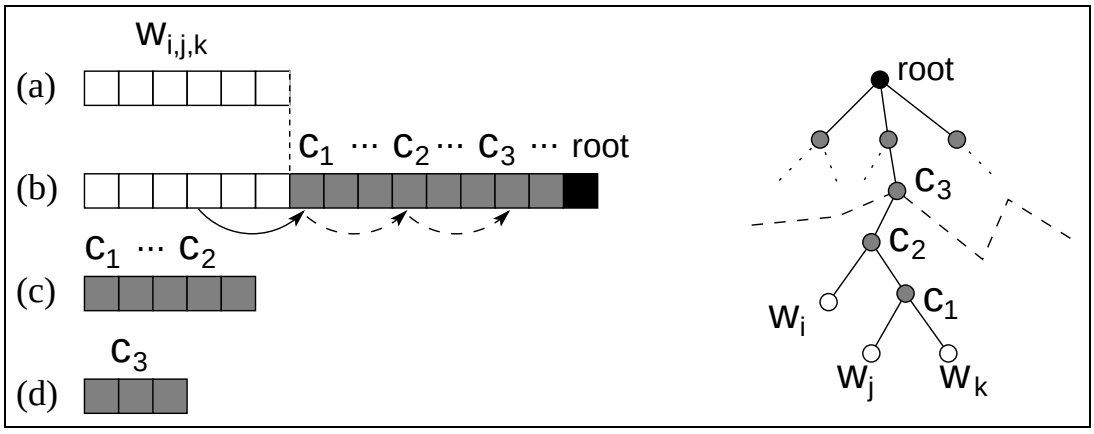


Figure 3.5*:* Différentes possibilités d'indexation en présence de connaissances hiérarchisées. Sur la gauche de la figure sont schématisés les vecteurs-profils produits par les différentes indexations, tandis que la partie droite représente l'organisation hiérarchisée par la ressource sémantique des termes d'index potentiels, c_i désignant des concepts et w_i des mots.

3.3 Indexation sémantique guidée par les données

Bien qu'il soit certainement souhaitable pour un système de recherche documentaire de prendre en compte un maximum d'informations, en particulier des informations de nature sémantique, la solution simpliste (cas (b) de la figure 3.5 [ci-dessus]) consistant à utiliser comme jeu d'indexation la totalité des termes disponibles (les termes « initiaux », présents dans les documents, ainsi que les termes « additionnels » issus du thésaurus), tel que proposé par Kiryakov et Simov [1999] se révèle en pratique clairement contre-productive (cf. 4.4.1 [page 64]). Ceci n'est guère surprenant : tenter de discriminer quelques documents parmi un ensemble sur la base d'un très grand nombre de critères est difficile à réaliser, la « distance » – généralement une similarité ou une dissemblance – entre chaque paire de documents tendant à devenir à peu près la même (effet connu sous le nom de « *curse of dimensionality* » – à ce propos, consulter (entre autres) Bellman [1961], Beyer *et al.* [1999], Pestov [1999]).

Kiryakov et Somov, qui avaient anticipé ce problème, proposent comme palliatif de limiter le nombre de termes hyperonymes adjoints à chaque terme de base en ne prenant pas en considération les hyperonymes trop abstraits, ou en demandant à l'utilisateur de définir le sous-ensemble d'hyperonymes à conserver. Malheureusement, outre l'arbitraire de cette mesure et son caractère non automatique, elle devient irréalisable en présence de thésaurus de grande taille tels qu'*EDR*, pour lesquels la majeur partie des hyperonymes sont précisément des concepts abstraits.

L'idée développée ici est d'utiliser la connaissance *a priori* des données à indexer (recherche documentaire classique) pour guider le choix du jeu d'indexation, en utilisant des critères statistiques ou issus de la théorie de l'information, dans le but d'obtenir une représentation des documents (par le biais de leurs profils vectoriels) permettant de les discriminer au mieux, compte tenu des particularités des ressources sémantiques à disposition. En effet, comme illustré par la figure 3.6 [page suivante], tant la précision de description de la ressource (ou de la combinaison de ressources) que la répartition des données des documents n'ont, en pratique, guère de chances d'être homogènes. Dès lors, choisir un degré de généralité pré-déterminé et constant ne se justifie pas.[7]

[7] Même en considérant le coût de l'indexation ou si les données à indexer ne sont pas connues *a priori*– à ce sujet, se rapporter aux conclusions de ce chapitre.

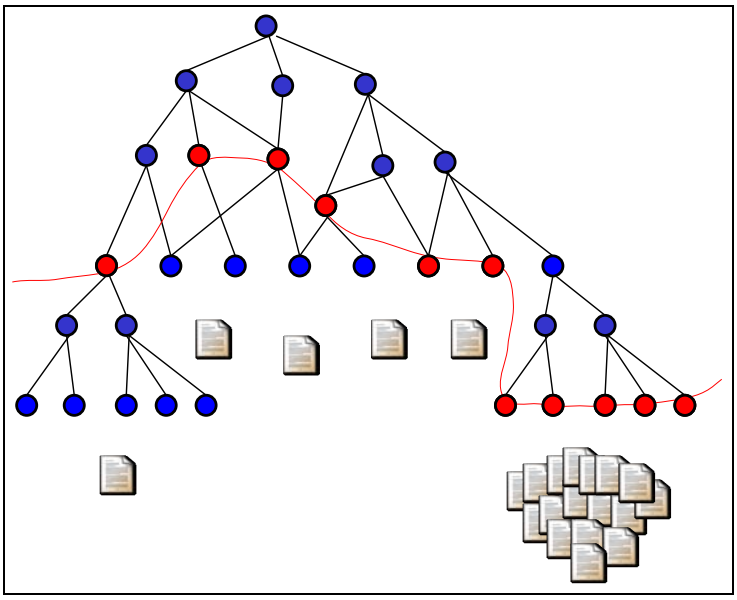


Figure 3.6*:* Thésaurus non homogène (granularité de description) et distribution non homogène des données à indexer.

En ce qui concerne la sélection des termes d'index en fonction des données, les différentes solutions envisagées pour automatiser cette sélection et présentées plus en détails dans la suite de ce chapitre ont été divisées de la façon suivante :

1. les critères de sélection « locaux » :
   les termes d'index sont choisis un à un sur la base d'une mesure locale (*i.e.* définie au niveau d'un sommet s), indépendamment les uns des autres, ou éventuellement en conditionnant le choix du $(i+1)^{\text{e}}$ terme par les i termes déjà choisi ;
2. les critères de sélection « globaux » :
   les termes sont choisi globalement (sélection de la totalité du jeu d'indexation), sur la base d'une mesure définie sur un ensemble de sommets, ce qui implique la définition d'une structure (Γ) sous-jacente déterminant un jeu d'indexation.
   (a) les critères « séparables » :
   pour lesquels la mesure induisant le critère est optimisable localement (*séparable*[8]),
   (b) les critères « non séparables » :
   pour lesquels la mesure est quelconque (non séparable).

   Remarque :

   Cette séparation prend tout son sens lorsque l'on considère les algorithmes à mettre en œuvre pour l'implémentation des critères.

Les critères sont construits sur la base de fonctions de sélection $\mathcal{F}(\mathcal{S}, \{0,1\}))$[9] prenant valeur dans l'ensemble $\{1,0\}$, selon que leur argument est ou non retenu pour l'indexation. La table 3.1 [page suivante] présente les différentes classes de mesures et fonctions de sélection et quelques-unes de leurs instanciations possibles, selon la famille de critère. Les fonctions de sélection des critères locaux s'appliquent à des sommets $s \in \mathcal{S}$ de G (où $\mathcal{S}$ est l'ensemble des sommets de G) et utilisent une mesure locale sur les sommets, notée $\mu(s)$. Les fonctions de sélection des

[8] Un critère c est *séparable* si tout sous-ensemble de c est *préférentiellement indépendant* dans c, par rapport à une relation de préférence globale ($\lesssim$) ; Bellman [1957, 2003].

[9] Où $\mathcal{F}(\mathcal{A}, \mathcal{B})$ représente l'ensemble des fonctions (applications) $f : \mathcal{A} \to \mathcal{B}$, faisant correspondre un et un seul élément de $\mathcal{B}$ à tout élément de $\mathcal{A}$.

	critères «locaux»	*critères «globaux»*	
		séparables	*non-séparables*
mesure locale : $\mu(s)$	$\mathsf{tf}(s), \mathsf{tfidf}(s), \chi^2(s), IM(s,d), P_{\mathsf{tf}}, P_{\mathsf{tfidf}}, \ldots$		
mesure globale séparable : $\lambda(\Gamma)$		$\sum_\Gamma \mu, E(\mu), dL(\Gamma), \ldots$	$H(\Gamma), \ldots$
mesure globale non-sép : $\Lambda(\Gamma)$			$R(\Gamma), \ldots$
fct de sélection locale : $F \in \mathcal{F}(\mathcal{S}, \{0,1\})$	$\in$ intervalle, k meilleures, ...		
fct de sélection globale : $F' \in \mathcal{F}'(\Upsilon, \{0,1\})$		$\max, \min$	

Table 3.1*:* Nomenclature des mesures et fonctions de sélection du jeu d'index, selon la famille de critère. $\mathcal{S}$ représente l'ensemble des sommets du thésauraus, s l'un de ces sommets et Γ un sous-ensemble de sommets (définissant une coupe) ; $\mu(s)$ est une mesure définie sur s, $\lambda(\Gamma)$ une mesure séparable définie sur Γ et $\Lambda(\Gamma)$ une mesure non séparable, également définie sur Γ. Les critères sont construits sur la base de fonctions de sélection $\mathcal{F}$.

critères globaux s'appliquent à des ensembles de sommets $\Gamma \in \Upsilon$, où Γ est un sous-ensemble de sommet correspondant au jeu d'index ; elles se résument principalement à minimiser ou maximiser une mesure globale ($\lambda(\Gamma)$ ou $\Lambda(\Gamma)$, selon que le critère est séparable ou non).

3.4 Critères « locaux »

La problématique de la sélection des termes d'indexation est largement répandue et étudiée depuis longtemps ; elle constitue en fait un domaine propre de recherche en apprentissage automatique (*feature selection*). Il existe tout un ensemble de critères susceptibles d'être mis en œuvre dans le cas qui nous intéresse et permettant de choisir les termes ayant la plus forte *valeur de discrimination* ; nous ne les détaillerons pas ici, mais en employerons quelques uns pour illustrer notre propos (le lecteur intéressé peut se reporter en 2.4.1.1 [page 11] pour une revue des principaux critères et quelques liens bibliographiques).

- D'une part, on peut considérer les critères de *filtrage* des termes non pertinents ; un exemple très simple est le filtrage fréquentiel, pour lequel seuls les termes dont la fréquence d'occurrence est comprise entre deux valeurs a et b sont conservés, ce qui permet, avec des bornes judicieusement choisies, d'éliminer les sommets extrêmes du graphe (proches de la racine et proches des feuilles). En utilisant le formalisme introduit avec la table 3.1, la mesure (locale) des sommets qui correspond à ce critère est :

$$\mu(s) = \mathsf{tf}(s) \tag{3.1}$$

  Et sa fonction de sélection :

$$F(s) = \chi_{[a,b]}(\mu(s)) \tag{3.2}$$

- D'autre part, on peut considérer de façon complémentaire la famille des critères de *sélection*, consistant globalement à ne retenir que les k meilleurs termes au sens de la mesure locale (μ). On pourrait par exemple ne retenir que les termes ayant le plus grand coefficient $\mathsf{tf.idf}$:

$$\begin{aligned} \mu(s) &= \mathsf{tf}(s) \cdot \mathsf{idf}(s) \\ F(s) &= \chi_{[0,k[}(rang(s, \mathcal{S}, \mu_{\geq}))^{10} \end{aligned} \tag{3.3}$$

Cependant, ces deux critères ne sont pas totalement satisfaisants : bien que les termes extrêmes (très généraux – donc sur-représentés – et très spécifiques – sous-représentés) puissent probablement être écartés, il y a fort à parier que les sommets conservés le soient avec leurs voisins directs dans le réseau (prédecesseurs et successeurs), ce qui n'est guère souhaitable dans la plupart des cas.

On peut cependant éviter cette situation en tenant explicitement compte de la topologie du réseau au moment de la sélection : plutôt que de sélectionner les termes indépendamment les uns des autres, il est préférable de conditionner le choix du $(i+1)^{\text{e}}$ terme par les i termes déjà sélectionnés ; par exemple en évitant de choisir des termes trop « proches » dans le réseau.

3.4.1 Algorithme

Une réalisation possible de sélection avec un tel conditionnement est donnée par l'algorithme 1, qui isole un ensemble d'indexation (d'au plus k éléments) en excluant la cohabitation de termes liés par une relation sémantique r donnée.

Algorithme 1 Sélection itérative : maximisation de μ et exclusion de termes sémantiquement liés.

Nécessite : les relations r à considérer, le nombre maximum k de termes à retenir.
Fourni : un jeu d'indexation I.

$I \leftarrow \varnothing$ # *Le jeu d'indexation*
$C \leftarrow \mathcal{S}$ # *Les sommets candidats*
tant que $(|I| < k)$ et $(C \neq \varnothing)$ **faire**
  $s \leftarrow \text{argmax}(\mu(C))$
  $I \leftarrow I \cup s$
  $C \leftarrow C \setminus \{s \cup s_r^{\Downarrow} \cup s_r^{\Uparrow}\}$
fin tant que
retourne I

Avec $s_r^{\Downarrow}$ l'ensemble des sommets résultant de la fermeture transitive de $s_r^{\downarrow}$, l'ensemble des sommets successeurs (dominés) de s via un arc de l'ensemble r. Similairement pour $s_r^{\Uparrow}$, mais avec les sommets prédécesseurs (dominants) de s.

Remarques :

- Pour être consistant, la mesure de similarité utilisée entre vecteurs profils devrait être en adéquation avec la représentativité des différentes composantes des vecteurs (importance relative des termes d'indexation) ; on peut assurer cette adéquation en pondérant les fréquences brutes dans les vecteurs profils, lors de l'indexation, ou en intégrant cette pondération dans la fonction de similarité, si elle s'y prête (cas de la similarité cosinus)[11].
- Les critères de cette famille sont en fait implémentés au moyen d'algorithmes *gloutons* (on part d'une solution partielle non réalisable, et on fixe à chaque étape une ou plusieurs variables de manière définitive, jusqu'à déterminer une solution réalisable).

[10] Avec $rang(x, E, \mu_r)$ le nombre de successeurs du minorant de l'ensemble E à parcourir pour atteindre l'élément $n \in E$, avec E un ensemble totalement ordonné par la relation d'ordre r sur la mesure μ (appliquée aux éléments de E).

[11] Naturellement, en plus du ratio entre les valeurs de μ sur les différents termes d'index, toutes les techniques de « lissage » peuvent être envisagées (normalisation linéaire, ordonnancement, etc.).

En alternative aux techniques construisant le jeu d'indexation par sélection ou filtrage des termes du thésaurus, on peut également envisager de mettre en œuvre des techniques d'*extraction de termes* (*réduction de dimensionalité*) :

- par le biais de regroupements (*clustering* : nuées dynamiques, *k-means*, classification hiérarchique, etc.)
- par le biais de techniques factorielles, de type LSI (cf. 2.4.1.2 [page 13]) ; dans ce cas, compte tenu du grand nombre de paramètres à prendre en compte, des techniques plus « légères » sont sans doute préférables, telles que *polynomial filtering* [Kokiopoulou et Saad, 2004], *random indexing* [Sahlgren, 2005] et peut-être *GLSA* [Matveeva *et al.*, 2005].

3.5 Critères globaux

Comme souvent avec les algorithmes de type glouton, le conditionnement local du choix d'un terme par les choix antérieurs peut être avantageusement remplacé par un conditionnement plus global (au prix cependant d'une plus grande complexité de traitements).[12] Plutôt que de choisir un à un les termes d'index, l'idée est ici de choisir le jeu d'index dans sa globalité, comparant pour cela plusieurs jeu d'index en compétition. On obtient ainsi un conditionnement global sur le choix des termes retenu pour l'indexation.

Pour disposer d'un tel conditionnement, il est nécessaire de définir une structure pour le jeu d'index (et de lui associer une mesure). Nous avons choisi d'utiliser à cette fin une structure de *coupe* dans le thésaurus, choix logique permettant de garantir une bonne couverture des mots présents dans les documents tout en excluant au maximum la redondance introduite par la présence de sommets voisins, comme cela est expliqué dans la section suivante.

3.5.1 Notion de coupe

Pour limiter, dans le jeu d'indexation, la redondance induite par certaines relations sémantiques définies dans le thésaurus, il convient de ne pas conserver simultanément deux sommets liés par cette relation. Dans le cas d'une relation inclusive,[13] il faut non seulement exclure les cas de liaisons directes, mais également celles obtenues par transitivité ; on cherche donc une structure qui interdise la sélection simultanée d'un sommet et de l'un de ses descendants ou ascendants. Une telle structure sera appelée une *coupe* dans le thésaurus.[14]

Dans le cas d'un arbre, on définira une coupe Γ comme un ensemble *minimal* (au sens de l'inclusion)[15] de sommets induisant une *partition*[16] des feuilles de l'arbre ; en d'autre terme, Γ est un ensemble de sommets tels que, pour toute feuille f de l'arbre, le chemin menant de f à la racine passe nécessairement par un et un seul sommet de Γ.

Pour généraliser cette notion de coupe au cas d'un DAG, où un même sommet peut avoir plusieurs prédecesseurs (il peut y avoir plusieurs chemins menant de f à la racine), il est nécessaire, pour garantir la maximalité de la *couverture* de la coupe, de relâcher la contrainte de partitionnement :

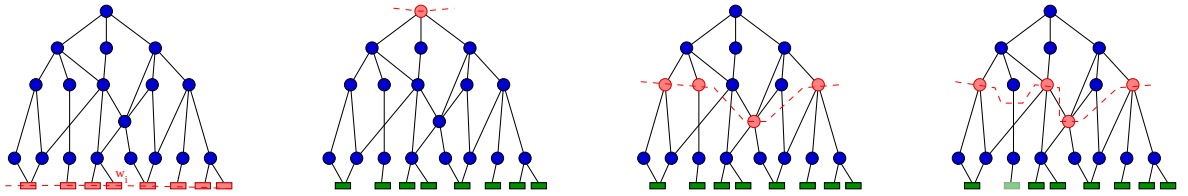
[12] En effet, on évite ainsi que des choix malheureux aient de fortes conséquences sur le reste de la sélection.

[13] Tels que le sont les relations « est-un » de *EDR* et d'hypo/hyperonymie de *WordNet*.

[14] Cette notion s'apparente aux « coupes » de la théorie des flots et réseaux de transport, mais elle est ici définie sur les sommets, et non les arcs.

[15] On entend qu'aucun sommet ne peut être retiré de la coupe sans en diminuer la couverture (*i.e.* conserver la couverture de l'ensemble des feuilles).

[16] Une *partition* d'un ensemble E étant une famille de sous-ensembles de E deux à deux disjoints tels que leur union est l'ensemble E.



(A) Une coupe sur les feuilles.

(B) Coupe sur la racine.

(C) Une coupe quelconque.

(D) Un ensemble de sommets qui n'est pas une coupe.

Figure 3.7*:* Exemples et contre-exemple de coupes dans un DAG.

Définition : coupe Γ

Une coupe Γ sur un DAG G est un ensemble *minimal* (au sens de l'inclusion)[15] de sommets *couvrant* la totalité des feuilles de G.

Exemple :

L'ensemble restreint à la « racine » de G, de même que l'ensemble des feuilles de G constituent tous deux des coupes de G (cas (A) et (B) de la figure 3.7).

Le cas (C) de cette même figure illustre une coupe quelconque correcte, tandis que les sommets marqués du cas (D) ne constituent par contre pas une coupe, et ce à double titre : la contrainte de minimalité n'est pas satisfaite (on peut retirer un sommet de l'ensemble sans diminuer le nombre de feuilles couvertes), pas plus que celle de la maximalité de la couverture (une des feuilles du DAG n'est pas couverte).

Remarques :

- On peut montrer que les jeux d'index produits par l'algorithme de sélection selon un critère local donné précedemment (1 [page 32]) correspondent en fait à des *coupes*.
- Pour un arbre complet d'arité b, avec N feuilles, le nombre de coupes possibles est de l'ordre de $\Theta(2^{N/b})$. Cependant, dans le cas général, ce nombre dépend fortement de la topologie interne du réseau (et pas uniquement de sa profondeur). Consulter Li [1998] pour la démonstration (et la formule permettant de déterminer par récurrence le nombre de coupes possibles, pour un arbre complet).

Pour être en mesure de déterminer la valeur d'un critère sur une coupe donnée, autorisant ainsi le choix d'une coupe parmi un ensemble, il est nécessaire d'associer aux sommets de la coupe une mesure reflètant la contribution du terme dans les données à indexer. On définit à cette fin un modèle probabilisé de coupe :

Définition : coupe probablisée M

Un modèle probabilisé de coupe $M = (\Gamma, \theta)$ est un couple associant une coupe Γ et une distribution de probabilité θ sur les éléments de la coupe Γ.

En général, θ est donné par un estimateur associé aux sommets de la coupe, de la forme :

$$\theta_s = P(s|\Gamma) \widehat{=} \frac{\mu'(s)}{\sum_{s_i \in \Gamma} \mu'(s_i)}, \tag{3.4}$$

avec μ' une mesure élémentaire définie sur un sommet.

Muni de notre structure définissant un jeu d'index (modèle de coupe probabilisée), nous pouvons maintenant examiner comment conditionner globalement le choix des termes d'index.

3.5.2 Critères globaux, séparables

Nous nous proposons d'examiner dans un premier temps la famille des critères (travaillant sur un jeu d'indexation complet) dont la fonction de sélection F' utilise une mesure *séparable* λ définie sur des coupes Γ, *i.e.* calculable par recomposition d'un calcul local (et donc optimisable localement, pour autant que l'on soit en présence d'une structure sémantique en arbre).

3.5.2.1 Critères « locaux » étendus à la structure de coupe

On peut dans un premier temps envisager de « rendre globaux » tout ou partie des critères de sélection locaux (3.4 [page 31]). En guise d'exemple, reprenons le critère de la fréquence absolue pondérée par la fréquence en documents ($\mathsf{tf.idf}$, 3.3 [page 31]) ; une généralisation naïve à un ensemble de sommets pourrait être :

$$\begin{aligned} \mu(s) &= \mathsf{tf}(s) \cdot \mathsf{tf.idf}(s) \\ \lambda(\Gamma) &= \sum_{s \in \Gamma} \mu(s) \\ F'(\Gamma) &= \delta\left(\Gamma, \operatorname*{argmax}_{\Gamma_i \in \Upsilon} (\lambda(\Gamma_i))\right) \end{aligned} \tag{3.5}$$

avec Υ l'ensemble des coupes possibles dans G.

Mais ce critère (donné en guise d'exemple) à en pratique peu de chance de produire des résultats intéressants. En effet, comme il s'agit de maximiser une somme, indépendamment du nombre de termes qu'elle comporte, il est fort probable que l'optimum soit la coupe constituée de l'ensemble des feuilles (celle admettant le plus de termes, exception faite de thésaurus avec une topologie vraiment particulière). Cette problématique se retrouve avec plusieurs des critères suggérés en 3.4 [page 31], notamment celui *d'entropie maximale* (voir 3.5.3.1 [page 40] pour une adaptation efficace de ce critère). C'est également le cas du critère visant à maximiser *l'information mutuelle* entre jeu d'indexation et données à indexer (consulter l'annexe B [page 131] pour une démonstration), et ceux qui lui sont apparentés.

Par ailleurs, compte tenu de la taille de l'ensemble des coupes possibles Υ, son évaluation complète en vue de trouver celle pour laquelle la mesure λ est maximum est impossible en pratique. L'algorithme d'implémentation du critère doit donc être adapté en conséquence (voir 3.5.2.3 [page 38]).

3.5.2.2 Critère *MDL* (Li et al.)

Dans un contexte simmilaire[17] mais pour une tâche quelque peu différente,[18] Li et Abe proposent [Li et Abe, 1998] de recourir au critère *MDL* (*description de longueur minimale* – « *Mi-*

[17] Coupe dans un réseau sémantique (*WordNet*) selon les occurrences des feuilles du réseau dans un jeu de données.

[18] Acquisition automatique de patrons pour relations prédicats-arguments (*case frame patterns*)

nimum Description Length »), introduit par Rissanen pour la compression de données et l'estimation statistique [Rissanen, 1989].

Étant donné un thésaurus $G = [\mathcal{S}, \mathcal{R}]$ et l'observation d'occurrences des feuilles $\mathcal{M}$ dans un jeu de données O, Li et Abe utilisent, pour représenter ces données, la coupe $\Gamma \in \mathcal{S}$ pour laquelle le critère *MDL* est minimum. Nous résumons ci-après la manière dont le critère est calculé ; pour une description complète ainsi que la justification de l'application du principe *MDL*, consulter Li [1998], Li et Abe [1998].

La *longueur de description* du modèle de coupe probabilisé $M = (\Gamma, \theta)$ et du jeu de données O représenté au moyen de M, $dL(M, O)$, est la somme des longueurs de *description du modèle*, $dL(\Gamma)$, de *description des paramètres*, $dL(\theta|\Gamma)$ et de *description des données*, $dL(O|\Gamma, \theta)$:

$$dL(M, O) = dL((\Gamma, \theta), O) = dL(\Gamma) + dL(\theta|\Gamma) + dL(O|\Gamma, \theta) \tag{3.6}$$

La longueur de description du modèle $dL(\Gamma)$ dépend du codage employé pour leur représentation ; sous l'hypothèse que les modèles de coupe sont *a priori* égaux (justifié par l'interprétation Bayésienne du critère), on admettra une longueur de codage identique pour chacun d'eux :

$$dL(\Gamma) = \log(|\mathcal{G}|) = K(\text{une constante, à } \mathcal{G} \text{ fixé}) \tag{3.7}$$

La longueur de description des paramètres, $dL(\theta|\Gamma)$ est donnée par :

$$dL(\theta|\Gamma) = \frac{k}{2} \cdot \log(|O|), \tag{3.8}$$

avec $k = |\Gamma| - 1$, le nombre de paramètres libres du modèle.

La longueur de description des données, $dL(O|\Gamma, \theta)$ est donnée par :

$$dL(O|\Gamma, \theta) = -\sum_{o \in O} \log(P(o)) \tag{3.9}$$

où $P(o)$ est obtenu par l'*estimateur de vraisemblance maximale* :

$$P(o) = \frac{P(s)}{|s^{\Downarrow} \in \mathcal{M}|} \text{ et } P(s) = \frac{f(s)}{|O|}$$

pour chaque sommet $s \in \Gamma$ et chaque occurrence o couverte par s, avec $f(s)$ le nombre d'occurrences du sommet s dans le jeu de données.

En reprenant la terminologie de la table 3.1 [page 31], on peut exprimer ce critère de la manière suivante :

$$\begin{aligned} \mu(s) &= -\frac{\log(|O|)}{2} \cdot f(s) \cdot \log(P(o)) \\ \lambda(\Gamma) &= \left(\sum_{s \in \Gamma} \mu(s) \right) - \frac{\log(|O|)}{2} \\ F'(\Gamma) &= \delta\left(\Gamma, \operatorname*{argmin}_{\Gamma_i \in \Upsilon} (\lambda(\Gamma_i)) \right) \end{aligned} \tag{3.10}$$

Comme d'autres l'ont constaté [Brockmann et Lapata, 2003, Clark et Weir, 2002, Tomuro, 2001], ce critère tend malheureusement à sélectionner des coupes trop générales en cas de représentation globalement équilibrée des différentes feuilles dans le jeu de données, ce qui est précisément le cas pour la tâche qui nous intéresse.[19] Nous ne donnons pas ici de démonstration formelle de cette faiblesse, mais l'illustrons empiriquement sur la base d'un exemple donné par

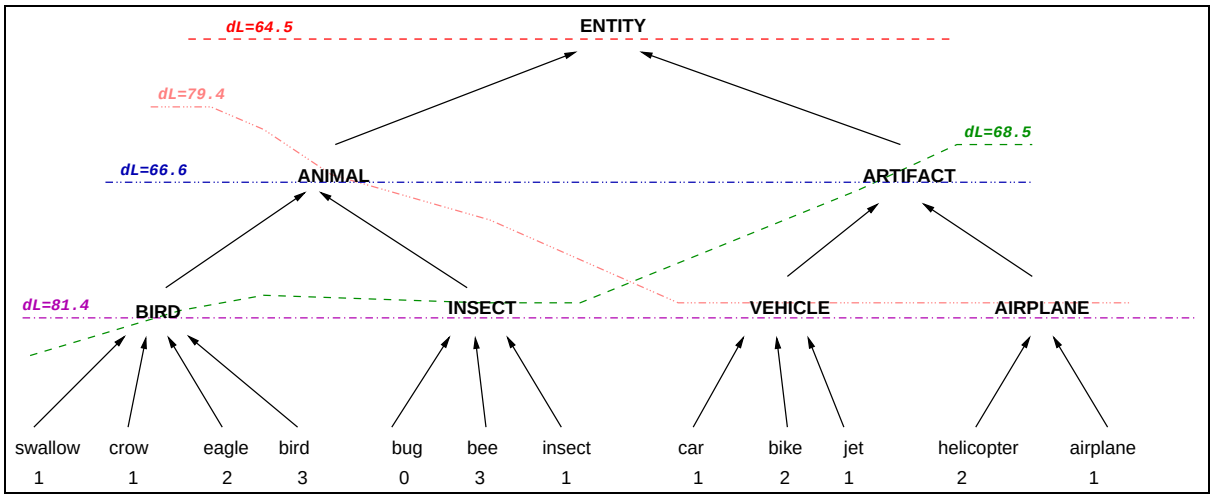


Figure 3.8*:* Critère *MDL* : longueur de description de différentes coupes.
Les fréquences d'occurrences (observées) des termes sont indiquées au-dessous de ceux-ci.

Li et Abe eux-mêmes dans leurs différentes publications, en modifiant simplement les fréquences d'occurrence des feuilles.

En considérant l'exemple de réseau sémantique donné en figure 3.8 [ci-dessus], et en admettant un jeu de données résultant en les fréquences d'occurrences des feuilles du réseau telles qu'indiquées sur la figure, on obtient, pour chaque sommet du réseau, les probabilités suivantes :

s	$f(s)$	$\|s\|$	$P(s)$	$P(o)$	$-f(s)\cdot\log(P(o))$[20]
Entity	18	12	1.0000	0.0833	64.5293
Animal	11	7	0.6111	0.0873	38.6963
Artifact	7	5	0.3889	0.0778	25.7915
Bird	7	4	0.3889	0.0972	23.5380
Insect	4	3	0.2222	0.0741	15.0196
Vehicle	3	2	0.1667	0.0833	10.7549
Airplane	4	3	0.2222	0.0741	15.0196
swallow	1	1	0.0556	0.0556	4.16993
crow	1	1	0.0556	0.0556	4.16993
eagle	2	1	0.1111	0.1111	6.33985
bird	3	1	0.1667	0.1667	7.75489
bug	0	1	0.0	0.0	0.0
bee	3	1	0.1667	0.1667	7.75489
insect	1	1	0.0556	0.0556	4.16993
car	1	1	0.0556	0.0556	4.16993
bike	2	1	0.1111	0.1111	6.33985
jet	1	1	0.0556	0.0556	4.16993
helicopter	2	1	0.1111	0.1111	6.33985
airplane	1	1	0.0556	0.0556	4.16993

[19] De fait, l'utilisation de ce critère pour sélectionner des coupes dans nos données de tests conduit systématiquement à la coupe réduite à la seule racine ; ce qui n'est bien entendu pas acceptable pour l'indexation des documents.

[20] La somme de cette valeur sur tous les sommets de la coupe correspond à la *longueur de description des données* $L(O|\Gamma,\theta)$; c'est pourquoi elle est indiquée ici.
Par ailleurs, on utilise ici le logarithme en base 2 ($\log = \log_2$).

En appliquant l'algorithme de recherche de coupe proposé par Li et Abe [1998] (dont l'algorithme 2 [page suivante] est une version équivalente), on obtient Γ = [Entity] comme coupe optimale. Ci-après, le détail du calcul du critère *CRM* pour cette coupe et quelques autres (parmi les 26 différentes coupes possibles du réseau) :

Γ	$L(\theta\|\Gamma)$	$L(O\|\Gamma,\theta)$	$L(M,O)$
[Entity]	0.0	64.5293	$K + 64.5293$
[Animal, Artifact]	2.0847	64.4878	$K + 66.5728$
[Bird, Insect, Artifact][21]	4.1699	64.3490	$K + 68.5190$
[Animal, Vehicle, Airplane]	4.1699	75.2427	$K + 79.4126$
[Bird, Insect, Vehicle, Airplane]	6.2549	75.1039	$K + 81.3588$

Ce critère n'est donc pas adapté à l'utilisation que l'on souhaite en faire.

3.5.2.3 Algorithme

Les fonctions de sélection de la famille des critères globaux nécessitent d'isoler une coupe parmi l'ensemble des coupes possibles. Cet ensemble étant de taille considérable, l'énumération de chaque solution n'est pas envisageable.

Pour la mise en œuvre de leur critère, Li et Abe proposent un algorithme de programmation dynamique [Bellman, 1957, 2003] (voir l'algorithme 2 [page suivante]). Ce paradigme de résolution consiste à plonger le problème à résoudre dans une famille de sous-problèmes de même nature mais de taille suffisamment petite pour qu'une solution optimale puisse être exhibée, puis de relier, par une relation de récurrence, les solutions de ces sous-problèmes et composer ainsi la solution optimale pour le problème d'origine [Gondran et Minoux, 1995].

Le fonctionnement de l'algorithme est relativement simple, et correspond en fait à une procédure d'exploration par séparation et évaluation (*branch and bound*, Land et Doig [1960]). L'algorithme optimise le critère localement, sur des portions données du thésaurus,[22] puis combine ces optimas locaux de sorte à couvrir la totalité du graphe. Plus précisément, il détermine pour chaque sommet s rencontré lors d'un parcours « en profondeur d'abord », l'ensemble d'indexation optimal pour le graphe partiel dominé par s, en comparant le critère calculé uniquement sur s à celui calculé sur la réunion des sous-ensembles optimaux d'indexation obtenu (récursivement) pour chaque $s^{\downarrow}$ sommets successeurs de s ; des deux ensembles d'indexation, le plus performant (au sens du critère) est retourné.

Pour obtenir le jeu d'indexation optimal sur l'ensemble du graphe, il suffit d'appliquer l'algorithme en partant de la racine du thésaurus.

S'il est effectivement efficace (linéaire en le nombre de sommets du graphe), cet algorithme souffre cependant d'une limitation majeure : il ne permet de trouver une coupe optimale que dans un arbre.

En effet, bien qu'il puisse être appliqué sur une structure DAG [23] – ce qui revient à transformer le DAG en arbre selon le principe esquissé en section 3.5.4.1 [page 44] – rien ne garantit plus

[21] Coupe *optimale* selon le critère de *redondance minimale*, présenté ci-après (cf. 3.5.3.1 [page 40]).

[22] Ce qui est justement possible en raison de la séparabilité du critère.

[23] Il faudra dans ce cas faire particulièrement attention à l'union de deux sous-jeux d'indexation, et un marquage des sommets visités pourrait être souhaitable.

Algorithme 2 FindCGS recherche de la coupe optimale (minimisant λ) d'un graphe partiel.

Nécessite : un sommet s.
Fourni : un ensemble de sommets (jeu d'indexation).

$\Gamma' \leftarrow \{s\}$ # *coupe indexant le sous-graphe par la racine*
$\Gamma" \leftarrow \varnothing$ # *combinaisons des sous-coupes*
si $s \in \mathcal{M}$ **alors**
  retourne Γ'
fin si
pour $s_i \in s^{\downarrow}$ **faire**
  $\Gamma" \leftarrow \Gamma" \cup \mathsf{FindCGS}(s_i)$
fin pour
retourne $\operatorname{argmin}(\lambda(\Gamma'), \lambda(\Gamma"))$

que la solution exhibée soit globalement optimale,[24] ni même qu'elle constitue une *coupe*, la contrainte de minimalité pouvant être violée.[25]

En présence d'un DAG, on pourra cependant améliorer la qualité de la coupe obtenue en utilisant (éventuellement après une première recherche au moyen de cet algorithme) un algorithme approché, tel que celui que nous proposons pour implémenter les critères non-séparables (cf. recherche par voisinage, algorithme 3 [page 43]).

Remarque :

Les critères consistant à minimiser une mesure globale séparable peuvent cependant être ramenés au cas largement étudié de la détermination de capacité maximale d'un réseau de transport (recherche d'une coupe minimale sur les arcs). Cela s'obtient en dédoublant chaque sommet, le binôme obtenu étant relié par un arc unique dont la capacité équivaut à la mesure λ sur le sommet initial, en dotant les autres arcs (relations sémantiques) d'une capacité infinie, et en ajoutant un sommet dominé par l'ensemble des feuilles, jouant le rôle de la source (la racine étant le puit, ou inversément). Des algorithmes performants (et moins limités que le « Ford Fulkerson » de 1956) ont été développés ces dernières années, notamment Stoer et Wagner [1997], et sont susceptibles d'être utilisés.

Une autre piste potentiellement intéressante est donnée par les travaux de Bienkowski *et al.* [2003], Harrelson *et al.* [2003], Räcke [2002], consistant à transformer un réseau unidirectionnel en arbre tout en capturant au maximum les capacités de transport du graphe originel.

3.5.3 Critères globaux, non séparables

Nous nous proposons enfin d'examiner la famille des critères travaillant sur un jeu d'indexation complet et dont la fonction de sélection F' utilise une mesure Λ *non-séparable*, définie sur des coupes Γ. Nous donnons également la forme générale d'un algorithme de recherche par voisinage permettant d'implémenter non seulement les critères de cette famille, mais également les critères séparables qui ne pourraient être calculés par programmation dynamique, en raison de la topologie du réseau sémantique (DAG).

[24] L'indépendance préférentielle des critères, nécessaire à l'application de la programmation dynamique, n'est plus vérifiée.

[25] Il est cependant possible de « rectifier » le jeu d'indexation de sorte à retrouver une coupe bien formée, en supprimant les sommets dominés par d'autres ; l'optimalité (au sens du critère) du jeu obtenu est cependant là aussi fortement remise en cause.

3.5.3.1 Critère de Redondance Minimale

L'idée sous-tendue par le critère que nous proposons ici est de chercher un jeu d'indexation pour lequel les probabilités d'occurrences des termes d'index dans la collection soient le plus uniforme possible (entropie maximale). Cette proposition a fait l'objet d'un certain nombre de publications [Seydoux et Chappelier, 2005*a*,*b*,*c*,*d*], dans lesquelles différentes variantes de calcul et différents thésaurus sont évalués. Nous en donnons ici une description reprenant la terminologie introduite pour les mesures et fonctions de sélection. Les résultats des évaluations sont détaillés dans le chapitre afférent (chapitre 4).

Cependant, la comparaison directe des entropies de différentes coupes n'a guère de sens,[26] l'entropie maximale possible étant dépendante de la taille de la coupe sur laquelle elle est mesurée.[27] Pour rendre comparable deux coupes, on considérera dès lors le rapport, pour une coupe donnée, entre l'entropie (observée) et l'entropie maximale possible pour cette coupe, ce qui conduit naturellement à la notion de redondance (au sens de la théorie de l'information – Shannon [1948]). Nous proposons alors d'utiliser comme jeu d'indexation une coupe dont la *redondance* est minimale.

L'entropie $H(X)$, qui mesure l'incertitude d'une variable aléatoire discrète X, admettant $\mathcal{X}$ comme alphabet, est donnée par :

$$H(X) \triangleq -\sum_{x\in\mathcal{X}} \underbrace{P(X=x)\cdot\log\big(P(X=x)\big)}_{h(x)}, \tag{3.11}$$

avec $h(x)=0$ lorsque $P(X=x)=0$.[28]

La *redondance d'information* $R(X)$ est la proportion entre l'entropie $H(X)$ et sa valeur *maximale possible*, $\log(|\mathcal{X}|)$ (laquelle est atteinte lorsque les événements sont équiprobables) :

$$R(X) \triangleq 1 - \frac{H(X)}{\log(|\mathcal{X}|)}. \tag{3.12}$$

On mesure la redondance d'un modèle de coupe probabilisé $M=(\Gamma,\theta)$ par :

$$\begin{aligned} R(M) &= 1 - \frac{H(M)}{\log(|M|)} \text{ , avec} \\ H(M) &= -\sum_{\theta_i\in\theta} \theta_i\cdot\log(\theta_i) \end{aligned} \tag{3.13}$$

Minimiser la redondance revient donc à maximiser le ratio entre l'entropie observée et l'entropie maximale possible, compte tenu du nombre d'éléments de la coupe. En reprenant la terminologie de la table 3.1 [page 31], on obtient le modèle suivant :

$$\begin{aligned} \mu(s) &= -\theta_s\cdot\log(\theta_s) \\ \lambda(\Gamma) &= H(\Gamma) = \sum_{s\in\Gamma}\mu(s) \\ \Lambda(\Gamma) &= 1 - R(\Gamma) = \frac{\lambda(\Gamma)}{\log(|\Gamma|)} \\ F'(\Gamma) &= \delta\left(\Gamma, \operatorname*{argmax}_{\Gamma_i\in\Upsilon}(\Lambda(\Gamma_i))\right). \end{aligned} \tag{3.14}$$

[26] Et ne permettrait probablement que d'isoler une coupe sur les feuilles.

[27] C'est là l'une des propriété de l'entropie : $0 \le H(X) \le \log(|\mathcal{X}|)$, avec $|\mathcal{X}|$ le nombre de valeurs pour X.

[28] Justifié par le prolongement par continuité en 0 : $\lim_{x\to 0^+} x\cdot\log(x) = 0$.

Exemple :

Pour illuster le fonctionnement de ce critère, examinons son application à la situation prenant en défaut le critère *MDL*, et synthétisée par la figure 3.8 [page 37].

On calculera la redondance en estimant θ uniquement en fonction du poids du sommet considéré ; le thésaurus étant un arbre, cette estimation ne dépend pas du reste de la coupe :[29]

$$\widehat{\theta_s} = P_{\mathsf{tf}}(s|\Gamma) = \frac{\mathsf{tf}(s)}{\sum_{s_i \in \Gamma} \mathsf{tf}(s_i)}$$

le « poids » $\mathsf{tf}(s)$ d'un sommet s correspond à la somme des fréquences d'occurrence de ses feuilles :

$$\mathsf{tf}(s) \triangleq \sum_{m \in (s^{\Downarrow} \cap \mathcal{M})} \mathsf{tf}(m),$$

où $\mathcal{M}$ représente le vocabulaire de la collection de documents.

La table suivante regroupe les valeurs nécessaires au calcul de la redondance des différentes coupes.

s	$\mathsf{tf}(s)$	$\widehat{\theta_s}$	$\mu(s)$ [30]
Entity	18	1.0000	0.0
Animal	11	0.6111	0.43419
Artifact	7	0.3889	0.52989
Bird	7	0.3889	0.52989
Insect	4	0.2222	0.48221
Vehicle	3	0.1667	0.43083
Airplane	4	0.2222	0.48221
swallow	1	0.0556	0.23166
crow	1	0.0556	0.23166
eagle	2	0.1111	0.35221
bird	3	0.1667	0.43083
bug	0	0.0	0.0
bee	3	0.1667	0.43083
insect	1	0.0556	0.23166
car	1	0.0556	0.23166
bike	2	0.1111	0.35221
jet	1	0.0556	0.23166
helicopter	2	0.1111	0.35221
airplane	1	0.0556	0.23166

En appliquant l'algorithme de recherche par proximité proposé ci-après (3 [page 43]), on trouve comme coupe optimale Γ = [Bird, Insect, Artifact] (l'optimum local correspondant ici à l'optimum global).

Ci-après, le détail du calcul du critère pour différentes coupes (parmi les 26 possibles) :

[29] Consulter 3.5.4 [page 44] pour plus de détails concernant l'estimation des probabilités des sommets des coupes probabilisées.

[30] En utilisant le logarithme en base 2.

Γ	$\lambda(\Gamma)$	$\Lambda(\Gamma)$	$R(\Gamma)$
[Bird, Insect, Artifact]	1.5420	0.9729	0.0271
[Animal, Artifact]	0.9641	0.9641	0.0359
[Bird, Insect, Vehicle, Airplane]	1.9251	0.9626	0.0374
[Animal, Vehicle, Airplane]	1.3472	0.8500	0.1500
[Entity]	0.0	0.0	1.0000

Remarque :

Le critère de *redondance minimale* ne permet pas, à lui seul, d'identifier une coupe optimale *unique*, mais un ensemble de coupes optimales. Il faut donc ajouter des conditions supplémentaires pour n'exhiber qu'une seule solution (lorsque on le souhaite) : choix au hasard, en fonction de la taille de la coupe, de la profondeur moyenne, etc. Étant attendu que les termes d'indexation ne doivent pas être trop généraux, nous proposons d'utiliser le conditionnement implicite fourni par l'algorithme de recherche approchée 3 [page ci-contre], en cherchant **la coupe optimale locale au voisinage des feuilles**, c'est-à-dire utiliser l'algorithme en débutant la recherche avec la coupe $\Gamma_i = \mathcal{M}$ (*i.e.* uniquement constituée des « mots »).

3.5.3.2 Algorithme

Le but étant de chercher l'optimum d'un critère non séparable ou sur une topologie de réseau ne se prêtant pas à une évaluation par séparation (cas traité par l'algorithme 2 [page 39]), nous proposons de recourir à un algorithme approché de *recherche par voisinage*. Son principe est relativement simple : l'algorithme part d'une solution réalisable et tente de l'améliorer par itérations successives. Il est dès lors évident que la solution exhibée n'est optimale que localement, par rapport au voisinage de la solution de départ ; il existe cependant diverses techniques permettant d'améliorer la recherche d'un optimum global, résumées dans les remarques ci-après.

L'algorithme 3 [page ci-contre] consiste à itérativement modifier la coupe initiale, en remplaçant à chaque itération *un* sommet s_i par l'un de ses prédécesseurs ($s_j \in s_i{}^{\uparrow}$) ou par l'ensemble de ses successeurs ($s_i{}^{\downarrow}$). Les sommets de la coupe sont examinés tour à tour, le remplaçement effectué étant celui ayant un impact maximal sur la mesure Λ ;[31] la recherche se termine lorsque aucun des remplaçements possibles ne permet d'améliorer la mesure Λ.

En raison de la topologie du graphe, ce remplacement peut impliquer d'autres sommets pour que la cohérence de la coupe soit maintenue ; lors du remplaçement par les $s_i{}^{\downarrow}$ sommets successeurs, il faut exclure ceux éventuellement dominés par d'autres sommets de la coupe – c'est-à-dire faire passer la nouvelle coupe par $s_i{}^{\downarrow} \setminus (\Gamma \setminus s_i{}^{\Downarrow})$, comme indiqué sur la figure 3.9(b) [page suivante]. De manière similaire lors du remplacement par un sommet prédécesseur $s_j \in s_i{}^{\uparrow}$, il faudra également retirer de la coupe les éventuels autres sommets dominés par s_j, soit $s_j{}^{\Downarrow}$ (figure 3.9(a) [page ci-contre]).

Remarques :

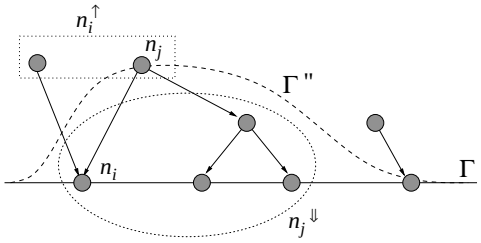
☛ Un avantage appréciable de cet algorithme est qu'il peut non seulement être interrompu à n'importe quelle itération (puisqu'à tout moment il travaille sur une coupe bien formée), mais également être poursuivi après interruption (la seule information spécifique à mémoriser étant la coupe trouvée au moment de l'arrêt).

[31] Pour obtenir une convergence plus rapide, il est possible d'opter pour une solution plus radicale consistant, lors de l'examen des sommets de la coupe, à effectuer immédiatement les remplacements positifs pour le critère ; le prix à payer étant de ne plus nécessairement converger vers l'un des optimums local au voisinage initialement spécifié.

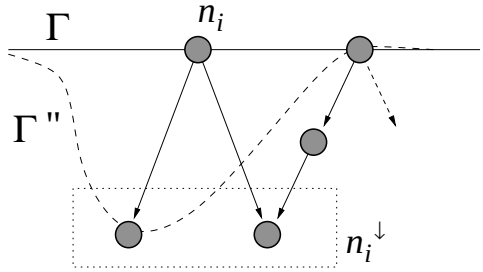
Algorithme 3 Vopt : recherche par voisinage d'un maximum local de Λ.

Nécessite : une coupe initiale Γ_0.
Fourni : une coupe Γ *minimisant* λ sur le voisinage de Γ_0.

$\Gamma \leftarrow \Gamma_0$ # *meilleure coupe rencontrée*
répéter
  $\Gamma' \leftarrow \varnothing$ # *un candidat.*
  $\widehat{\Gamma'} \leftarrow \varnothing$ # *le meilleur des candidats.*
  continue $\leftarrow$ *faux* # *drapeau de contrôle de la recherche.*
  pour tout $s_i \in \Gamma$ **faire**
    # *Évaluation des successeurs :*
    $\Gamma' \leftarrow (\Gamma \setminus \{s_i\}) \cup \left(s_i{}^{\downarrow} \setminus (\Gamma \setminus \{s_i\})^{\Downarrow}\right)$
    $\widehat{\Gamma'} \leftarrow \operatorname{argmax}\left(\Lambda(\Gamma'), \Lambda\left(\widehat{\Gamma'}\right)\right)$
    # *Évaluation de chaque prédécesseur :*
    pour tout $s_j \in s_i{}^{\uparrow}$ **faire**
      $\Gamma' \leftarrow (\Gamma \cup \{s_j\}) \setminus s_j{}^{\Downarrow}$
      $\widehat{\Gamma'} \leftarrow \operatorname{argmax}\left(\Lambda(\Gamma'), \Lambda\left(\widehat{\Gamma'}\right)\right)$
    fin pour
  fin pour
  si $\Lambda(\Gamma) < \Lambda\left(\widehat{\Gamma'}\right)$ **alors**
    $\Gamma \leftarrow \widehat{\Gamma'}$ # *mémoriser la meilleure coupe.*
    continue $\leftarrow$ *vrai* # *poursuivre la recherche.*
    # *chien de garde ou arrêt sur timer éventuels*
  fin si
jusqu'à ce que continue $=$ *vrai*
retourne Γ



(A) Remplaçement par un prédecesseur : tous les sommets dominés doivent être retirés.



(B) Remplaçement par les successeurs : uniquement inclure ceux non dominés par ailleurs.

Figure 3.9*:* Remplaçement d'un sommet d'une coupe en recherche par voisinage.

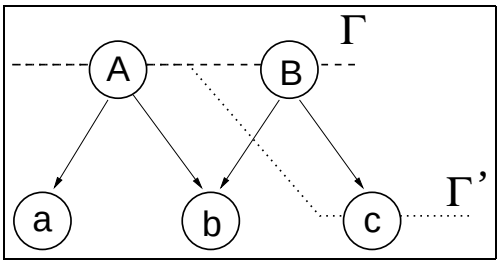


Figure 3.10*:* Polysémie dans un thésaurus (DAG).

- Plusieurs pistes peuvent être envisagées pour améliorer la globalité de l'optimum trouvé par l'algorithme. Traditionnellement, on améliore en probabilité ce type d'algorithmes itératifs en réalisant plusieurs recherches, à partir de solutions initiales différentes (dont une partie au moins devrait, en toute généralité, être tirée au hasard). Dans le cas présent, une solution consisterait à démarrer la recherche en utilisant la coupe issue d'une première exécution de l'algorithme de recherche par programmation dynamique (algorithme 2 [page 39]).[32] Notons encore l'existence de variantes plus élaborées, permettant d'éviter à l'algorithme de stopper sur le premier optimum local rencontré, telles que la *recherche Tabou* [Glover *et al.*, 1993, Kernighan et Lin, 1970] ou le *recuit simulé* [Kirkpatrick *et al.*, 1983].
- Selon les cas, il devrait être possible d'assimiler l'optimisation de critères non séparables à la problématique relativement bien connue du *ratio minimum*, et utiliser les techniques d'optimisation afférentes (cf. Gondran et Minoux [1995], annexe 5) :
  soit un ensemble X et deux fonctions réelles $f(x)$ et $g(x)$ définie sur X, avec $g(x) > 0, \forall x \in X$, le problème du ratio minimum se pose comme la recherche de $x \in X$ minimisant $f(x)/g(x)$. C'est en particulier le cas de la redondance minimale :

$$\min_{x \in X} \left(\frac{f(x)}{g(x)} \right) = \min_{\Gamma \in \Upsilon} \left(\frac{\log(|\Gamma|) - H(\Gamma_i)}{\log(|\Gamma|)} \right)$$

  Nous n'avons pas approfondi cette analogie, mais il y a peut-être là de quoi améliorer les algorithmes implémentant des critères de cette sorte.

3.5.4 Différentes estimations de θ

L'estimation des mesures de probabilités associées aux sommets d'une coupe peut elle aussi être envisagée de différentes manières. Avant d'examiner les choix de mesures les plus évidents, commençons par étudier les fondements possibles pour de telles mesures.

3.5.4.1 Poids des occurrences

On admet qu'il y a *occurrence* du sommet s lorsqu'il y a occurrence (notée o) de l'un des $m_i \in s^{\Downarrow}$ mots dominés par s. Le « poids » de l'occurrence d'un sommet n'est cependant pas nécessairement le même que celui de son mot associé (o), selon que l'on prend ou non en compte l'éventuelle polysémie présente dans le thésaurus.

[32] Une modification de cet algorithme est cependant requise pour gérer le cas des mesures nulles ou non définies sur un sommet unique.

Exemple :

La figure 3.10 illustre le cas d'une relation « polysémique ». Le « mot » b est partagé entre les concepts A et B (on peut également retrouver ce cas de figure entre concepts uniquement). Dès lors, comment une occurrence de ce mot doit-elle être comptabilisée au niveau de A et B ? Doit-il y avoir une différence selon que le jeu d'indexation considéré est Γ ou Γ' ?

Avec D l'ensemble des documents à indexer, et O l'ensemble des occurrences d'un terme dans les documents à indexer, on définira de la manière suivante le « poids » d'une occurrence :

Définition : poids $\mathcal{W}(s)$ d'un sommet s

Le **poids** (total) d'un sommet s, noté $\mathcal{W}(s)$, est donné par la somme des **contributions** $w(s,o)$ **de chaque occurrence** o de s, et correspond à la fréquence du terme, $\mathsf{tf}(s)$, « pondérée » par le partage de sens induit par l'éventuelle polysémie :[33]

$$\mathcal{W}(s) = \sum_{d \in D} \sum_{o \in d} w(s,o) = \sum_{o \in O} w(s,o)$$

En présence d'un thésaurus non polysémique (*i.e.* possédant une structure d'arbre), le poids d'un sommet est simplement donné par la somme des occurrences des m_i feuilles qu'il domine :

$$\mathcal{W}(s) = \sum_{m \in (s^{\Downarrow} \cap \mathcal{M})} \mathsf{tf}(m) \triangleq \mathsf{tf}(s)$$

Remarque :

L'hypothèse de non-polysémie est rarement vérifiée en pratique : même si le thésaurus possède effectivement une structure d'arbre, il demeure généralement des ambiguïtés lors de la mise en correspondance entre une forme de surface (dans les données) et une feuille du thésaurus[34] dues à la segmentation, aux variations flexionnelles, ou même à la casse.[35]

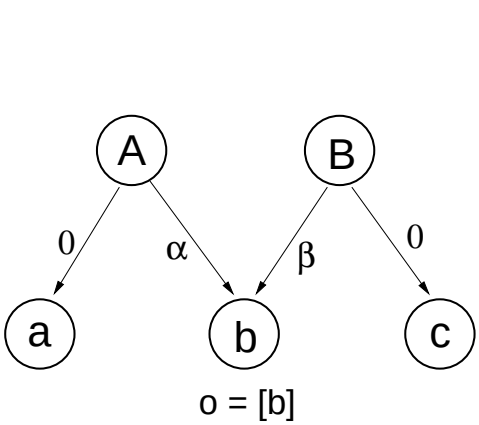
Prise en compte de la polysémie

De manière générale, on considérera qu'à chaque relation $r_{\mathsf{i,j}}$ entre les sommets i et j est associé un coefficient $v_{ij}(o) \in [0,1]$ indiquant, pour une occurrence o donnée, la *vraisemblance* de la relation $r_{\mathsf{i,j}}$, comme illustré par la figure 3.11(a) [page suivante]. Lorsque survient une occurrence $o = \mathsf{b}$, la vraisemblance des relations $r_{\mathsf{b,A}}$, respectivement $r_{\mathsf{b,B}}$, est donnée par $v_{r(\mathsf{b,A})}(o) = \alpha$, respectivement $v_{r(\mathsf{b,B})}(o) = \beta$. Cela revient en fait à considérer la situation de la figure 3.11(b) [page suivante], où l'occurrence $o = \mathsf{b}$ est décomposée en occurrences de mots artificiels $\mathsf{b'}$ et $\mathsf{b''}$ (non polysémique), ces occurrences étant respectivement pondérées par les coefficients de vraisemblance α et β.

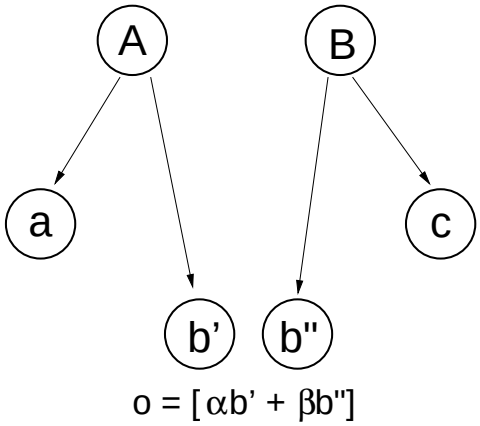
[33] Cette pondération est intégrée dans les contributions w des occurrences.

[34] Dans ce cas, on considère le thésaurus étendu d'un niveau supplémentaire sur ses feuilles, lequel invalide la structure en arbre.

[35] On constate par exemple qu'avec l'ontologie *EDR*, les formes « dog » et « Dog » ne sont pas équivalentes, la première concernant l'animal tandis que la seconde faisant référence à la constellation !



(A) Vraisemblance des relations pour une ocurrence.



(B) Arbre équivalent.

Figure 3.11*:* «Suppression» (artifcielle) de la polysémie (transformation d'un DAG en arbre).

Définition : contribution $w(s,o)$ d'une occurrence o d'un mot m

La **contribution** $w(s,o)$ d'une occurrence o d'un mot (sommet feuille) $m \in \mathcal{M}$ au poids d'un sommet s est donnée par la somme, sur tous les chemins entre m et s (où $\mathcal{C}$ est cet ensemble de chemins), de la pondération du poids propre de l'occurrence, $w(o)$, par le produit des vraisemblances des relations rencontrées :

$$w(s,o) = w(o) \cdot \sum_{c \in \mathcal{C}} \prod_{r \in c} v(r)$$

Le poids de l'occurrence étant, en toute généralité, conditionné par la probabilité du document (d) dans lequel elle survient, on a donc de plus :

$$w(o) = 1 \cdot p(d)$$

(Avec une collection de documents équiprobables, on admettra que $w(o) = 1$, ce qui augmente uniformément le poids des sommets d'un facteur $|D|$.)

Deux cas de figures sont à envisager :

1. Le poids $w(o)$ de l'occurrence $o = \mathsf{b}$ est *réparti* entre A et B ; *i.e.* :

   $$\alpha + \beta = 1$$

   (les coefficients de vraisemblance v sont en fait des probabilités).
   Sans autres mesures compensatoires, les occurrences d'un sommet polysémique non complètement couvert par la coupe auront un poids moindre que les autres occurrences (cas de la coupe Γ' de la figure 3.10 [page 44]).

2. Le poids $w(o)$ de l'occurrence $o = \mathsf{b}$ est *reporté* totalement ou partiellement sur A et B, *i.e.* :

   $$\alpha + \beta > 1$$

   Sans autres mesures compensatoires, les occurrences d'un sommet polysémique totalement couvert par un sommet unique (*i.e.* dominant la réunion des chemins) auront un poids supérieur aux autres occurrences.

Remarques :

- ☛ La situation d'une polysémie d'un mot vers deux concepts décrite ici est bien entendu généralisable au cas de plus de deux concepts, et aux « polysémies » entre concepts eux-mêmes.
- ☛ La valuation des vraisemblances des relations peut s'obtenir de différentes manières :
  1. Dans le cas idéal, on bénéficie d'un module de *désambiguisation sémantique* (WSD contextuelle) ; quelque soit son mode de fonctionnement (exhibation d'une relation, simple ordonnancement ou affectation de probabilité) il sera possible d'en tirer une estimation des vraisemblances v.
  2. En l'absence de désambigüisation sémantique en contexte, le palliatif usuel (pour autant que l'information soit disponible) est d'utiliser une « probabilité absolue » de ces relations (estimée sur de très grands corpus, reflètant l'usage courant dans la langue). Par exemple, *WordNet* permet de classer les associations (graphie, `synsets`) des plus fréquentes (usuelles) aux plus rares, et *EDR* donne les fréquences des associations (graphie, `concept`) pour un très grand corpus. Cependant, cette solution ne permet en général que de valuer les relations (mots-concepts), c'est-à-dire les relations entre les feuilles et le reste du thésaurus ; pour les autres, il faudra se rabattre sur l'option présentée ci-après.
  3. Finalement, en l'absence de toute information utilisable, la seule option raisonnable est de choisir une vraisemblance identique pour chacune des relations :
     - on peut d'une part considérer que les relations sont *exclusives* entre elles ; la vraisemblance est alors divisé par le nombre de relations à considérer pour un sommet donné (dans le cas de la figure 3.11 [page précédente], on aurait $\alpha = \beta = 1/2$) ;
     - d'autre part, on peut considérer que les relations sont *concomitantes* et toujours réalisées ;[36] la vraisemblance est alors 1, « l'excès de poids » (au total) dû à la polysémie étant compensé lors de l'estimation de θ par le dénominateur (cf. équation 3.4 [page 35]), comme cela est montré dans la section suivante.

3.5.4.2 Estimations de θ

Si l'on souhaite équilibrer les probabilités d'apparition des termes de la coupe dans les données indexées, on prendra le « poids » comme mesure élémentaire : $\mu' = \mathcal{W}$ ($\approx$ *term frequency*)

$$\theta_s \widehat{=} P_{\mathsf{tf}}(s|\Gamma) = \frac{\mathcal{W}(s)}{\sum_{s_i \in \Gamma} \mathcal{W}(s_i)} \tag{3.15}$$

Dans le cas d'un thésaurus non polysémique, et avec une collection de documents équiprobables, cette probablilité devient simplement :

$$P_{\mathsf{tf}}(s|\Gamma) = \frac{\mathsf{tf}(s)}{|O|}$$

avec O l'ensemble des occurrences à indexer, et $\mathsf{tf}(x)$ la *fréquence* de x (*i.e.* nombre d'occurrences).

D'autres distributions de probabilité peuvent naturellement être utilisées pour $\widehat{\theta}$. On pourra par exemple préférer pondérer la « fréquence » d'un terme par le coefficient de fréquence inverse en

[36] Comme par exemple avec la proposition « Qui veut aller loin ménage sa monture », où « monture » désigne à la fois l'animal *et* le moyen de transport.

document (cf. 2.4.2 [page 13]) ; dans ce cas, on prendra comme mesure élémentaire $\mu' = \mathcal{W} \cdot \mathsf{idf}$, soit :

$$\theta_s \widehat{=} P_{\mathsf{tfidf}}(s|\Gamma) = \frac{\mathcal{W}(s) \cdot \log\left(\frac{|D|}{\mathsf{df}(s)}\right)}{\sum_{s_i \in \Gamma}\left(\mathcal{W}(s_i) \cdot \log\left(\frac{|D|}{\mathsf{df}(s_i)}\right)\right)} \tag{3.16}$$

avec $\mathsf{df}(x)$ la *fréquence en document* de x.

Dans le cas d'un thésaurus non polysémique, et avec une collection de documents équiprobables, cette probablilité devient simplement :

$$P_{\mathsf{tfidf}}(s|\Gamma) = \frac{\mathsf{tf}(s)}{|O|} \cdot \log\left(\frac{|D|}{\mathsf{df}(s)}\right)$$

3.6 Documents supplémentaires

La problématique des « documents supplémentaires » (en réalité, le problème se pose avec les termes « supplémentaires », c'est-à-dire nouveaux) ne peut être laissée de côté. En effet, excepté dans le cadre d'une évaluation sur la base d'un corpus de référence, les requêtes soumises à un système de recherche documentaire ne sont jamais connues *a priori*, et constituent donc toujours des « documents supplémentaires » (au regard de la sélection du jeu d'index).

En introduisant une dépendance entre données à indexer et manière dont est réalisée l'indexation, on rend délicate la prise en compte ultérieure de documents *supplémentaires* : que faire en effet de l'information nouvelle qu'ils apportent, tant sur le plan terminologique (nouveaux mots) que sur celui de la distribution des termes dans la collection ?

Remarquons que ce problème ne concerne pas uniquement l'aspect de mise en œuvre pratique d'une méthode d'indexation (par exemple l'indexation d'une base globalement fixe, à laquelle s'ajoutent et se retirent ponctuellement des documents) : dans le cas d'une recherche documentaire classique, les requêtes sont à considérer comme autant de documents supplémentaires.

Dans le cadre de LSI par exemple (cf. 2.4.1.2 [page 13]), les « individus » supplémentaires sont généralement *projetés* dans l'espace factoriel réduit, sans apporter de contribution à la constitution de cet espace (les documents supplémentaires sont simplement indexés, mais l'information qu'ils portent sur le vocabulaire n'est pas utilisée pour influencer le jeu d'indexation).

En indexation sémantique, outre la solution coûteuse d'un nouveau calcul de coupe (impliquant une réindexation), on peut également choisir de ne pas utiliser l'information apportée, mais uniquement d'indexer au mieux ces nouveaux documents. Dans ce cas, différentes attitudes peuvent être adoptées selon la couverture par la ressource sémantique de ces documents. La figure 3.12 [page suivante] en donne une vision schématique ; la partie gauche de la figure représente le réseau sémantique lors de l'indexation de la collection initiale, et la partie droite l'extension de ce réseau due à un document supplémentaire.

Le traitement de ces termes additionnels peut se faire (en particulier) selon les approches suivantes :

1. Les mots absents du thésaurus (par exemple le mot a de la figure) seront indexés par eux-mêmes (comme pour les autres documents), ou ignorés si le document supplémentaire ne doit pas être intégré à la base (cas d'une requète).
2. Les mots (nouveau) mais déjà couverts par l'actuelle coupe Γ sont indexés usuellement, par les sommets de la coupe qui les couvrent (situation non représentée sur la figure)

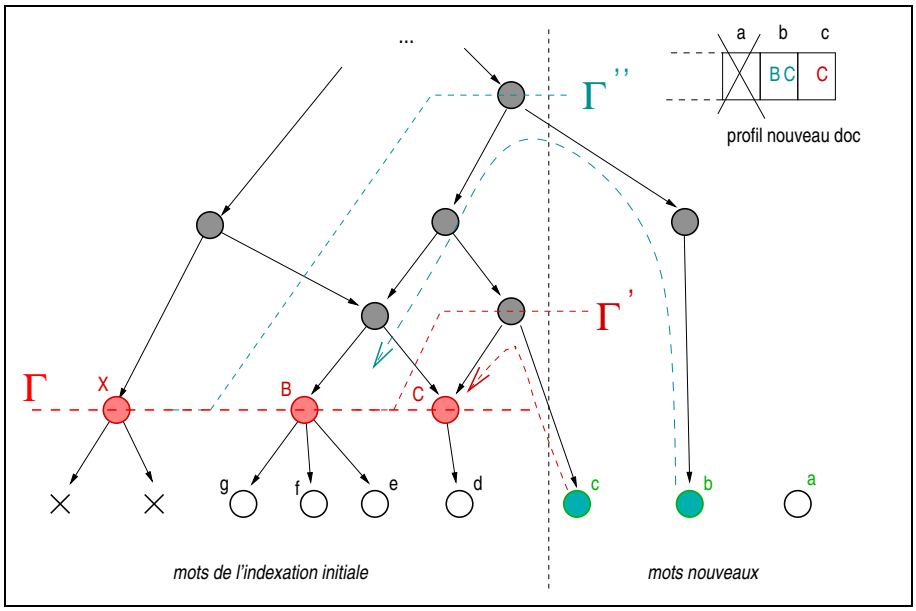


Figure 3.12*:* Stratégies de prise en compte de nouveaux documents en indexation sémantique

3. Finalement, pour les mots non couverts par l'actuelle coupe Γ, mais dont les prédecesseurs (racine exclue) dominent une partie de celle-ci (mots b et c), on peut choisir de les indexer par eux-mêmes (équivalent au cas de mots absents du thésaurus) ou par la partie du jeu d'indexation dominée par leurs prédécesseurs. Bien entendu, plus il faut remonter « haut » dans la hiérarchie des prédecesseurs pour en trouver un qui domine un segment de coupe, et plus la pertinence de l'indexation du mot par ce segment dominé sera faible : le mot c de la figure pourra ainsi être indexé par le concept [C], alors que le mot b devra lui être indexé par les concepts [B, C].

Remarques :

- Remarquons qu'une réindexation complète peut cependant se faire à moindre frais (surtout si l'on compare à LSI) au moyen de l'algorithme approché de recherche par voisinage (3.5.3.2 [page 42]), proposé pour l'implémentation des critères non-séparable : après expansion du réseau (pour la prise en compte des éventuels mots nouveaux) et mise à jour des fréquences d'occurrences, probabilités, etc. (linéaire en le nombre de sommets), il « suffit » de poursuivre la recherche d'un optimum local en partant de la coupe précedemment trouvée. Bien entendu, le résultat de cette recherche peut ne pas correspondre à l'optimum local au voisinage initial (*i.e.* défini lors de la première recherche).
- L'indexation d'un mot nouveau par la partie de la coupe dominée par l'un de ses prédécesseurs peut être vue comme une modification dynamique (uniquement pour le mot considéré) de la coupe : dans ce cas, on obtient le même résultat que si la coupe était passée par le prédécesseur en question ; ceci est illustré sur la figure 3.12 par les coupes Γ′ et Γ”. Cette façon de procéder est donc cohérente avec le concept de coupe utilisé pour représenter un jeu d'indexation, et s'inscrit parfaitement dans le cadre général introduit pour l'indexation sémantique.

3.7 Utilisation de thésaurus indépendamment des données

L'information contenue dans un réseau sémantique peut également être utilisée indépendamment de l'ensemble des données à traiter ; on peut par exemple vouloir adapter statiquement

le thésaurus (selon sa profondeur et son degré de branchage), de manière à pouvoir effectuer une indexation sémantique dans un autre cadre que celui de la recherche documentaire traditionnelle, typiquement lorsque les données ne sont pas connues à l'avance (tâche de filtrage ou de classification), mais également si l'on souhaite pouvoir contrôler le degré de généralité des termes d'indexation.

Ce cas de figure est succinctement abordé dans la section suivante. Remarquons que la famille de critères donnée ci-après n'a pas été évaluée dans le cadre d'une recherche documentaire ; ne constituant pas un élément majeur de notre travail, nous n'avons pas cherché à confronter cette famille de critères à ceux obtenus en prenant en compte les données à indexer. Le lecteur intéressé pourra néanmoins se rapporter à Rajman *et al.* [2005], dans lequel sont donnés les résultats d'une évaluation des performances de cette technique dans le cadre d'une tâche d'identification/extraction de thèmes.

3.7.1 Critère d'élagage du thésaurus selon la profondeur et le branchage

Bien que nous ayons argué qu'utiliser en indexation une ressource sémantique en pré-déterminant le niveau de généralité n'est pas une bonne idée, cela peut cependant être nécessaire dans le cas de certaines applications spécifiques. En particulier, on peut désirer avoir la possibilité de disposer d'un contrôle plus précis du degré de généralité, notamment lors de processus de classification (*clusterisation*) ou de processus à la frontière entre résumé automatique et extraction de *topics*.

Pour cela, nous présentons ici une méthode visant à atténuer les différences de finesse de description dans un thésaurus, indépendamment des données à indexer, mais surtout en donnant à l'expérimenteur la possibilité d'ajuster la précision de la représentation (consulter également Rajman *et al.* [2005]).

Avec la plupart des réseaux sémantiques, les « concepts » définis au niveau du réseau possèdent également une ou plusieurs formes littérales en language naturel, explicitant le concept (il peut tout à la fois s'agir d'un ou plusieurs mots le définissant, ou de propositions en dressant les contours)[37]. Le but est alors de choisir un jeu d'indexation permettant, au travers des définitions des termes d'un document, de construire une représentation contrôlée du document.

Tout comme en indexation guidée par les données, on choisi ici de contraindre l'ensemble des termes d'indexation à des structures de coupes (cf. section 3.5.1 [page 33]) mais dont les éléments ne peuvent *qu'être des concepts*.

Le critère de sélection que l'on utilise est un critère local, mais obtenu à l'aide d'un algorithme de recherche globale (algorithme 4 [page 52]), et pas d'un algorithme de type Tabou. Remarquons également que les mesures utilisées ici ne s'appliquent qu'à des concepts (puisque les coupes ne portent que sur ces derniers).

La mesure locale à chaque concept représente la composition de deux autres mesures (locales), G– mesurant le « degré de généralité » du concept – et I– mesurant le « caractère informatif »[38] du concept – il n'y a pas de mesure globale à proprement parler, la décision entre conserver un concept ou ses sucesseurs étant prise en comparant la moyenne arithmétique des mesures locales.

[37] Remarquons cependant qu'avec certaines ressources, il arrive que certains de ces concepts soient totalement « anonymes » ; c'est en particulier le cas avec la partie anglaise de *EDR*– certains concepts de cette ressource semblant en effet n'avoir de signification qu'en japonnais. On considérera ici que les concepts « anonymes » sont retirés de la hiérarchie.

[38] Mais pas au sens de la théorie de l'information.

3.7.1.1 Degré de généralité G

Il paraît relativement intuitif que dans un thésaurus (*i.e.* un réseau sémantique structuré par une relation inclusive, cf. 3.2 [page 24]), un concept soit plus « général » que ses sous-concepts ; par exemple, le concept ANIMAL est moins spécifique que le concept CHIEN. Dans un même temps, plus un concept est général (donc « haut » dans le thésaurus), et plus le nombre de feuilles qu'il domine est grand. Par exemple, si chat, chien et serpent sont des feuilles d'un thésaurus, le concept VERTEBRÉ ne couvrira que les deux feuilles chat et chien, tandis que son super-concept ANIMAL couvrira (au moins) les trois.

Au regard de cette observation, on en déduit que la mesure de la « généralité » $G(c)$ d'un concept c doit être en relation avec la proportion de feuilles du thésaurus qu'il domine. Cette mesure devrait donc satisfaire les contraintes suivantes :

- la « généralité » d'un concept qui ne couvre qu'une feuille est nulle :

$$\text{si } |c^{\Downarrow} \cap \mathcal{M}| = 1, \text{alors } G(c) \triangleq 0;$$

- la « généralité » d'un concept qui couvre toutes les feuilles est maximale :

$$\text{si } |c^{\Downarrow} \cap \mathcal{M}| = |\mathcal{M}|, \text{alors } G(c) \triangleq 1;$$

En faisant l'hypothèse d'une linéarité de cette notion, on définira alors cette mesure comme suit :

$$G(c) \triangleq \frac{|c^{\Downarrow} \cap \mathcal{M}| - 1}{|\mathcal{M}| - 1} \tag{3.17}$$

3.7.1.2 Caractère informatif I

Si l'on ne prenait en compte que la généralité, les coupes sélectionnées se limiteraient aux concepts directement reliés aux feuilles ou à la racine. Il est donc important de considérer également la quantité d'information relative aux feuilles préservée par les concepts d'une coupe. Pour quantifier cette préservation d'information, on définit une seconde mesure, I, pour laquelle on décide ici que la valeur de cette mesure doit, pour un concept c donné, être linéairement dépendante de la longueur moyenne normalisée de tous les chemins entre le concept et les feuilles qu'il domine. On ajoute de plus à cette mesure les contraintes suivantes :

- le « caractère informatif » de la racine est nul :

$$\text{si } c^{\Downarrow} = (\mathcal{S} \setminus c), \text{alors } I(c) \triangleq 0;$$

- le « caractère informatif » d'un concept ne dominant que des feuilles est maximal :

$$\text{si } c^{\Downarrow} \subset \mathcal{M}, \text{alors } I(c) \triangleq 1;$$

On définira alors cette mesure comme suit :

$$I(c) \triangleq 1 - \frac{\frac{1}{|c^{\Downarrow} \cap \mathcal{M}|} \cdot \sum_{m_i \in (c^{\Downarrow} \cap \mathcal{M})} \bar{d}(m_i, c)}{\underbrace{\frac{1}{|\mathcal{M}|} \cdot \sum_{m_j \in \mathcal{M}} \bar{d}(m_j, root)}_{\text{constante}}} \tag{3.18}$$

où $root$ est le sommet racine, et $\bar{d}(m, c)$ est la moyenne des longueurs de tous les chemins entre $m \in \mathcal{M}$ et $c \in \mathcal{C}$.

3.7.1.3 Mesure locale d'un concept

La sélection des termes d'indexation est basée sur une mesure μ, locale à chaque concept, obtenue par combinaison de G et I au sein d'une moyenne géométrique pondérée :

$$\mu(c) = I(c)^{\alpha} \cdot G(c)^{1-\alpha} \tag{3.19}$$

où $\alpha \in [0,1]$ est un paramètre permettant de contrôler la « précision » du jeu d'indexation (de même que le nombre de termes) : une valeur élevée favorisera le caractère informatif des concepts, et le jeu d'indexation sera proche des feuilles ; une valeur faible favorisera la généralité des concepts, et le jeu d'indexation sera proche de la racine.

3.7.1.4 Algorithme

L'algorithme de sélection d'une coupe donné ici est une adaptation de l'algorithme de sélection des critères séparables (section 2 [page 39]), fonctionnant sans mesure globale, et maximisant *en moyenne* la mesure μ. Il doit également être lancé depuis la racine, pour obtenir une coupe couvrant la totalité du thésaurus.

Algorithme 4 FindCut recherche, dans un arbre, d'une coupe optmisant λ.

Nécessite : un concept racine c.
Fourni : un ensemble de concepts (jeu d'indexation).

$\Gamma' \leftarrow \varnothing$ *# combinaisons des sous-coupes*
$F \leftarrow c^{\downarrow} \cap \mathcal{C}$ *# l'ensemble des sous-concepts*
pour $c_i \in F$ **faire**
  $\Gamma' \leftarrow \Gamma' \cup \mathsf{FindCut}(c_i)$
fin pour
$\widehat{\mu'(F)} \leftarrow {}^{1}/_{|F|} \cdot \sum_{c_i \in F} \mu'(c_i)$
si $\mu(c) > \widehat{\mu'(F)}$ **alors**
  $\mu' \leftarrow \mu(c)$ *# le concept reste en lice*
  retourne $\{c\}$
sinon
  $\mu' \leftarrow \widehat{\mu'(F)}$ *# le concept n'est pas conservé*
  retourne Γ'
fin si

Remarque :

Cet algorithme souffre des mêmes limitations que celui dont il est issu : dans le cas d'un thésaurus avec une structure de DAG, le jeu d'indexation produit ne correspondra pas forcément à une coupe, et nécessitera éventuellement[39] une rectification ultérieure.

[39] En effet, selon l'application, le fait de ne pas avoir de coupe « bien formée » peut être acceptable.

3.8 Conclusion

Dans ce chapitre, nous avons étudié en détail le principe de l'*indexation sémantique*, permettant d'intégrer des informations sémantiques explicites issues de ressources externes dans l'espace de représentation des documents (les profils d'indexation).

Nous avons notamment présenté plusieurs manières permettant de mettre à profit la connaissance des données à indexer dans le but d'extraire des ressources externes les informations les plus pertinentes, en sélectionnant le degré de précision de description de ces informations le plus en adéquation avec les données devant être traitées.

Nous avons structuré notre présentation en proposant un cadre général permettant de décrire différentes techniques envisageables pour réaliser des indexations sémantique. Nous avons utilisé ce cadre pour détailler trois familles de critères (quatre, si l'on considère la sélection du jeu d'index indépendamment des données) utilisables pour l'indexation sémantique, en donnant à chaque fois les algorithmes spécifiques permettant leur mise en œuvre.

Les deux premières familles nous servent à considérer, dans une optique d'indexation sémantique, plusieurs critères déjà connus en sélection de termes ; nous montrons par ailleurs qu'un certain nombre d'entre eux ne sont en fait que peu efficaces pour la tâche considérée. La troisième famille nous a permis d'introduire le critère *CRM*, que nous proposons comme critère privilégié pour l'*indexation sémantique* (notons que l'évaluation des performances de ce critère et la comparaison avec d'autres techniques plus anciennes d'indexation sémantique est détaillée dans le chapitre suivant). Nous avons en outre proposé une solution permettant de répondre au problème posé par les données additionnelles, non connues au moment de la sélection du jeu d'index.

Bien que nous n'ayons pas abordé le problème de la désambigüisation sémantique, ou plus précisément, nous n'avons pas tenté d'y apporter de réponse,[40] nous avons néanmoins tenu compte du phénomène de la polysémie dans l'élaboration de notre cadre, de manière à pouvoir facilement y intégrer les informations obtenues par un module de désambiguisation.

En plus de l'élaboration de jeu d'indexation sémantique adapté aux données à traiter, nous avons également proposé une méthode permettant d'élaguer une ressource sémantique (et par-là même également obtenir un jeu d'index) indépendante des données, et donc utilisable lorsque celles-ci ne sont pas connues *a priori*.

Pour terminer, remarquons que les techniques proposées ici permettent de facilement combiner plusieurs ressources sémantiques, en particulier des ressources spécialisées dans différents domaines (il est simplement requis de fusionner les racines des différentes ressources).

[40] Cette problématique étant bien trop vaste, nous avons délibérément choisi de ne pas nous y attarder.

Chapitre 4

Validation de l'utilisation de thésaurus en RD

Résumé

Ce chapitre définit le cadre dans lequel les techniques d'indexation sémantique décrites au chapitre précédent ont été évaluées.

L'essentiel des évaluations a été réalisé en considérant la problématique de la recherche documentaire et en prenant en compte la relation sémantique fortement structurante d'hyper/hyponymie « est-un » pour un modèle vectoriel de recherche documentaire.

La pertinence des différentes techniques est évaluée sur la base des performances obtenues sur un ensemble de collections de références, et mesurées par le biais de deux indicateurs usuels : la précision moyenne et les courbes précision-rappel.

Les résultats que nous avons obtenus lors de ces évaluations nous permettent d'affirmer qu'il existe un potentiel certain pour l'utilisation de thesaurus sémantiques, qu'ils soient généralistes ou spécialisés.

Ces résultats montrent en outre que notre critère *CRM* que nous proposons d'utiliser pour déterminer l'index « sémantique » est effectivement prometteur et donne des résultats sensiblement supérieurs aux autres.

Dans un premier temps, nous donnons une description succincte des principes de base sur lesquelles se fondent l'essentiel des évaluations en recherche documentaire, les principales campagnes d'évaluations et conférences internationales du domaine, ainsi que les mesures permettant d'estimer la qualité ou le succès d'un processus de recherche documentaire. Nous présentons ensuite brièvement la tâche qui nous sert de cadre d'évaluation, ainsi que les collections de références et les ressources sémantiques (réseaux sémantiques *EDR* et *WordNet*) utilisées pour ces évaluations. Une description détaillée de la chaîne des traitements ayant permi de réaliser les évaluations est donnée, suivie de la présentation des résultats obtenus, selon différentes configuration de cette chaîne de traitement (avec ou sans pondérations des index lors de la recherche, en utilisant différents thésaurus sémantiques, ou encore différentes techniques de prise en compte des documents « dynamiques » que constituent les requêtes). En fin de chapitre, nous portons un regard critique sur la manière dont nos évaluations ont été menées, puis nous concluons en résumant les points principaux et en ajoutant quelques remarques sur le processus d'évaluation tel que conduit dans la communauté de la recherche documentaire.

4.1 Méthodes d'évaluations en RD

De manière générale, les applications des domaines du traitement de la langue sont difficiles à évaluer.

Chacune des tâches nécessite l'élaboration de protocoles d'évaluation (et l'obtention d'un concensus de la part de la communauté sur la pertinence de ce protocole), consistant le plus souvent à examiner le comportement du système à évaluer lorsqu'on lui soumet des *données de référence*, pour lesquelles les réponses « admissibles » (devant être considérées comme correctes) sont « connues ».[1] En recherche documentaire, on parlera de « corpus d'évaluation », combinant un ensemble de données (base documentaire) et un *référentiel* (ensemble de requêtes et réponses admissibles correspondantes).

Mais la constitution de tels corpus d'évaluation n'est pas aisée. Le volume des données devant être réunies est considérable (pour être statistiquement représentatif d'une part, et correspondre à la réalité de l'utilisation des systèmes de traitement de la langue d'autre part) ; le choix même des données à considérer est délicat (tant sur le plan de leur forme – articles de presse, correspondance, conversations, etc – que de leur domaine de provenance – généraliste ou par branche).[2] Enfin, l'obtention du référentiel est également une tâche ardue : elle nécessite en effet la constitution d'une collection de requêtes réalistes, et la confrontation de ces requêtes avec les documents de la base ; réalisé de manière essentiellement manuelle, le référentiel est non seulement coûteux à produire en raison du volume des données mais également en raison de la subjectivité intrinsèque de la tâche (qu'il s'agisse d'étiquetage morphosyntaxique ou de recherche documentaire, le taux d'accord entre deux experts humains n'est pas aussi élevé que notre *a priori* le laisserait supposer – l'expérience rapportée dans Gull [1956] est assez révélatrice à ce sujet).

Pendant de nombreuses années, les évaluations en matière de recherche documentaire ont ainsi été conduites sur la base de corpus d'évaluation de petites tailles, en appliquant un protocole d'évaluation (mesure effective de la performance d'un système) également spécifique, rendant très difficile toute comparaison.[3]

Dans le but de disposer de corpus d'évaluation de plus grande taille d'une part, et de faciliter le transfert technologique entre centres de recherche et produits d'autres part, le NIST[4] a, sous l'impulsion de Donna Harman, mis sur pied dans le début des années 1990 une conférence annuelle, batisée *TREC* (*Text REtrieval Conference*) organisée sous forme de campagne d'évaluation. En fournissant des corpus de référence de tailles nettement plus conséquente que ce qui existait jusqu'alors[5] et un protocole d'évaluation (*benchmark*), cette conférence permet aux groupes qui y participent de disposer d'un cadre commun pour l'évaluation et la valida-

[1] C'est le cadre classique pour l'évaluation des performances intrinsèques d'un modèle, mais cette approche « boîte-noire » ne dit rien sur l'ergonomie du système, et représente souvent une mesure de trop haut niveau pour évaluer correctement l'un des composants du système.

[2] À tel point que l'on considère généralement comme dépendant de la tâche la forme et la provenance des données.

[3] Remarquons que bien que la situation se soit quelque peu améliorée (notamment grâce aux campagnes d'évaluations), de telles comparaisons restent quasiment impossible à réaliser avec des évaluations individuelles de systèmes, décalées dans le temps.

[4] National Institute of Stnadards and Technology, Maryland, USA.

[5] Les premières éditions de la conférence ont utilisé les données collectées dans le cadre du programme TIPSTER.

tion de leurs systèmes.[6] Au cours des années, d'autres conférences du même genre ont vu le jour ; citons notamment la conférence asiatique NTCIR (dès 1998) et la conférence Européenne CLEF (dès 2000). D'autres campagnes ont également été ponctuellement conduites (Amaryllis, INEX), et ont permis d'enrichir la palette des corpus d'évaluation. (Un historique de la technique d'évaluation en RD, avec en particulier une présentation du paradigme *Cranfield* ainsi qu'une évaluation de la stabilité d'un certain nombre de ces techniques est présenté dans Voorhees [2001].)

Évaluation d'un système de recherche documentaire

Mesurer la performance d'un outil de recherche documentaire est difficile à réaliser : de nombreux facteurs peuvent être considérés, selon la nature et le contexte de la tâche (temps de réponse du système, importance de la justesse des résultats par rapport à leur exhaustivité, possibilités de « corriger » de manière itérative la recherche, etc.).

Un certain nombre de principes standards sont néanmoins couramment employés pour conduire des évaluation globales, indépendamment de l'implantation effective du système dans un environnement de fonctionnement. Deux indicateurs sont principalement utilisés pour rendre compte de la qualité d'un processus de recherche documentaire réalisé sur la base d'un corpus d'évaluation : la **précision**, qui correspond au ratio entre le nombre de documents pertinents retournés par le système et le nombre total de document retournés, et permet de mesurer la *justesse* du système,[7] et le **rappel**, correspondant au ratio entre le nombre de documents pertinents retournés et le nombre total de documents pertinents, et permettant de mesurer l'*exhaustivité* du système.[8] Dans l'idéal, on cherche naturellement à maximiser ces deux indicateurs.

Pour pouvoir tirer des conclusions fondées sur la base de ces indicateurs, il est nécessaire de les mesurer, pour un même système, sur un grand nombre de requêtes, afin d'en obtenir une moyenne fiable [Buckley et Voorhees, 2000, Voorhees, 2001].

Bien qu'ils permettent d'obtenir une indication relativement synthétique des performances d'un système, qui plus est, basée sur une évaluation de type « boîte-noire » (donc idéale quand il s'agit de comparer les performances de plusieurs systèmes, typiquement dans le cadre de campagnes d'évaluation), ces deux indicateurs posent néanmoins quelques problèmes. En premier lieu, on constate que la mesure du *rappel* s'applique bien dans le cadre de recherche en base de donnée (cadre dans lequel s'inscrivent les systèmes booléens), il n'en va pas de même en recherche vectorielle (ie. en présence de systèmes retournant une mesure de (dis)similarité, valuable sur tout couple <document,requête>). On constate également que l'ordre des documents n'est pas

[6] Une des limites de la conférence *TREC* est que les ressources produites sont à exploiter avec prudence hors du cadre de la conférence ; cela tient à la manière dont le référentiel est obtenu : en raison de la taille des bases documentaires, un jugement de pertinence manuel sur chaque document serait prohibitif. Une stratégie de *pooling* est mise en place pour limiter le nombre de documents examinés manuellement : n'entrent en jeu pour le jugement de pertinence que les n (typiquement 100) premiers documents rapportés par chaque concurrent. Le référentiel est donc en partie déterminé par les systèmes participant à la campagne ; l'utilisation après-coup du référentiel pour évaluer la performance d'un système est dès lors passablement approximative, les « mauvais » documents retournés par le système ne l'étant pas forcément [de Loupy, 2000]. Remarquons que l'influence de ces documents mal catégorisés est cependant faible (d'après Voorhees [2001]), si la procédure de *polling* est au point et qu'une mesure de performance « stable » est utilisée.

[7] On utilise parfois la notion de **bruit**, qui correspond au pourcentage de documents non pertinents retournés : $bruit = 1 - précision$.

[8] On utilise aussi la notion de **silence**, correspondant au pourcentage de documents pertinents « manqués » par le système : $silence = 1 - rappel$.

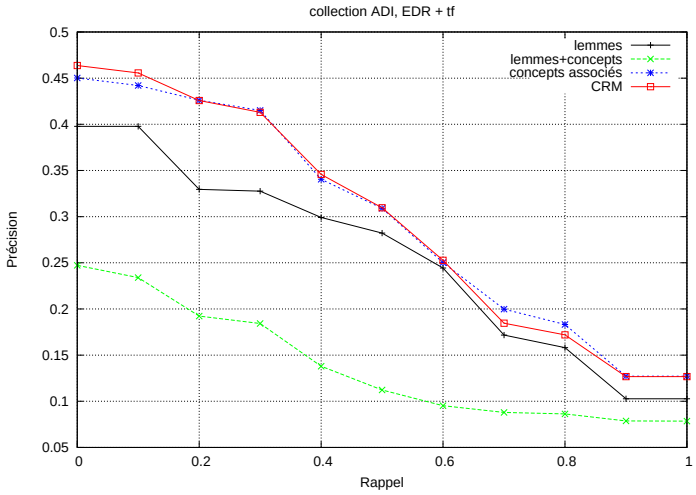


Figure 4.1: Exemple de courbes *Précision-Rappel*.

pris en compte, pas plus que « l'intensité » de la pertinence[9]. Il s'agit en fait du même problème, mais décomposé pour des raisons historiques. Avec un système retournant une réponse sous forme de liste ordonnée, la position dans la liste (plutôt vers la tête ou vers la queue) des documents effectivement pertinents est une indication importante sur la qualité du système – en particulier si potentiellement tous les documents de la collection peuvent être retournés. Les premiers systèmes de RD, basés sur le modèle booléen, retournaient une réponse sous forme de liste non ordonnée : le problème ne se posait pas. Une amélioration importante de ce modèle fut de permettre de retourner les résultats sous forme de liste ordonnée, les documents « les plus pertinents » (du moins du point de vue du système) étant en tête de liste : le problème de la mesure de l'ordre des documents effectivement pertinents s'est alors posé. Avec l'apparition du modèle vectoriel, la réponse (brute) du système consiste en une mesure de (dis)similarité pour l'ensemble des documents de la collection : bien que l'on puisse toujours le considérer comme un problème d'ordre, il s'agit plutôt d'un manque de mesure sur l'intensité des pertinences relevées (pour en tenir compte, il est nécessaire de disposer d'un référentiel valuant l'intensité des pertinences).

Un grand nombre de critères (tous plus ou moins basés sur ces notions de précision et de rappel) ont été mis au point pour tenir également compte de ces éléments lors de l'évaluation des perfomances d'un système de RD : *3/11pt-precision*, *R-precision*, rappel après N documents, etc.[10]

Bien que l'intérêt pour une mesure unique, synthétique, de la « performance » d'un système de RD soit marqué actuellement, une des représentations les plus informatives reste néanmoins la visualisation sous forme de courbes *précision/rappel* : pour une ensemble de requête, on mesure la précision en différents points de rappel (typiquement tous les 10%), et l'on reporte les valeurs moyennes obtenues sur un graphe, en plaçant la précision en ordonnée et le rappel en abscisse, tel qu'illustré par la figure 4.1. Il s'agit donc formellement de courbes de type *précision*(*rappel*), la précision étant calculée pour un rappel fixé ; au point de rappel 0, la précision usuellement

[9] Le faible taux d'accord entre experts humains pour cette tâche est en partie dû au fait qu'un document peut être plus ou moins pertinent, pour une requête donnée ; bien que la majorité des référentiels des corpus d'évaluation utilisent une valuation binaire de la pertinence d'un couple <document,requête> donné, il paraît évident qu'une valuation plus fine (sur 3 à 5 niveaux typiquement) permetterai de mieux rendre compte de la réalité de ces pertinences (p.ex. conf. NTCIR).

[10] Consulter à ce propos le « manuel » accompagnant l'outil d'évaluation (`trec_eval`) mis au point pour la campagne *TREC*, proposant actuellement un peu plus d'une quarantaine de critères.

admise correspond au maximum de précision relevé en considérant chacun des documents pertinents retournés par le système.

Relevons encore que ce type de courbe est particulièrement adapté aux systèmes à base de modèle vectoriel, pour lesquels il est facile de varier le rappel (et d'atteindre un rappel de 100%).

4.2 Tâche, corpus d'évaluation et ressources

Pour évaluer l'impact des techniques d'indexation proposées en 3 [page 21], nous avons pris le parti de considérer la tâche « historique » (et fondatrice) de la recherche documentaire, à savoir la *tâche ad hoc*.

Définition : tâche *ad hoc*

En **tâche *ad hoc***, on confronte à une base de documents fixe (et connue au préalable) un ensemble de requêtes *nouvelles*.[11]

Les corpus d'évaluation issus des campagnes du type *TREC*, bien qu'extrêmement intéressant à de nombreux égards, souffrent malgré tout d'un certain nombre d'inconvénients, en particulier lorsqu'il s'agit de réaliser des tests en profondeurs d'une technique donnée. Outre les inconvénients liés à la méthode de *pooling* utilisée pour constituer les listes de pertinence, la taille considérable des bases en question est elle-même problématique : bien qu'assurant des mesures statistiquement stables, les phénomènes localisés de sur ou sous-performances (généralement riches d'enseignements) sont totalement gommés ou noyés dans la masse des résultats. Par ailleurs, l'utilisation même de ces corpus nécessite un investissement en temps considérable, tant au niveau de leur mise en place que pour la réalisation des tests.

Nous avons donc choisi d'utiliser comme corpus d'évaluation des collections de plus petites tailles, fréquemment utilisées par le passé (et donc pour lesquelles il existe dans la littérature une base de comparaison – certes imparfaite, notamment en raison des pré-traitements généralement pas ou mal décrits, et donc présentant des résultats difficilement reproductibles). De l'ensemble des collections assemblées par Fox à l'institut polytechnique de Virginie (disponibles notamment en complément du système Smart) et brièvement décrites ci-après, nous avons utilisé les corpus *ADI*, *TIME*, *MED*, *CACM* et *CISI* ; en plus de l'intérêt présenté par leur petite taille, les bases de cette collection ont l'avantage de constituer des entités homogènes spécifiques, tant sur le plan du vocabulaire (conceptuel) que dans leur forme.[12] Utiliser plusieurs collections de nature différente permet en outre de limiter le risque de conclusions erronées, basées sur des résultats faussés par les caractéristiques sous-jacentes de la collection utilisée.

Ressources sémantiques additionnelles

Les évaluations ont été conduites en utilisant deux ressources sémantiques à couverture large : le dictionnaire électronique *EDR* d'une part, et le réseau sémantique issu du projet *WordNet* d'autre part.

[11] Du moins admises comme tel, en particulier lors d'une évaluation.

[12] D'autres particularités peuvent également être intéressantes, comme des informations de nature non textuelles (auteurs, dates, références bibliographiques, etc.), mais n'ont pas été prises en compte ici.

Collection	*Thématique*	*# doc.*	*# req.*	*# voc*	$\overline{occ/doc}$	$\overline{occ/req}$	$\overline{pert/req}$
ADI	(e) science de l'information	82	35	1632	58.146	13.800	4.857
TIME	(a) tous sujets	425	83	29999	561.082	14.108	3.904
CISI	(r) *library science*	1460	112	17471	123.097	77.143	27.804
MED	(r) médical	1033	30	15917	146.563	19.233	23.200
CACM	(r) informatique	3204	64	16702	56.963	20.828	12.438

Table 4.1*:* Description succincte et quelques données statistiques sur les collections de tests assemblées par Fox [1983] (voir également [Baeza-Yates et Ribeiro-Neto, 1999, pg. 91–97]) ; entre parenthèses est précisée la nature des documents, avec (a) pour *article*, (r) pour *résumé* et (e) pour *extrait* ; sont comptabilisés les mots (identifiés par l'outil de segmentation), sans filtrage d'aucune sorte.

4.2.0.5 Dictionnaire Électronique EDR

Le dictionnaire électronique *EDR* Miyoshi *et al.* [1996] est une ressource[13] bilingue Japonais–Anglais, développée entre 1986 et 1994 par le *Japan Electronic Dictionnary Research Institute* (créé à cet occasion) et actuellement maintenue par le *National Institute of Information and Communications Technology*. Son financement a été assuré par le *Japan Key Technology Center*, ainsi que différentes entreprises actives dans le secteur informatique.

Le but visé par la création de cette ressource était de disposer d'une infrastructure permettant des traitements avancés en langage naturel, au moyen d'ordinateurs et d'outils de traitements de connaissances.

Les principales caractéristiques du dictionnaire *EDR* sont :

1. une « couverture large » du vocabulaire utilisé de manière courante ;
2. une structuration et un format n'imposant pas d'applications ou d'algorithmes particuliers (usage non restreint) ;
3. la mise à disposition des connaissances nécessaires pour de réelles analyses sémantiques ;
4. un degré élevé d'objectivité, obtenu en se basant sur des corpus textuels conséquents ;
5. un contenu fondamental très général, indépendant de la langue ou du domaine.

Notons que bien qu'*EDR* traite pour l'essentiel de la langue dans son usage courant, une partie de la ressource cependant est spécialisée (le dictionnaire « technique », portant sur le domaine du traitement de l'information et des domaines connexes : informatique, électronique, etc).

La ressource est constituée de dictionnaires de quatre types différents (à peu près indépendant les uns des autres), complétés par une série de corpus :

Les dictionnaires de mots : ces dictionnaires définissent le vocabulaire couvert par la ressource ; ils comportent les dictionnaires : des mots anglais ($\approx 240'000$ entrées), des mots japonnais ($\approx 250'000$), des termes techniques anglais ($\approx 90'000$) et des termes techniques japonnais ($\approx 120'000$).

Chaque dictionnaire regroupe les informations morphologiques (graphie, prononciation, découpage syllabique, inflexion, accent, adjacence, ...) et syntaxique (cms, dénombrabilité, flexions, ...) des « mots » qu'ils définissent – principalement des lemmes (du moins en anglais) mais on trouve également un grand nombre[14] de multi-termes : mots composés, expressions idiomatiques, etc. – et permet de lier ces mots à des entrées du dictionnaire des concepts.

[13] Payante !

[14] Plus de $113'000$ en anglais, le plus long « mot » étant : « *want something as much as one want a hole in the head* ».

Les dictionnaires bilingues : qui mettent en correspondance les « mots » des dictionnaires des deux langues, tout en reprenant une partie des informations syntaxiques et grammaticales des dictionnaires de mots, de même que les liens vers le dictionnaire des concepts. Le dictionnaire « japonais → anglais » contient environ 230′000 entrées, tandis que « l'anglais → japonais » en compte lui un peu plus de 160′000. À ces deux dictionnaires s'ajoutent les deux traitant des termes techniques.

Les dictionnaire des concepts : constitués de deux sous-ensemble de dictionnaires, l'un pour les termes généraux et l'autre pour les termes techniques. Chaque sous-ensemble comporte trois dictionnaires distinct :

- le *Headconcept Dictionary* donne une description de chaque concept, comprenant un mot illustrant la signification du concept et une phrase explicitant le concept (dans un grand nombre de cas cependant, le « mot-sens » et/ou la phrase explicative n'existe qu'en japonais) ; environ 420′000 concepts « généraux » et 90′000 dédiés aux concepts « techniques » sont ainsi définis ;
- le *Concept Classification Dictionary* organise les concepts au moyen d'une relation inclusive (*est-un*) ; à chaque concept est associé un ensemble de sur et sous-concepts ($\tilde{\text{h}}$yperonyme/hyponymes) ; près de 510′000 relations (binaires) sont ainsi définies entre les différents concepts, constituant 17 niveaux pour les concepts « généraux » et 13 niveaux pour les concepts « techniques » ;
- le *Concept Description Dictionary* finalement, qui définit un ensemble de relations sémantiques binaires additionnelles (« objet », « agent », « but », « emplacement », etc) entre les concepts co-occurrant dans les phrases du corpus associés ; au total, un peu moins de 510′000 relations binaires sont ainsi définie, uniquement sur les concepts « généraux » (pas de CDD pour les concepts « techniques »).

Remarquons qu'une large part des concepts (plus de 205′000) ne sont associés à aucun mot[15] ; ces concepts ne peuvent être définis et appréhendés qu'au travers de leurs relations avec les autres concepts.

Les dictionnaires des co-occurrences : ces dictionnaires (quatre au total, selon la langue et l'aspect général ou technique) décrivent les informations de collocations de mots identifiées dans les corpus, sous forme de relations binaires. Le dictionnaire « général » japonais décrit environ 900′000 « phrases », et près de 460′000 pour le dictionnaire anglais.

Les corpus : contiennent des informations linguistiques obtenues en collectant un large ensemble de textes et en analysant leur contenu morphologique, syntaxique et sémantique ; pour chaque phrase (environ 200′000 pour le corpus japonais et 120′000 pour le corpus anglais), on dispose ainsi du découpage en mots, de l'étiquetage morphosyntaxique et de la racinisation, de l'arbre de dérivation syntaxique et finalement des informations sémantiques comprenant le découpage logique et bien évidemment les concepts mis en œuvre.

Dans notre cas, nous nous sommes restreints à l'utilisation du **dictionnaire des mots** anglais (généraux et techniques) et du **dictionnaire des concepts** (en nous limitant à la structure données par le *Concept Classification Dictionary*). Les feuilles de notre *réseau sémantique* (cf. 3.2 [page 24]) correspondent donc aux entrées du dictionnaire des mots, les autres sommets étant eux donnés par les entrées du dictionnaire des concepts ; les arcs entre les feuilles et les sommets sont donnés par les concepts possibles pour un mot donné, les arcs entre les autres sommets étant issus des relations binaires définies dans les dictionnaires de classification (*Concept Classification Dictionary*) et de description (*Concept Description Dictionary*) des concepts, seul le premier étant utilisé dans nos expériences (restriction à un thésaurus).

[15] Et pour la plupart, n'ont pas « d'explication » (mot-sens ou phrase explicative) en anglais.

4.2.0.6 Projet WordNet

WordNet [Miller *et al.*, 1990] est un thésaurus généraliste pour l'anglais. À l'origine, ce thésaurus, dont la constitution a débuté en 1985, visait à permettre un parcours conceptuel plutôt qu'alphabétique d'un dictionnaire, en regroupant les mots par classes de synonymes. Au fil du temps, le projet a pris de l'ampleur, et de nombreuses relations sémantiques sont maintenant définies entre les mots de la ressource (synonymie, hyponymie/hyperonymie, antonymie, méronymie, métonymie, implication, causalité, etc.). La différence majeure entre *WordNet* et *EDR* tient à l'organisation même du thésaurus (en plus du fait que *WordNet* est une ressource monolingue) : avec *WordNet*, les concepts (appelés *synset*) sont toujours reliés à des mots (organisés en classes de synonymes dont le synset est le représentant). Le nombre de type de relations sémantiques définies est également plus élevé, mais la taille globale du thésaurus (généraliste uniquement) est plus faible que celle d'*EDR* : *WordNet*, dans sa version actuelle, définit $\approx$ 200000 mots, organisés en $\approx$ 115000 classes de synonymes (synsets), et comporte $\approx$ 100000 relations d'hypo/hyperonymie. Finalement, notons que ce thésaurus ne comporte, en entrée du dictionnaire (ie. les mots définis par *WordNet*), que des formes lemmatisées de noms, verbes, ajectifs et adverbes.

Pour nos expériences, nous avons utilisé le portage *MySQL* de la version 2.1 de ce thésaurus.

4.3 Chaîne de traitements

Pour conduire nos expériences, nous avons mis en place la chaîne de traitements suivante, appliquée à chacune des collections de test précédemment évoquées :

Une phase de *pré-traitments* est tout d'abord effectuée, de manière à identifier les entités utiles à l'indexation :

1. En premier lieu, pour chaque document et chaque requête, l'ensemble des informations textuelles disponibles (titre et contenu) sont agrégées ; les informations d'autre nature (auteurs, sources, dates, etc.) sont elles ignorées.
2. Les contenus textuels sont alors segmentés, lemmatisés et étiquettés, au moyen d'un outil additionnel.[16] On obtient ainsi, pour chacun des « mots » de la collection, le triplet {*graphie*, *lemme*, *catégorie morpho-syntaxique (CMS)*}.
3. Les séquences de termes sont ensuite filtrées, sur la base de leurs CMS (ne sont conservés que les noms, adjectifs, verbes et adverbes).

L'indexation au moyen du thésaurus est ensuite réalisée dans une seconde phase, répétée pour chaque expérience :

1. Pour chaque document, on établit les correspondances entre les « mots » des documents et les « mots » (feuilles) du thésaurus ; on tente d'établir en priorité une correspondance avec :
   - la graphie et la catégorie morphosyntaxique ;
   - la graphie uniquement ;
   - le lemme et la catégorie morphosyntaxique ;
   - lemme uniquement.

   Les « termes » sans correspondance sont (artificiellement) ajoutés au réseau, comme des termes isolés.
2. On sélectionne ensuite dans le réseau sémantique les « concepts » immédiatement reliés aux « mots » identifiés à l'étape précédente ; selon l'expérience, on sélectionne :

[16] Sylex 1.7, © 1993-98 Decan Ingénia [Constant, 1995].

- la totalité des concepts possibles ;
- le concept « le plus probable » (dans l'absolu) pour le mot considéré – information fournie par les ontologies utilisées ;

La sélection du concept le plus fréquent en moyenne nous sert de pis-aller de désambigüisation sémantique.

3. La structure du réseau sémantique est ensuite parcourue en partant de chacun des sommets précédemment sélectionnés et en suivant les relations sémantiques considérées pour l'expérience, soit dans notre cas, la relation d'hyperonymie (« est-un ») ; on exhibe ainsi une sous-partie du réseau sémantique, restreinte à la couverture des documents de la collection à indexer.
4. Les termes d'index sont alors choisis (dans le réseau restreint), selon le critère adopté pour l'expérience (cf. chapitre 3 [page 21]).
5. L'indexation des documents est alors réalisée,[17] au moyen du jeu déterminé à l'étape précédente. Pour ce faire, on substitue les « termes » constitutifs des documents par les éléments du jeu d'index auxquels ils sont rattachés, *en fonction de la nature des relations sémantiques considérées*. Dans notre cas, compte tenu de la nature *inclusive* de la relation d'hyperonymie, les substitutifs des « termes » sont l'ensemble de leurs subordonnants[18] dans le jeu d'indexation.
6. L'indexation des requêtes est réalisée de manière similaire à celle des documents, à ceci près que, le jeu d'indexation étant construit uniquement à partir des données des documents (les requêtes étant supposées non connues *a priori*), des « termes » peuvent ne pas avoir de substitut dans le jeu d'indexation. Selon les expériences, on utilisera alors l'une ou l'autre des techniques proposées pour la prise en compte « d'individus supplémentaires » (cf. § 3.6 [page 48]).[19]

Lorsque l'indexation documents-requêtes est obtenue, la recherche documentaire à proprement parler est effectuée, au moyen du logiciel *Smart* (cf. § 2.2.2 [page 9]).

4.4 Résultats

Plusieurs expériences ont été réalisées, dans le but de tester différentes techniques d'indexation sémantique. Pour éviter que les paramètres du processus de recherche proprememnt dite ne compliquent par trop l'évalutation, nous nous sommes limités à une configuration relativement minimaliste (mais cependant tout à fait standard) ; la similarité *cosinus* (c.f. 2.4.3 [page 15]) est systématiquement utilisée pour valuer la « proximité » de deux vecteurs, et deux schémas de pondération sont appliqués pour projeter les documents dans l'espace vectoriel : la *fréquence d'occurrence* absolue par document (`tf`) et la *fréquence inverse en document* (`tf.idf`) – voir 2.4.2 [page 13].

Le programme `trec_eval`,[20] utilisé en sortie du système *Smart* permet d'obtenir les différentes valeurs de *précision* et *rappel* permettant d'évaluer les performances du processus de recherche. Nous ne retiendrons ici que la mesure de *précision* (pour être exact, la « précision moyenne globale » MAP[21]) et les courbes *précision*(*rappel*) (la précision moyenne en fonction d'un rappel

[17] Plus précisément, dans le cadre de nos expériences, les « termes » sont, à ce stade, uniquement « substitués » ; la phase d'indexation proprement dite est réalisée ultérieurement.

[18] Avec la convention qu'un élément se subordonne lui-même.

[19] En pratique, les « termes » non couverts par le réseau sémantique sont simplement ignorés, puisque nous n'utilisons pas dans nos expériences de techniques susceptibles d'apporter une contribution à ces éléments (sémantique distributionnelle, latente, etc.).

[20] `trec_eval` v.7.3, disponible sur le site web de la conférence *TREC* (rubrique « results »).

[21] « *Mean Average Precision* », soit la moyenne des précisions de chaque document pertinent retourné par le système pour une requête, puis moyennées en tenant compte de l'ensemble des requêtes.

interpolé).[22] Les mesures retenues présentent l'avantage d'être extrêmement stables (contrairement à d'autres tels que le rappel ou la précision à N documents par exemple), permettant de déduire des tendances même en présence de faibles différences [Voorhees, 2001].[23]

4.4.1 Évaluation de différentes méthodes d'indexation, modèle de base

Pour évaluer les performances d'une indexation sémantique, en particulier utilisant le critère de redondance minimale, *CRM*, introduit au chapitre 3 (cf. § 3.5.3.1 [page 40]), nous avons choisi de confronter trois techniques d'indexation utilisant l'information additionnelle d'un réseau sémantique et une indexation traditionnelle, basée sur les mots (plus précisément sur les lemmes).

indexation « traditionnelle », par les *lemmes* :
: indexation standard, utilisant les lemmes des « termes » des documents comme jeu d'index – les informations apportées par le réseau sémantique ne sont pas utilisées ;

lemmes+concepts :
: *expansion* du jeu d'indexation standard, chaque « terme » étant complété par l'ensemble des « concepts » auxquels ils sont reliés au travers des relations sémantiques considérées, transitivement ; dans notre cas, par l'ensemble des concepts hyperonymes directs ou indirects du « terme » ; (correspond à l'indexation *onto-matching* de Kiryakov et Simov [1999] – cf. § 2.5.2.2 [page 19]) ;

concept associé :
: *substitution* de chaque « terme » par le ou les « concepts » associés – dans *WordNet*, ensemble des *synsets* du « terme » (correspond à l'indexation par classes de synonymes (Gonzalo *et al.* [1998*b*], Voorhees [1993] – c.f. 2.5.2.2 [page 19]) ; cette technique peut être vue comme choisir une coupe de hauteur pré-déterminée.

coupe de redondance minimale :
: *subsitution* des « termes » par un ou plusieurs « concepts » appartenant à la coupe de redondance minimale, tel que décrit précédemment.

La table 4.2 [page 66] présente les tailles d'index et les valeurs du critère de précision moyenne MAP obtenues avec les différents schémas d'indexation, en utilisant la ressource sémantique *EDR*. Pour chacune des bases d'évaluation, la ou les techniques d'indexation donnant le meilleur résultat (sur la mesure considérée) est reportée en gras.

4.4.1.1 Taille des index

À la lecture de cette table, on remarque sans surprise que l'expansion par l'ensemble des concepts conduit à une augmentation considérable des index (d'un facteur 4 à 8), la subsitution

[22] Nous avons volontairement renoncé à l'emploi d'indicateurs supplémentaires, tel qu'une mesure unique de rappel (difficilement compatible avec le classement des documents tels que réalisé par le système vectoriel, en l'absence de définition précise d'une tâche), ou des mesures plus récentes, telles que la *F-Measure*, dont le détail de construction, en particulier dans les utilisations reportées dans la littérature, sont pour le moins peu explicites – type de précision utilisée (moyenne, interpolée, etc), ratio entre précision et rappel, etc.

[23] Par ailleurs, le choix de mesures plus spécifiques n'est à nos yeux justifié que dans le cas où la mise en œuvre de l'outil de recherche est connue : importance de l'exhausitvité des documents pertinents (recherche en jurisprudence v.s. recherche sur internet), importance de la pertinence des premiers documents retrouvés (interaction vocale ou par sms avec l'utilisateur v.s. filtrage en entrée d'un système automatique de réponse à des questions).

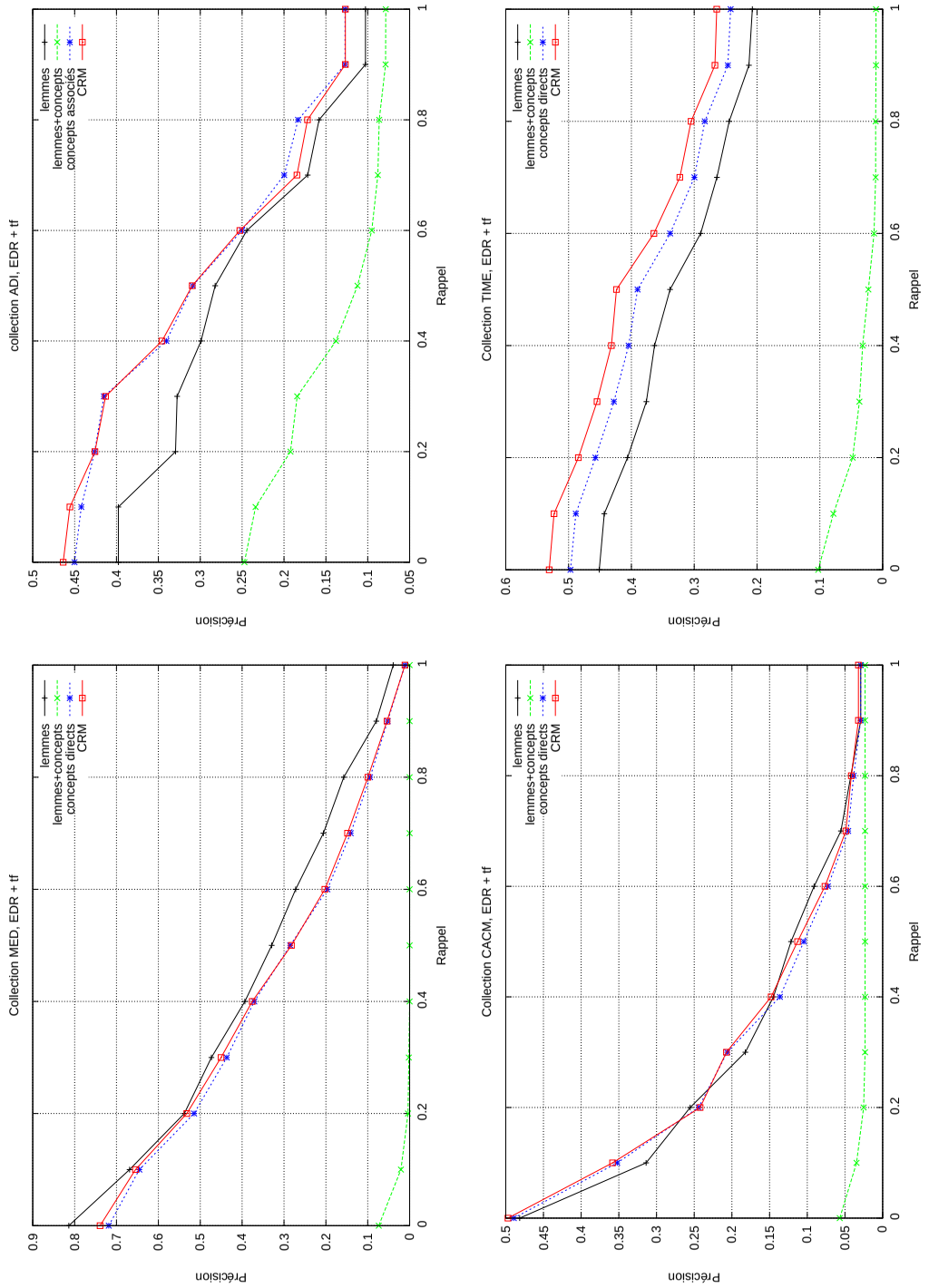


Figure 4.2*:* Courbes *précision–rappel (interpolé)* (PR) de différentes méthodes d'indexations – pas de repondération (tf), thésaurus *EDR*, concept le plus fréquent. Chaque graphique présente, pour une collection donnée, les courbes PR de l'indexation au moyen : des mots uniquement (base de comparaison), des mots augmentés de tous les concepts (expansion), des concepts directement associés aux mots (factorisation), et des termes issus de la coupe de redondance minimale.

		lemmes	**lemmes & concepts**		**concepts associés**		**coupe *CRM***	
		seuls	*sens mltp*	*+fréquent*	*sens mltp*	*+fréquent*	*sens mltp*	*+fréquent*
ADI	*|index|*	1800	14664	5254	10080	2891	5516	1740
	précision	0.2409	0.1085	0.1251	0.2343	**0.2828**	0.2310	**0.2819**
TIME	*|index|*	21815	92117	53142	69354	31583	18949	18404
	précision	0.3133	0.0298	0.0307	0.2185	0.3576	0.3625	**0.3852**
MED	*|index|*	11893	56091	30305	41742	18113	10206	10306
	précision	**0.3404**	0.0069	0.0061	0.1677	0.2930	0.1521	0.3014
CISI	*|index|*	10019	53453	26278	39544	14998	8404	8114
	précision	0.0507	0.0142	0.0108	0.0407	0.0617	0.0579	**0.0633**
CACM	*|index|*	10053	51712	25208	38524	14696	8410	8339
	précision	0.1442	0.0143	0.0262	0.1425	0.1423	0.1264	**0.1472**

Table 4.2*:* Comparaison de différents mode d'indexation : taille du jeu d'index et précision MAP – recherche sans pondération (`tf`), réseau sémantique *EDR*.

par les concepts associés aux termes augmente, elle, le jeu d'index d'une facteur 2 à 3, tandis que la substitution par les éléments de la coupe *CRM* permet d'obtenir un jeu d'index de taille comparable à celui constitué des termes, et même légèrement plus petite (env. 85%).

Sans surprise également, le fait de limiter la perplexité du système, en choisissant un concept parmi les candidats susceptibles de correspondre à un terme, permet d'obtenir un jeu d'index 2 à 3 fois plus petit qu'en conservant l'ensemble des candidats (colonnes *sens mltp* vs. *+fréquent*). On constate néanmoins que les tailles d'index obtenus au moyen de la coupe *CRM* ne semblent pas réellement être affectées par cette réduction de perplexité, à l'exception notoire de la collection *ADI* (voir ci-après) – on peut même constater l'effet inverse (cas de la collection *MED*), à savoir une augmentation de la taille de l'index. Ce phénomène, surprenant au premier abord, s'explique cependant très bien lorsque l'on considère le fonctionnement de l'algorithme : en effet, rien n'empêche qu'un plus fort degré de généralisation soit privilégié lorsque les extrémités des branches du DAG sont plus peuplées (*sens mltp* vs. *+fréquent*). De plus, il faut garder à l'esprit que l'algorithme utilisé procède par recherche de proche en proche, en dirigeant l'exploration vers le haut du DAG (généralisation) dans un premier temps, avant d'éventuellement redescendre là où cela peut amener un gain ; la recherche étant interrompue après un certain laps de temps, il n'est pas impossible qu'une exploration sur un nombre plus faible de chemins (cas *+fréquent*, limitation la perplexité) arrive plus rapidement à la phase « descendante » de la recherche que dans le cas où sont conservés dans le DAG tous les concepts possibles pour un même terme (cas *sens mltpl*), produisant par-là même un jeu d'index plus fourni (plus spécifique).

La collection *ADI* présente certaines particularités : tout d'abord, l'expansion du jeu d'index résulte en une très forte augmentation de sa taille (d'un facteur 8, contre 4 en moyenne pour les autres collections) ; on constate également qu'avec cette collection, la limitation de perplexité sur les concepts associés à un terme à un impact important sur la taille du jeu d'index issu de la coupe *CRM*, contrairement aux autres collections. Deux phénomènes concomitant sont vraisemblablement à l'origine de cela : d'une part, la taille très faible – dans l'absolu – de la collection, qui rend proportionnellement important le surcoût de termes lié à l'utilisation du réseau sémantique, et d'autre part le fait que le réseau utilisé consiste en la réunion de deux réseaux sémantiques (vocabulaire général pour l'un, technique (informatique, etc) pour l'autre), partageant fréquemment les mêmes termes ; le vocabulaire spécifique de la collection étant sur-représenté parmi ces termes définis doublement (et donc pour lesquels la perplexité est singulièrement plus élevée qu'en moyenne).

Finalement, notons également un point qui pourrait prêter à confusion : on constate en effet qu'avec l'indexation par les lemmes, la taille du jeu d'index (1800) est *supérieure* au nombre de

mots identifés pour cette collection (1632) et rapporté dans la table 4.1 [page 60]. L'explication est que les termes d'index sont ici définis par l'outil de recherche documentaire (*Smart*), la segmentation n'étant pas la même (la table 4.1 rapportant un nombre de « mots » tels que définis par la segmentation de l'outil d'étiquetage morphosytaxique (*Sylex*, plus habile que *Smart* dans la détection des mots composés).

4.4.1.2 Performance des techniques d'indexation

Si l'on effectue une comparaison brutale des performances des différentes techniques d'indexation (mesurées à l'aune de la précision moyenne MAP, table 4.2 [page précédente]), en les classant de la meilleure à la plus mauvaise pour chaque collection, on obtient les rangs moyens suivants :

Technique	*Rang moyen*
Coupe *CRM*, sens le plus fréquent	1.4
Concepts directs, sens le plus fréquent	2.6
Indexation par les lemmes (standard)	2.8
Coupe *CRM*, sens multiples	4.0
Concepts directs, sens multiples	4.2
Lemmes et concepts, sens le plus fréquent	6.4
Lemmes et concepts, sens multiples	6.6

Nous constatons que :

- la réduction de perplexité (« désambigüisation sémantique »), bien que réalisée de manière très grossière, améliore notablement les performances, et est presque systématiquement préférable à l'alternative consistant à conserver l'ensemble des sens possibles (absence de désambigüisation sémantique) ;
- l'indexation par expansion des termes est la pire méthode, dégradant notablement les performances de la technique standard ; cela n'est guère étonnant, en particulier en l'absence de pondération modérant l'inflation des termes d'index (dont un nombre important se retrouvent dans tous les documents).
- Les techniques d'indexation sémantique par substitution des termes d'index montre un potentiel intéressant, avec un avantage certain pour l'indexation au moyen de la coupe *CRM*.
- Le potentiel des indexations sémantiques par substitution est probant dans le cas de collections admettant un vocabulaire très général (*TIME*) – bien couvert par le réseau sémantique – ou admettant un vocabulaire spécifique (*ADI*) particulièrement bien couvert par le réseau. *A contrario*, lorsque la collection présente un vocabulaire (spécifique) peu ou mal couvert par le réseau (*MED*), l'indexation sémantique n'apporte pas d'amélioration des performances.

Pour obtenir un portrait un peu plus précis des performances de ces différentes techniques, il est intéressant d'examiner les courbes de *Précision(Rappel)* présentées en figure 4.2 [page 65] [24] pour les quatres bases *ADI*, *TIME*, *MED* et *CACM*. En plus de valider les constats précédents, ces courbes nous permettent de voir que les deux techniques d'indexation sémantique qui arrivent en tête au niveau de la mesure MAP sont, selon les collections, soit meilleures que l'indexation standard, soit relativement équivalente.

Afin de vérifier si le gain en performance des techniques d'indexation sémantique présent lors d'une recherche de base est conservé avec une recherche plus sophistiquée, en particulier si

[24] Les courbes présentées admettent toutes la suppression de la polysémie par la sélection du concepts le plus fréquent en moyenne.

		lemmes seuls	**lemmes + concepts**		**concepts directs**		**coupe *CRM***	
			sens mltp	*+fréquent*	*sens mltp*	*+fréquent*	*sens mltp*	*+fréquent*
ADI	*tf*	0.2409	0.1085	0.1251	0.2343	**0.2828**	0.2310	**0.2819**
	tf.idf	0.3431	0.2871	0.3840	0.3140	**0.4071**	0.2976	**0.4071**
TIME	*tf*	0.3133	0.0298	0.0307	0.2185	0.3576	0.3625	**0.3852**
	tf.idf	0.5349	0.3908	0.4999	0.4284	0.5442	**0.5476**	**0.5471**
MED	*tf*	**0.3404**	0.0069	0.0061	0.1677	0.2930	0.1521	0.3014
	tf.idf	**0.4470**	0.2532	0.3984	0.2704	0.4284	0.3679	0.4226
CISI	*tf*	0.0507	0.0142	0.0108	0.0407	0.0617	0.0579	**0.0633**
	tf.idf	0.1535	0.0875	0.1413	0.0959	**0.1632**	0.1300	0.1489
CACM	*tf*	0.1442	0.0143	0.0262	0.1425	0.1423	0.1264	**0.1472**
	tf.idf	0.2814	0.1245	0.2324	0.1869	0.2774	0.2532	**0.2950**

Table 4.3*:* Performance (précision MAP) des différents mode d'indexation sans pondération et en pondération tf.idf (réseau sémantique *EDR*).

ce gain ne disparaît pas en présence de pondération du poids des termes d'index, nous avons réitéré l'expérience précédente en pondérant les termes d'index par leur fréquence inverse en document (tf.idf) – modèle ususel, connu pour donner des performances raisonnables[25] – les résultats (précision moyenne MAP) de cette expérience sont reportés en table 4.3 [ci-dessus].

En réalisant à nouveau la moyenne des rangs des différentes techniques, on obtient le classement suivant :

Technique	*Rang moyen*
(sens +fréquent) Coupe *CRM* **et** concepts directs (*ex æquo*)	2.0
Indexation par les lemmes (standard)	2.6
Coupe *CRM* (sens multiples) **et** Lemmes+concepts (sens +fréquent) (*ex æquo*)	4.2
Concepts directs, sens multiples	5.8
Lemmes et concepts, sens multiples	7.0

On constate, malgré un certain tassement, que les techniques d'indexation sémantique par substitution restent optimales. L'essentiel des constatations précédentes demeure valable, même si, comme l'on pouvait s'y attendre, la technique d'expansion des index devient un peu plus performante (avec désambigüisation sémantique).

4.4.2 *EDR* vs. *WordNet*

La table 4.4 [page ci-contre] permet de comparer les performances obtenues avec deux réseaux sémantiques, *EDR* et *WordNet*, tous deux étant à large couverture mais de structures sensiblement différentes. Dans l'ensemble, les résultats obtenus sur chacun de ces réseaux sont relativement comparables ; bien que la mesure de précision MAP soit plus souvent meilleure avec *WordNet* qu'*EDR*, lorsque ce dernier est supérieur, il l'est de manière un peu plus prononcée.

Si l'on se concentre sur les résultats obtenus avec la pondération tf.idf, on observe qu'*EDR* est préférable à *WordNet* sur les collections *ADI* et, dans une moindre mesure, *CACM* ; sur les collections *MED* et, dans une moindre mesure, *TIME*, c'est au contraire *WordNet* qui est supérieur ; sur *CISI* enfin, les deux réseaux sont globalement au même niveau. Le vocabulaire

[25] Même si des pondérations plus évoluées, notamment les pondérations à pivot (par exemple *Lnu*), semblent encore plus performantes, du moins sur des corpus volumineux.

			lemmes seuls	concepts directs		coupe *CRM*	
				sens mltp	*+fréquent*	*sens mltp*	*+fréquent*
ADI	*tf*	*EDR*	0.2409	0.2343	**0.2828**	0.2310	0.2819
		WordNet		0.2494	0.2552	0.2463	**0.2834**
	tfidf	*EDR*	0.3431	0.3140	**0.4071**	0.2976	**0.4071**
		WordNet		0.3370	0.3858	0.3353	0.3766
TIME	*tf*	*EDR*	0.3133	0.2185	0.3576	0.3625	**0.3852**
		WordNet		0.2458	0.3445	0.2858	0.3323
	tfidf	*EDR*	0.5349	0.4284	0.5442	0.5476	0.5471
		WordNet		0.4927	0.5408	0.5492	**0.5594**
MED	*tf*	*EDR*	0.3404	0.1677	0.2930	0.1521	0.3014
		WordNet		0.2575	**0.3471**	0.1732	0.3141
	tfidf	*EDR*	0.4470	0.2704	0.4284	0.3679	0.4226
		WordNet		0.4051	**0.4611**	0.3330	0.4535
CISI	*tf*	*EDR*	0.0507	0.0407	0.0617	0.0579	0.0633
		WordNet		0.0428	**0.0853**	0.0327	0.0591
	tfidf	*EDR*	0.1535	0.0959	**0.1632**	0.1300	0.1489
		WordNet		0.1158	0.1581	0.0688	0.1520
CACM	*tf*	*EDR*	0.1442	0.1425	0.1423	0.1264	0.1472
		WordNet		**0.1502**	0.1327	0.1027	**0.1499**
	tfidf	*EDR*	0.2814	0.1869	0.2774	0.2532	**0.2950**
		WordNet		0.2277	0.2607	0.1808	0.2852

Table 4.4*:* Comparaison des performances (précision MAP) des réseaux sémantiques *EDR* et *WordNet*.

spécifique des différentes collections explique sans doute en grande partie ces observations. La partie spécialisée du thésaurus d'*EDR* correspond bien au vocabulaire présent dans les collections *ADI* et *CACM* (termes techniques), mais ne sert pas dans les autres cas. La collection *CISI* est connue pour les mauvaises performances qu'obtiennent avec elle les systèmes de RD ; ceux que nous testons ne font pas exception.[26] A l'inverse, *TIME* permet d'obtenir le plus souvent des performances élevées[27] ; la nature très généraliste du vocabulaire de cette collection s'accorde visiblement mieux avec *WordNet* (les différences sont néanmoins faibles).

Cependant, en comparant cette fois les résultats sans pondération (tf) et avec (tf.idf), ou entre la conservation de la polysémie et sa suppression au profit du sens le plus fréquent, on constate que ces résultats sont pour une large part remis en cause (sauf sur *CISI*). Cela tient vraisemblablement au fait que la suppression de la polysémie concomitante avec la pondération tf.idf permet de réellement utiliser la partie spécialisée du thésaurus *EDR*. Dans ce cas en effet, le sens spécialisé du vocabulaire réellement spécifique est conservé, et le poids de ces termes est, compte tenu de leur rareté, rehaussé par la pondération ; en l'absence de pondération particulière, ou lorsque tous les sens sont conservés. ces termes se retrouvent noyés, et n'ont plus d'influence.

Finalement, on notera que le haut degré de polysémie présent dans *EDR* dessert ce thésaurus si son utilisation n'est pas accompagnée d'un processus de désambigüisation.

[26] Dans cette collection en effet, les documents pertinents pour une requête donnée ne partagent le plus souvent aucun termes avec cette requête ; il semble qu'il en va de même avec les hyperonymes – il semble effectivement que pour cette collection, les informations qui permettent de lier un document à une requête sont plutôt à trouver dans les champs additionnels accompagnant les documents (auteurs, co-citations, etc.).

[27] Grâce à la longueur des documents de cette collection, et au nombre élevés de termes partagés par les requêtes (elles aussi relativement longues) et par les documents qui leur correspondent.

			lemmes seuls	c.directs (+fréq)	coupe *CRM* sens mltp	coupe *CRM* +fréquent
ADI	*tf*	*standard*	0.2409	0.2828	0.2310	0.2819
		concept			0.2307	**0.2834**
		hyperonyme			0.2246	0.2630
	tfidf	*standard*	0.3431	0.4071	0.2976	**0.4071**
		concept			0.2973	0.4058
		hyperonyme			0.2785	0.3466
TIME	*tf*	*standard*	0.3133	0.3576	0.3625	**0.3852**
		concept			0.3623	**0.3852**
		hyperonyme			0.3623	**0.3852**
	tfidf	*standard*	0.5349	0.5442	0**.5476**	**0.5471**
		concept			0.5469	0.5468
		hyperonyme			0.5469	0.5468
MED	*tf*	*standard*	**0.3404**	0.2930	0.1521	0.3014
		concept			0.1520	0.3014
		hyperonyme			0.1511	0.2934
	tfidf	*standard*	**0.4470**	0.4284	0.3679	0.4226
		concept			0.3667	0.4226
		hyperonyme			0.3375	0.3831
CISI	*tf*	*standard*	0.0507	0.0617	0.0579	**0.0633**
		concept			0.0596	**0.0633**
		hyperonyme			0.0587	0.0626
	tfidf	*standard*	0.1535	**0.1632**	0.1300	0.1489
		concept			0.1408	0.1488
		hyperonyme			0.1324	0.1422
CACM	*tf*	*standard*	0.1442	0.1423	0.1264	**0.1472**
		concept			0.1257	0.1439
		hyperonyme			0.1242	0.1427
	tfidf	*standard*	0.2814	0.2774	0.2532	**0.2950**
		concept			0.2529	**0.2946**
		hyperonyme			0.2342	0.2674

Table 4.5*:* Comparaison des performances (précision MAP) selon le traitement des termes supplémentaires présents dans les requêtes (réseau *EDR*. Dans l'expérience « *standard* », ces termes sont simplement ignorés ; avec « *concept* », les termes pris en compte sont ceux dont les concepts dominent une partie de l'index ; « *hyperonyme* » permet de prendre en compte les termes dont les concepts ou les hyperonymes directs de ces concepts dominent une partie de l'index (cf. § 3.6 [page 48])

4.4.3 Évaluation de l'indexation des requêtes

Les différentes manières d'indexer les individus « supplémentaires » (le vocabulaire présent dans la ressource sémantique mais non rencontré lors de l'indexation des documents), exposées en section 3.6 [page 48], ont également fait l'objet d'une évaluation.La table 4.5 [ci-dessus] permet de comparer les trois types de prise en compte des mots absents de l'index mais pouvant y être rattachés : aucun traitement (correspondant à l'expérience désignée « standard »), apparentement aux termes d'index subordonnés par les concepts associés à ces mots (expérience « concept »), et apparentement aux termes d'index subordonnés par les concepts reliés à ces mots et/ou leurs hyperonymes directs (expérience « hyperonyme »).

Si la technique consistant à remonter jusqu'aux hyperonymes des concepts associés aux termes supplémentaires est clairement la moins performante, les choses sont un peu moins nettes en ce qui concerne les stratégies consistant à remonter jusqu'aux concepts ou à ignorer systémati-

	$\widehat{\theta_s}$	sens multiples		sens +fréquent	
		tf	*tfidf*	*tf*	*tfidf*
ADI	P_{tf}	0.2310	0.2976	0.2819	0.4071
	P_{tfidf}	**0.2891**	**0.3415**	0.2813	**0.4155**
TIME	P_{tf}	0.3625	0.5476	**0.3852**	0.5471
	P_{tfidf}	**0.3797**	**0.5491**	0.3818	**0.5501**
MED	P_{tf}	0.1521	0.3679	0.3014	0.4226
	P_{tfidf}	**0.3141**	**0.4241**	**0.3138**	**0.4300**
CISI	P_{tf}	0.0579	0.1300	0.0633	0.1489
	P_{tfidf}	**0.0651**	**0.1525**	**0.0650**	**0.1540**
CACM	P_{tf}	0.1264	**0.2532**	0.1472	0.2950
	P_{tfidf}	**0.1388**	0.2303	**0.1489**	**0.3018**

Table 4.6*:* Comparaison des performances (précision MAP) des techniques d'estimation de la probabilité d'une coupe P_{tf} et P_{tfidf} (sélection du jeu d'index selon le critère *CRM*, ressource sémantique *EDR*).

quement ces termes, bien que cette dernière semble (en moyenne sur les collections considérées) malgré tout légèrement préférable.

On remarque en outre que l'influence de ces traitements se fait surtout sentir (de manière négative) sur les collections *ADI* et *CACM*, alors que leur contribution est très faible dans les autres cas. Cela s'explique par la proportion nettement plus élevée de termes concernés par ces appariements avec ces deux collections qu'avec les autres.

4.4.4 $\widehat{\theta_s}$ via $P_{\mathsf{tf}}(s|\Gamma)$ vs $P_{\mathsf{tfidf}}(s|\Gamma)$

Nous avons également réalisé une série d'évaluations en changeant la mesure élémentaire qui nous permet de comptabiliser les « poids » des occurrences des termes dominés par les sommets d'une coupe en vue d'estimer leur probabilité dans les documents de la collection (cf 3.5.4.2 [page 47]).

À la mesure employée jusqu'alors, $\mu' = \mathcal{W}$ (soit $\theta_s \widehat{=} P_{\mathsf{tf}}(s|\Gamma)$), nous avons choisi de confronter la mesure résultant de la pondération par la fréquence inverse en document, $\mu' = \mathcal{W} \cdot \mathsf{idf}$ (soit $\theta_s \widehat{=} P_{\mathsf{tfidf}}(s|\Gamma)$), ce schéma de pondération donnant de bons résultats pour la « capture » de l'importance d'un terme en fonction de sa distribution dans la collection de documents. L'idée est ici de favoriser, dans le jeu d'index, les termes présentant non pas une probabilité d'occurrence (moyenne) uniforme[28] mais une « distribution » uniforme dans la collection.[29]

La table 4.6 reporte les performances obtenues avec ces deux estimations, sur les expériences sans désambiguisation (« sens multiples ») et avec (« sens +fréquent »), et sans re-pondération lors de l'indexation (« indexation : tf »), ou avec une pondération inverse en document (« indexation : tfidf »). On constate que l'estimation selon la mesure P_{tfidf} donne de meilleurs résultats dans pratiquement tous les cas. L'utilisation de ce même mode de pondération en phase d'indexation augmente dans une proportion à peu près similaire les performances des deux techniques d'estimation, et l'on constate de plus que l'apport de la désambigüisation sémantique rudimentaire est globalement plus fort sur l'estimation via P_{tf} que sur P_{tfidf}.

On peut donc en conclure que l'estimateur utilisé pour attribuer une probabilité aux éléments des coupes est un élément important (ce qui n'est en fin de compte guère surprenant), et que le

[28] Cas de $\mu' = \mathcal{W}$, ne distinguant pas un terme fortement sur-représenté dans une petit nombre de documents (et absents des autres) d'un terme faiblement représenté mais occurrant dans un grand nombre de document.

[29] Dans la limite du reflet de la distribution des occurences d'un terme sur les documents apportée par la fréquence inverse en document.

recours à une mesure différente de celle de la probabilité intrinsèque d'apparition d'un concept dans le corpus, reflétant mieux la valeur informative de ce concept, doit permettre de choisir des jeux d'index avec un pouvoir discrimant encore plus élevé.

4.5 Conclusion

Ce chapitre a porté sur l'évaluation, dans le cadre d'une tâche traditionnelle de recherche documentaire (tâche *ad hoc*), d'un certain nombre de techniques d'*indexation sémantique* présentées au chapitre 3 [page 21], permettant d'intégrer, dans l'espace de représentation des documents d'une collection, des informations de nature sémantique issues de ressources externes. Après avoir donné le cadre général utilisé pour effectuer ces évaluations (comprenant la méthodologie employée, les collections de références ainsi qu'une brève description des ressources externes utilisées), nous avons ensuite détaillé les traitements effectués pour conduire nos évaluations, et avons terminé par une revue des différentes expériences effectuées, en donnant et commentant les principaux résultats de ces expériences.

Bien que partielles et menées sur des collections de petites tailles, il ressort au final de ces évaluations qu'il existe un potentiel certain pour la mise en application du principe d'*indexation sémantique* en recherche documentaire.[30]

Le point primordial pour que ces techniques offrent un intérêt est évidemment la bonne adéquation entre cette (ou ces) ressource et les données à indexer. On observera néanmoins qu'avec une méthode adaptative du jeu d'index, telle que proposée au travers du critère *CRM*, non seulement on autorise l'exploitation du potentiel de cette information sémantique là où elle est pertinente, mais de plus on gomme les portions qui ne seraient pas adéquates ; cette propriété est particulièrement intéressante dans l'optique d'une combinaison de ressources, typiquement généralistes d'une part (*EDR*, *WordNet*) et très spécialisées d'autre part (par exemple *Mesh*).

Il reste néanmoins de nombreux problèmes à surmonter, le plus conséquent d'entre eux étant certainement une désambigüisation sémantique suffisamment performante pour assurer une utilisation correcte de la ressource sémantique.

On peut en effet émettre un certain nombre de réserves à l'encontre de notre mise en œuvre des indexations sémantiques, et notamment de la chaîne de traitement présentée en section 4.3 [page 62]. Avant de les énumerer, remarquons que ces réserves n'impliquent pas que les résultats obtenus par les indexations sémantiques (et en particulier les indexations de type *CRM*) aient été surévalués, ou que ces indexations aient été favorisées, mais au contraire qu'il est assurément possible d'améliorer notablement leur mise en œuvre, et donc obtenir des résultats probablement encore meilleures pour ces différentes techniques.

Nous retiendrons les éléments suivants :

- Pour des raisons liées à la manière dont la chaîne de traitement a progressivement été mise en place, **trois segmentation différentes** s'y cotoyent au final : la segmentation initiale des données par l'étiqueteur morphosyntaxique, lors des pré-traitements ; celle de l'outil de recherche documentaire (*Smart*) ; et finalement la segmentation (implicite) induite par le réseau sémantique. Si les deux premières peuvent sans peine être fusionnées, les choses sont un

[30] Il serait indéniablement souhaitable de mettre en œuvre un test statistique (de type « test du signe », Goutte et Gaussier [2005], Savoy [2006]) afin de s'assurer que les différences de performances observées sont effectivement statistiquement significatives. On peut cependant observer que la stabilité de la mesure employée d'une part, et la multiplication des évaluations menées d'autre part conforte les conclusion auxquelles nous arrivons, et ce tant sur le plan de la signification statistique qu'en ce qui concerne la fragilité induite par les déficiences des collections de références utilisées.

peu plus compliquées en ce qui concerne la segmentation due au réseau sémantique (incontournable, et sur laquelle nous ne pouvons avoir d'influence). La seule solution pour unifier ces différentes segmentations consiste donc à n'utiliser que celle induite par la ressource sémantique. Les avantages d'une telle approche sont multiples : d'une part on peut maximiser la couverture des données par la ressource (en particulier lorsque l'on est en présence de multi-termes), et d'autre part on bénéficie ainsi d'une segmentation établie avec soin (lors de la construction de la ressource sémantique), et pertinente par rapport au(x) domaine(s) auquel(s) se rapporte la ressource. La connaissance des flexions des différents termes du réseau (ainsi qu'une grammaire minimaliste pour les expressions idiomatiques – deux informations présentes dans le réseau *EDR*) rend envisageable une telle homogénéisation des segmentations, au prix de l'entraînement d'un parseur combiné à une segmentation de « dernier recours », fonctionnant sur des données non couvertes par la ressource.

- Lors de l'indexation proprement dite, il conviendrait de **repondérer les termes d'index** associé à l'occurrence d'un mot. En effet, du fait de la polysémie présente dans la ressource sémantique (que ce soit au niveau des mots ou des concepts), et sans traitement particulier, l'occurrence d'un mot polysémique se traduit par une « expansion » de cette occurrence en autant de termes que le mot admet de sens dans le jeu d'index. Pour éviter que la « contribution » de cette occurrence au document ne soit surévaluée dans la représentation indexée, il faut normaliser la contribution de plusieurs termes induits par une occurrence unique. On notera au passage qu'une telle correction de « poids » doit être réalisée en tenant compte de la mesure de similarité utilisée entre documents et requêtes.[31] Cette absence de correction de poids est certainement une cause de la faiblesse observée des critères conservant l'ensemble des sens d'un mot, les documents et requêtes partageant des mots fortement polysémiques (donc ambigu) étant rapprochés de manière exagérée par la mesure de similarité.
- **Absence de désambigüisation sémantique** : la sélection des « concepts » associés à chaque occurrence d'un terme devrait prendre en compte le contexte de l'occurrence du terme (désambiguisation sémantique) ; idéalement, il convient de pondérer le poids accordé à ces concepts en fonction de leur vraissemblance en contexte, tel que proposé en section 3.5.4.1 [page 44]. En réalisant une telle pondération, non seulement lors de la sélection du jeu d'index, mais également lors de l'indexation elle-même, la pondération des polysèmes induits par une même occurence (mentionnée au point précédent) n'est plus nécessaire.
- Dans les expériences menées au cours de ce travail, seule la relation d'hypo/hyperonymie a été utilisée ; il serait probablement préférable de tenir également compte **d'autres relations sémantiques** (antonymie, méronymie, etc), en particulier lors de la propagation des « poids » des occurrences, tout en conservant l'hypo/hyperonymie comme relation structurante. Avec le recul, l'intérêt de pouvoir mettre ainsi en relation des éléments liés sémantiquement par d'autre type de relations que la seule hypo/hyperonymie semble fort en pratique (en guise d'exemple, considérons les mots « savon », « douche » et « salle de bain », sans aucun lien d'hypo/hyperonymie entre eux).

[31] Par exemple, en supposant que l'occurrence d'un mot donné induise l'occurrence de k termes d'indexation, il faudrais, avec la similarité cosinus, donner à chacun des k termes le poids $1/\sqrt{k}$ (et non $1/k$ comme on serait tenté de le croire de prime abord) pour contrebalancer cette expansion. En effet, sans pondération, la contribution excessive de ces k termes se retrouve comme suit dans la mesure de similarité entre un document d et une requête r :

$$\frac{\ldots + k \cdot d_i \cdot r_i}{\sqrt{(\ldots + k \cdot {d_i}^2)\,(\ldots + k \cdot {r_i}^2)}}$$

En normalisant d_i et r_i par $1/\sqrt{k}$, on annule l'effet multiplicateur :

$$\frac{\ldots + \cancel{k} \cdot \frac{d_i \cdot r_i}{\cancel{\sqrt{k}} \cdot \cancel{\sqrt{k}}}}{\sqrt{(\ldots + \cancel{k} \cdots \frac{{d_i}^2}{\cancel{\sqrt{k}}\cancel{\sqrt{k}}})\,(\ldots + \cancel{k} \cdots \frac{{r_i}^2}{\cancel{\sqrt{k}}\cancel{\sqrt{k}}})}} = \frac{\ldots + d_i \cdot r_i}{\sqrt{(\ldots + {d_i}^2)\,(\ldots + {r_i}^2)}}$$

Chapitre 5

Amélioration de l'interaction avec l'utilisateur

Résumé

Lorsque l'on envisage dans sa globalité la problématique d'une tâche de recherche documentaire automatisée, il devient évident que le déroulement de la recherche elle-même est au moins aussi importante (pour son succès) que la manière dont le système associe des données aux désirs supposés de l'utilisateur (notamment par l'indexation et les mesures de similarité entre documents). Il est important d'une part que l'utilisateur ait à sa disposition des moyens efficaces pour exprimer et préciser son désir d'information – cela est particulièrement vrai pour les utilisateurs occasionnels ; et d'autre part, le système doit être en mesure de renseigner au mieux l'utilisateur, c'est-à-dire lui retourner l'information désirée, sous une forme intelligible et adaptée à la nature même de cette information.

En d'autres termes, il est également important de porter attention aux *modalités de l'interaction* entre un utilisateur (supposé humain) et un système de recherche d'information.

Dans cette optique, nous présentons dans ce chapitre notre contribution à deux projets visant chacun une réalisation ciblée portant sur l'interaction entre l'utilisateur et la machine. Le premier de ces projets porte sur *l'interaction vocale, en language naturel*, entre un utilisateur et un système de recherche d'information, alors que le second est relatif à la *visualisation de données textuelles* dans un but d'extraction d'information. En d'autres termes, ce chapitre aborde la problématique de l'accès à l'information dans une perspective centrée non plus sur le système de recherche, mais sur l'utilisateur.

Bien que la généralisation de portails d'accès multicanaux à des bases documentaires soit encore en phase de gestation, cette approche a un avenir certain (tant dans une optique de marché – conception d'un unique système d'accès aux données – que dans celle d'améliorer l'accessibilité à ces bases d'informations pour les personnes malvoyantes) ; les travaux présentés ici montrent par ailleurs qu'il est possible de pallier en grande partie les faiblesses des reconnaisseurs de parole actuels, en mettant en œuvre des stratégies adaptées dans le cadre de la gestion du dialogue (les aspects de prototypage rapide et d'évaluation de tels systèmes ne sont abordés que superficiellement ici, mais des références sont données). Au vu des résultats obtenus, on constate en outre qu'il est en grande partie possible de masquer à l'utilisateur le processus de recherche proprement dit, en particulier en se focalisant sur la finalité de cette recherche ; l'intéret est évident pour toute une série de tâches du quotidien.

En matière de visualisation de grandes collections de données textuelles, nous présentons une application de l'analyse des correspondances (permettant de mettre en évidence des similitudes ou des oppositions entre différents groupes, construits sur la base des traits additionnels) au cas de données issues de bases de brevets, permettant de déterminer, pour divers groupes (pays, sociétés, etc), les éléments spécifiques communs à certains de ces groupes (similitudes), ou au contraire les opposant (différences). Nous proposons par ailleurs une méthode basée sur le principe de réplication *bootsrap* permettant de déterminer un intervalle de confiance pour les positionnements relatifs des différents groupes, de manière à juger immédiatement de la fiabilité des similitudes ou oppositions visuellement apparentes. Ces outils sont utilisables dans le cadre d'une méthodologie d'analyse de bases de brevets, permettant de réaliser des comparaisons multi-critères de l'activité « d'innovation » de différents pays, de différents secteurs d'activité ou encore de grandes compagnies ; ils présentent également un intérêt pour l'identification de concurrents dans un secteur donné, ou l'étude des interactions pouvant exister entre différents domaines d'activité technologique ou différents pôles d'innovation à l'intérieur de ces domaines.

5.1 Interactions en langage naturel : gestion de dialogue en interaction vocale

En matière d'interaction homme-machine, le langage naturel représente un élément d'importance croissante ; en raison de l'immixtion de plus en plus forte de systèmes informatiques complexes dans la vie d'individus non nécessairement au fait des technologies mises en œuvre, la possiblité offerte à l'utilisateur d'utiliser son propore langage pour interagir avec une machine offre de nombreux avantages. Cette méthode d'interaction permet d'une part une communication qui ne requiert pas de la part de l'utilisateur (de plus en plus souvent occasionnel) d'expertises particulières, celles-ci étant transférées (du moins censées l'être) vers la machine. D'autre part, l'utilisateur peut concentrer ses efforts sur la tâche qu'il souhaite accomplir, la surcharge cognitive induite par la communication entre l'utilisateur et la machine étant considérablement réduite. Dans le cadre de l'accès aux informations d'une base de données, le language naturel, en particulier au moyen d'interactions vocales, permet notamment de libérer l'utilisateur de la connaissance de l'organisation interne des données et, dans une moindre mesure, de la terminologie utilisée pour la description de ces données.

Cependant, pour que des tâches d'une relative complexité puissent être réalisées, il ne suffit pas de pouvoir commander quelques actions au moyen de la voix ; la machine et l'utilisateur doivent être en mesure de produire une série d'actes de nature conversationnel, s'apparentant à un dialogue.

Les obstacles sont hélas nombreux (et de taille) ; en effet, les actes de dialogues pouvant exister entre humains sont difficilement codifiables (ils sont le fruit d'un apprentissage) et les modalités mises en œuvre sont délicates à appréhender au moyen d'une machine (nombreuses références anaphoriques et temporelles, désignation construite en cours de tâche – introduction de nouveaux concepts, nouveaux énoncés, etc.). La reconnaissance même de la parole (reconnaissance vocale) est une technologie encore peu fiable,[1] de même que la synthèse vocale, qui lorsqu'elle n'atteint pas des performances suffisantes (fluidité, vocalisation et intonation notamment) peut se révéler très perturbante pour l'utilisateur humain, avec comme conséquence une forte diminution de la qualité du « dialogue », pouvant même être rédhibitoire vis à vis de la tâche à accomplir.

Nous présentons ici le fruit des travaux menés dans le cadre d'un projet semi-industriel,[2] en nous focalisant sur la problématique de haut niveau[3] que constitue la gestion de l'interaction entre l'homme et la machine, au moyen d'un gestionnaire de dialogue (et qui constitue l'un de nos principaux apports dans le projet sus-nommé).

5.1.1 Présentation du projet

L'objectif principal du projet *InfoVox* était la spécification et validation d'une méthodologie permettant la *conception rapide*[4] de modèles de dialogue qui puissent, pour une tâche donnée, être employés d'une façon convaincante dans un contexte industriel pour la conception de *modèles de dialogue finalisés* (i.e. des modèles ciblés, spécifiques à la tâche considérée), compatible avec la réalisation de portail multi-canaux, vocal et textuel [Van Kommer *et al.*, 2000].

[1] En particulier dans le cas de l'application qui nous intéresse ici, une reconnaissance multi-locuteurs par le biais d'un canal téléphonique (incluant les appareils de type GSM, donc susceptible de véhiculer un signal accompagné d'un important bruit de fond), avec peu de ressources adaptées pour l'entraînement du modèle.

[2] Le projet *InfoVox*, partiellement financé par la Commission Suisse pour la Technologie et l'Innovation (CTI - OFFPT), et réalisé conjointement par l'EPFL, l'IDIAP, ainsi que les sociétés Swisscom, Omedia et VoxCom.

[3] Par rapport à des tâches de nature plus technique, comme la reconnaissance de parole ou la synthèse vocale.

[4] Par *conception rapide*, on entend que, pour une tâche simple, le modèle de dialogue initial doit pouvoir être produit en l'espace de quelques heures.

En d'autres termes, l'objectif n'était pas la production d'un modèle « générique » de gestion de dialogues, mais plutôt permettre la production rapide et systématique de modèles extrêmement spécialisés. Dans cette optique, l'idée sous-jacente de la méthodologie de prototypage obtenue est que le modèle de dialogue visé est, pour l'essentiel, un modèle à états finis, facilement et systématiquement dérivable à partir d'une modélisation de la tâche elle-même.

Pour obtenir cette méthodologie, le projet a consisté en l'élaboration d'un prototype de serveur vocal, pour une tâche alibi – accès téléphonique aux informations d'une base de données sur les restaurants de la ville de Martigny (Suisse), à la manière du projet *Berp* [Jurafsky *et al.*, 1994], mais dans une optique de conseil – suivi de son évaluation au cours d'un *field-test*. Une fois le prototype réalisé et évalué, la méthodologie de développement a pu être formalisée ; elle a ensuite été mise en œuvre (et affinée) dans le cadre de deux autres projets de recherche[5] auxquel nous n'avons pris part que marginalement (nous n'en parlerons donc pas ici).

5.1.2 Description du prototype

Les contraintes imposées au portail multi-accès étaient :

Dialogues en français Destiné à un « marché » francophone, le système doit naturellement être capable d'interagir dans cette langue. Cette contrainte reste assez forte par rapport à des systèmes destiné à l'anglais – en particulier dans le cadre de projet de faible envergure[6] – vu la grande disparités des ressources disponibles (modèles acoustiques et de langage, thésaurus, etc.)[7] Remarquons que la problématique multilingue n'a pas été abordée dans ce projet.

Interaction intuitive, adaptée à l'expertise de l'utilisateur Un utilisateur novice doit être à même de pouvoir utiliser le système, en énonçant ses désirs d'une manière naturelle ; l'utilisateur expérimenté doit lui pouvoir énoncer ses désirs d'une manière directe. Ceci implique que le système fonctionne en langage non contraint, et que l'acquisition d'informations visant à la réalisation de la tâche doit pouvoir se faire par de multiples biais (l'utilisateur novice doit être guidé par le système tandis que l'utilisateur expérimenté doit pouvoir orienter le dialogue comme il l'entend – l'initiative mixte (ci-après) est un excellent moyen pour cela).

Initiative mixte (limitée) Le « contrôle du dialogue » doit pouvoir être pris librement par l'utilisateur ou par le système ; à l'inverse de nombreux systèmes comparables [par exemple Albesano *et al.*, 1997, Allen *et al.*, 2001], cette initiative mixte ne doit pas se limiter à des demandes de clarifications ou de corrections, mais doit également porter sur la manière dont la tâche est réalisée (tout en restant dans le domaine ciblé par l'application – qui peut cependant recouvrir différents contextes, comme c'est le cas dans les projets Inspire et MDM, et dans le cadre des systèmes transactionnels).

---

[5] Le projet européen Inspire [Möller et Skowronek, 2003], dont le but était le contrôle vocal de différents appareils ménagers (lumières, TV, Vidéo, ...) par le biais de dialogues spécifiques dans le cadre du « Smart Home Environment », et le projet MDM (*Multimodal Dialogue Management*) intégré au sein du programme suisse de recherche IM2 (*Intelligent Management of Multimodal Information*, Armstrong *et al.* [2003]) dont le but est la création des mécanismes d'interaction (basés sur le dialogue) avec une base de données contenant des transcriptions multimodales de réunions.

[6] I.e. des projets insuffisamment dotés (temps et force de travail) pour autoriser la création de ressources spécifiques.

[7] Et pourtant, le français est loin de faire partie de la catégorie des langues « peu dotées », pour lesquelles les ressources électroniques sont pour ainsi dire inexistantes.

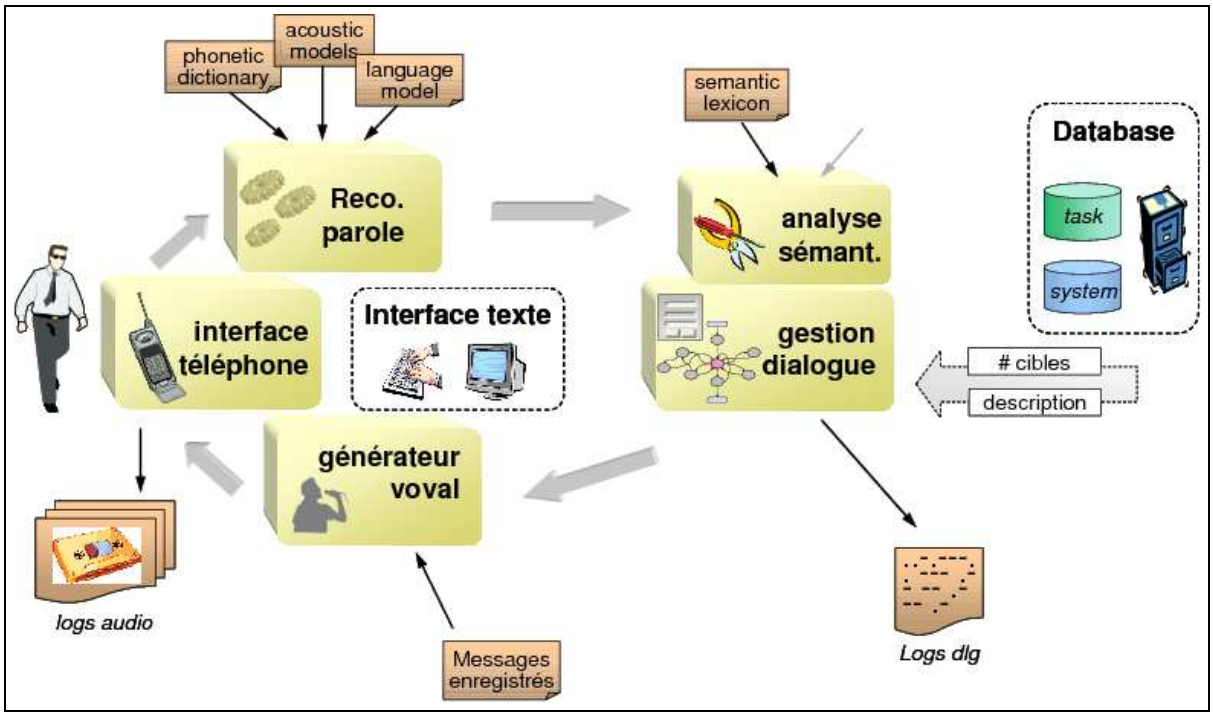


Figure 5.1*:* Architecture générale du système *InfoVox* ; l'accès textuel (par exemple via une page web) court-circuite simplement l'interface téléphonique, le reconnaisseur de parole et la génération de signal.

Robustesse Le gestionnaire de dialogue doit être à même de fonctionner en dépit d'erreurs de reconnaissance (point particulièrement crucial dans le prototype du projet *InfoVox*, au vu de la faible qualité du reconnaisseur de parole) et s'assurer de plus de la compréhension mutuelle des deux interlocuteurs au cours du dialogue. Les problèmes d'incompréhension notamment doivent être traités (résolu) dans le cadre du dialogue autant que possible, et ce d'une manière la plus naturelle possible (le recours au DTMF[8] est par exemple exclu).

Le prototype a été élaboré sur la base des observations faites et en utilisant les ressources constituées lors d'une phase d'initiation du système, l'expérience *Wizard of Oz* (WoZ, cf. 5.1.3 [page 81]). Au cours de cette phase, un certain nombre de dialogues tests ont eu lieu entre des utilisateurs cobayes et un système minimaliste, piloté par un opérateur humain remplissant les tâches de chacun des composants (encore inexistant à ce stade) du système. Après quoi, l'opérateur à progressivement été remplacé, jusqu'à obtenir un prototype complet, dont l'architecture générale est schématisée dans la figure 5.1 ; on y trouve les composants suivants :

Interface téléphonique Il s'agit là d'un élement essentiellement technique, permettant d'interfacer une carte téléphonique (pouvant supporter un certain nombre de lignes, ie d'appels simultanés – typiquement 32) au reste du système : détection des appels et des terminaisons intempestives, acquisition et restitution de signal audio (détection de signal en entrée, restitution du signal synthétisé en sortie), etc. La possibilité de signaler à l'utilisateur que le système travaille (« sablier » sonore) peut se révéler intéressante, si le délai de traitement est supérieur à la seconde.[9]

[8] *Dual-tone multi-frequency* : la combinaison de fréquences tonales utilisée pour la téléphonie moderne (la technologie qui a remplacé la sélection par impulsions). Ces codes ont permis la création des premiers services vocaux.

[9] Les évaluations du système ont en effet montré qu'il y avait une plus grande acceptation du délai de réponse en présence d'un tel dispositif.

De multiples traitements du signal audio auraient idéalement dû être réalisés au niveau de cette interface, notamment la suppression des bruits ambiants (important pour l'analyseur d'énergie chargé de détecter les débuts et fin de signaux d'entrée) et la correction des distorsions dues à la ligne téléphonique (important pour le reconnaisseur vocal). Par ailleurs, pour permettre une fluidité accrue dans le dialogue, il est usuel que la détection de signal entrant prenne le pas sur la génération du signal sortant (*barge-in*, permettant à l'utilisateur de « couper la parole » au système) ; ce n'était pas le cas dans le système *InfoVox*, en raison de limitations au niveau de la carte téléphonique (*half-duplex*) ; les tours de parole étant marqués au moyen d'un indicateur sonore.

Reconnaisseur de parole Le rôle du reconnaisseur de parole est (bien entendu) d'identifier, dans le signal audio d'entrée, les « mots » prononcés par l'interlocuteur du système ; plus précisément, le système identifie en premier lieu une suite de phonèmes, et les assemble pour produire la séquence (textuelle) de sortie. Idéalement, le reconnaisseur de parole fournit de plus une estimation de la fiabilité de la reconnaissance de chaque terme ou séquence ; cette information pouvant être utilisée pour décider s'il y a lieu ou non de demander à l'utilisateur une confirmation explicite de l'une ou l'autre des informations recueillies.

Dans le cas du projet *InfoVox*, le reconnaisseur mis au point est un système multi-locuteurs, en parole continue. Il a été mis au point en adoptant une approche dite « hybride », pour laquelle l'estimation de la probabilité des ditributions de phonèmes est réalisée par un réseau de neurones, tandis que le décodage des phonèmes est effectué au moyen d'un modèle de Markov caché [Bourlard et Morgan, 1995, Morgan et Bourlard, 1995]. En dépit du fait que cette approche soit généralement considérée comme robuste et performante, l'évaluation du reconnaisseur a révélé que, dans le cas d'*InfoVox*du moins, la qualité de la reconnaissance était mauvaise[10] ; le modèle de dialogue (et son gestionnaire) a donc dû être adapté en conséquence, de manière à pallier autant que possible les faiblesses de la reconnaissance vocale.

Analyseur sémantique Le rôle de l'analyseur sémantique est d'obtenir la représentation contrôlée des éléments potentiellement utiles au dialogue trouvé dans les énoncés fournis par le reconnaisseur de parole. Il s'agit donc d'identifier ces éléments, d'en uniformiser la représentation, et d'éventuellement procéder à l'expansion des éléments porteurs de plusieurs informations dans la représentation contrôlée (par exemple, à une question du type « Quelle jour souhaitez-vous aller manger ? » l'énoncé-réponse « dans vingt minutes » donne à la fois une information temporelle – le jour choisi pour le repas – et une information sur la nature de ce repas – de midi ou du soir, typiquement – toutes deux devant de plus être interprétées, car données relativement à l'instant auquel le dialogue a lieu). La représentation contrôlée correspond pour l'essentiel au lexique utilisé pour étiqueter les informations de la base de données, augmenté d'un certain nombre de valeurs logiques (oui,non, etc.) et d'actions. En examinant les résultats issus du *Wizard of Oz* (5.1.3 [page ci-contre]), il est apparu qu'une discrimination suffisante du vocabulaire relatif aux différentes parties du dialogue existait, et qu'une liste de séquences clefs pré-établies (mots et expressions à trous) était suffisante pour cette tâche (les quelques ambigüités subsistants dans la représentation contrôlée pouvant être levée grâce à la connaissance du contexte dans lequel l'énoncé est produit, et déterminé par les questions fermées posées par le système).

[10] Deux raisons expliquent ce manque de performances : (1) le modèle acoustique utilisé pour entrainer le réseau de neurones n'était pas adapté à l'environnement d'utilisation du système (notons cependant que des évaluations faites avec le modèle acoustique lui-même ont montré de faibles performances d'estimation – l'entraînement du système est donc également à mettre en cause) ; (2) le modèle de langage (qui permet de passer d'une suite de phonèmes aux mots) a été constitué sur la base de données récoltées lors du *Wizard of Oz* – voir ci-après – (15′000 mots couvrant un lexique d'environ 1′000 termes), données clairement insuffisantes pour permettre d'en dériver un modèle de langage fiable.

Gestionnaire de dialogue Le gestionnaire de dialogue est l'élément central du système ; les tâches qui lui incombent sont [Gaussier et Stéfanini, 2003, pg. 223] :

- la construction d'un univers sémiotique partagé et l'échange des connaissances ;
- l'organisation du dialogue (gestion des tours de parole, des échanges, des interventions) ;
- le choix des stratégies de dialogue ;
- la réparation des erreurs de communication et son maintien ;
- l'aide dans la tâche et dans la conduite des activités.

Relevons que dans notre cas, l'univers sémiotique est en grande partie construit par l'analyseur sémantique (et naturellement l'étiquetage en vocabulaire contrôlé des données de la base). L'architecture et le fonctionnement du gestionnaire de dialogue d'*InfoVox* sont décrit plus en détails dans la suite de ce chapitre.

Génération de signal Le module de génération de signal est reponsable de traduire en signal audio les actes de dialogues produits par le système. Il s'agit habituellement d'une synthèse vocale, qui offre ainsi toute latitude pour produire dynamiquement des actes de dialogues ; dans le cas du système *InfoVox*cependant, la génération de signal est obtenue au moyen d'un ensemble de séquences audio pré-enregistrés, certains représentant des messages entiers, tandis que d'autres ne sont que des segments, dynamiquement assemblés pour obtenir un message complet. La problématique du coût (que ce soit en temps de développement ou le coût financier représenté par l'acquisition d'un système commercical) n'est pas l'unique motivation de ce choix ; il a en effet été établi que la qualité de la synthèse vocale est un élément crucial, qui peut à lui seul changer la perception que l'utilisateur à du système (synthétique, voire mécanique, ou au contraire « naturel »). Compte tenu de la faible diversité et complexité des actes de dialogues émis par le système, la stratégie des messages préenregistrés a été choisie, offrant l'avantage d'humaniser fortement le système, en gommant cet aspect mécanique.[11]

Le système complet joue donc le role d'un opérateur effectuant une navigation dans une base de données, sans que l'utilisateur n'en soit conscient ; plus précisément, le système restreint peu à peu les entrées de la base de données en utilisant les informations fournies par l'utilisateur, et ce jusqu'à isoler un ensemble suffisemment petit pour qu'il puisse être soumis à l'utilisateur.

5.1.3 Initiation du modèle – WoZ

La création d'un modèle initial de dialogue cohérent et adapté à la tâche à réaliser est une étape importante pour l'obtention d'un système convaincant ; pour valider la modélisation initiale, la tenue d'une expérience *Wizard of Oz* (WoZ) est un moyen idéal, permettant à la fois d'affiner le modèle et d'acquérir de précieuses données pour l'élaboration ultérieure d'un processus automatisé.

Une expérience WoZ [Fraser et Gilbert, 1991] peut être définie comme la simulation d'une interaction humain-machine, au cours de laquelle un utilisateur est confronté à un système qu'il pense totalement automatique, alors qu'en fait un opérateur humain caché (le magicien) réalise tout ou partie des tâches non encore implémentées. Les objectifs usuels d'une telle expérience sont l'évaluation du comportement des utilisateurs (les chemins utilisés pour atteindre le but fixé), voire leur degré d'acceptation du système, ainsi que l'évaluation de l'ergonomie du système (c'est également un moyen idéal pour présenter à de potentiels clients ou investisseurs la technologie en cours de développement). Outre l'aspect de validation, une expérience de ce type

[11] Naturellement, il faut pour cela apporter un soin particulier aux différences d'intonation et d'intensité sonore, notamment au niveau des raccords.

constitue une opportunité pour acquérir des données expérimentales « réalistes », notamment concernant le comportement effectif des futurs utilisateurs ; il est nécessaire pour cela d'enregistrer soigneusement les interactions pendant l'expérience, et d'analyser les données recueillies pour améliorer la modélisation du système [Boyce et Gorin, 1996, Daly-Jones *et al.*, 1999].

Dans le cadre du projet *InfoVox*, l'expérience WoZ a duré 38 jours en tout (une des difficultés étant naturellement de maintenir un niveau de « fonctionnement » le plus uniforme possible, indépendamment de la fatigue de l'opérateur). Une centaine de personnes ont participé à l'expérience, permettant d'enregistrer plus de 250 dialogues [Rajman *et al.*, 2004, 2003].

5.1.4 Gestionnaire de dialogue

Le gestionnaire de dialogue que nous avons conçu est bâti sur une architecture mélangeant les techniques *finite state script*, *frame-based* et *sets of contexts* [Allen *et al.*, 2001].

Concrètement, le gestionnaire de dialogue est une machine à état fini, formé de noeuds d'actions purs et de noeuds génériques de dialogue (GDN – voir 5.1.4.3), le rôle de ces derniers étant de remplir les champs d'un questionnaire. Chaque questionnaire correspond à une partie « atomique » de la tâche (*i.e.* qui ne peut être réalisée de manière entrelaçée avec d'autres ; par exemple, il faut nécessairement avoir proposé un premier restaurant à l'utilisateur avant de lui demander s'il souhaite s'en voir proposer un autre).

Plus précisément, la tâche est modélisée sous la forme d'un ensemble de questionnaires dont les champs, associés chacun à un contexte, représentent les différents attributs devant être informés pour que la tâche puisse être réalisée. Par exemple, pour l'application de recherche de restaurants, les champs suivants ont été utilisés pour modéliser la sous-tâche de recherche : type de cuisine, tranche de prix, localisation, jours et heures d'ouverture.

Les autres sous-tâches sont essentiellement associées à des questions de type 'oui/non' (correspondant donc à un formulaire trivial), et permettent de fournir à l'utilisateur l'information désirée, dans une seconde partie de dialogue (cf. figure 5.2 [page suivante]).

Concrètement, la machine d'états est utilisée pour séquencer le dialogue à la fois globalement (les différents sous-dialogues) et de manière très précise (noeud générique de dialogure) ; entre ces deux extrêmes, c'est le formulaire des sous-tâche qui permet de structurer le dialogue, et notamment les questions adressées à l'utilisateur.

Un avantage de cette approche est que l'interface pour l'accès web peut être dérivée directement du modèle de dialogue ; en plus des formulaires, une ligne de saisie peut être utilisée pour l'entrée de texte libre,[12] directement connectée à l'entrée du module d'extraction de contrainte.[13]

Principes directeurs : Pour atteindre les objectifs imposés au système, un certain nombre de principes, décidés à l'issue de l'examen des résultats du WoZ, ont été utilisés pour guider les choix de développement :

1. permettre l'initiative mixte (limitée) ;
2. éviter les répétitions et les lourdeurs ;

[12] Il faut naturellement prévoir de plus une zone d'affichage des messages, et le changement de focus pour les différents formulaires.

[13] De fait, la partie téléphonique ainsi que les reconaissance et synthèse vocale sont simplement court-circuitées pour l'accès web, fonctionnant en parallèle de ces trois modules, avec la même application de gestion du dialogue que pour une interaction téléphonique.

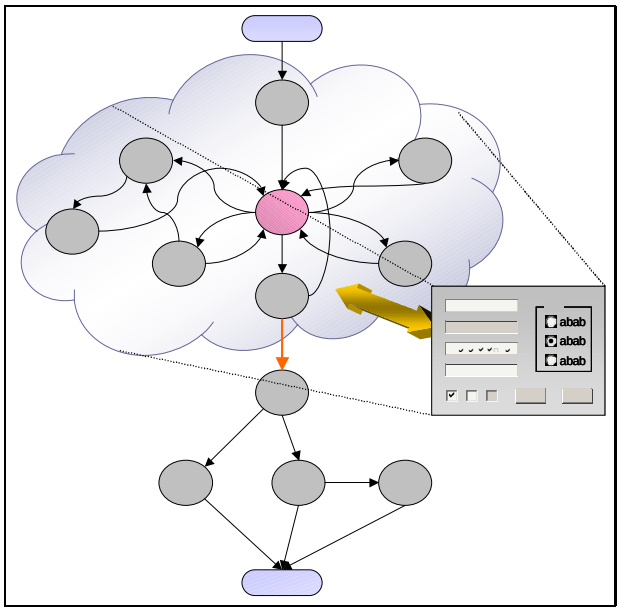


Figure 5.2*:* Vue simplifiée de la machine d'état (contrôle haut et bas niveau) ; à un premier groupe d'états (centrés autours d'un nœud générique de dialogue) correspond un formulaire qui contrôle le fonctionnement du nœud de dialogue ; les états suivants permettent d'informer l'utilisateur du résultat de la recherche ; composés de questions « oui/non », ils ne sont qu'implicitement associés à un formulaire (trivial).

3. traiter aussi naturellement que possible les « réparations » de dialogue, tant en ce qui concerne les situations d'incompréhension[14] que les ambigüités ou incohérences dans les réponses données par l'utilisateur (par exemple dues à un changement d'avis en cours de dialogue) ;
4. tendre à minimiser la durée des dialogues ;
5. informer l'utilisateur sur l'état du système (feedback) ;

La manière dont ces différents principes ont été traduit dans la réalisation du prototype est décrite dans les sections suivantes.

5.1.4.1 Interaction à initiative mixte (limitée)

Le système tel qu'il a été conçu ne permet de fait qu'une *initiative mixte limitée* [Allen *et al.*, 2001] consistant pour l'utilisateur en la possibilité d'une part de rompre le flux du dialogue imposé par le système, en demandant la répétition ou une explication de la dernière question posée (cas assimilés aux *dialog repairs*), et d'autre part à anticiper les futures questions en fournissant par avance des éléments de réponses.[15]

Remarquons cependant que l'utilisateur peut réellement choisir de ne pas répondre à une question qui lui est posée : d'une part, il se peut que des informations supplémentaires fournies par anticipation permettent d'accomplir la tâche en cours, et donc de progresser dans le graphe de la machine d'états ; d'autre part, l'attribution d'un champ à un noeud de dialogue étant

[14] *Dialog repairs* (par analogie avec les *speech repairs*) désignant ici les actes de dialogue portant sur le dialogue lui-même et visant à gérer des situations « d'erreurs » de communication (« *breakdown and repairs sequences on the dialogue* », ou « *effective grounding* » Allen *et al.* [2000], Hirst *et al.* [1994])

[15] C'est typiquement le cas avec les questions (semi-)ouvertes d'accueils, initiant le dialogue en laissant une large lattitude à l'utilisateur pour exprimer ses souhaits.

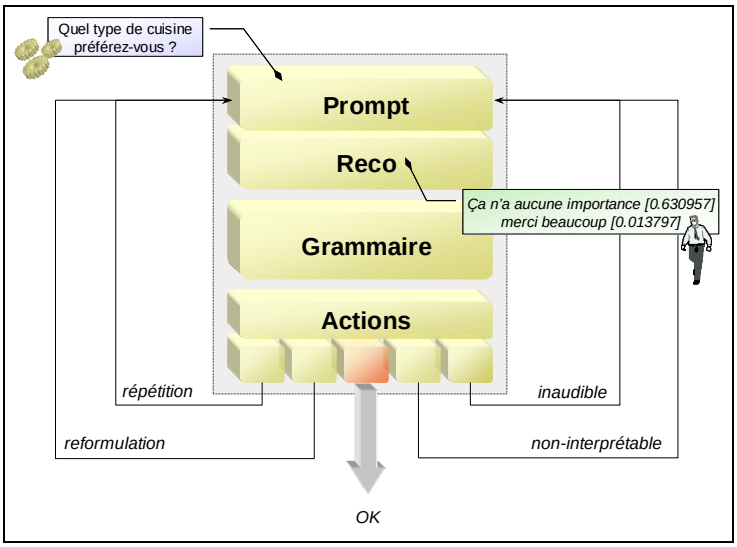


Figure 5.3*:* Nœud générique de dialogue ; le *prompt* détermine (en fonction de l'état du formulaire) la question à adresser à l'utilisateur, et ses éventuelles reformulations ; le résultat de la reconnaissance (avec la mesure de confiance) est ensuite interprêté (deux types de « grammaires » étant utilisées, l'une spécifique à la question posée, et l'autre généraliste – toutes deux en relation avec le modèle de langage du reconnaisseur vocale), et une série d'actions peuvent en découler ; s'il n'y a pas de réparation de dialogue, on progresse dans le graphe de la machine d'état.

réalisée dynamiquement, lors de chaque visite du noeud, un retours sur ce noeud ne conduira pas nécessairement le système à reposer la même question.

L'utilisateur novice sera donc, comme dans le cas d'un dialogue entre humain et pour une tâche similaire, guidé par le système, et se contentera de répondre à ses sollicitations, tandis que l'utilisateur expérimenté pourra fournir d'emblée l'ensemble des éléments qu'il sait pertinents pour le système ([Larsen, 1997], à la différence qu'il n'est ici pas nécessaire de prévoir et définir explicitement les raccourcis dans la structure du dialogue, comme c'est le cas avec le système de Larsen).

5.1.4.2 Éviter les répétitions d'énoncés

Pour éviter que l'aspect « mécanique » et limité du système ne transparaisse par trop, il a semblé important d'éviter au maximum les répétitions de messages. Pour cela, différentes formulations alternatives, plus ou moins équivalentes, ont été établies pour chaque message ; par ailleurs, un petit nombre d'entre eux ont été contextualisés (par exemple, les messages d'accueil et de prise de congé, suivant l'heure et le jour du dialogue : « bonjour/bonsoir », « excellente journée/après-midi/soirée/week-end »). Lorsqu'un message doit être joué plusieurs fois au cours d'un dialogue, le système optera pour une formulation différente de celles des précédentes occurences du message. Cette technique présente de plus l'avantage de pouvoir être mise en oeuvre comme mécanisme de désambigüisation, lorsque l'utilisateur signale qu'il ne comprend pas un message.

5.1.4.3 Réparation et nœud générique de dialogue

Pour une réponse de l'utilisateur à une question du système, on distinguera les cas de figures suivants, sur la base des valeurs et contextes présents dans la réponse, ainsi que le contexte dans lequel cette réponse a été donnée :

Ok et Initiative :

L'interlocuteur répond à la question (on trouve dans sa réponse des éléments d'informations –valeurs– correspondant au contexte de la question), ou du moins répond quelque chose que le système est à même d'interpréter, éventuellement de manière partielle, comme la réponse à une autre question, qu'elle ait été posée ou non (initiative de l'utilisateur).

Dans ce cas, il n'y a pas de réparation de dialogue et le traitement se poursuit par la mise à jour des champs du formulaire.

Répétition :

L'utilisateur demande explicitement au système de répéter son dernier message.

La dernière sollicitation du système est alors répétée. La répétition est d'abord effectuée sans reformulation, celle-ci intervenant dans le cas ou l'utilisateur demande consécutivement plusieurs reformulation (3 dans notre cas).

Incompréhension utilisateur :

L'utilisateur signale, de manière plus ou moins explicite,[16] *qu'il n'a pas compris la sollicitation du système.*

Une formulation alternative de la dernière sollicitation est utilisée dans un premier temps ; en cas de nouvelle incompréhension manifestée par l'utilisateur, le système bascule sur une *demande d'assistance*. Dans le cas où le message concerné est une question ouverte et que l'incompréhension perdure, le système bascule vers un dialogue guidé, en demandant à l'utilisateur d'informer un et un seul des champs du formulaire.

Demande d'assistance :

L'utilisateur ne voit pas comment répondre à la question.

Le système indique à l'utilisateur comment répondre, en fournissant à titre d'exemple quelques valeurs admissibles pour le champs concerné (« Vous pouvez choisir par exemple : cuisine chinoise, française, pizzeria, brasserie »), ou en fermant (partiellement du moins) la question, s'il s'agissait d'une question ouverte (« Vous pouvez indiquer le type de cuisine qui vous plairait ou quel jour vous souhaitez aller manger »).

Échéance de temporisateur :

L'utilisateur ne dit rien pendant plusieurs secondes, et l'enregistrement (interface téléphonique) cesse sur un time-out.

Le système demande à l'utilisateur de parler un peu plus fort, et d'attendre la sollicitation sonore avant de répondre ; il bascule ensuite sur la *répétition*.

Incompréhension système ou hors contexte :

Aucune information utilisable ne peut être extraite de la réponse de l'utilisateur (problème de reconnaissance vocale, réponse trop complexe, onomatopées, etc.), ou lorsque l'utilisateur répond quelque chose sans relation avec la tâche[17]*.*

Le système indique à l'utilisateur qu'il ne l'a pas compris, et demande la répétition de la réponse. Si la situation se reproduit une seconde fois, la *demande d'assistance* est invoquée ; comme dans le cas de l'*incompréhension utilisateur*, le système peut basculer sur un dialogue guidé, voir éventuellement invoquer un mécanisme exceptionnel de sortie de boucle (déterminé par un chien de garde général –voir ci-après– mesurant la progression du dialogue).

[16] Un certain nombre d'onomatopées peuvent effectivement révéler une incompréhension côté utilisateur.

[17] Identifier ce cas nécessite de disposer d'une grammaire généraliste à large couverture, ce qui n'est pas toujours bénéfique (complexe à construire et gérer, augmente la perplexité du reconnaisseur de parole, etc.)

5.1.4.4 Minimiser la durée des dialogues

La minimisation de la durée des dialogues (aussi bien en terme de nombre de tours de parole que de temps total écoulé), sans restriction imposée sur la tâche effectuée est en fait plus un but à atteindre qu'un moyen permettant d'améliorer le dialogue.[18] Bien entendu, beaucoup de facteurs influent sur la durée du dialogue ; les recommandations suivantes sont néanmoins applicables :

- Pour les messages longs susceptibles d'être joués plusieurs fois au cours du dialogue, il peut être avantageux de disposer, en plus des formulations alternatives (généralement longues elles-aussi), de reformulations plus concises (elliptiques). Les multiples formulations longues sont utilisées si nécessaire pour expliciter le message, et les reformulations résumées lorsque, plus tard au cours du dialogue, le même message doit être rejoué.
- L'initiative mixte, permettant à l'utilisateur de préciser des informations d'emblée en début de dialogue, est également un moyen efficace.
- D'une manière générale, il faut éviter les cycles dans le dialogue. Ceci n'est pas nécessairement trivial à réaliser, certains cycles étant nécessaires (par exemple lorsque l'utilisateur demande explicitement au système de répéter ce qu'il vient de dire, et ce plusieurs fois).
  La solution retenue dans le cadre du projet *InfoVox* a été de mettre en place des chiens de garde à différents niveaux :
  - en cas d'incompréhensions répétées (utilisateurs ou système) sur une question ouverte, basculement vers un dialogue guidé ;
  - dans le même cas, mais avec une question du dialogue guidé, pour un formulaire admettant plusieurs champs non encore tous renseignés, abandon du champ au profit d'un autre non-encore examiné ;
  - toujours en cas de cycle sur les incompréhensions, mais avec des questions oui/non : sélection, par le système, du choix induisant le moins de conséquences (valeur "par défaut"), et information de l'utilisateur, par un message expliquant les raisons d'une telle décision (par exemple, pas de prise de réservation, si le système n'arrive pas, plusieurs fois de suite, à déterminer la prise de réservation est voulue ou non par l'utilisateur) ;
  - chien de garde global, déclenché en cas de non progression suffisante du dialogue (lorsque aucun champ n'est modifié ou renseigné au cours des n dernières interactions, répétitions strictes exclues) ; dans ce cas, suivant l'application, différentes actions peuvent être envisagées : soit on propose à l'utilisateur de lui fournir les informations déjà disponibles, soit on lui propose d'avorter le dialogue, en le mettant éventuellement en contact avec un opérateur humain, etc.[19]

En plus de ces éléments, on peut de plus mettre à profit la nature de la tâche à réaliser, soit pour choisir « intelligemment » les questions à poser (l'ordre des champs à informer en priorité), soit, en cas de problème, pour terminer de manière anticipée la tâche, en fournissant tout de même une information (ou un service) partielle :

Service de renseignements *Recherche d'information sur un ou plusieurs éléments a priori présents dans la base de données.*

Pour ce type de tâche, on peut choisir, lors du dialogue guidé, d'informer en priorité les champs ayant le plus fort potentiel discriminant, minimisant ainsi, en moyenne, la durée de chaque dialogue (un algorithme basé sur ID3 est proposé ci-après – 5.1.5 [page 89] – consulter également Quinlan [1986, 1993]).

[18] Pour un même résultat, un dialogue plus court est nécessairement plus « efficace » ; on limite le risque de susciter l'ennui chez l'utilisateur, ainsi que d'éventuelles situations d'incompréhension.

[19] Évidemment, les choses se gâtent lorsque des incompréhensions surviennent également à cet instant. Une bonne précaution dans ce cas est de disposer d'un dialogue de secours minimaliste par DTMF (utilisant les touches du téléphone).

Un avantage supplémentaire de cette technique est qu'elle offre la possibilité de pondérer les éléments de la base en fonction de leur popularité, permettant de minimiser un peu plus encore la longueur moyenne des dialogues.[20]

Service de conseils *Attente de propositions (« conseil ») ; on peut toujours le considérer comme une recherche d'information, mais sur des éléments qui ne se trouvent pas nécessairement dans la base de données.*

C'est l'orientation qui a été retenue pour le projet *InfoVox*(le système effectue des propositions de restaurants, sur la base de préférences indiquées par l'utilisateur). Dans ce cas, le calcul du potentiel discriminant d'une question n'a plus vraiment de sens (il ne s'agit pas nécessairement d'isoler un groupe d'éléments de la base de données, du moins dans un premier temps), mais il devient possible de soumettre à l'utilisateur des éléments ne satisfaisant pas l'ensemble des critères énoncés par celui-ci.

Cette technique est relativement simple à mettre en œuvre : l'utilisateur indique ses choix au cours du dialogue avec le système. Lorsque plus aucun élément de la base ne satisfait les choix énoncés, plutôt que d'abandonner le dialogue, de proposer à l'utilisateur de recommencer ou de modifier ses choix, on peut lui demander s'il accepte des propositions ne satisfaisant qu'une partie de ses préférences. Il suffit alors de relaxer les dernières contraintes spécifiées,[21] éventuellement de manière successive (attention dans ce cas à ne pas soumettre plusieurs fois les mêmes propositions).

Avant de proposer à l'utilisateur de recourir à ces « solutions approchées », il faut cependant prendre garde à ce que la relaxation conduise à des propositions qui conservent un sens ; si elle est mise en œuvre alors que l'utilisateur n'avait précisé qu'un petit nombre de contraintes, la relaxation d'une seule peut suffir à remettre en lice un grand nombre d'éléments de peu d'intérêt pour l'utilisateur. Dans le cadre du prototype *InfoVox*, avant de proposer des tels éléments à l'utilisateur, on s'assure lors de chaque relaxation que le nombre de cibles réintroduites reste raisonnable en regard du nombre de cibles encore exclues. Dans le cas où cette proportion devient trop grande, les mécanismes énoncés précédemment (demander à l'utilisateur s'il souhaite modifier sa requête, en soumettre une autre ou abandonner) sont appliqués.

5.1.4.5 Informer l'utilisateur sur l'état du système (*feedback*)

Dans le cas de dialogue entre humains, la progression du dialogue est fluidifiée par des actes d'acquittements des interactions de chaque participant. Sur la base d'un corpus de conversations spontanées d'environ 200′000 interactions, Stolcke *et al.* [2000] ont comptabilisé qu'environ 20% des actes de dialogues étaient des acquitements (*backchannel/acknowledge*).

De tels acquittements ne sont malheureusement pas réalisables avec un système tel que celui utilisé pour *InfoVox*(absence de *barge-in*, reconnaissance des acquittements très aléatoire, délai de réaction du système, etc.).

Pour éviter que de mauvaises interprétations côté système (rendues nombreuses en raison des erreurs de reconnaissance) ne conduisent à fournir à l'utilisateur des informations sans aucun rapport avec ses désirs, il est nécessaire de détecter (et corriger) ces situations ; la seule possibilité pour cela est, comme dans le cas de dialogues entre humains, que le système informe l'utilisateur des éléments qu'il a retenus.

[20] Il faut cependant veiller à éviter les ordonnancements trop « surprenants » (risquant de déstabiliser l'utilisateur) ; par exemple, demander à l'utilisateur la couleur de l'article qu'il recherche, avant de le questionner sur la nature de cet article. Remarquons également que la fonctionnalité d'initiative mixte oblige à réaliser le calcul de la question la plus discriminante de manière dynamique ; cela peut être source de problème (délais trop longs) si la base est grande et que les critères sont nombreux et admettent de nombreuses modalités.

[21] En tablant sur le fait que l'utilisateur précisera en priorité les éléments qui lui semblent les plus importants.

Une première manière pour le système de fournir un tel *feedback* serait qu'avant d'accomplir la finalisation d'une sous-tâche (par exemple en sortie d'un formulaire), le système produise une synthèse des différents éléments retenus (les valeurs des champs nouvellement informés depuis la dernière synthèse), auquel l'utilisateur aurait loisir de réagir.

Cette technique à elle seule n'a toutefois pas été jugée suffisante (la correction de valeurs erronées peut donner lieu à des dialogues fastidieux, les utilisateurs ne sont pas nécessairement attentifs et ne réalisent pas sur le coup que le système a compris autre chose que ce qu'ils avaient demandé, en particulier lorsque le nombre de champs est important, etc.).

Il a donc été décidé de tenter de réaliser une forme d'acquitement au cours du dialogue, en insérant en préambule de chaque question fermée une indication sur **un**[22] des champs précédemment informés.

Un exemple de message composé d'un préfix de confirmation et d'une question visant à informer un champ pourrait être : « Pour votre repas de ce soir, avez-vous une préférence sur la région où vous souhaitez aller manger ? »

En cas de non protestation de la part de l'utilisateur, le système marque le champs comme étant « confirmé » ; il ne pourra dès lors plus être modifié par le mécanisme de l'initiative mixte. Si l'utilisateur proteste, soit par une réponse négative, soit en indiquant une valeur compatible avec le champ confirmé, mais différente de celle déjà présente, le mécanisme de traitement des informations conflictuelles est amorcé. Le système demande alors à l'utilisateur sa préférence pour le champ en question, tout en indiquant les valeurs en conflit.

Cette solution ne permet pas de confirmer l'ensemble des informations avec lesquelles le système va travailler, comme c'est le cas dans Danieli [1997], Danieli *et al.* [1997], en particulier si l'utilisateur anticipe fortement les questions du système, mais se révèle équilibrée ; l'impact négatif sur la qualité du dialogue est limité, et d'ultérieures erreurs d'interprétation du système ne remettrons pas en cause les valeurs déjà acquises et confirmées (une modification explicite de ces valeurs reste possible, l'utilisateur devant forcer le système à passer par une question fermée sur le champs concerné).

5.1.4.6 Traitement des informations conflictuelles

Lorsqu'au cours du dialogue ou d'une interaction, l'utilisateur spécifie deux valeurs incompatibles pour un même champ, un sous-dialogue visant à déterminer la valeur désirée pour le champ en question est amorcé (le système demandant alors à l'utilisateur de choisir une valeur pour ce champ, en indiquant de plus les éléments en conflit).

Pour éviter que des erreurs de reconnaissance ne provoquent trop de situations de ce type, les informations identifiées par le système sont fortement filtrées avant la mise à jour des champs du formulaire :

- Si la réponse de l'utilisateur est donné alors qu'un contexte est imposé par le dernier message du système, ou que celui-ci contenait en préambule un feedback de confirmation, seuls les éléments compatibles avec l'un ou l'autre de ces contextes sont conservés. Dans le cas où il

[22] Il y a plusieurs possibilités pour le choix du champ en question : le plus anciennement informé, le plus récemment informé, etc. Le critère finalement retenu pour le projet *InfoVox* est de choisir les champs de la réponse la plus récente, selon l'ordre d'énonciation des valeurs. Naturellement, il faut de plus tenir compte du contexte de la question pour le choix du champ à confirmer, en évitant si possible les couples ayant des modalités partagées (la valeur identifiée dans sa réponse est-elle relative à la confirmation ou constitue-t-elle la réponse à la question posée – le but étant d'éviter autant que possible d'introduire des ambigüités sur l'interpretation de la réponse de l'utilisateur).

n'existerait pas de tels éléments, le filtrage est annulé et la totalité des éléments de la réponse est prise en compte. [23]
- Si l'analyse de la réponse de l'utilisateur indique la présence d'au moins une valeur non conflictuelle, les éléments de la réponse en conflit sont filtrés, à l'exception de ceux dont le contexte est compatible avec l'éventuelle feedback de confirmation.

5.1.5 Champ le plus discriminant

Les différents attributs caractérisant les cibles et utilisés pour la sélection sont autant de critères permettant un **partitionnement** en différentes classes de l'espace de recherche.

Par exemple :
- *tous* les restaurants *près de la gare* ;
- *tous* les restaurants *chers ouverts le dimanche.*

Selon la répartition des valeurs de ces attributs sur l'ensemble des cibles non encore écartées, certains champs peuvent offrir un intérêt plus grand que d'autres à être informés en priorité. Ainsi, en Suisse, où quasiment tous les restaurants chinois sont chers, la connaissance de la tranche de prix du restaurant recherché est de moindre intérêt qu'un autre champ, comme par exemple la localisation de ce restaurant, si l'utilisateur a déjà indiqué qu'il cherchait un restaurant chinois. Dès lors, on voit bien que, pour autant que l'élément cherché soit présent dans la base, un choix judicieux dans l'ordre de renseignement des champs peut permettre d'arriver plus rapidement à une solution, ou un ensemble de solutions potentielles suffisemment petit pour être énuméré à l'utilisateur.

Du point de vue de la théorie de l'information, on dira que la valeur informative de la connaissance de chacun des champs est différente. Pour minimiser le nombre d'interactions avec l'utilisateur, il est donc préférable d'informer en premier lieu les champs avec la plus grande valeur informative, c'est-à-dire les champs engendrant un partitionnement des cibles qui en minimise le « désordre » (l'entropie).

Le *gain informatif* sur l'ensemble des cibles potentielles Ω (i.e. les cibles non encore écartées) apporté par la connaissance de l'attribut γ est mesuré par la réduction de l'incertitude $I(\Omega;\gamma)$:

$$I(\Omega;\gamma) = H(\Omega) - H(\Omega \mid \gamma) \tag{5.1}$$

> Pour maximiser le gain informatif, on choisira d'informer l'attribut γ qui minimise l'entropie $H(\Omega \mid \gamma)$[24]

avec :

$$H(\Omega \mid \gamma) = \sum_{m \in \gamma} P_\gamma(m) \cdot \underbrace{H(\Omega \mid \gamma{=}m)} \tag{5.2}$$

$$H(\Omega \mid \gamma{=}m) = -\sum_{\omega \in \Omega} P_{\Omega|\gamma}(\omega \mid m) \cdot \log\big(P_{\Omega|\gamma}(\omega \mid m)\big) \tag{5.3}$$

L'algorithme 5 [page suivante] – adapté au cas de données stockées dans une base indépendante – permet de déterminer, sur la base du critère simple précédent, le champ à informer en priorité.

[23] Un tel filtrage est relativement brutal, et limite fortement les possibilités d'initiative des utilisateurs. Une solution préférable serait de ne filtrer que les éléments pour lesquels l'indice de confiance est trop faible, permettant ainsi de court-circuiter le filtrage pour les interactions via le web.

[24] Pour éviter que des attributs avec beaucoup de modalités ne soient systématiquement choisi (par exemple les identifiants – au sens d'une base de données – ont une incertitude nulle, mais sont peu intéressants), on préférera en général minimiser $H(\Omega|\gamma)/|\gamma|$.

Algorithme 5 BestChoice(Ω, Γ) : identification du champ le plus discriminant

Nécessite : Ω l'ensemble des cibles compatibles avec les champs déjà informés ;
Nécessite : Γ l'ensemble des champs non encore informés ;
Fourni : $c \in \Gamma$ le champ le plus discriminant de Γ

 # Étape 1 : initialisation
$H_{min} \leftarrow \infty$
pour $\gamma \in \Gamma$ # **faire**
  pour $m \in \gamma$ # **faire**
    $\Upsilon[\gamma][m] \leftarrow 0$
  fin pour
fin pour
 # Étape 2 : examen des données de la base
pour cible $\omega \in \Omega$ **faire**
  pour champ $\gamma \in \Gamma$ **faire**
    pour modalité m présente dans $\omega \Rightarrow \gamma$ **faire**
      $\Upsilon[\gamma][m] += 1$
    fin pour
  fin pour
fin pour
 # Étape 3 : calcul du critère
pour champ $\gamma \in \Gamma$ **faire**
  $H_\gamma \leftarrow 0$
  pour modalité m possible pour γ **faire**
    $H_\gamma += \log_2(\Upsilon[\gamma][m]) \cdot \Upsilon[\gamma][m] \cdot |\Omega|^{-1}$
  fin pour
  si $H_\gamma < H_{min}$ **alors**
    $H_{min} \leftarrow H_\gamma$
    $c \leftarrow \gamma$
  fin si
fin pour
retourne c

5.1.6 Conclusion

Nous nous sommes focalisés ici sur les moyens permettant de mettre en œuvre un gestionnaire de dialogue ; nous avons de plus donné des principes (validés dans le cadre du projet *Info-Vox*) permettant de pallier, au niveau du dialogue, les faiblesses de la reconnaissance vocale, et cela sans nécessiter de ressources linguistiques importantes. Il est évident qu'une reconnaissance robuste serait préférable ; mais comme aucun système automatique d'interaction vocale en langage naturel ne peut compter sur une reconnaissance vocale fiable à 100% (même le « reconnaisseur humain » n'atteint pas cette performance), les principes décrits ici restent valables même avec un reconnaisseur de parole plus fiable que celui d'*InfoVox* (par ailleurs, il y a fort à parier que les applications visant à permettre une interaction via le téléphone, fixe ou cellulaire, seront encore longtemps confrontées à un fort taux d'erreurs de reconnaissance).

Bien évidemment, ces solutions ont été appliquées dans le cadre d'une tâche simple (vocabulaire bien discriminant) ; l'étape suivante est naturellement de déployer cette architecture et ces solutions dans le cadre d'applications plus complexes, comme c'est le cas avec les projet *INSPIRE* et *IM2*.

Remarquons encore que, contrairement à d'autres projets du même type réalisés par le passé, c'est bien ici la méthodologie de conception/développement qui est « générique » (i.e. réinstanciable), et non les systèmes de dialogue produits. Cette approche permet des conceptions tenant compte au mieux des spécificités et du contexte de l'application, tout en assurant une certaine cohérence dans les dialogues obtenus.

5.2 Visualisation de bases de grandes tailles

Mesurer et évaluer l'innovation technologique est une tâche particulière, mais qui intéresse plusieurs acteurs de la société, chacun dans une finalité qui lui est propre. La demande pour de tels indicateurs s'est considérablement accrue au cours des dernières décennies, et ce aussi bien dans le but de conduire des analyses micro et macro-économiques que pour la mise en place de politiques de décision à différents niveaux. Parmi les acteurs intéressés, on trouve au premier chef les offices de statistiques nationaux, ayant comme mission de fournir aux décideurs politiques les données nécessaires pour conduire les politiques publiques en la matière ; on y trouve également des fonds d'investissement et risqueurs de capitaux, intéressés eux à détecter les secteurs porteurs au niveau d'un continent ou d'un pays, mais également au niveau des entreprises (estimation de leur « valeur ») ; finalement les entreprises elles-mêmes peuvent être intéressées par ce type d'indicateurs, par exemple dans le but de déterminer leur stratégie de développement (segment de marché à investir, politique d'acquisition, etc). Les indicateurs existant de « l'inventivité » étant essentiellement construits sur la base de l'activité en matière de brevet (d'invention), la mise au point de méthodologies efficaces pour l'analyse de l'information contenue dans les brevets et relative à l'innovation technologique, est une nécessité [Comanor et Scherer, 1969, Dou, 1995, Narin, 1995]. Les bases de brevets constituant de larges collections de documents, structurés au moyen de multiples descripteurs (notamment bibliographiques), mais essentiellement porteurs d'une information textuelle, nous présentons dans ce mémoire de thèse une partie de nos travaux sur le sujet, réalisés dans le cadre du projet européen *Sting*,[25] en nous focalisant sur les techniques proposées pour représenter de manière visuelle les informations pertinentes pour l'analyse et automatiquement extraites du contenu textuel des documents de la collection.

Nous présentons les outils d'analyse multidimensionnelle de l'information « cachée » dans les collections de brevets, et notamment la mise en œuvre de *l'analyse des correspondances* [Benzécri, 1992, Johnson et Wichern, 1998], sur laquelle repose (en partie) la méthodologie proposée dans le cadre du projet *Sting* et visant à produire des indicateurs sur les tendances en matière d'innovations technologiques [Guellec et van Pottelsberghe, 1998]. Plus précisément, cette méthodologie devait permettre d'analyser les progrès et l'innovation technologique et scientifique en Europe, et ce à différents niveaux ; les outils employés devaient donc permettre de conduire ce genre d'analyse aussi bien pour des secteurs technologiques considérés isolément (par exemple un secteur dans un pays donné) que pour un ensemble de secteurs, au niveau mondial.

On peut résumer les objectifs à remplir comme suit :
- permettre des comparaisons multidimensionnelles entre l'activité d'innovation des pays, secteurs ou compagnies ;
- identifier les concurrents en compétition sur un secteur donné ;
- finalement, capturer les interactions qui peuvent exister entre différents domaines d'activité ou pôles d'innovations à l'intérieur de ces domaines.

Remarquons que bien que les techniques d'indexation sémantique présentées dans le reste de ce document n'aient pas été mises en œuvre dans le cadre du projet *Sting*, la problématique qui y était traitée et les méthodes employées constituent indéniablement un champ d'application idéal pour l'indexation sémantique.

5.2.1 Description générale de la méthodologie

L'un des points saillant de la méthodologie développée est l'utilisation conjointe de l'information textuelle présente dans les brevets et des descripteurs additionnels attachés aux documents,

[25] *Evaluation of Scientific & Technological Innovation and Progress in Europe through Patents*, EU IST99-20847, Computer Technology Institute (Patras).

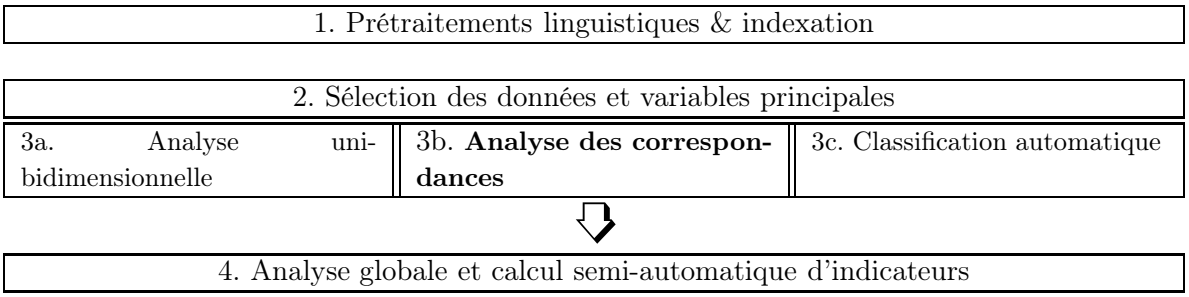


Figure 5.4: Vue schématique de la méthodologie d'analyse *Sting*

notamment leur classification dans la hiérarchie IPC (*International Patent Classification*). Pour exploiter ces informations et en dériver des résultats à la fois fiables et pertinents, des techniques d'analyses statistiques et textuelles ont été combinées ; en particulier, on effectue une analyse des correspondances d'une part et une classification automatique d'autre part, en utilisant dans les deux cas le contenu textuel des brevets, éventuellement agrégés sur la base des valeurs d'un ou plusieurs descripteurs ; les variables additionnelles (typiquement les compagnies, inventeurs, date de dépôts, etc.) sont ensuite utilisées pour enrichir la représentation obtenue.

La figure 5.4 représente de manière synoptique les différentes étapes constituant la méthodologie *Sting*.

La première étape de l'analyse consiste en une série de pré-traitements permettant de « nettoyer » les contenus textuels qui sont ensuite indexés ; cette étape est réalisée une fois pour toute, sur l'ensemble de la collection.

La seconde étape constitue le premier pas de la méthodologie d'analyse proprement dite. Cette étape est importante, car elle détermine en grande partie le sujet de l'étude qui va être conduite. Concrètement, l'utilisateur doit d'une part sélectionner le sous-ensemble des documents de la collection sur lequel il souhaite travailler (par exemple, le secteur d'activité sur lequel va porter son étude, ou bien les zones géographiques à considérer), et d'autre part il doit définir le niveau de granularité de son étude, c'est-à-dire déterminer ce que seront les individus élémentaires des analyses statistiques ; là encore, il le fera par le biais des descripteurs attachés aux brevets (par exemple, une étude centrée sur l'activité des entreprises, ou sur l'évolution au cours des années de l'activité).

Différentes analyses sont ensuite proposées ; la première catégorie regroupe les analyses standards (mono ou bivariées essentiellement), ne portant que sur les descripteurs. Avec l'analyse des correspondances, on propose de visualiser une partie de l'information présente dans la collection, en particulier les positionnements respectifs des individus sur lesquels porte l'étude, et la possibilité d'expliquer ces positionnements (résultant principalement en appariements ou oppositions). Le troisième type d'analyse utilise une classification automatique des individus (toujours sur la base de leur contenu textuel), et permet entre autres d'identifier des groupes partageant des technologies, de mesurer la densité des connections entre ces groupes (et par-là même, identifier les secteurs nouveaux, en mutation ou sur le point de disparaître[26]).

5.2.2 Préparation des données et analyse factorielle

Pour illustrer les différentes étapes de l'analyse présentées ici, nous avons utilisé une collection de 23′750 brevets, obtenus à partir de la base ESPACE/EPA, Vol. 2000/006 (constituant en fait une simple mise à jour de la base ESPACE/EPA) ; pour des raisons d'uniformité de langue, seuls les titre et les résumés (toujours disponible en anglais) ont été employé en tant que contenu textuel.

[26] Il faut pour cela procéder à une analyse par tranches temporelles.

f_{min}	–	5	10	10	**15**	15	30	50	100	200	500
f_{max}	–	–	–	4000	**4000**	2000	2000	2000	2000	2000	2000
|voc.|	39'939	14'778	9'877	9'848	**7'853**	7'807	5'113	3'670	2'252	1'271	475

Table 5.1*:* Taille de vocabulaire selon différents filtrages fréquentiels.

5.2.2.1 Pré-traitements linguistique et indexation

Pour pouvoir utiliser le contenu textuel des documents lors des analyses statistiques, il est nécessaire d'indexer la collection. Cette opération est en tout point similaire à l'indexation réalisée en recherche documentaire, dont il est question dans le reste de ce document ; la finalité est un peu différente, en ce qu'on ne s'intéresse pas nécessairement ici aux vecteurs profils des documents, bien qu'ils constituent le matériau de base des analyses effectuées (mais le but n'est pas de les comparer entre-eux[27]).

Pour que cette indexation ait une chance de capturer l'information intéressante présente dans les documents, il est nécessaire de pré-traiter les contenus textuels de ces documents, afin de retirer des données le bruit induit par des éléments peu pertinents (la ponctuation, les mots vides, les variations flexionnelles des mots, etc). Dans le cas présent, il faut également arriver à une taille d'index suffisamment faible pour permettre une analyse factorielle.

> Exemple :
> La chaîne de pré-traitements linguistiques que nous avons appliquée consiste tout d'abord à segmenter en mots (graphies) les contenus textuels[28], puis à attribuer à chaque mot ainsi identifié son lemme et sa catégorie morphosyntaxique (CMS) – cf. 4.3 [page 62], le même outil étant utilisé pour ces différentes étapes.
> Pour obtenir le jeu d'index, nous avons tout d'abord fusionné les mots sur la base de leurs lemmes, puis avons filtré les termes obtenus de manière à ne conserver que les catégories de mots porteuses d'un contenu sémantique important (noms, verbes, adjectifs et CMS indéterminée). Le sous-ensemble ainsi obtenu avec notre collection de test comportant encore près de 40′000 éléments, un filtrage supplémentaire s'est avéré nécessaire[29]. Pour cela, nous avons calculé l'impact sur la taille du vocabulaire de différentes bornes minimales et maximales sur la fréquence d'occurrence de ces termes (table 5.1).
> Nous avons alors choisi de filtrer selon les fréquences $f_{min} = 15$ et $f_{max} = 4000$ (soit $\chi_{15,4000}\,(\mathtt{tf})$). Finalement, à l'aide d'un anti-dictionnaire et d'une série d'expressions régulières, nous avons retiré les termes correspondant à des mots « vides » (tels que *be*, *have*, *can*, *may*) ainsi que les formes non lexicales (telles que *p1, st11, 6-12c*).
> De la sorte, nous avons isolé un jeu d'index comprenant 7′724 lemmes ; au vu des filtrages aggressifs appliqués, cette taille indique que la variété lexicale présente dans les titres et résumés des brevets considérés est importante.

5.2.2.2 Analyse des factorielle des correspondances

L'analyse factorielle est une technique permettant de réaliser une approximation linéaire (avec un minimum de distortions) d'un espace de haute dimensionalité en des espaces de dimensionalité réduite, plus facilement visualisables. Plus exactement, le rôle de l'analyse factorielle

[27] Même si cela est effectivement réalisé lors de la classification automatique.

[28] Les documents étant déjà segmentés en sections (titre, résumé, revendications, contenu, références, etc.) et en paragraphes (même si nous n'avons pas utilisé cette information).

[29] Une des raisons étant de pouvoir effectuer l'analyse des correspondances en un temps raisonnable.

US	DE	JP	GB	FR	IT	SE	FI	NL	AU	KR	CH	AT	DK	IL	ES
8'534	4'113	2'908	1'531	1'275	545	430	272	240	215	204	187	153	142	125	111

Table 5.2*:* Nombre de brevets par pays (dans la collection de test).

est de trouver les sous-espaces de faible dimensionalité qui permettent d'approximer au mieux une distribution de points dans l'espace d'origine (*i.e.* les données). Remarquons cependant que cette réduction du nombre de dimensions ne peut s'obtenir sans la perte d'une certaine information.

L'analyse factorielle est appliquée à des tables de contingence (ou de dénombrement) et permet d'étudier les relations qui existent entre différentes variables nominales au sein d'une collection d'individus (les données) ; les modalités de ces variables sont associées aux colonnes et aux lignes de la table, tandis que les valeurs des cellules correspondent au dénombrement des occurrences simultanées au sein de la collection des modalités associées aux lignes et colonnes respectives. Dans le cas de données textuelles, une table de contingence particulière est utilisée, où l'une des dimensions (ligne ou colonne) représente les mots (lemmes) tandis que l'autre représente les individus, le plus souvent des documents. Dans le cas où les individus sont les documents, la table de contingence correspond à la réunion des vecteurs-profils des documents (pour être exact, dans le cas où la « contribution » des termes d'index « aux documents » est mesurée par leur fréquence d'occurrence, `tf`). Avant de procéder à l'analyse, il est usuel de standardiser les valeurs de dénombrement : on procédera d'une part au *centrage* du tableau, consistant à soustraire, pour chaque terme d'un individu, la *masse* moyenne[30] de tous les termes de cet individu (le but étant de rendre comparable entre eux les profils des individus, même si leur taille en terme de contenu textuel est très différente). La *réduction* des données, usuellement employée pour limiter l'impact du choix d'une unité de mesure plutôt qu'une autre (les valeurs sont, après centrage, divisées par l'écart-type de la modalité considérée), n'a pas lieu d'être.

Pour une description détaillée des techniques factorielles (analyse et classification), de leurs fondements mathématiques et de leur utilisation en analyse des données, consulter par exemple [Escofier et Pagès, 1998, Lebart *et al.*, 2000, 1998].

Exemple : constitution de la table de contingence

Après que les données textuelles ont été pré-traitées et indexée, la première étape de la méthodologie d'analyse consiste à choisir les données sur lesquelles portera l'étude. Pour notre exemple, nous choisissons d'examiner l'activité en matière de propriété intellectuelle des 16 pays les plus actifs en la matière ; c'est-à-dire que les individus de notre analyse seront les pays, et que l'on ne retiendra que les documents correspondant à des brevets déposés dans l'un ou l'autre des 16 pays les plus actifs. La table 5.2 présente, pour chacun de ces pays, le nombre de brevets disponibles dans notre collection.

Avec l'index retenu précédemment et les individus retenus pour l'analyse, la table de contingence représentant nos données brutes est une matrice 16×7724. On associera un pays à chacune des lignes de cette matrice, et un terme d'indexation (lemme) à chacune de ses colonnes. Pour déterminer rapidement le contenu des cellules, il suffit d'aggréger les vecteurs-profils des documents correspondant à un individu (*i.e.* pour chaque terme d'index, somme des fréquences de ce terme sur tous les documents constituant l'individu).

Avec une telle configuration, le but de l'analyse des correspondances est de permettre l'exploration des dépendances non aléatoires qui peuvent être observées

[30] Dans le jargon de l'analyse factorielle, la masse correspond à la fréquence d'occurrence.

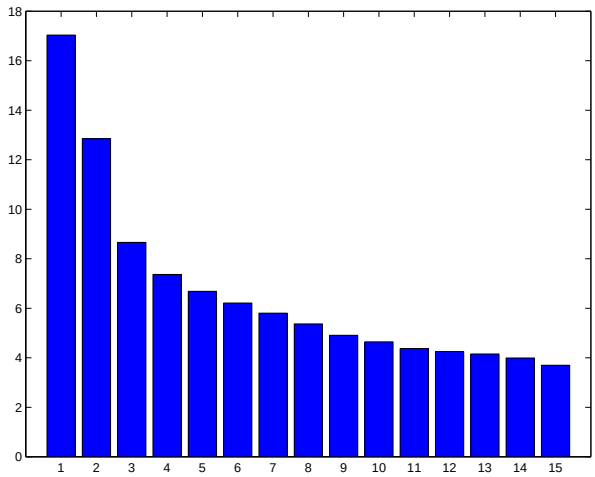


Figure 5.5*:* Proportion de l'inertie expliquée par chacun des axes factoriels (pays ⊗ vocabulaire).

> entre les activités en matière de brevets des pays retenus. Plus précisément, l'analyse des correspondances produit un nouvel espace vectoriel (appelé espace factoriel) dans lequel les similarités observées entre lignes et colonnes de la matrice de contingence (et mesurée par la distance du χ^2) peuvent être visualisées sous forme de proximités géométriques.

Une fois les axes factoriels déterminés (c'est-à-dire que l'analyse factorielle des correspondances proprement dite est réalisée), on peut procéder à la visualisation et à l'analyse des résultats.

La première difficulté rencontrée est de décider quels axes factoriels (*i.e.* quels facteurs) doivent être examinés (en particulier quelle combinaison des axes pour une représentation en 2 ou 3 dimensions). Pour cela, plusieurs mesures (coefficients de contribution) peuvent être employées. En premier lieu, on peut déterminer pour chaque facteur une **contribution globale** à l'espace factoriel complet, relativement aux autres facteurs (correspond en fait à la proportion de l'*inertie* présente dans le nuage de points *expliquée* par le facteur, et donne ainsi une mesure de l'importance du facteur par rapport aux données d'origine). En second lieu, on peu examiner, pour chaque individu, la **contribution relative** (désignée cos^2) de l'individu à la construction de l'axe, et qui permet, pour un individu donné, de connaître les axes qui le représentent le mieux (minimum de distortion).

> Exemple : choix des axes factoriels à considérer
>
> La figure 5.5 représente l'histogramme des *contributions globales* de chacun des 15 axes factoriels de notre exemple. La décroissance relativement régulière indique qu'il n'y a pas de sous-ensemble de facteurs qui pourraient expliquer l'essentiel des dépendances observées) ; on constate en effet que même les deux premiers axes, qui se détachent un peu du reste, ne contribuent ensemble qu'à 30% de l'inertie globale.
>
> En examinant la *contribution relative* des individus à chacun des axes factoriels, on peut néanmoins isoler un ensemble d'axes pertinents pour l'analyse. Par exemple, pour les quatres pays contribuant le plus fortement à un axe, on observe les valeurs de cos^2 suivantes :
>
> **JP :** Facteur 2 (0.768), Facteur 1 (0.209), ...
>
> **US :** Facteur 1 (0.951), Facteur 5 (0.021), ...
>
> **DE :** Facteur 1 (0.476), Facteur 2 (0.159), Facteur 7 (0.146), Facteur 5 (0.106), ...

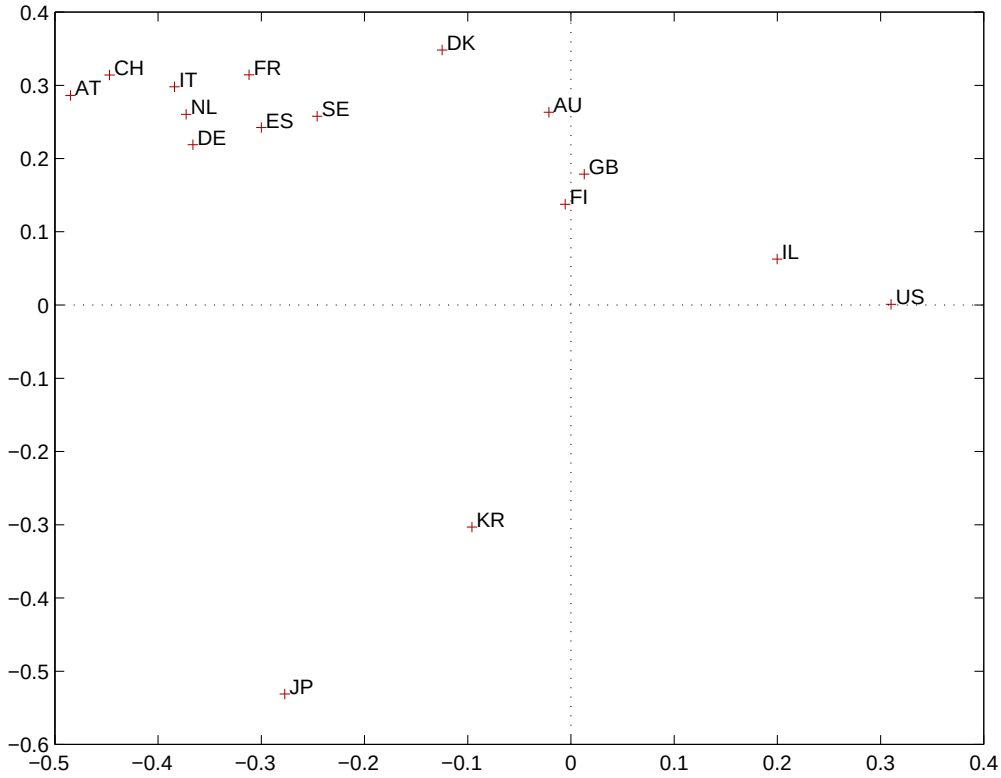


Figure 5.6*:* Projection des pays sur les axes factoriel 1 et 2.

FI : Facteur 3 (0.852), Facteur 5 (0.058), ...

Sur la base des contributions relatives, et en admettant comme seuil la valeur 0.150, les facteurs 1 et 2 ne peuvent être valablement utilisés que pour examiner les individus US, JP et DE, tandis que le facteur 3 n'est lui réellement pertinent que pour l'individu FI. En appliquant la même procédure de sélection à chaque individu et chaque axe, on obtient les résultats suivants[31] :

- facteurs 1 et 2 : individus JP (0.997), US (0.951) et DE (0.605)
- facteurs 5 et 6 : individus DK (0.854) et GB (0.719)
- facteurs 7 et 9 : individus IT (0.709) et SE (0.501)
- facteurs 11 et 12 : individus IL (0.764) et AT (0.629)
- facteurs 13 et 14 : individus ES (0.824) et AU (0.771)

Les autres facteurs sont essentiellement spécifiques à un unique individu (facteur 3 : IT(0.852), facteur 4 : FR(0.691), facteur 8 : KR(0.727), facteur 10 :NL (0.435), et facteur 15 : CH (0.826)).

En figure 5.6, on peut voir la projection de l'ensemble des individus sur les axes 1 et 2, qui, rappelons-le, ne sont réellement pertinents que pour l'analyse des positions relatives de US, JP et DE (dans la suite de nos exemples, nous nous limiterons par soucis de concision à l'examen des facteurs 1 et 2).

Visualiser les positions relatives des individus présente bien évidemment un certain intérêt, en particulier en présence de proximités ; néanmoins, une méthode plus précise de la lecture des axes est nécessaire pour permettre une interprétation qui puisse réellement apprendre quelque chose sur l'information cachée dans les données. En particulier, il est important de pouvoir caractériser les proximités ou éloignements observés, notamment au moyen du vocabulaire

[31] Où la valeur entre parenthèse correspond à la somme des cos^2 sur l'ensemble des facteurs indiqués.

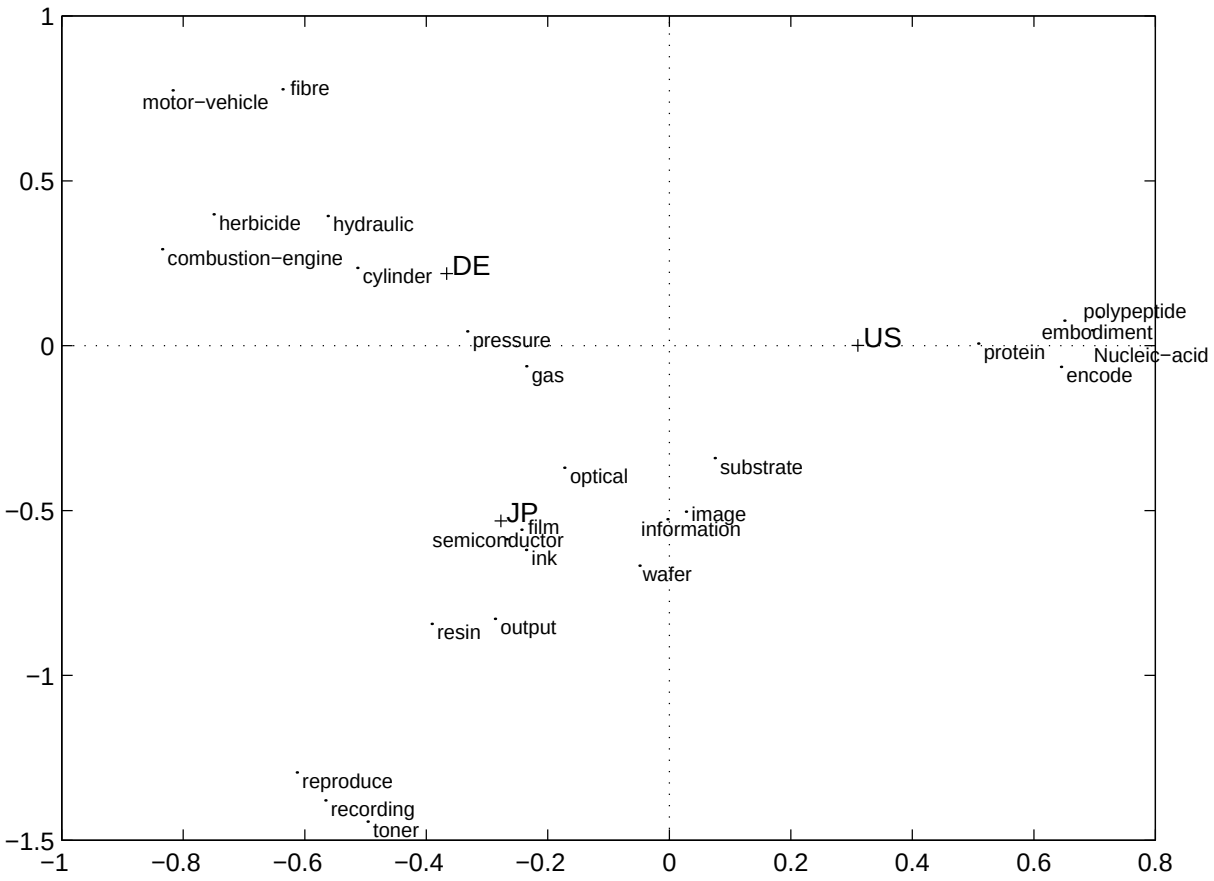


Figure 5.7*:* Projection simultanée des lemmes et des pays, facteur 1 et 2.

présent dans les brevets. Comme le nombre de modalité pour cette variable est grand (taille de l'index), un effort supplémentaire doit être fait pour automatiser la sélection des lemmes les plus intéressants pour l'interprétation des tracés. Pour réaliser cela, on examine, pour un tracé donné (*i.e.* deux facteurs), chaque paire d'individus ; on examine, sur le tracé concerné, la variation de distance entre les individus de la paire consécutive à la suppression d'un des lemmes dans la représentation (on ne refait pas l'analyse factorielle, mais on examine les positions qu'auraient ces individus en tant qu'individus supplémentaires). Seuls les lemmes induisant les variations les plus fortes sont affichés sur le tracé (représentation simultanée), pour aider à caractériser les positions relatives.

Exemple : interprétation des facteurs

En appliquant la technique proposée à notre exemple, on obtient le tracé donné en figure 5.7.

On peut constater que l'heuristique de sélection des mots permet effectivement de relativement bien interpréter les positions relatives des individus US, JP et DE. En effet, les lemmes *Nucleic-acid*, *protein*, *encode*, ... caractéristiques pour le facteur 1, qui oppose US à JP et DE, suggèrent nettement un positionnement marqué des brevets US dans le domaine des biotechnologies, tandis que les brevets JP et DE sont eux plus axés sur l'ingénieurie traditionnelle. On observe également que l'opposition sur le second axe entre des lemmes tels que *reproduce*, *recording*, *toner*, *ink*, *semiconductor*, *wafer*, ... du coté JP et des lemmes tels que *cylinder*, *combustion-engine*, *motor-vehicle*, *pressure*, ... du coté DE laisse à penser que JP

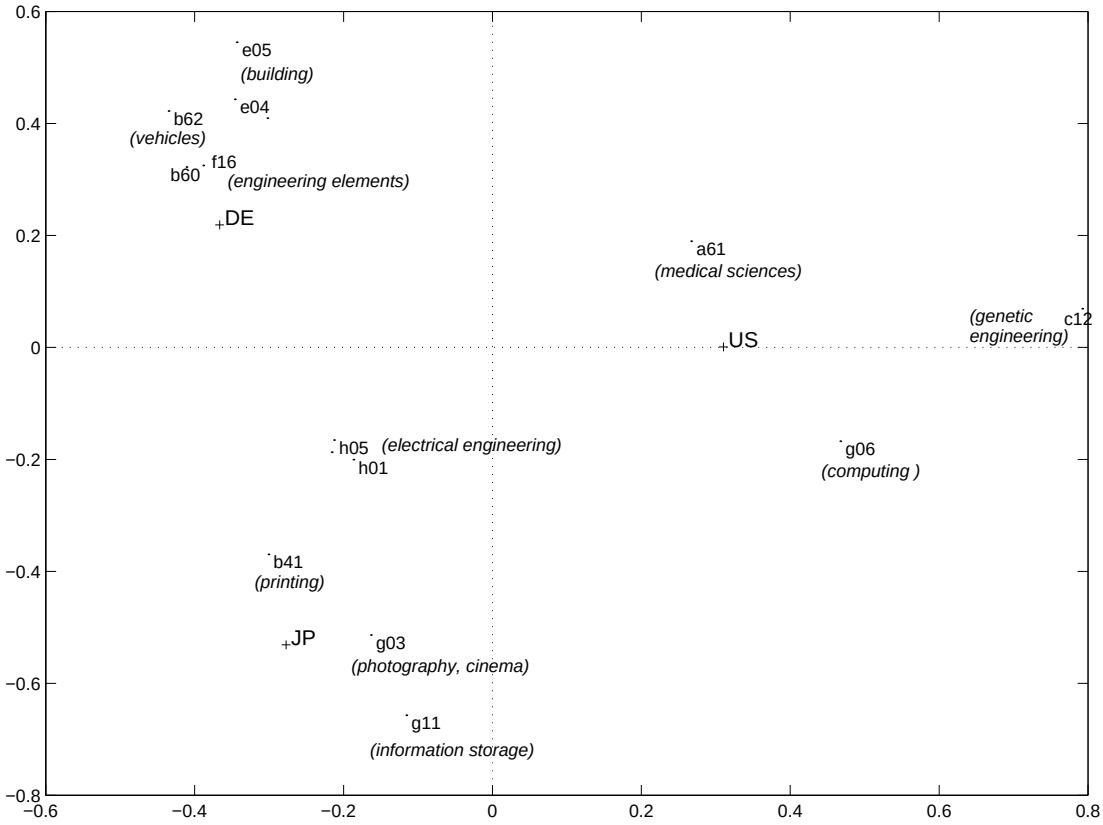


Figure 5.8*:* Projection simultanée des codes IPC et des pays, facteurs 1 et 2

se démarque par rapport à DE en matière d'activité en électronique et industrie d'impression, DE étant plus centré sur l'ingénieurie mécanique.

On peut dès lors enrichir encore la représentation, par exemple en visualisant (en tant qu'individus supplémentaires) de regroupements effectués au moyen d'autres variables. Par exemple, en prenant les premières branches de la classification IPC, et en appliquant la même technique de sélection de l'information affichée que celle employée pour les lemmes (cos^2), on obtient le tracé donné en figure 5.8.

Il est intéressant de constater que les codes qui expliquent le mieux la position de US (`a61`, `g06`, `c12`) dans son opposition à JP et DE sont respectivement associés aux domaines des *sciences médicales*, de l'*informatique*[32] et des *micro-organismes*. Sur le second axe, les codes caractérisant la position de DE sont essentiellement relatifs à l'ingénieurie traditionnelle (avec des codes tels que `b`, représentant *industries techniques diverses*) tandis que ceux associé à JP concernent les domaines de l'*impression* et le *stockage d'informations*.

5.2.3 Estimation de stabilité

Il peut être important – voire nécessaire – pour celui qui interprète les résultats d'une classification ou d'une analyse factorielle d'avoir une indication sur la marge d'erreur qui entache les résultats observés. Dans ce but, il est nécessaire de déterminer dans un premier temps quelles sont les sources possibles d'erreurs, et dans un deuxième temps comment en mesurer l'importance :

[32] À ce propos, on ne manquera pas de noter qu'en Europe, les brevets en matière de logiciels restent interdits, malgré la politique laxiste des différents offices nationaux et de l'OEB sur le sujet.

- Observe-t-on réellement quelque chose ? les données ont-elles une structure, ou de simple fluctuations dans l'échantillonnage peuvent-elles expliquer les résultats obtenus ?
- A-t-on découvert une classe pré-existante, ou a-t-on simplement coupé une réalité continue en classes ?
- Sommes-nous en présence d'une configuration stable, considérant ce que l'on sait à propos de la précision des données, de la nature de leur codage et de l'importance relative des différentes variables ? Quelle serait l'influence sur les résultats d'une modification dans le jeu de données (addition ou suppression de variables ou d'individus, modification de leur codage, etc.) ?

Les deux première questions portent sur les méthodes employées pour réduire la dimensionalité de l'espace des données et synthétiser celles-ci ; il s'agit en fait de conduire une évaluation « *glass box* » sur la pertinence des différentes étapes, et les gardes-fous à vérifier. Dans le cadre de l'analyse des correspondances, l'examen des cos^2 est typiquement un moyen pour s'assurer de la pertinence des observations. La troisième question en revanche porte directement sur les données que l'on cherche à analyser, sans considération pour la nature de la méthode (évaluation « *black box* ») : quelle est la stabilité des résultats observés en regard des perturbations pouvant affecter ces données ?

Les résultats produits par une méthode factorielle ne sont pas des assertions, mais des représentations, c'est-à-dire des objets complexes, sur lesquels s'appliquent difficilement les techniques usuelles en statistique de mesure d'information.

Comment peut-on valider une représentation dans un plan factoriel ?

- au moyen d'une procédure externe, similaire à celle utilisée pour déterminer le nombre d'axes factoriels à conserver (utilisation de connaissances *a priori*, position de variables supplémentaires, etc.) ;
- au moyen d'un calcul de stabilité adapté (exploration du voisinage des données, guidée par une mesure d'incertitude sur ces données) ;
- par la construction de zones de confiance sur le positionnement des variables (profils lignes ou colonnes), soit au moyen d'une méthode analytique basée sur des hypothèses probabilistes, soit à l'aide de méthodes empiriques basées sur des techniques de ré-échantillonnage.

5.2.3.1 Stabilité et sensibilité

Les calculs de stabilité et de sensibilité sont probablement les procédures de validation les plus convaincantes [Lebart *et al.*, 2000]. L'élément central de ces opérations consiste en un test de la stabilité d'une configuration après que différentes distorsions aient été appliquées aux données initiales.

Plusieurs recherches ont été réalisées sur la stabilité, d'un point de vue théorique, des résultats d'une analyse en composantes principales et des correspondances [Escofier et Roux, 1972], et en particulier l'étude de la variation maximale des facteurs et des valeurs propres lorsque l'on apporte aux données des modifications bien déterminées : ajout ou suppression d'un élément dans la table, influence du regroupement de différents éléments, influence de la pondération et de la métrique.

Les sous-espaces correspondant au haut du spectre (maximum de l'inertie expliquée) sont les plus stables en regard des perturbations possibles sur les données de la matrice à inverser. En fait, cette matrice est moins sensible aux fluctuations d'échantillonnage [Tanaka et Huba, 1984] que ne le sont des paramètres comme la moyenne, le pourcentage (moments du premier ordre)

On peut montrer d'une part que l'essentiel des variations de valeurs propres ne dépendent pas de variations des vecteurs propres. D'autre part, les variations des constitutants des vecteurs propres dépendent des distances entre la valeur propre correspondante et les autres valeurs

propres, c'est-à-dire du degré d'isolement de la valeur. Ainsi, c'est la distance entre les valeurs propres délimintant le sous-espace qui définit la stabilité de ce sous-espace. Dans le cas d'un sous-espace engendré par les premiers facteurs, l'élément majeur est la distance entre la dernière valeur propre correspondant au sous-espace et celle qui lui est immédiatement consécutive ; l'angle entre le sous-espace construit sur les données initiales et son homologue issu des données perturbées sera d'autant plus petit que la distance entre ces valeurs propres est grande.

5.2.3.2 Estimation de la stabilité globale au moyen de techniques de ré-échantillonnage

Pour estimer la stabilité des résultats d'une analyse (factorielle, classification, etc.), il est naturellement nécessaire de connaître la stabilité des données qui sont analysées. Dans le cas de données relatives aux brevets d'invention, deux éléments sont à considérer :

- l'exhaustivité des données considérées, impossible à atteindre dans le cas d'une analyse non rétrospective (analyse sur des instantanés)[33] ;
- l'influence du vocabulaire.

La méthode privilégiée pour prendre en compte ces éléments – pour lesquels nous ne disposons pas d'hypothèses probabilistes fiables – est la méthode empirique du ré-échantillonnage, qui consiste à produire un ou plusieurs jeux alternatifs de données en ajoutant un certain niveau de bruit à celui des données observées, et à comparer le résultat final de l'analyse sur chacun de ces jeux.

Il y a plusieurs possibilités pour produire ces jeux supplémentairs de données (*i.e.* plusieurs méthodes pour « bruiter » les données), mais dans le cas où l'on s'intéresse à la stabilité d'une analyse des correspondances, la technique de *bootstrap* est pratiquement l'une des seules à pouvoir être employée.[34]

5.2.3.3 Ré-échantillonnage *bootstrap*

Présentation

La technique de réplication *bootstrap* [35] que l'on doit à Bradley Efron (1979), consiste à simuler à partir d'un jeu de données observées, m échantillons (réplications) constitués chacun du même nombre d'individus que l'échantillon initial, et obtenus par tirages successifs (avec remise)[36] dans cet échantillon d'origine. Les individus réellement observés se retrouvent donc, selon les échantillons, avec un poids relatif rehaussé (plusieurs tirages), identique (un tirage) ou absent (non tiré).

Cette méthode peut être employée pour analyser la variabilité de paramètres statistiques simples, en produisant des intervalles de confiance de ces paramètres ; mais elle est surtout appliquée lorsque l'on ne peut pas estimer analytiquement la variabilité d'un paramètre, comme

[33] Remarquons que ce point couvre une partie des erreurs affectant les descripteurs (sélection incomplète des données).

[34] Remarquons que la technique du *jacknife* [Efron, 1982], qui consiste à retirer un ou plusieurs individus au hasard doit être utilisée avec précautions (voire évitée) lorsqu'il s'agit d'estimer des paramètres pour lesquels la valeur théorique est dépendante du nombre d'individus dans le jeu de données ; les techniques de permutations (Fisher, Pittman, Dwass, ...) ne sont ici d'aucun secours.

[35] Ainsi appelée parce qu'utiliser les données disponibles pour générer de nouvelles données n'est pas sans analogie avec l'astuce employée par le personnage fictif du Baron de Munchausen qui, alors qu'il s'enlisait dans la vase, se sauva de la noyade en tirant sur ses propres bottes.

[36] Dans le cas général, les individus observés étant supposés être indépendants statistiquement les uns des autres, le tirage se fait selon une distribution équiprobable

c'est en général le cas avec les méthodes multidimensionnelles, où les hypothèses de multinormalité sont rarement vérifiées.

Le *bootstrap* n'est rien d'autre qu'une technique de simulation particulière, fondée sur la distribution empirique de l'échantillon des observations. Efron et Tibshirani utilisent le terme de « *bootstrap* non paramétrique » pour désigner ce type de simulation, et parlent de « *bootstrap* paramétrique » dans le cas de simulations mettant en jeu une distribution théorique et des paramètres calculés à partir de l'échantillon observé (simulation classique). Au contraire de la méthode du *jacknife*, le *bootstrap* n'est pas déterministe, car il ne fait pas intervenir l'échantillon de façon symétrique. La méthode donne dans la plupart des cas une bonne image de la précision statistique de l'estimation sur un échantillon.

Les recherches théoriques, en particulier menées par Efron montrent que pour de nombreux paramètres statistiques, l'intervalle de confiance correspondant à la distribution simulée (par *bootstrap*) et celui correspondant à la distribution réelle sont généralement de même amplitude. Notons cependant que la méthode échoue lors de l'estimation des bornes d'un intervalle pour une loi uniforme sur cet intervalle.[37]

Pour une description complète de la technique et de ses applications, se référer (entre autres) à Davison et Hinkley [1997], Efron et Tibshirani [1994].

Bootstrap d'une analyse factorielle

Pour estimer la variabilité des valeurs propres et des taux d'inertie des facteurs produits par une analyse factorielle (analyse en composantes principales, analyse des correspondances, etc.), il faut reproduire l'analyse (*i.e.* inverser la matrice) pour chaque échantillon simulé, et calculer la distribution des fréquences de chacun des paramètres. Dans le cas où l'on ne s'intéresse finalement qu'aux coordonnées factorielles des profils lignes et/ou colonnes (c'est-à-dire les projections des centres de masse des modalités des variables mises en correspondance), l'intérêt de la méthode de validation par réplication *bootstrap* est qu'elle ne requiert pas de répéter l'analyse factorielle (l'inversion de la matrice) ; outre l'économie appréciable de temps de calcul, on évite ainsi l'écueil de la comparaison de projections d'individus dans différents plans factoriels.[38] Dans ce cas en effet on peut se contenter, dans l'espace factoriel issu des données observées, de projeter en tant qu'individus supplémentaires les profils lignes ou colonnes obtenus par réplications successives. On se servira dès lors de la distribution des positions de chaque profil pour construire et visualiser une zone de confiance pour le positionnement du profil, cette zone pouvant être simplement l'enveloppe convexe des positions répliquées, ou un ellipsoïde de confiance (construit de sorte à englober k % des points répliqués, où k vaut typiquement 90%, 95% ou 98%).

> Exemple :
>
> Soit la table de contingence (imaginaire) *brevets*×*mots* 5.3 [page suivante], dans laquelle les brevets sont identifiés par le pays qui leur est associé :

[37] Il est en effet évident dans ce cas que l'estimation donnée par les valeurs extrêmes ne sera pas améliorée par des tirages à l'intérieur de l'échantillon de base.

[38] Comparaison délicate à réaliser, les axes pouvant non seulement changer d'orientation (même en l'absence d'altération des données), mais aussi changer de rang ou d'inclinaison.

C_{ij}	ceramic	electronic	barium	plasma	neuron	nuclear	$\sum C_{\cdot j}$
US1	48	59	1	35	25	8	176
JP1	61	68	5	38	24	20	216
US2	96	92	37	0	51	39	315
UK1	34	21	0	27	22	5	109
JP2	6	17	24	58	0	37	142
JP3	0	57	0	47	0	0	104
US3	27	25	2	0	1	1	56
US4	2	5	1	7	57	0	72
$\sum C_{i\cdot}$	274	344	70	212	180	110	1′190

Table 5.3*:* Table de contingence *brevets*×*mots*.

Réplication (élémentaire) complète

Pour réaliser un *bootstrap* «élémentaire» sur l'échantillon supposé à l'origine de cette table de contingence, il convient de se souvenir que l'individu élémentaire dénombré dans cette table est l'occurrence d'un mot dans un énoncé textuel. On est donc en présence de 1′190 individus élémentaires, équiprobables et indépendants entre eux ; chaque réplication doit donc consister en 1′190 tirages avec remise.

En d'autres termes, pour chaque réplication, l'occurrence d'un individu élémentaire correspond à une expérience de Bernouilli $B(1, p = 1/n = 1190^{-1})$, et le nombre d'occurrences de cet individu suit une distribution binomiale $B(n, 1/n)$. Comme au final c'est la fréquence d'occurrences du terme qui nous intéresse, plutôt que de simuler,$m \cdot 1'190$ valeurs aléatoires, et rechercher pour chacune d'elle à quelle cellule (i, j) de la table $T_{r\in[1:m]}$ correspond le tirage (afin d'incrémenter la valeur de dénombrement), il est possible de générer directement les $m \cdot 8 \cdot 6$ valeurs de dénombrements C_{ij}, en utilisant l'approximation par la normale.

Approximation par la loi normale

Comme chaque cellule de la table de contingence correspond à un ensemble d'individus élémentaires, la probabilité qu'un individu de la réplication r soit assigné à la cellule C^r_{ij} (soit que le mot j apparaisse dans le brevet i) est donnée par :

$$P(C^r_{ij}) = \frac{C_{ij}}{\sum C_{ij}} = \frac{C_{ij}}{n}$$

Dans le cas de l'échantillon de notre exemple, la probabilité d'avoir le mot ceramic dans le brevet US1 est de $48/1190$.

Comme on peut approximer la distribution binomiale[39] par la loi normale :

$$B(n, \frac{C_{ij}}{n}) \approx N(C_{ij}, C_{ij} \cdot (1 - \frac{C_{ij}}{n})$$

la valeur de la cellule C^r_{ij} peut être directement obtenue au moyen de cette distribution.[40]

[39] Ou une multinomiale, si l'on considère toutes les cellules de la table.

[40] Des algorithmes efficaces permettant la simulation de variables normales ou de Poisson (ainsi que de nombreuses autres distributions) peuvent être trouvée dans [Ripley, 1987, pg. 78–92].

C_{ij}	ceramic	electronic	barium	plasma	neuron	nuclear	$\sum C_{\cdot j}$
JP1	61	68	5	38	24	20	216
US4	2	5	1	7	57	0	72
JP1	61	68	5	38	24	20	216
US2	96	92	37	0	51	39	315
US4	2	5	1	7	57	0	72
US1	48	59	1	35	25	8	176
JP2	6	17	24	58	0	37	142
US2	96	92	37	0	51	39	315
$\sum C_{i\cdot}$	372	406	111	183	289	163	1′524

Table 5.4*:* Contingence d'un échantillon hypothétique (réplication au niveau des brevets).

Remarque :

En toute rigueur, si C_{ij} est petit (< 10), il est préférable d'utiliser l'approximation de Poisson :

$$B(n, \frac{C_{ij}}{n}) \approx P(\lambda = C_{ij}).$$

Réplication d'individus non-indépendants

L'hypothèse d'indépendance des individus élémentaires n'étant assurément pas vérifiée dans le cas de données textuelles,[41] nous avons choisi de conserver ces dépendances lors de la réplication des échantillons (en considérant comme individu élémentaire l'occurrence d'un brevet).

Exemple :

En reprenant la situation de l'exemple précédent, mais en considérant cette fois que l'on se trouve en présence de 8 individus seulement – chacun d'eux ayant la même probabilité (1/8) de survenir, une réplication *bootstrap* revient alors à constituer un jeu de 8 brevets, en tirant ceux-ci parmi les brevets de l'échantillon initial.

On pourrait ainsi avoir comme réplication le jeu [JP1,US4,JP1,US2,US4,US1,JP2,US2], auquel correspond la table de contingence 5.4.

Dans le cas où l'on agrège ensuite les profils des brevets (par exemple, pour avoir une contingence *pays* × *mots*), cette technique peut être utilisée pour estimer la position des individus (les pays) dans l'espace factoriel, en mesurant l'impact de l'absence d'exhaustivité dans le jeu initial, ainsi que d'assignations erronées dans les descripteurs. Par ailleurs, si l'on ne s'intéresse qu'à un sous-ensemble des individus, il est possible de limiter les calculs pour générer une réplication en utilisant l'approximation par la loi normale présentée précédemment, mais basée cette fois sur les masses relatives de chaque individu plutôt que sur les valeurs de dénombrement des mots reportées dans la table de contingence.

En appliquant la technique de la réplication *bootstrap* sur la collection de 23′750 brevets issus de la base ESPACE/EPA, Vol. 2000/006, et en considrant comme indivisible les contenus textuels

[41] Il n'est pas choquant de considérer que les contenus textuels des documents représentent des entités indivisibles, en particulier lorsque l'on tente d'estimer la stabilité vis-à-vis de l'exhaustivité des observations, et non vis-à-vis du vocabulaire.

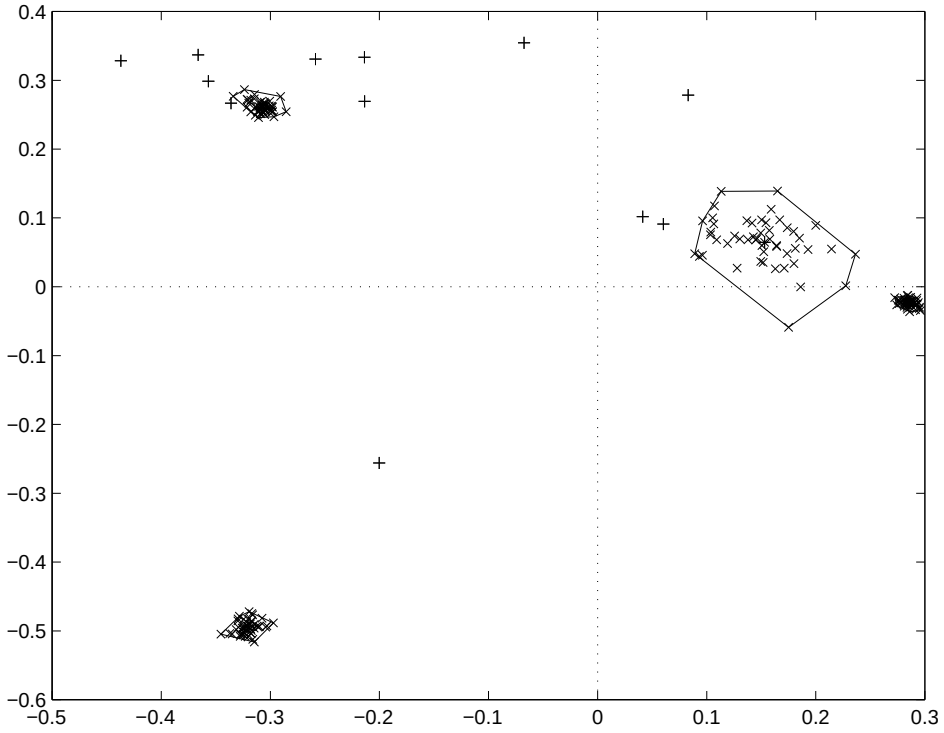


Figure 5.9*:* Projection de (50) réplications des pays US,JP,DE et IL, facteurs 1 et 2.

des brevets, nous avons réalisé 50 réplications des individus US, JD, DE et IL (ce dernier à des fins de comparaison). La figure 5.9.présente la projection, en tant qu'individus supplémentaires, de ces réplications, avec le travé de l'enveloppe convexe (90%).

On observe que les 3 individus US, JP et DE sont particulièrement stables, tandis qu'IL l'est nettement moins. Le nombre de brevets dans chacune de ces classes (plus faible pour IL que pour les trois autres) explique en partie cette différence de stabilité, l'autre partie de l'explication étant que les axes visualisés (facteurs 1 et 2) sont adaptés pour ces trois individus, mais pas pour IL.[42]

5.2.4 Conclusion

Tenter de mesurer l'innovation par le biais de l'analyse textuelle des brevets d'invention déposés est délicat : d'une part, le très grand nombre de données complique pasablement les opérations, mais le problème principal réside dans le caractère faiblement informatif de ces descriptions textuelles (du moins en ce qui concerne le résumé et les revendications).[43]

Néanmoins, tout en gardant à l'esprit que les techniques d'interprétations basées sur le vocabulaire restent essentiellement exploratoires, et que les conclusions qui peuvent en ressortir doivent être validées par d'autres analyses (le retour au texte est également important), nous avons montré qu'il est possible, grâce aux outils développés pour l'analyse des bases de brevets (dans le cadre du projet européen *Sting*) et présentés ici, de faire ressortir un ensemble pertinent d'informations extraites d'une collection documentaire, permettant de se faire rapidement une idée des structures et relations autrement noyées dans la masse documentaire.

[42] Néanmoins, ces deux points sont liés, la masse des classes US, JP et DE induisant les axes considérés.

[43] L'utilisation de bases réécrites en vocabulaire contrôlé– par ex. la base Derwent (Thomson Scientific) – permettent néanmoins d'améliorer la situation (mais avec un coût financier certain).

Chapitre 6

Conclusion

Les travaux présentés dans ce mémoire de thèse s'inscrivent dans le domaine de la recherche documentaire (RD). Une large part est consacrée à la thématique centrale de la RD, à savoir l'identification, dans une collection de documents, d'un sous-ensemble pertinent pour la « requête » d'un utilisateur. L'interaction proprement dite avec cet utilisateur a également été considérée par le biais de l'étude d'un système d'accès à une base de données à l'aide d'une interface multimodale (vocale et textuelle) en langage naturel. Finalement, la problématique de la visualisation et de l'analyse de bases documentaires de grande taille a également été abordée et un certain nombre de techniques utilisables dans cette optique ont été présentées.

En ce qui concerne l'interaction homme-machine, nous avons présenté un schéma général permettant de construire de manière rapide et systématique des interfaces de dialogue à initiative mixte, laissant à l'utilisateur humain une large latitude dans la conduite de l'interaction.[1] Ce schéma est à la fois utilisable dans des applications typiques de recherche d'information dans une base de données,[2] pour lesquelles l'utilisateur connaît assez précisément l'information qu'il désire, ou est du moins capable de la décrire avec précision, et dans des applications de conseil, pour lesquelles l'utilisateur n'a pas nécessairement d'idée précise sur ce qu'il désire, et attend de la part du système non seulement une aide pour préciser ses souhaits, mais également un ensemble de propositions comme résultat final. Nous avons en particulier mis l'accent sur les techniques permettant d'obtenir une interface robuste, capable, par le dialogue, de pallier dans une large mesure les erreurs de reconnaissance vocale. Les bons résultats obtenus par les travaux menés dans ce cadre ont d'ores et déjà conduit à leur reprise et poursuite par d'autres équipes, notamment dans le cadre des projets *Inspire*[3] et MDM[4].

Pour ce qui est de la visualisation de grandes collections de données textuelles, nous avons présenté l'application de l'analyse des correspondances, permettant de mettre en évidence des similitudes ou des oppositions entre différents regroupements de documents, construits sur la base de traits additionnels. Appliquées au cas de données issues de bases de brevets[5], on peut ainsi déterminer, pour divers groupes (pays, sociétés, etc) leurs éléments spécifiques communs (similitudes), ou au contraire leurs différences. Nous avons par ailleurs proposé une méthode, basée sur le principe de réplication *bootstrap*, permettant de déterminer un intervalle de confiance

[1] Travaux réalisés dans le cadre du projet InfoVox, « *Interactive Voice Servers for Advanced Computer Telephony Application* », CTI n° 4247.1, EPFL & IDIAP.

[2] Laquelle est totalement masquée à l'utilisateur.

[3] Inspire project : INfotainment managment with SPeech Interaction via REmote microphones and telephone interface, IST-2001-32746.

[4] MDM : Multimodal Dialogue Managment, projet réalisé dans le cadre du Centre National Suisse de Compétences en Recherche IM2.

[5] Projet sting, « *Evaluation of Scientific & Technological Innovation and Progress in Europe through Patents* », EU IST99-20847, Computer Technology Institute (Patras).

pour les positionnements relatifs des différents groupes, et ainsi de juger de la fiabilité des similitudes ou oppositions visuellement apparentes.

Mais le cœur de notre travail a concerné la RD proprement dite. L'idée centrale sur laquelle nous avons axés nos travaux repose sur le principe *d'indexation sémantique*, et a visé une amélioration de la représentation des documents textuels par la biais de termes d'index intégrant des informations de nature sémantique. Plutôt que d'inférer ces informations sur la base de statistiques établies sur les données à traiter (comme par exemple dans le cas l'indexation sémantique latente), nous nous sommes intéressés à l'utilisation de ressources sémantiques externes (de type réseaux sémantiques ou thésaurus) comme support de l'information apportée au jeu d'index.

En nous appuyant sur plusieurs expériences rapportées dans la littérature, nous avons imaginé puis étudié une méthode novatrice pour réaliser cette indexation, en ne sélectionnant, dans la ressource externe, que les connaissances réellement pertinentes au regard des données à indexer. Plus exactement, nous avons proposé de choisir automatiquement le degré de généralité le plus adapté aux données, au moyen d'un critère original, baptisé « *coupe de redondance minimale* » (*CRM*). Construit sur la base de la théorie de l'information, ce critère permet d'obtenir des termes d'index ayant des probabilités d'occurrences les plus équilibrées possibles dans la collection de documents, condition nécessaire pour maximiser en moyenne le pouvoir discriminant de chaque terme d'index, et par là même améliorer les recherches de documents ainsi indexés.

Afin de mieux présenter nos travaux, nous avons proposé un cadre général permettant de décrire différentes techniques envisageables pour réaliser des indexations sémantiques, en adaptant si possible la richesse des descriptions issues des ressources externes aux données à représenter. Nous avons utilisé ce cadre pour détailler trois familles de critères utilisables pour l'indexation sémantique, en donnant les algorithmes permettant leur mise en œuvre. Les deux premières familles permettent de considérer, dans une optique d'indexation sémantique, plusieurs critères déjà connu en sélection de termes. Nous montrons par ailleurs qu'un certain nombre d'entre eux ne sont en fait que peu efficaces pour la tâche considérée. La troisième famille nous a permis d'introduire notre critère *CRM*.

Les évaluations conduites en employant deux ressources sémantiques différentes, *EDR* et *WordNet*, sont prometteuses et permettent de conclure que notre technique offre un réel potentiel pour les tâches de RD. En effet, en confrontant les résultats obtenus, et en les comparant aux performances d'une indexation traditionnelle (utilisant les lemmes des mots des documents comme espace de représentation), nous avons pu, d'une part, montrer la pertinence de l'indexation sémantique en général, et d'autre part, illustrer la qualité indéniable du critère *CRM* que nous proposons pour cette indexation.

Perspectives

Il reste cependant un certain nombre de questions encore ouvertes.

Tout d'abord, il est nécessaire de valider les techniques proposées sur des collections de références de plus grande taille (telles les collections issues des campagnes TREC, CLEF ou NTCIR). En effet, bien que présentant des avantages certains pour la mise au point et l'évaluation fine d'une technique de recherche documentaire, les collections utilisées pour les évaluations restent trop limitées pour exhiber des tendances lourdes sur l'utilisation en moyenne de ces techniques à plus grande échelle. Relevons cependant à nouveau que l'utilisation a posteriori de collections de référence telles que celles issues de *TREC* permet de valider la technique, mais ne permet pas de l'évaluer : il est pour cela en effet nécessaire de participer à la campagne.

L'utilisation combinée de ressources sémantiques nettement plus spécialisées serait également intéressante à examiner (typiquement MeSH ou UMLS, en *Genomics Track*).

Il serait également intéressant de confronter les résultats obtenus par indexation sémantique et ceux obtenus avec d'autres techniques, comme l'indexation sémantique latente, ou l'utilisation des ressources sémantiques lors du calcul des similarités entre documents. Cette dernière technique en particulier devrait permettre d'au moins atteindre les résultats de l'indexation sémantique – le changement d'espace de représentation opéré par l'indexation sémantique peut très bien être effectué *à la volée*, lors du calcul de similarité – et probablement les dépasser. Un plus grand degré de liberté est en effet ainsi offert, et des modulations plus fines peuvent parfaitement être envisagées (telle par exemple qu'une prise en compte de liens sémantiques conditionnée par le contenu de la requête).

Remarquons que, bien qu'un certain nombre de ces techniques aient déjà été appliquées aux collections que nous avons nous-même utilisées, et que les résultats obtenus soient disponibles dans la littérature, la comparaison des performances sur la base d'expériences menées par des équipes différentes reste quasi impossible en pratique. Bien que les campagnes d'évaluations aient permis de dresser un cadre global pour les mesures de performances d'un système de recherche documentaire, les pratiques en matière de publication sont malheureusement trop « exotiques » pour autoriser de telles comparaisons a posteriori. Relevons notamment l'absence quasi systématique de résultats obtenus dans une configuration standard de référence (*baseline*) sans prétraitement particuliers (ces derniers n'étant pratiquement jamais décrits, du moins avec suffisamment de détails pour en permettre la reproduction), une description souvent elliptique des mesures utilisées (il existe une quantité de manières différentes de calculer une précision ou un rappel, il suffit de regarder les résultats fournis par `trec_eval` pour s'en convaincre), et rarement l'identification précise (en particulier l'origine) des collections de référence utilisées, ces dernières étant malheureusement souvent disponibles en de multiples versions plus ou moins divergentes.[6]

Les expériences réalisées dans le cadre de ce travail n'ont porté sur l'utilisation que d'une seule relation sémantique présente dans les ressources, la relation d'hyponymie/hyperonymie. Mais si l'on considère les termes « médecin, hôpital, médicament », et bien que le lien entre ces termes soit évident, la relation d'hyponymie/hyperonymie ne permet en aucun cas de les rapprocher les uns des autres. On peut donc raisonnablement espérer que la prise en compte de relations supplémentaires (méronymie, antonymie, etc.) permette d'améliorer encore les performances, même si cette prise en compte n'est pas totalement triviale à réaliser (en particulier la valuation

[6] Comparer les données Fox [1983] (table 1 et 2), de Baeza-Yates et Ribeiro-Neto [1999] (table 3.4, page 95), de Mandala *et al.* [1998] (table 1) et les statistiques de ces mêmes collections données sur le site internet de l'université de Glasgow (http://ir.dcs.gla.ac.uk/resources/test_collections/) ; on notera même que la version (actuellement) disponible sur ce même site de la collection *TIME* présente un certain nombre de défauts (documents tronqués et manquants). Les collections constituées dans le cadre des campagnes d'évaluation permettent heureusement d'échapper à ce genre de problèmes.

du poids des occurrences des termes ainsi atteints, et la propagation de ces occurrences malgrés les cycles du graphe, et selon des relations du second, troisième, ... ordre).

Finalement, la chaîne de traitement que nous avons utilisée pour réaliser l'indexation sémantique souffre d'un certain nombre de faiblesses dont l'impact final sur les performances de la recherche est délicat à estimer. Notamment notre mise en œuvre pourrait être améliorée en matière de mise en correspondance des informations ressources-documents : par exemple en utilisant une segmentation adaptée à la ressource, afin d'identifier des séquences plus longues – multi-termes, voire expressions idiomatiques[7] – mais également en mettant en œuvre de réelles techniques de désambiguïsation sémantique (en utilisant par exemple le principe de la *densité conceptuelle* [Agirre et Rigau, 1996], qui semble donner de bons résultats en moyenne [Hotho *et al.*, 2003]), et limiter ainsi les appariements erronés dus à la polysémie.

Par ailleurs, l'importance exagérée accordée aux termes ambigus par le calcul de similarité mériterait sans doute d'être contrebalancé, afin d'éviter de trop fort rapprochements uniquement sur la base de termes ambigus communs.

Relevons pour finir que la technique de l'indexation sémantique, et en particulier le critère que nous avons proposé ici, peut être appliquée à bien d'autres tâches, telle que la classification automatique (supervisée ou non) ; en fait, à toute tâche nécessitant une phase d'indexation, y compris la visualisation et l'analyse de bases documentaires au moyen d'une analyse des correspondances, tels qu'abordée en fin de ce document.

[7] Et donc réaliser des appariements plus pertinents : il est bien connu qu'il est préférable d'identifier et mettre en relation les occurrences du multi-terme unique « pomme de terre » que des termes disjoints « pomme » et « terre ».

Annexe A

Détails des résultats en indexation sémantique

Nous présentons dans cette annexe le détails des évaluation des modèles à indexation sémantique obtenues par le biais d'expériences de RD sur les collections de références mentionnées au chapitre 4 [page 55].

Comme avec les résultats partiels présentés dans ce même chapitre 4, nous n'avons retenus ici que la mesure de « précision moyenne globale » (MAP), ainsi que les courbes *Précision-Rappel*.

Les désignations des expériences réalisées sont, sauf mention explicite, similaires à celles utilisées au chapitre 4.

A.1 Indexation par les lemmes (traditionnelle)

Pondération	*ADI*	*TIME*	*MED*	*CISI*	*CACM*
tf	0.2409	0.3133	0.3404	0.0507	0.1442
tf.idf	0.3431	0.5349	0.4470	0.1535	0.2814

A.2 Indexation par les lemmes et les concepts (expansion)

Pondération	*Ressource*	*WSD*	*ADI*	*TIME*	*MED*	*CISI*	*CACM*
tf	*EDR*	sans	0.1085	0.0298	0.0069	0.0142	0.0143
tf	*EDR*	+fréquent	0.1251	0.0307	0.0061	0.0108	0.0262
tf.idf	*EDR*	sans	0.2871	0.3908	0.2532	0.0875	0.1245
tf.idf	*EDR*	+fréquent	0.3840	0.4999	0.3984	0.1413	0.2324

Avec « *WSD* sans » la conservation de tous les sens des mots et concepts polysémiques, et « *WSD*+fréquent » le pis-aller de désambigüisation sémantique, ne conservant pour chaque mot que le concept associé au sens le plus fréquent, dans l'absolu.

A.3 Indexation par les concepts associés aux mots (synsets)

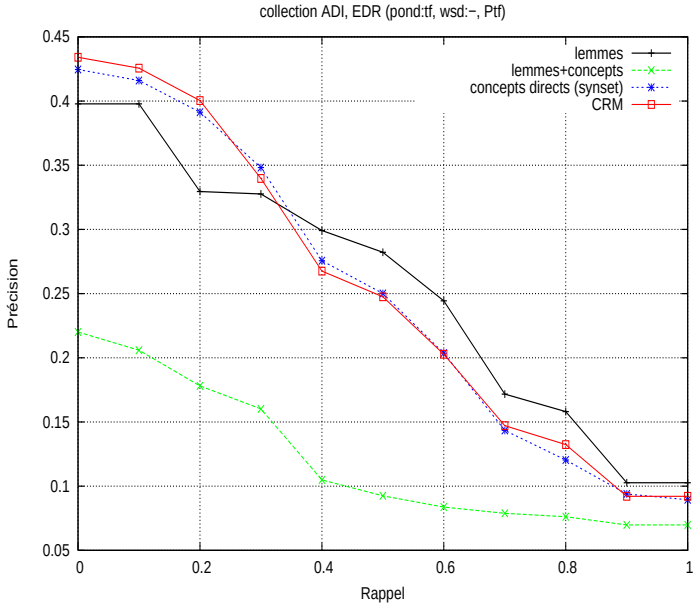
Pondération	*Ressource*	*WSD*	*ADI*	*TIME*	*MED*	*CISI*	*CACM*
tf	*EDR*	sans	0.2343	0.2185	0.1677	0.0407	0.1425
tf	*EDR*	+fréquent	0.2828	0.3576	0.2930	0.0617	0.1423
tf	*WordNet*	sans	0.2494	0.2458	0.2575	0.0428	0.1502
tf	*WordNet*	+fréquent	0.2552	0.3445	0.3471	0.0853	0.1327
tf.idf	*EDR*	sans	0.3140	0.4284	0.2704	0.0959	0.1869
tf.idf	*EDR*	+fréquent	0.4071	0.5442	0.4284	0.1632	0.2774
tf.idf	*WordNet*	sans	0.3370	0.4927	0.4051	0.1158	0.2277
tf.idf	*WordNet*	+fréquent	0.3858	0.5408	0.4611	0.1581	0.2607

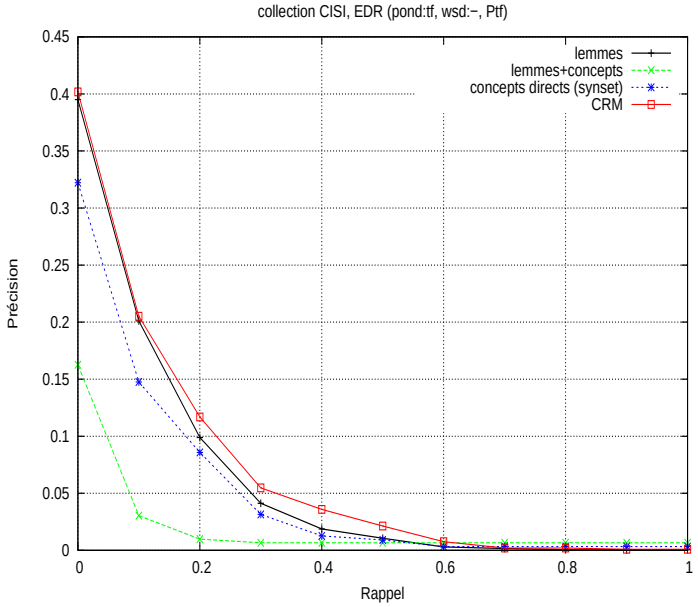
A.4 Indexation par la coupe de redondance minimale (*CRM*)

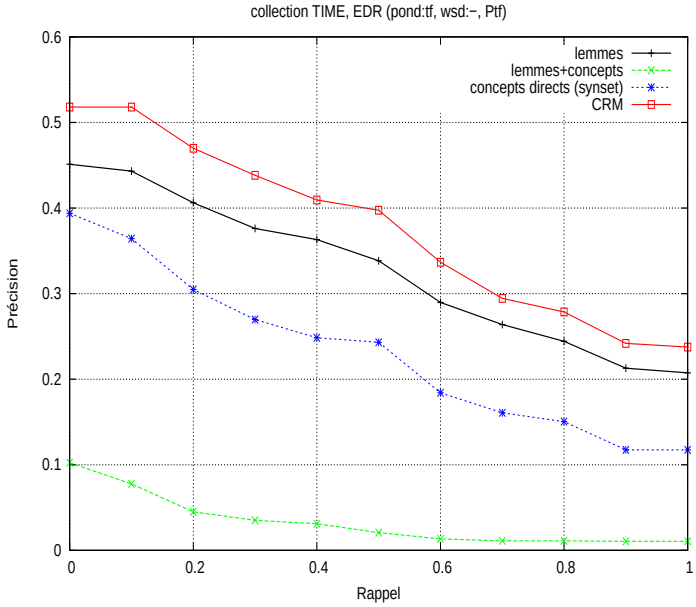
Pond	*Ressource*	*WSD*	*termes suppl.*	$\widehat{\theta_s}$	*ADI*	*TIME*	*MED*	*CISI*	*CACM*
tf	*EDR*	*sans*	*standard*	$P_{\mathsf{tf}}(s\|\Gamma)$	0.2310	0.3625	0.1521	0.0579	0.1264
tf	*EDR*	*sans*	*standard*	$P_{\mathsf{tfidf}}(s\|\Gamma)$	0.2891	0.3797	0.3141	0.0651	0.1388
tf	*EDR*	*sans*	*concept*	$P_{\mathsf{tf}}(s\|\Gamma)$	0.2307	0.3623	0.1520	0.0596	0.1257
tf	*EDR*	*sans*	*concept*	$P_{\mathsf{tfidf}}(s\|\Gamma)$	0.2958	0.3795	0.3143	0.0739	0.1377
tf	*EDR*	*sans*	*hyperonyme*	$P_{\mathsf{tf}}(s\|\Gamma)$	0.2246	0.3623	0.1511	0.0587	0.1242
tf	*EDR*	*sans*	*hyperonyme*	$P_{\mathsf{tfidf}}(s\|\Gamma)$	0.2948	0.3795	0.3041	0.0721	0.1399
tf	*EDR*	*+fréquent*	*standard*	$P_{\mathsf{tf}}(s\|\Gamma)$	0.2819	0.3852	0.3014	0.0633	0.1472
tf	*EDR*	*+fréquent*	*standard*	$P_{\mathsf{tfidf}}(s\|\Gamma)$	0.2813	0.3818	0.3138	0.0650	0.1489
tf	*EDR*	*+fréquent*	*concept*	$P_{\mathsf{tf}}(s\|\Gamma)$	0.2834	0.3852	0.3014	0.0633	0.1439
tf	*EDR*	*+fréquent*	*concept*	$P_{\mathsf{tfidf}}(s\|\Gamma)$	0.2828	0.3818	0.3138	0.0650	0.1456
tf	*EDR*	*+fréquent*	*hyperonyme*	$P_{\mathsf{tf}}(s\|\Gamma)$	0.2630	0.3852	0.2934	0.0626	0.1427
tf	*EDR*	*+fréquent*	*hyperonyme*	$P_{\mathsf{tfidf}}(s\|\Gamma)$	0.2762	0.3818	0.3086	0.0647	0.1447
tf	*WordNet*	*sans*	*standard*	$P_{\mathsf{tf}}(s\|\Gamma)$	0.2463	0.2858	0.1732	0.0327	0.1027
tf	*WordNet*	*sans*	*concept*	$P_{\mathsf{tf}}(s\|\Gamma)$	0.2499	0.2858	0.1733	0.0327	0.1026
tf	*WordNet*	*sans*	*hyperonyme*	$P_{\mathsf{tf}}(s\|\Gamma)$	0.2494	0.2859	0.1746	0.0327	0.1021
tf	*WordNet*	*+fréquent*	*standard*	$P_{\mathsf{tf}}(s\|\Gamma)$	0.2834	0.3323	0.3141	0.0591	0.1499
tf	*WordNet*	*+fréquent*	*concept*	$P_{\mathsf{tf}}(s\|\Gamma)$	0.2858	0.3323	0.3158	0.0591	0.1500
tf	*WordNet*	*+fréquent*	*hyperonyme*	$P_{\mathsf{tf}}(s\|\Gamma)$	0.2831	0.3323	0.3145	0.0592	0.1478
tf.idf	*EDR*	*sans*	*standard*	$P_{\mathsf{tf}}(s\|\Gamma)$	0.2976	0.5476	0.3679	0.1300	0.2532
tf.idf	*EDR*	*sans*	*standard*	$P_{\mathsf{tfidf}}(s\|\Gamma)$	0.3415	0.5491	0.4241	0.1525	0.2303
tf.idf	*EDR*	*sans*	*concept*	$P_{\mathsf{tf}}(s\|\Gamma)$	0.2973	0.5469	0.3667	0.1408	0.2529
tf.idf	*EDR*	*sans*	*concept*	$P_{\mathsf{tfidf}}(s\|\Gamma)$	0.3477	0.5484	0.4189	0.1580	0.2303
tf.idf	*EDR*	*sans*	*hyperonyme*	$P_{\mathsf{tf}}(s\|\Gamma)$	0.2785	0.5469	0.3375	0.1324	0.2342
tf.idf	*EDR*	*sans*	*hyperonyme*	$P_{\mathsf{tfidf}}(s\|\Gamma)$	0.3506	0.5484	0.3761	0.1468	0.2223
tf.idf	*EDR*	*+fréquent*	*standard*	$P_{\mathsf{tf}}(s\|\Gamma)$	0.4071	0.5471	0.4226	0.1489	0.2950
tf.idf	*EDR*	*+fréquent*	*standard*	$P_{\mathsf{tfidf}}(s\|\Gamma)$	0.4155	0.5501	0.4300	0.1540	0.3018
tf.idf	*EDR*	*+fréquent*	*concept*	$P_{\mathsf{tf}}(s\|\Gamma)$	0.4058	0.5468	0.4226	0.1488	0.2946
tf.idf	*EDR*	*+fréquent*	*concept*	$P_{\mathsf{tfidf}}(s\|\Gamma)$	0.4144	0.5498	0.4300	0.1539	0.3015
tf.idf	*EDR*	*+fréquent*	*hyperonyme*	$P_{\mathsf{tf}}(s\|\Gamma)$	0.3466	0.5468	0.3831	0.1422	0.2674
tf.idf	*EDR*	*+fréquent*	*hyperonyme*	$P_{\mathsf{tfidf}}(s\|\Gamma)$	0.3589	0.5498	0.3938	0.1484	0.2767
tf.idf	*WordNet*	*sans*	*standard*	$P_{\mathsf{tf}}(s\|\Gamma)$	0.3353	0.5492	0.3330	0.0688	0.1808
tf.idf	*WordNet*	*sans*	*concept*	$P_{\mathsf{tf}}(s\|\Gamma)$	0.3272	0.5484	0.3337	0.0688	0.1809
tf.idf	*WordNet*	*sans*	*hyperonyme*	$P_{\mathsf{tf}}(s\|\Gamma)$	0.3163	0.5541	0.3359	0.0688	0.1789
tf.idf	*WordNet*	*+fréquent*	*standard*	$P_{\mathsf{tf}}(s\|\Gamma)$	0.3766	0.5594	0.4535	0.1520	0.2852
tf.idf	*WordNet*	*+fréquent*	*concept*	$P_{\mathsf{tf}}(s\|\Gamma)$	0.3875	0.5592	0.4500	0.1519	0.2853
tf.idf	*WordNet*	*+fréquent*	*hyperonyme*	$P_{\mathsf{tf}}(s\|\Gamma)$	0.3709	0.5592	0.4414	0.1517	0.2755

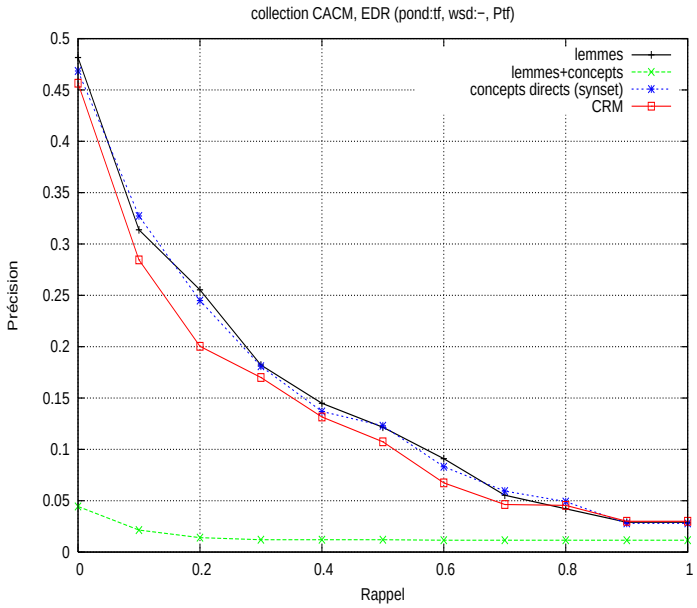
A.5 Courbes PR, cmp. des indexations, paramètres de base

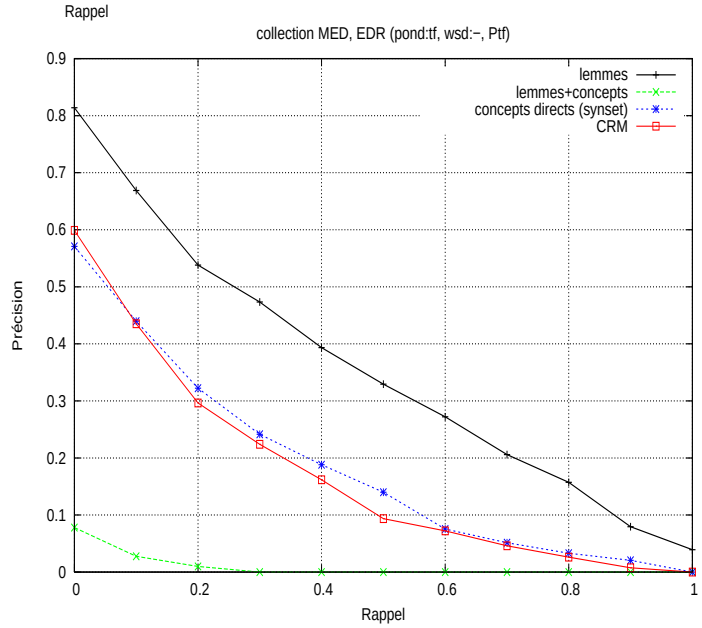
Sans pondération



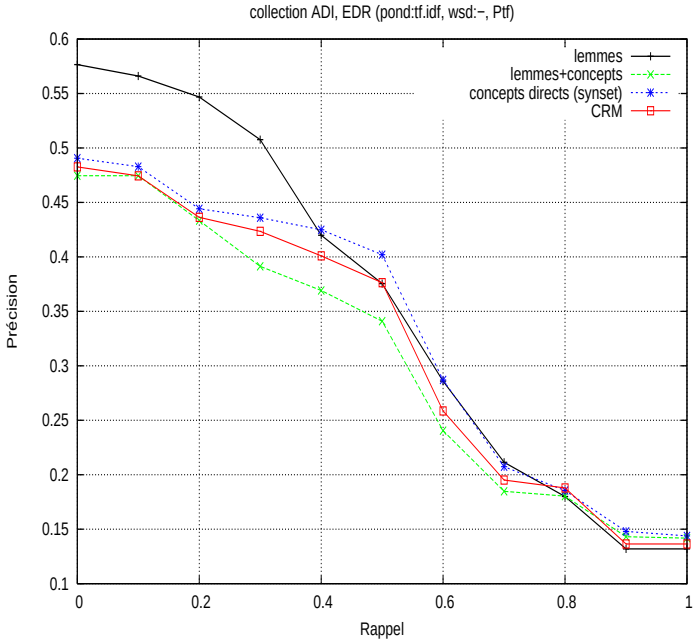


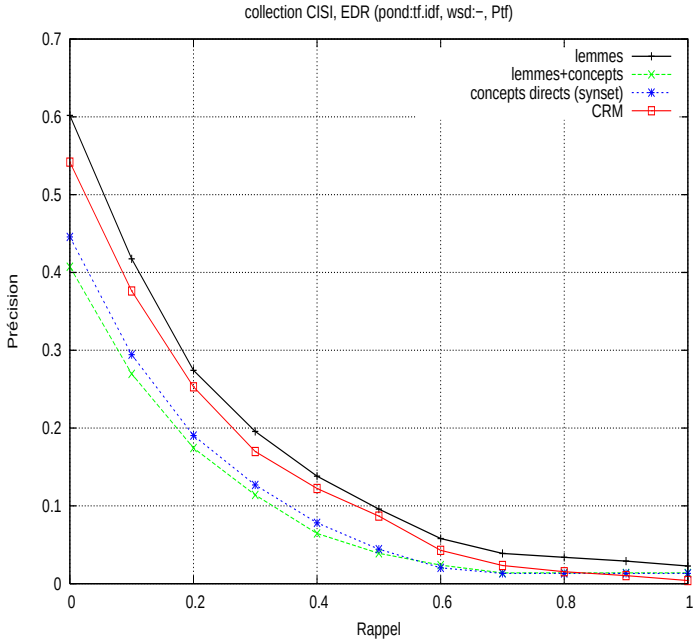


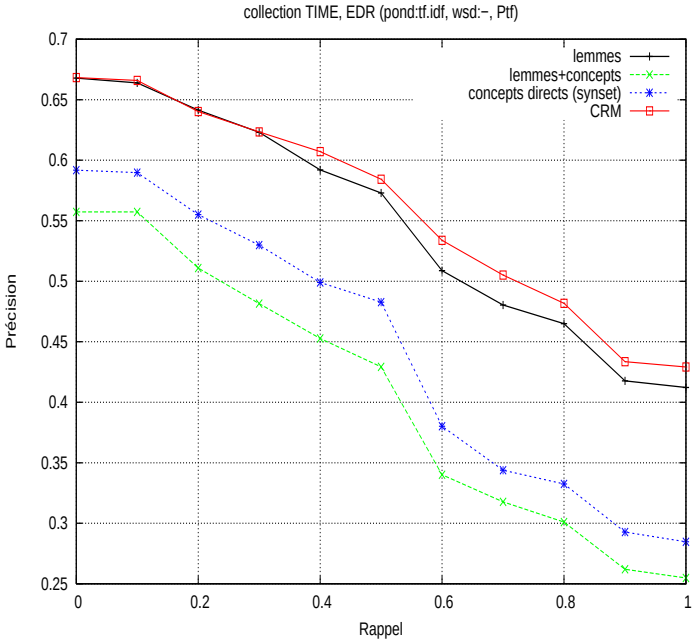


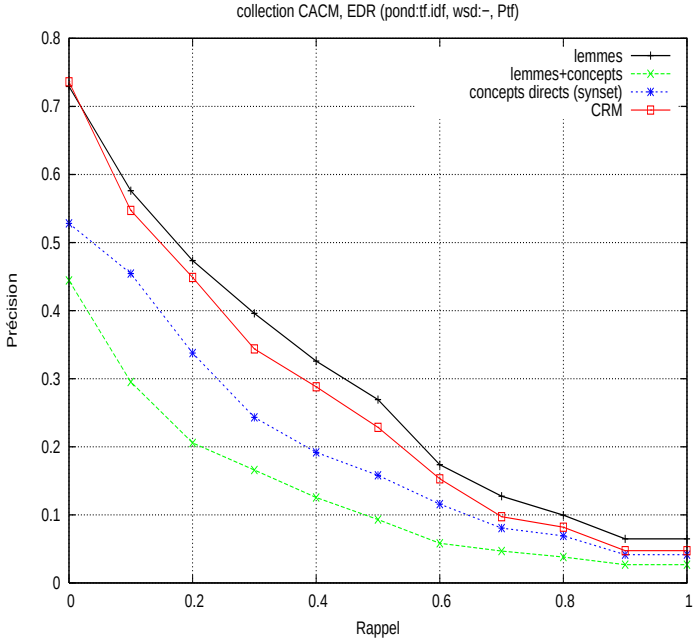


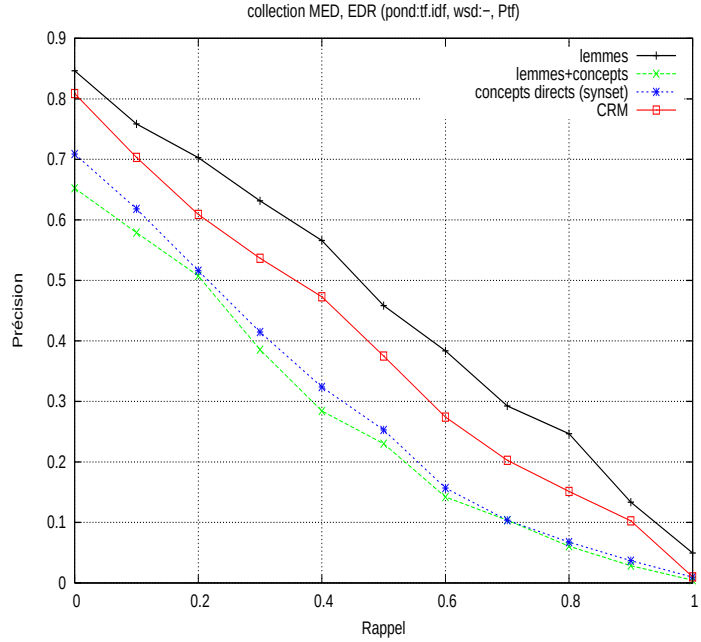
Pondération inverse en documents





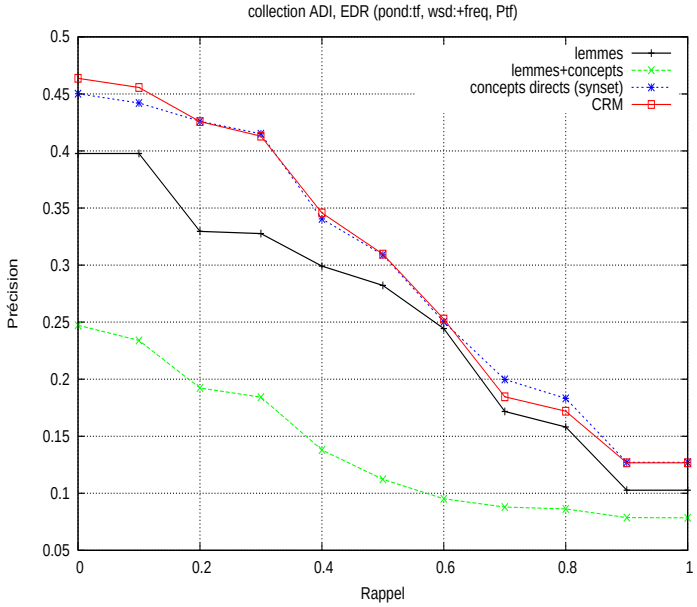


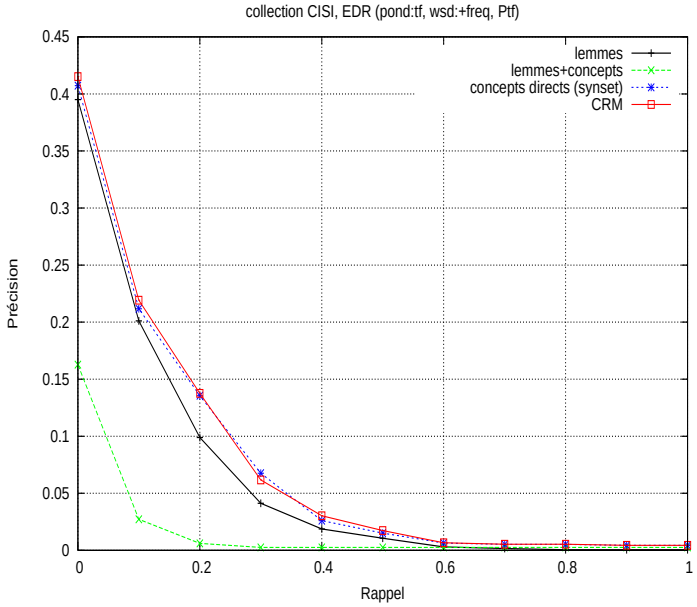


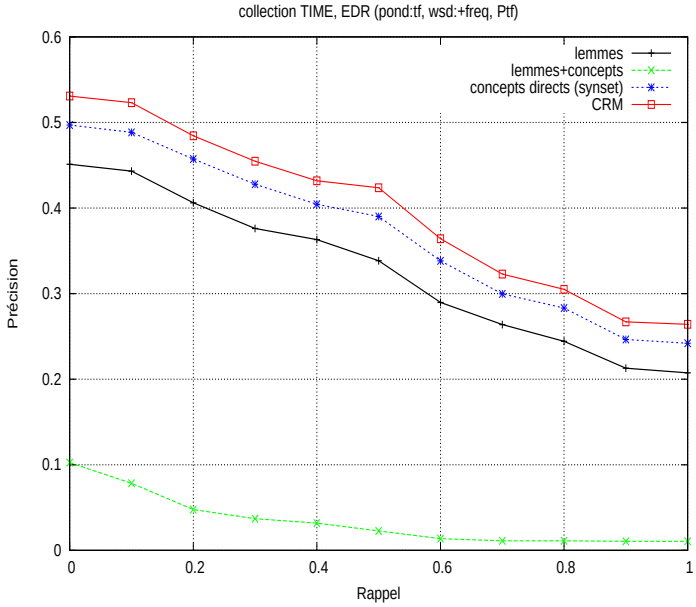


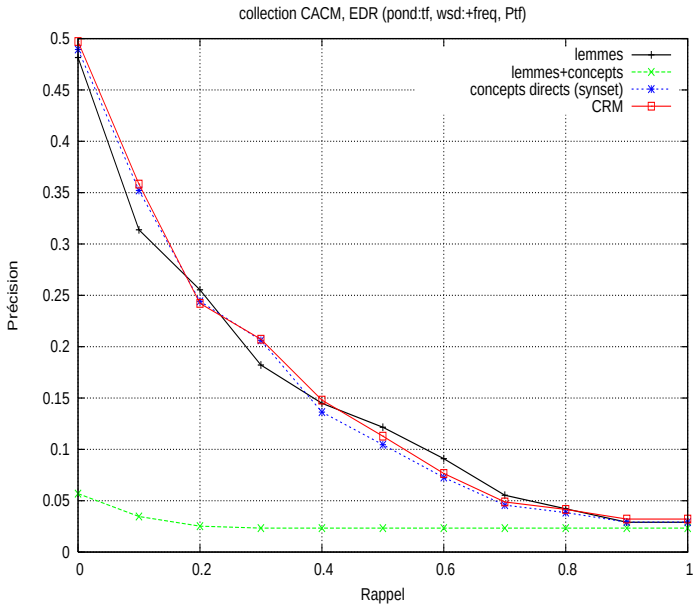
A.6 Courbes PR, cmp. des indexations, sens le plus fréquent

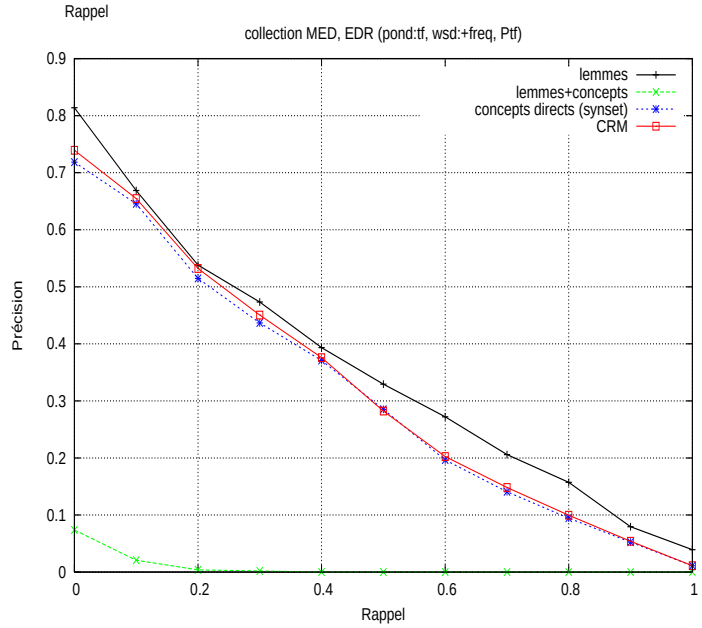
Sans pondération



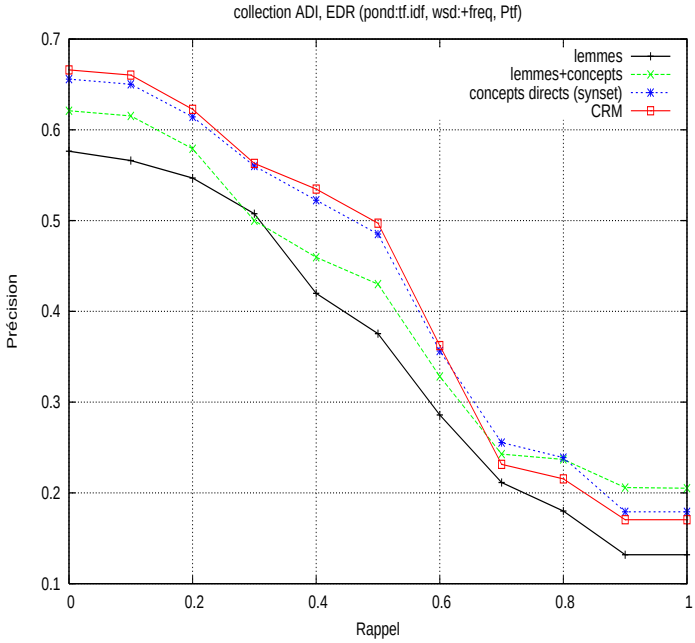


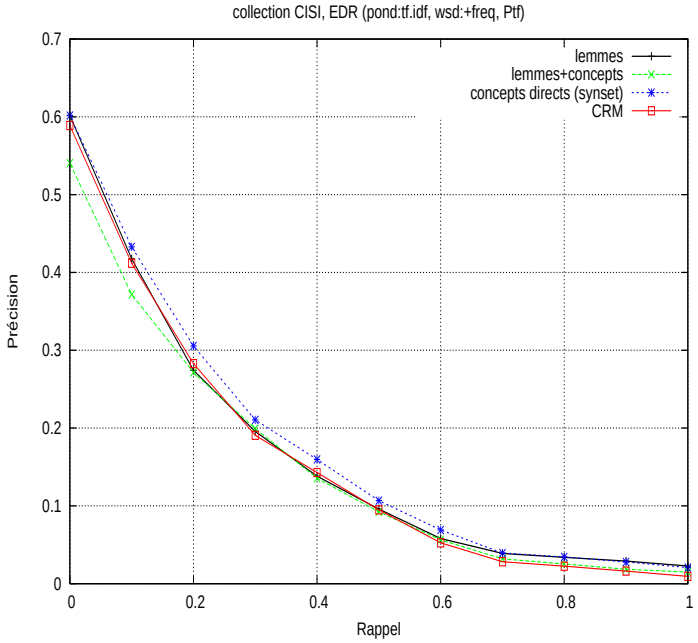


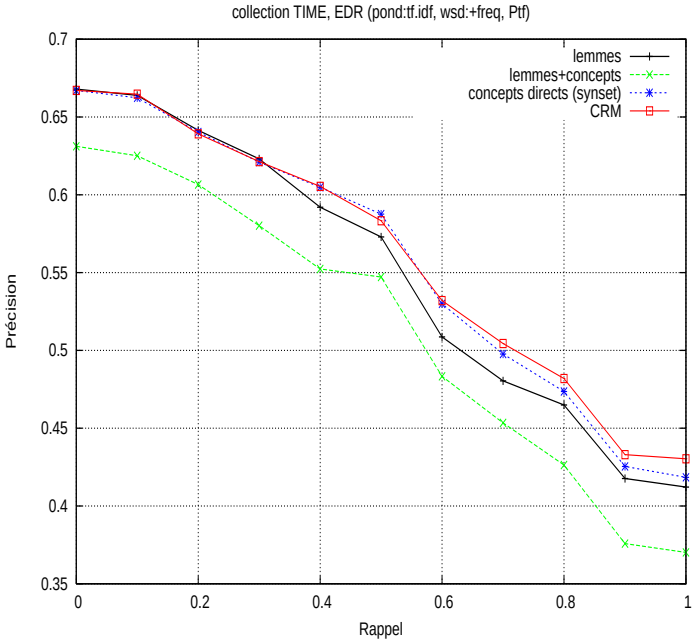


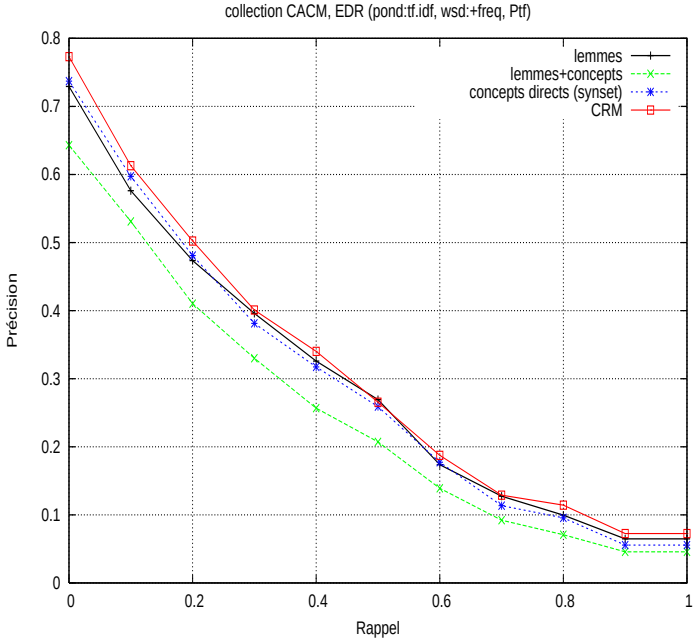


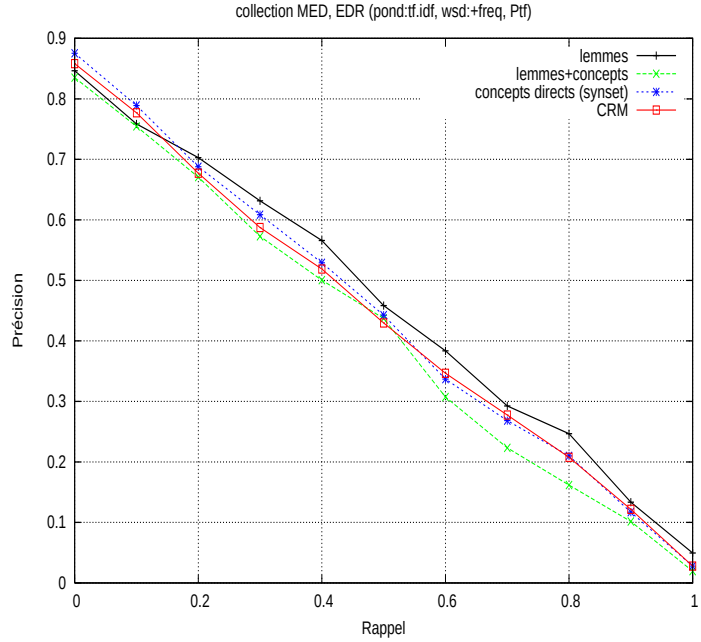
Pondération inverse en documents





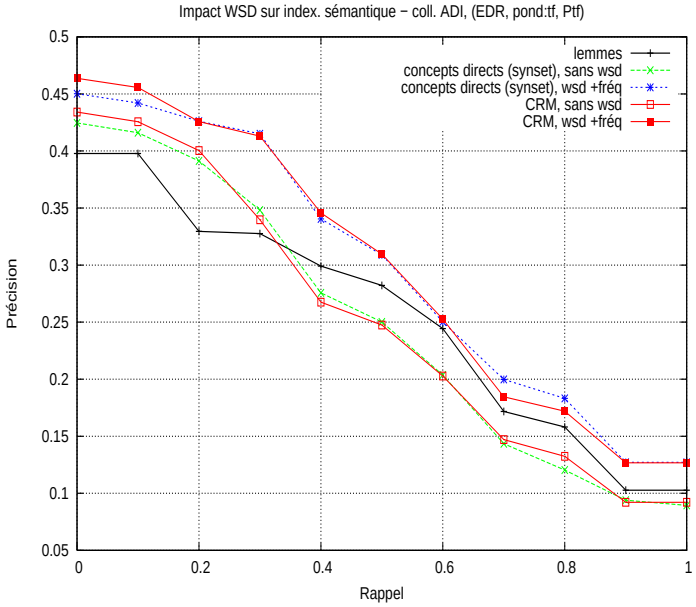


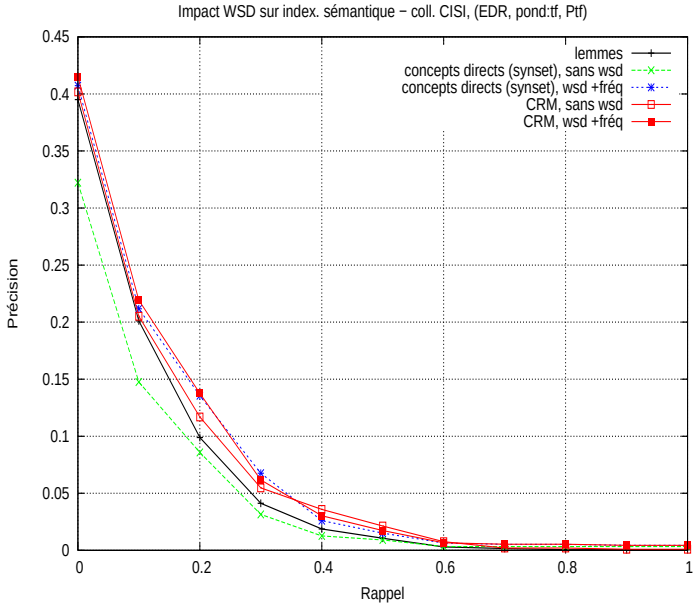


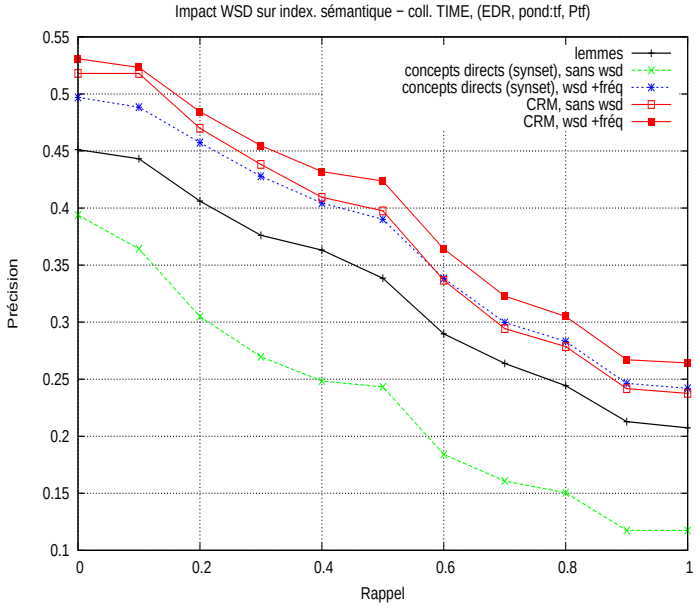


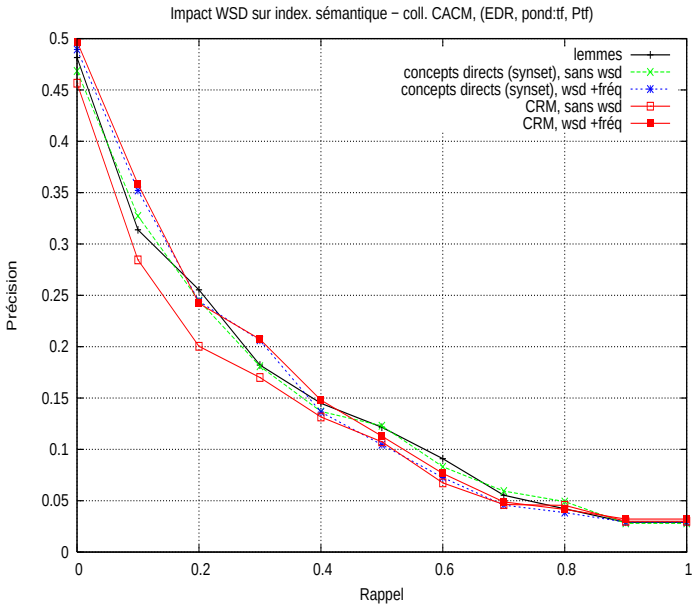
A.7 Courbes PR, influence de la polysémie

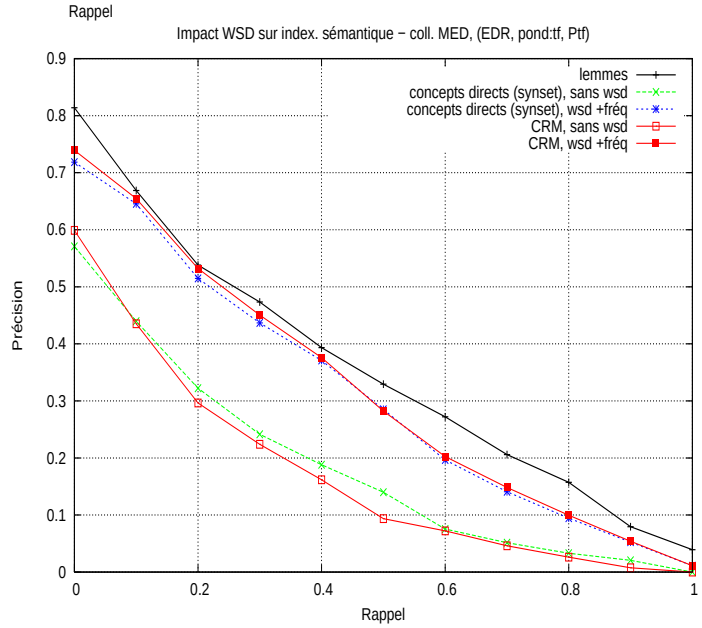
Sans pondération



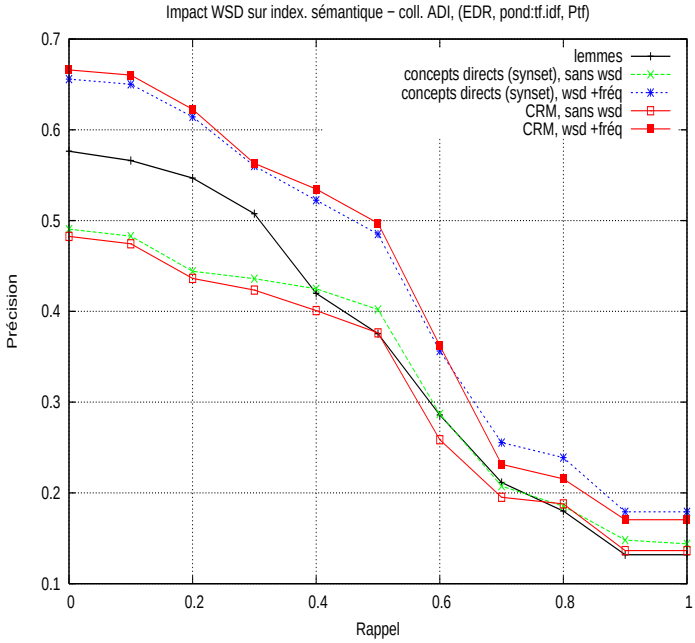


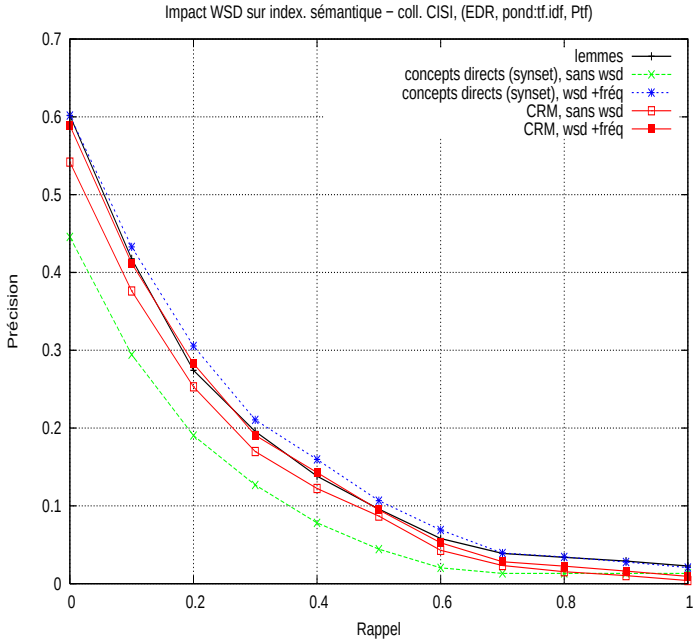


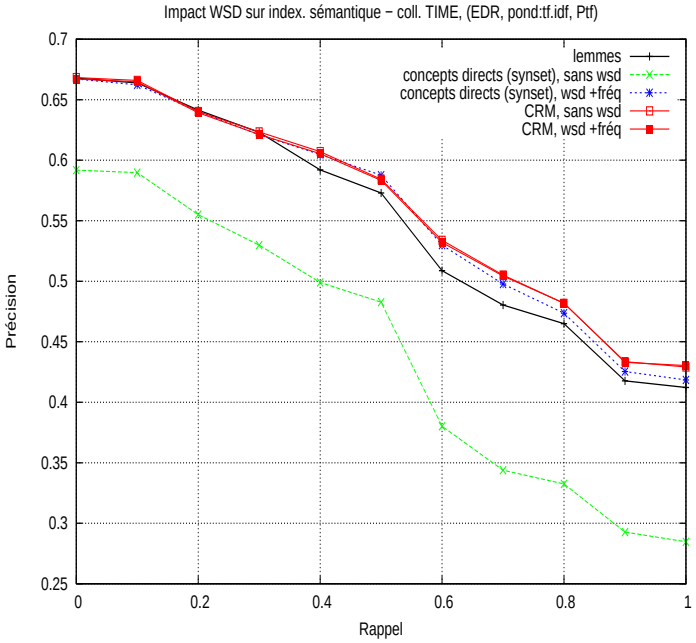


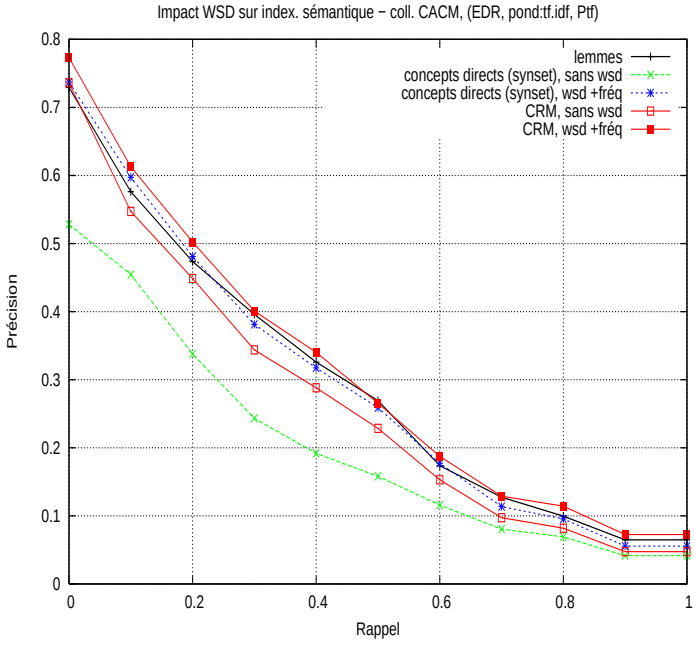


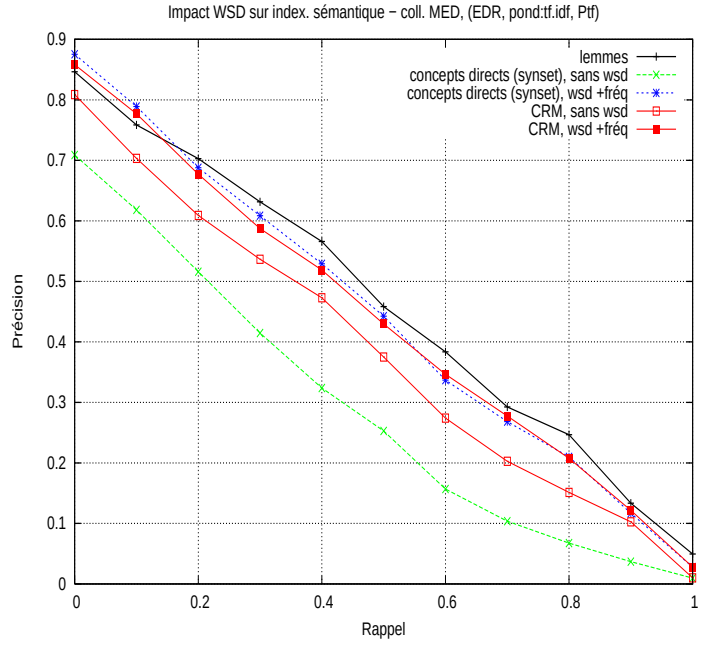
Pondération inverse en documents





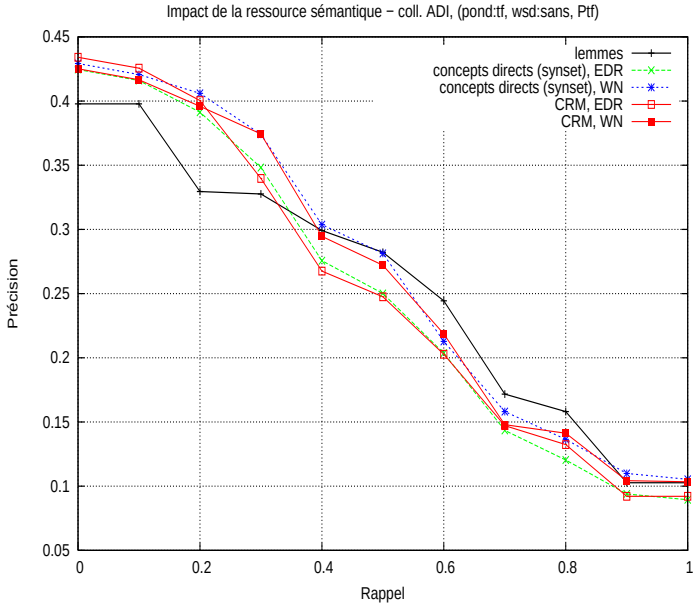


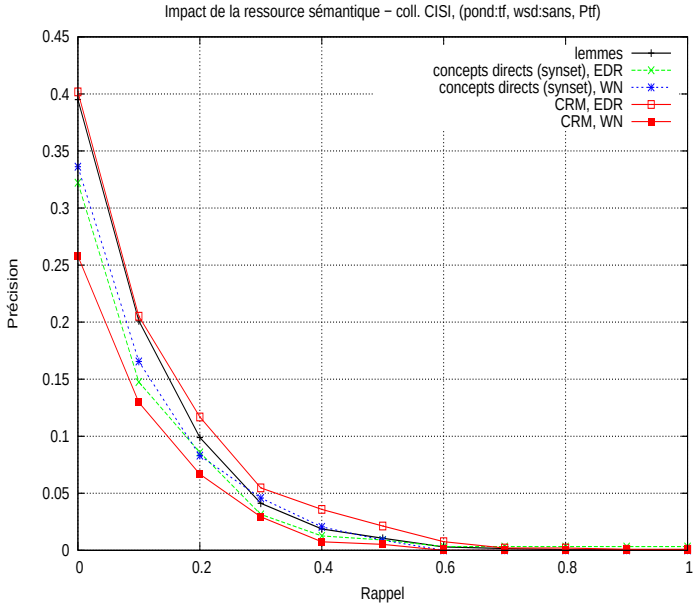


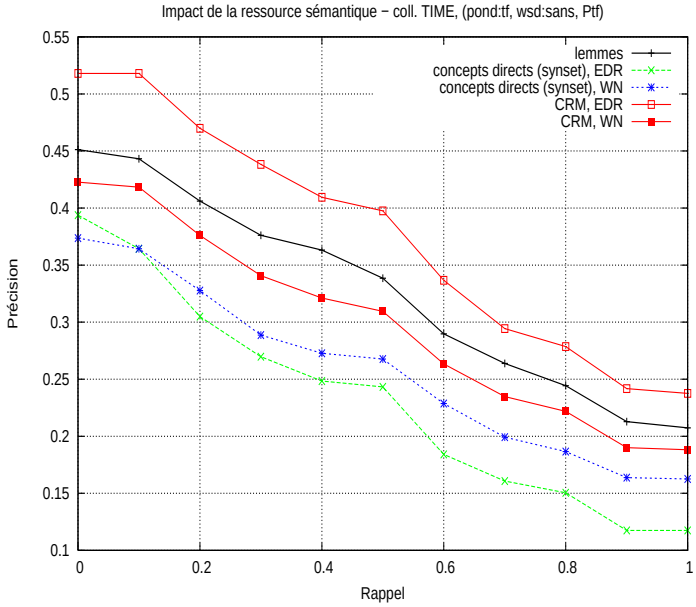


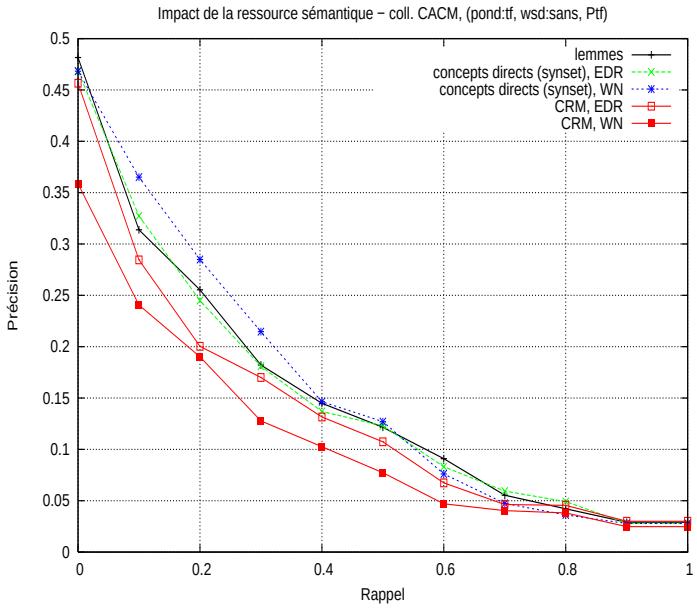
A.8 Courbes PR, influence de la ressource sémantique

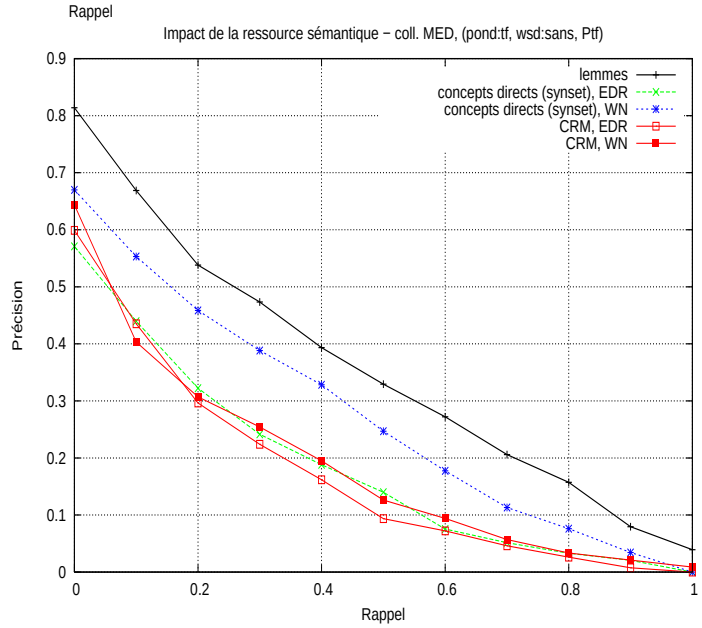
Sans pondération, conservation de la polysémie



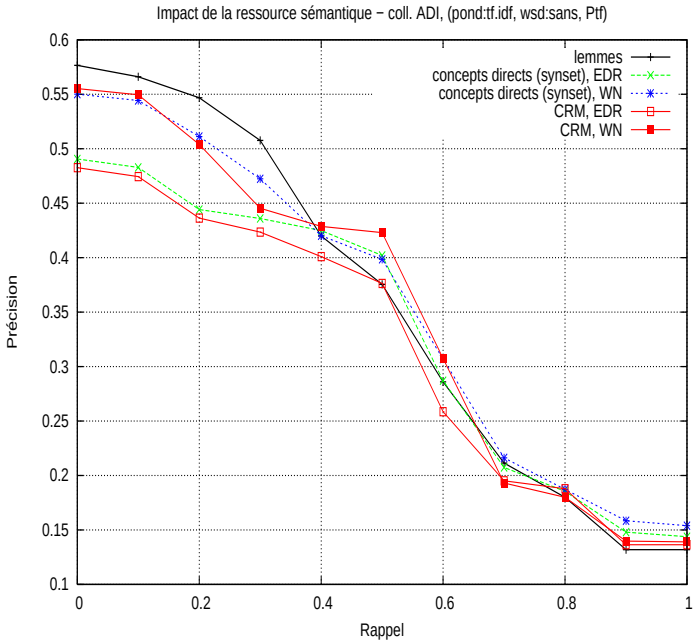


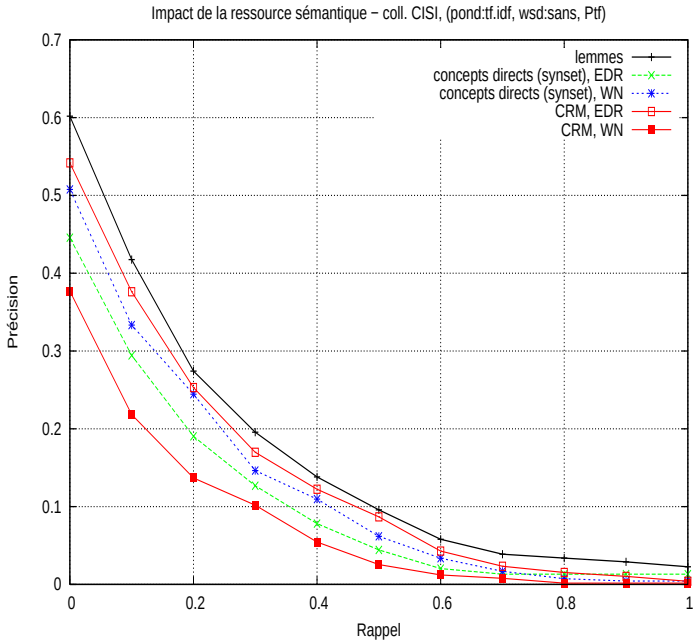


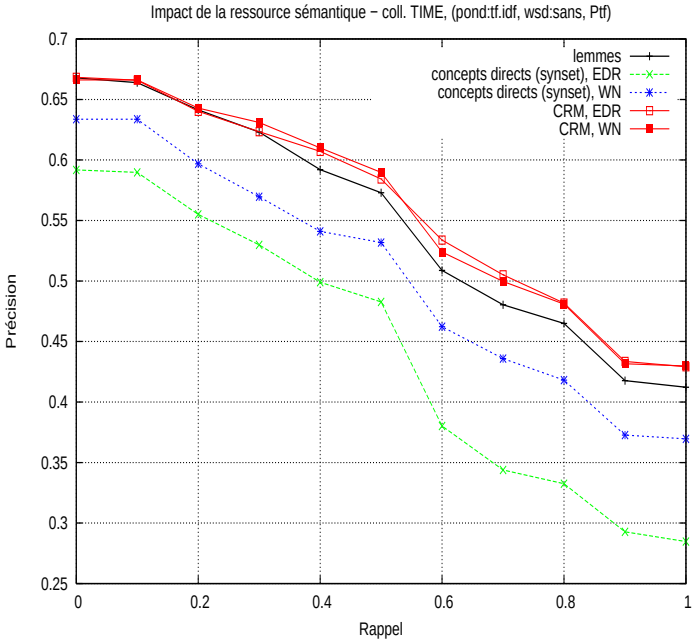


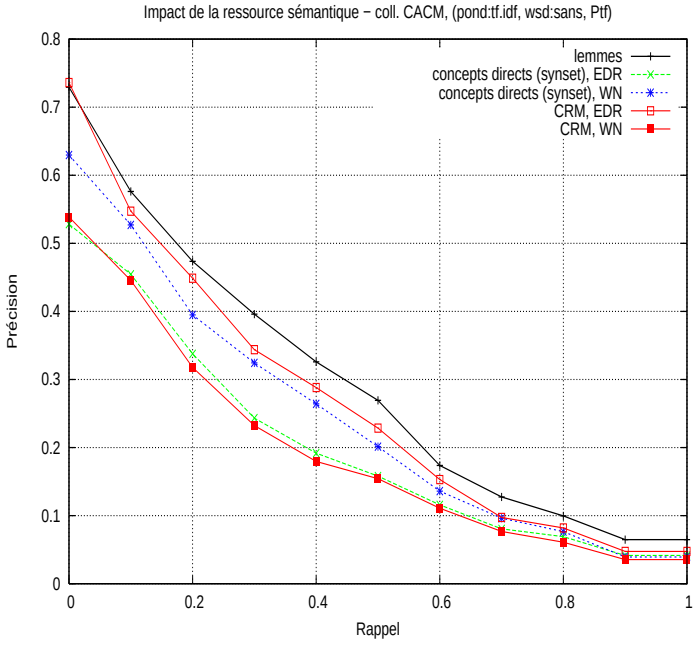


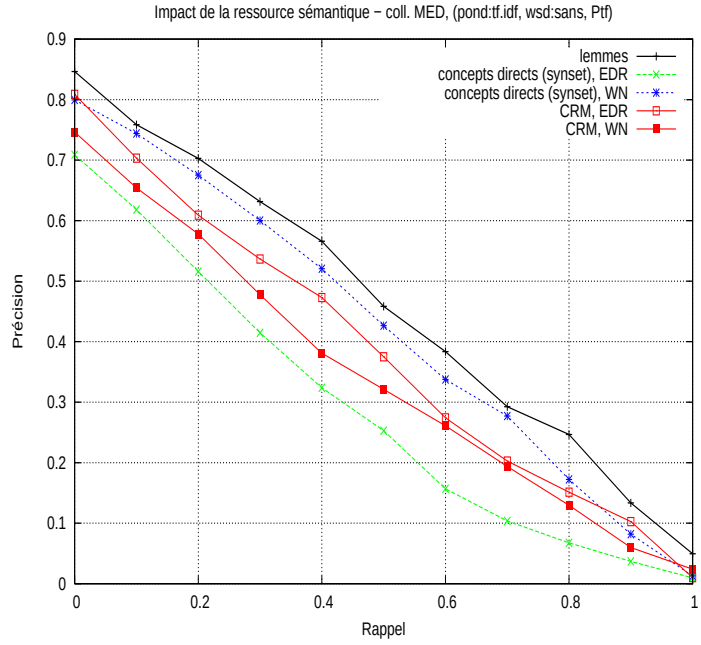
Pondération inverse en document, conservation de la polysémie



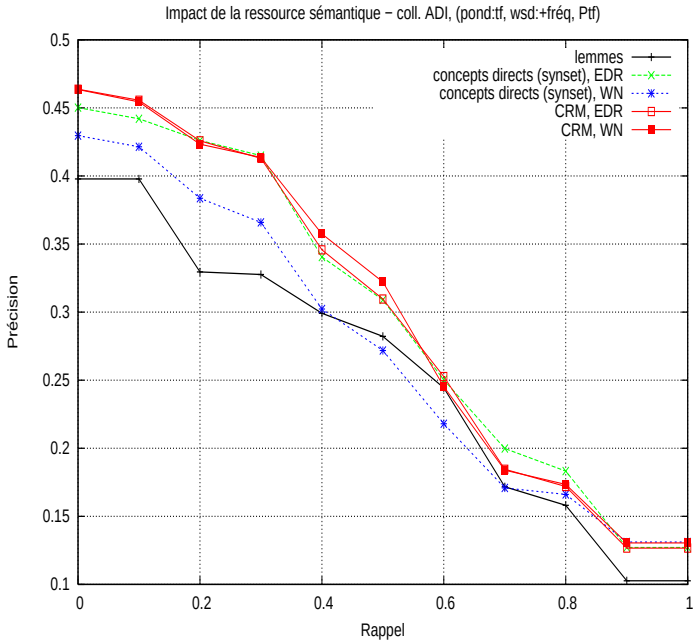


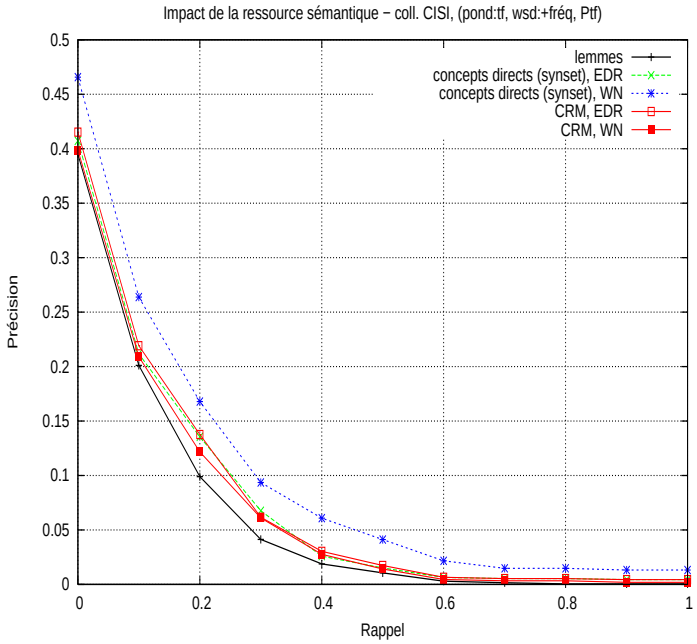


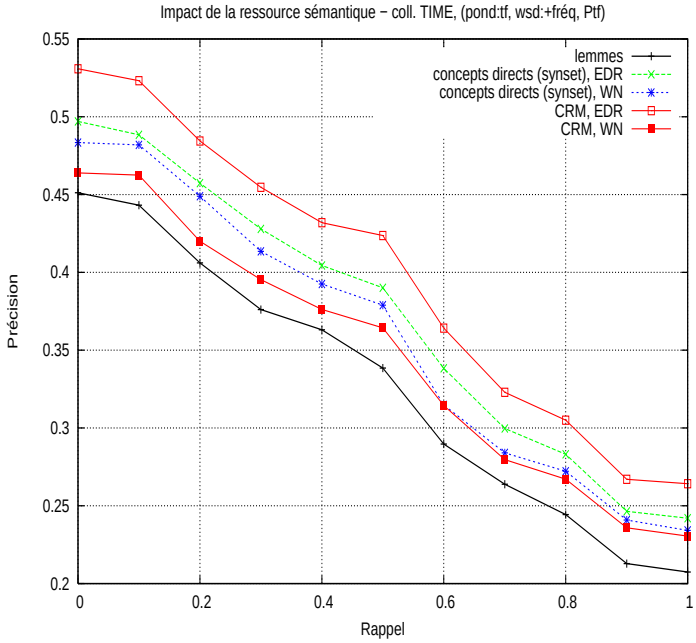


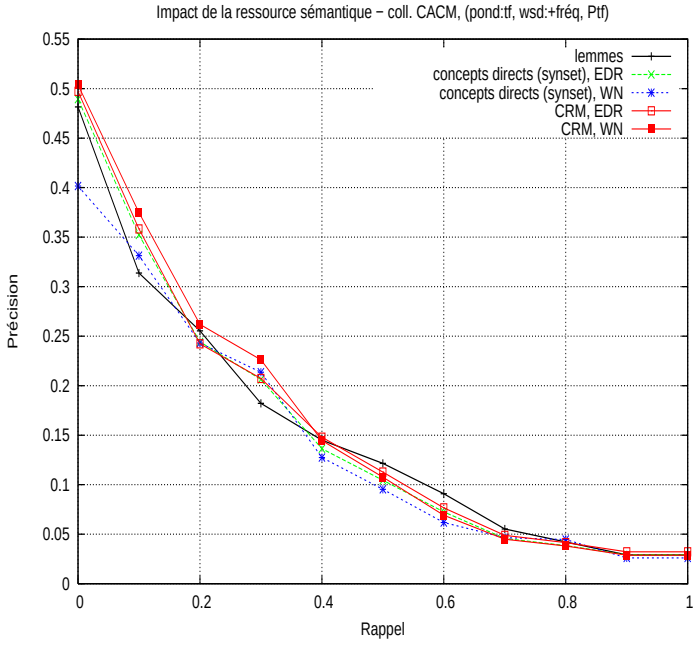


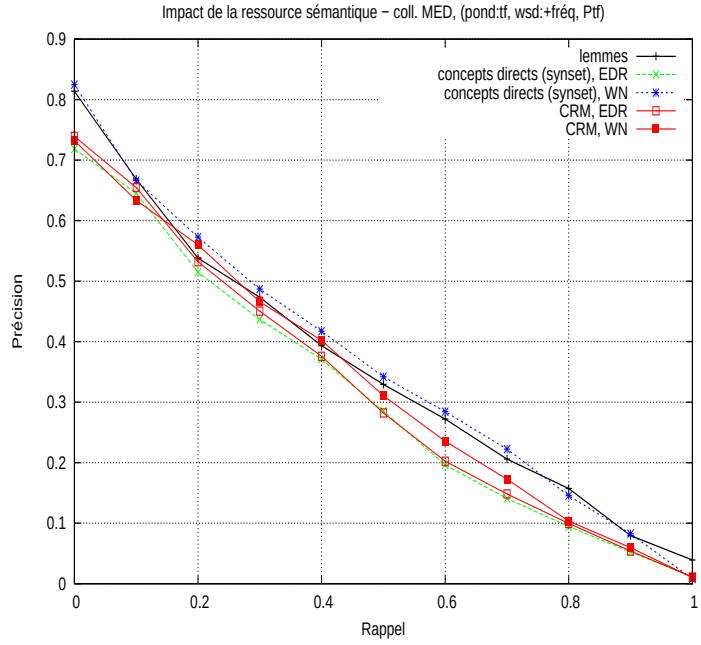
Sans pondération, suppression de la polysémie



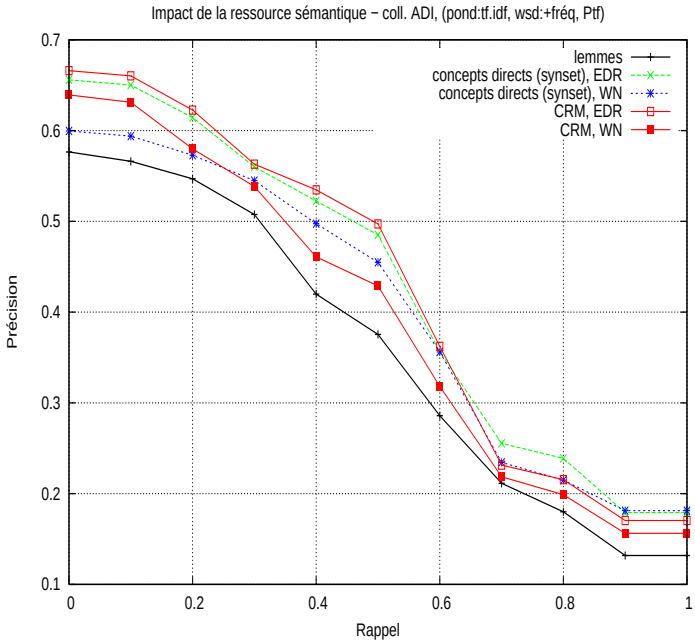


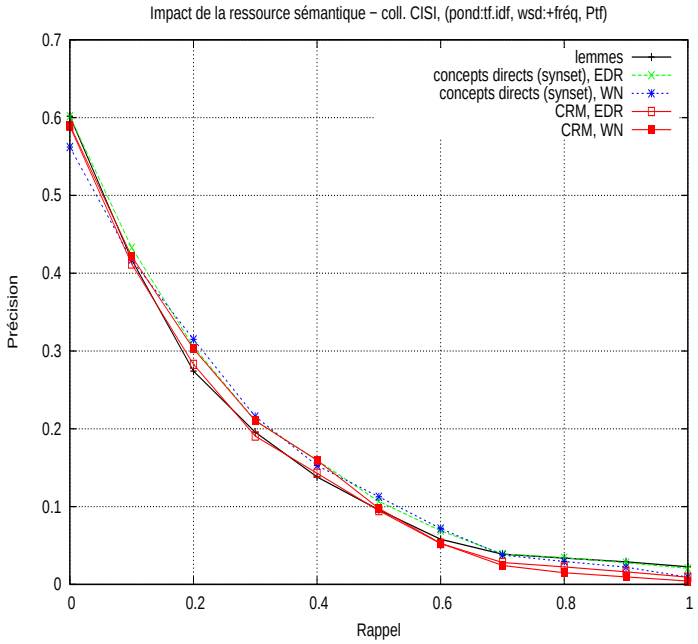


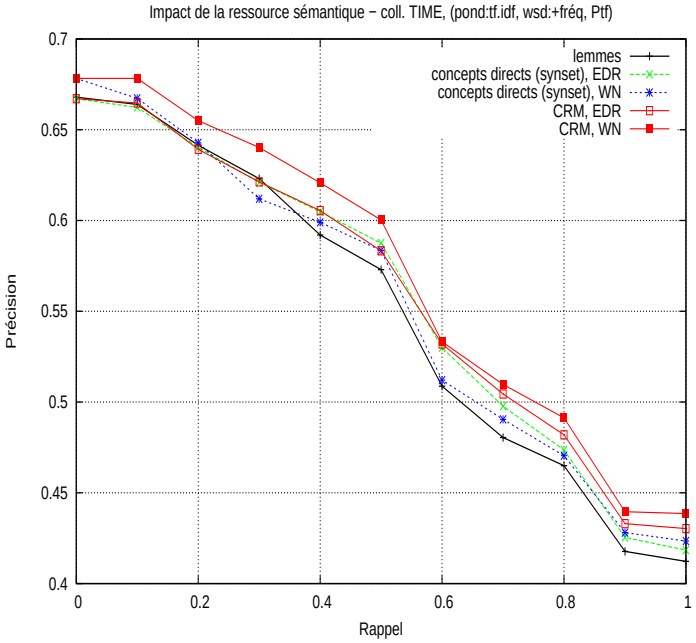


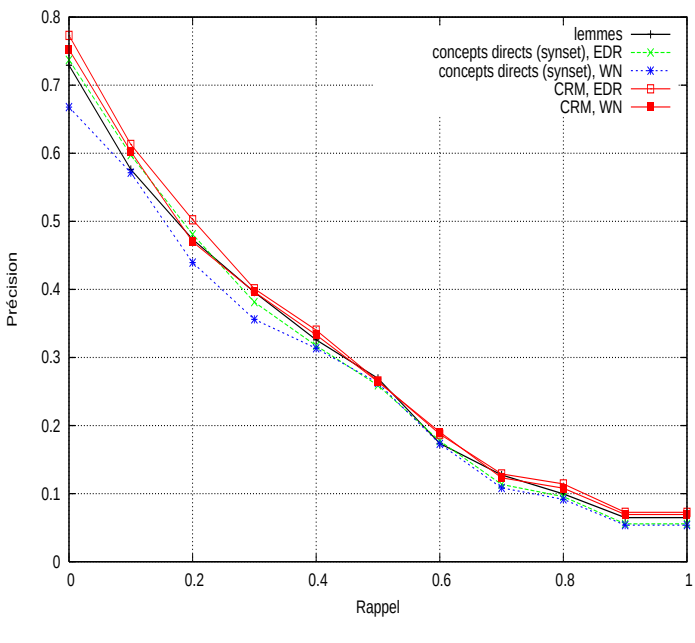


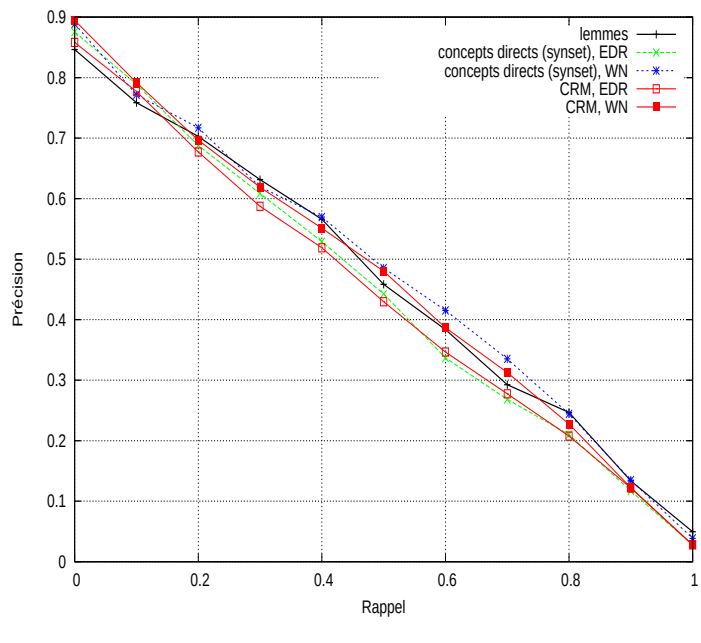
Pondération inverse en document, suppression de la polysémie





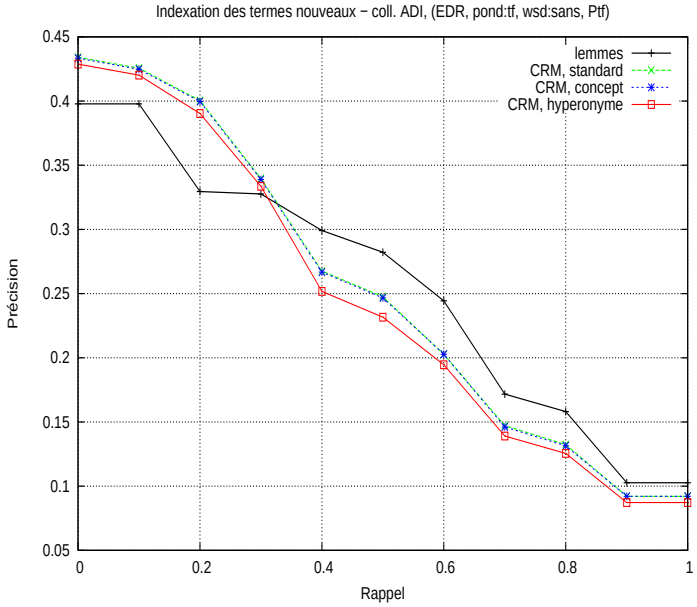


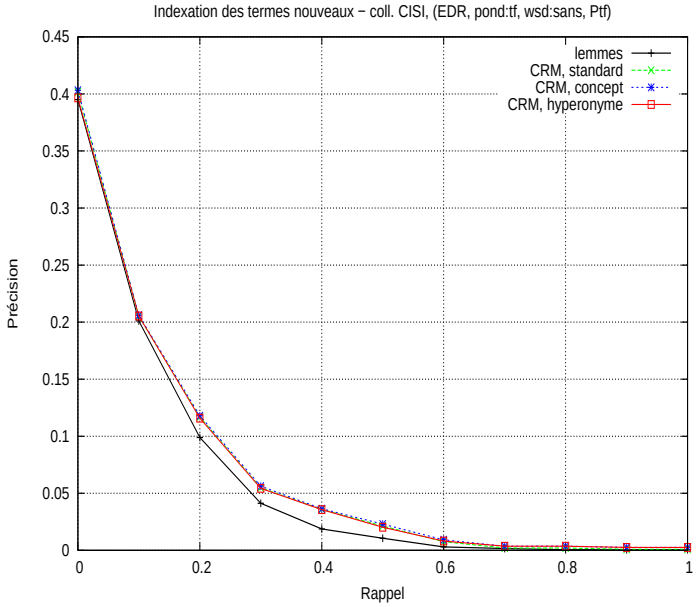


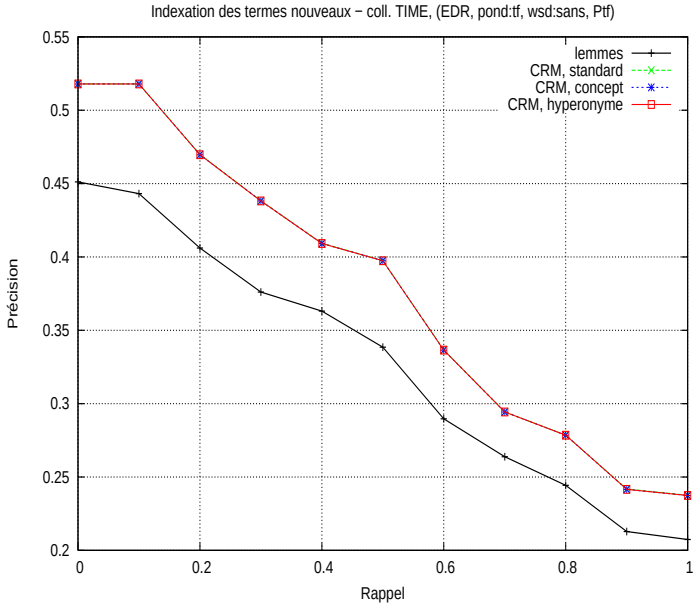


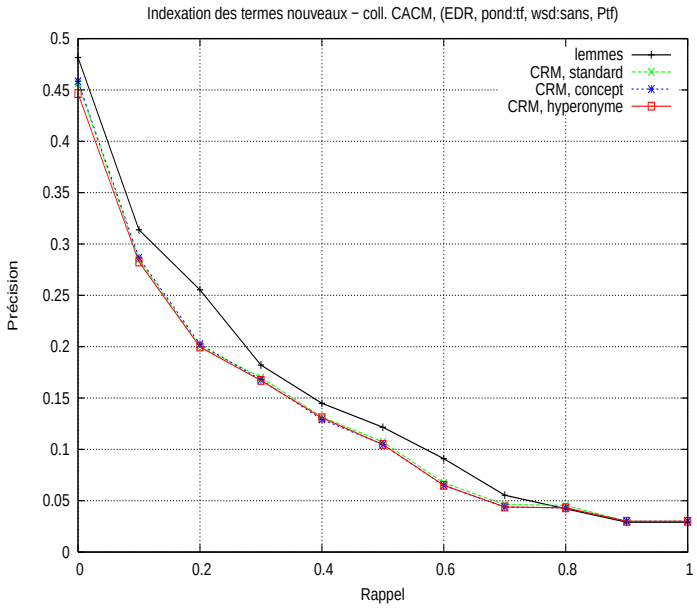
A.9 Courbes PR, indexation des termes nouveaux

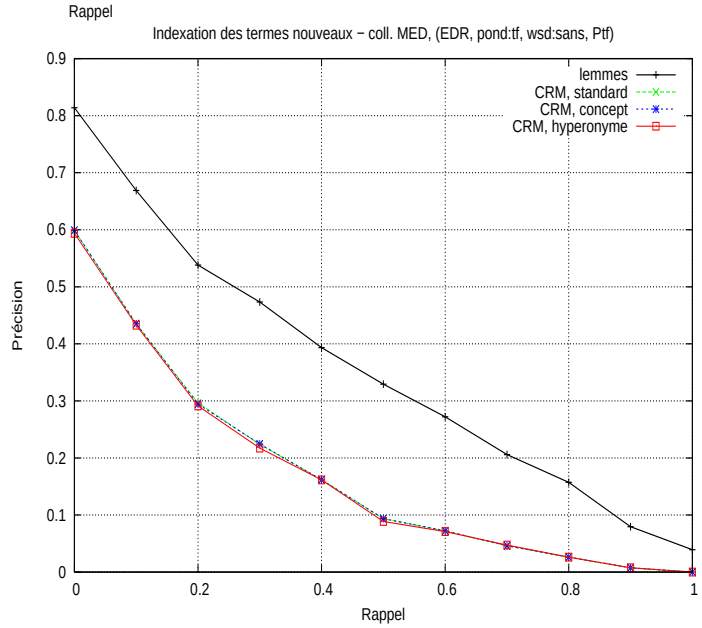
Sans pondération, conservation de la polysémie



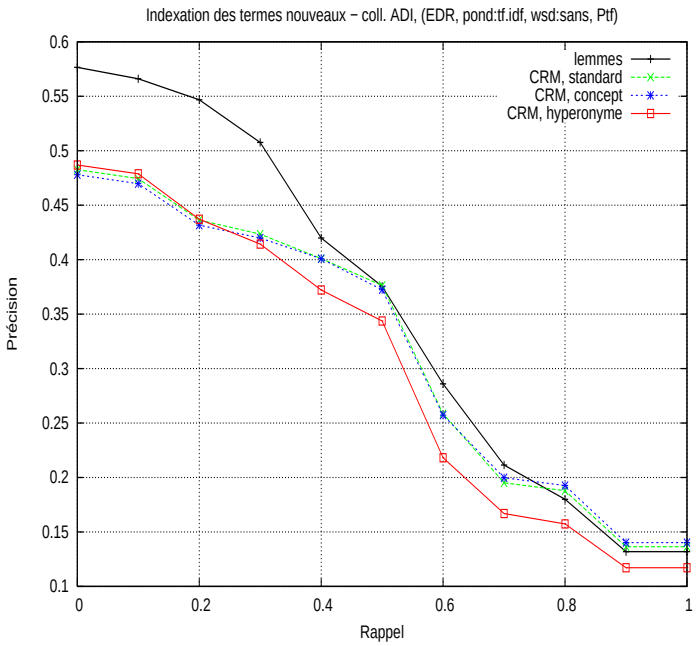


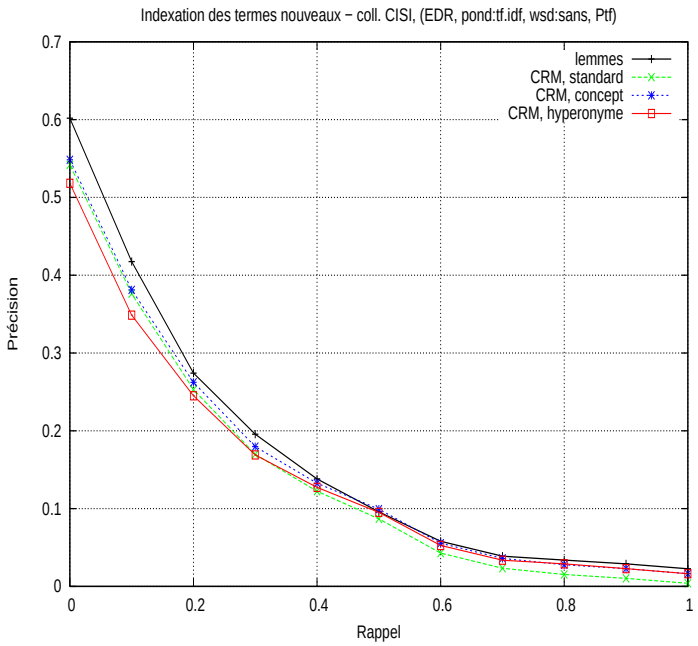


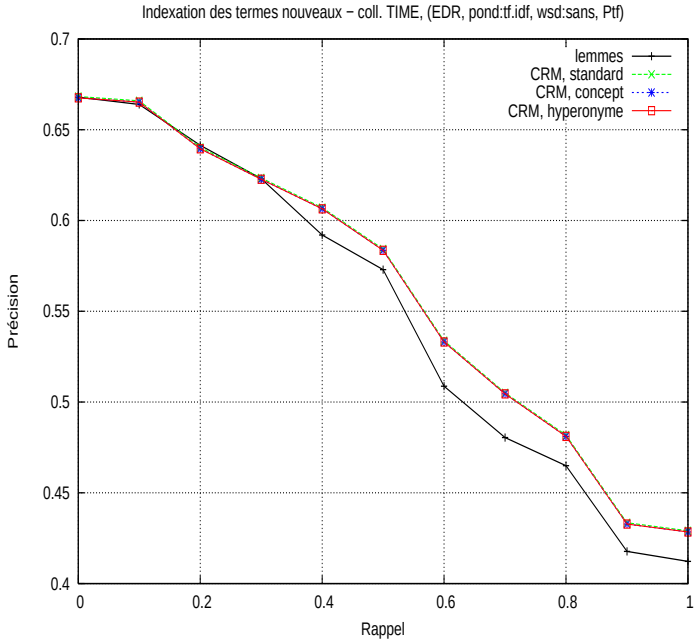


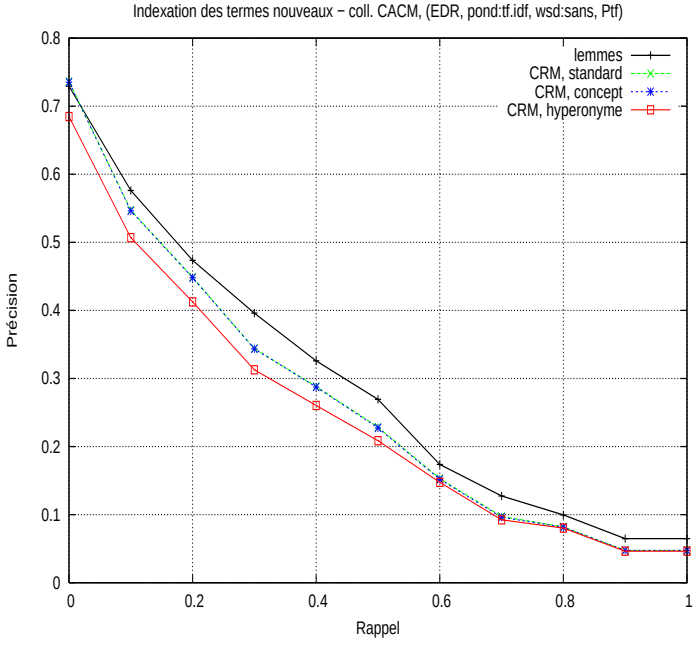


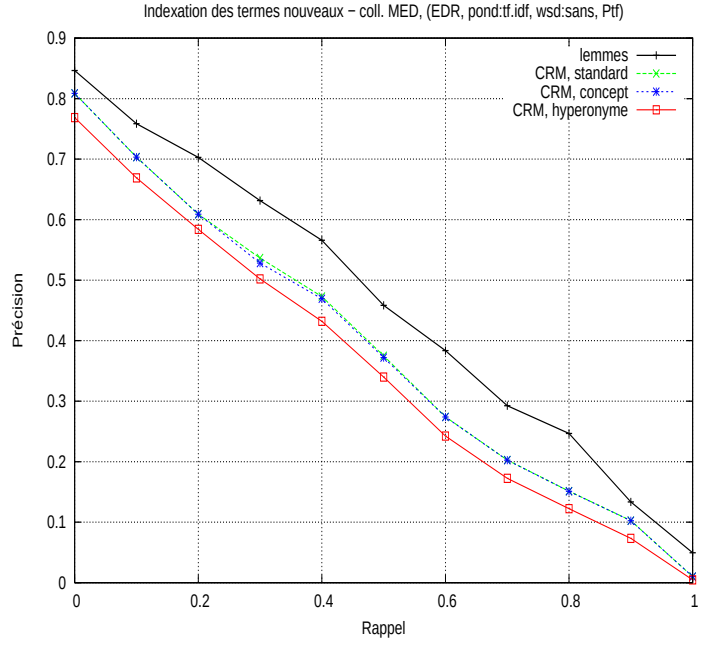
Pondération inverse en document, conservation de la polysémie



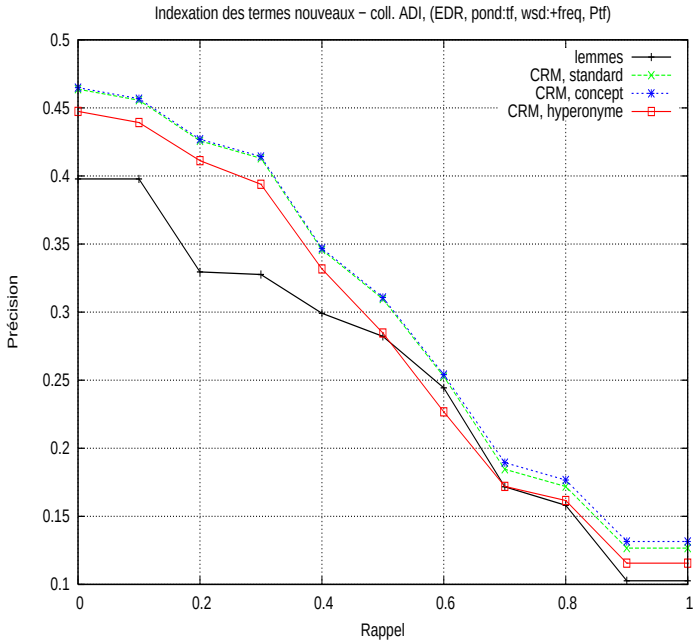


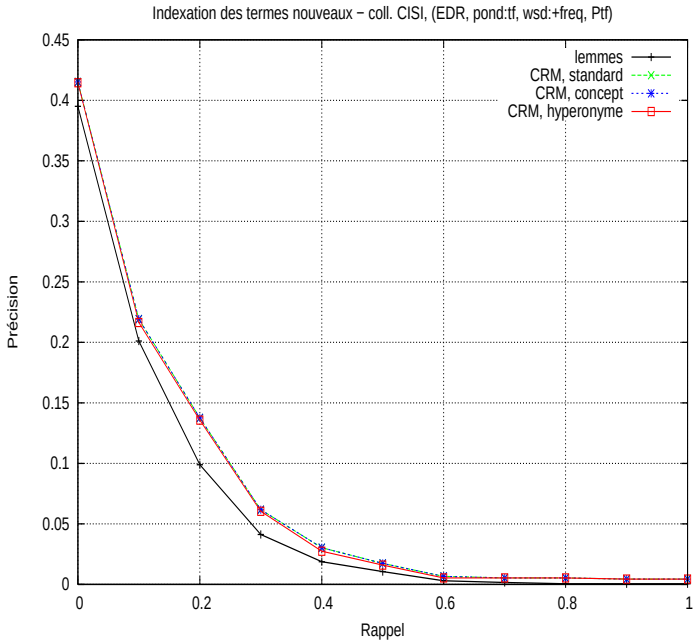


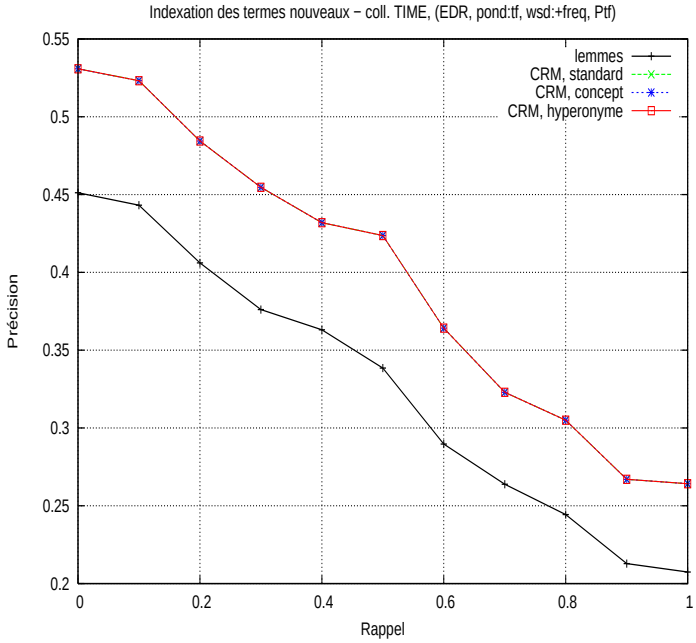


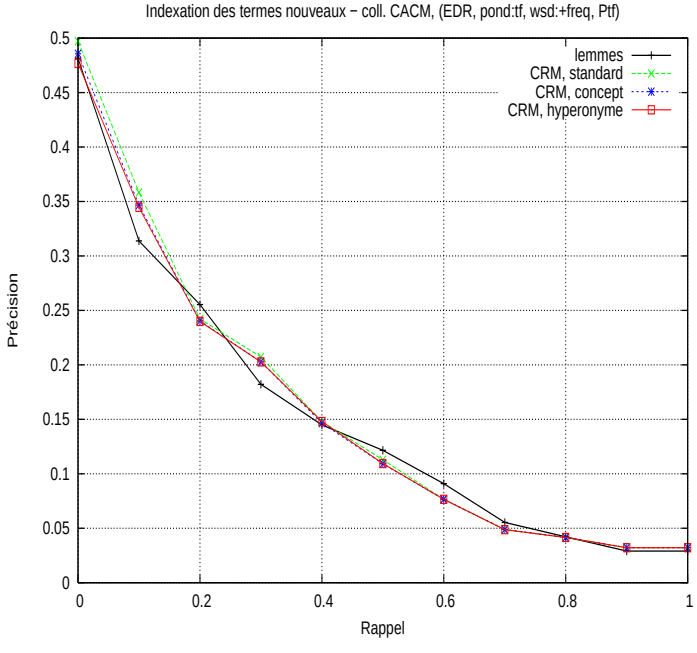


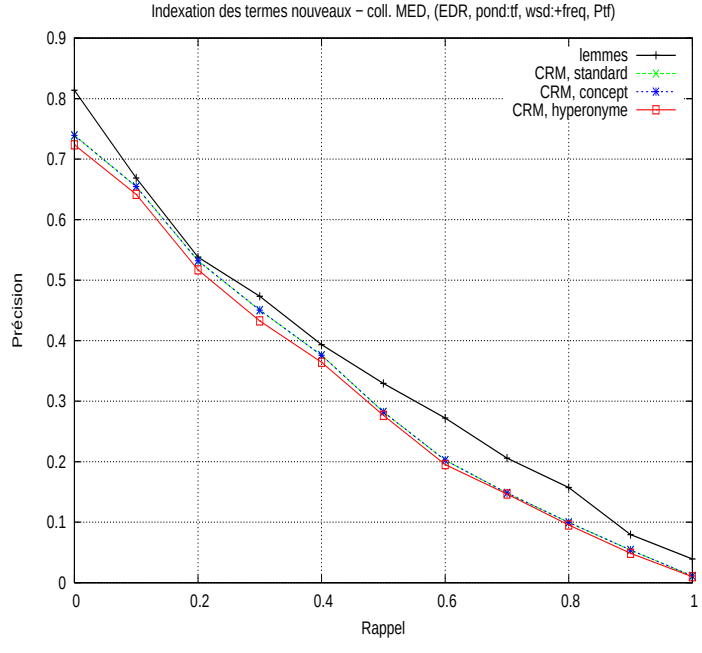
Sans pondération, suppression de la polysémie



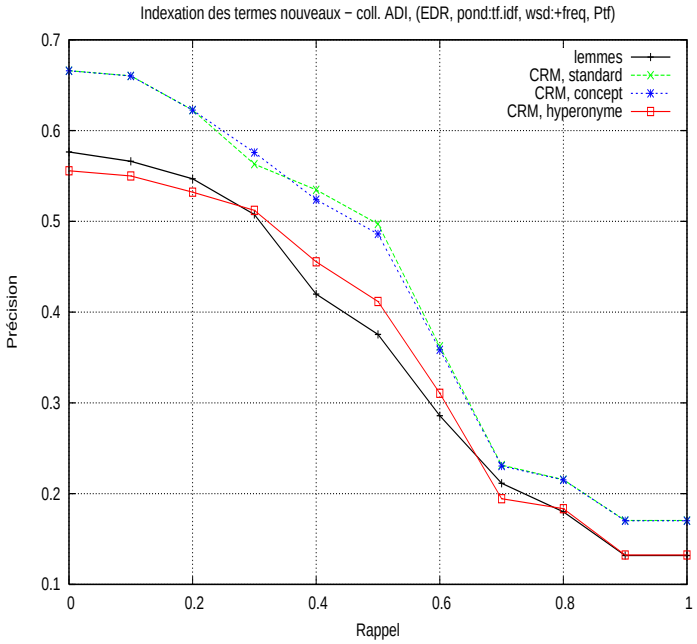


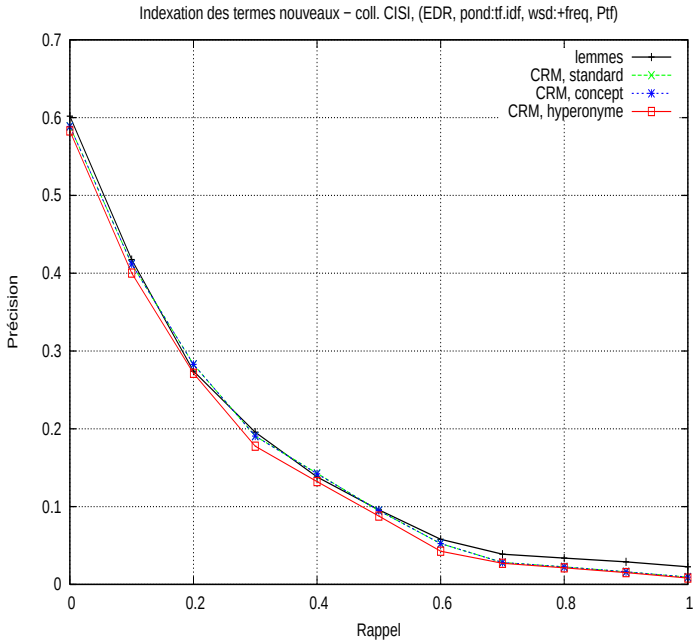


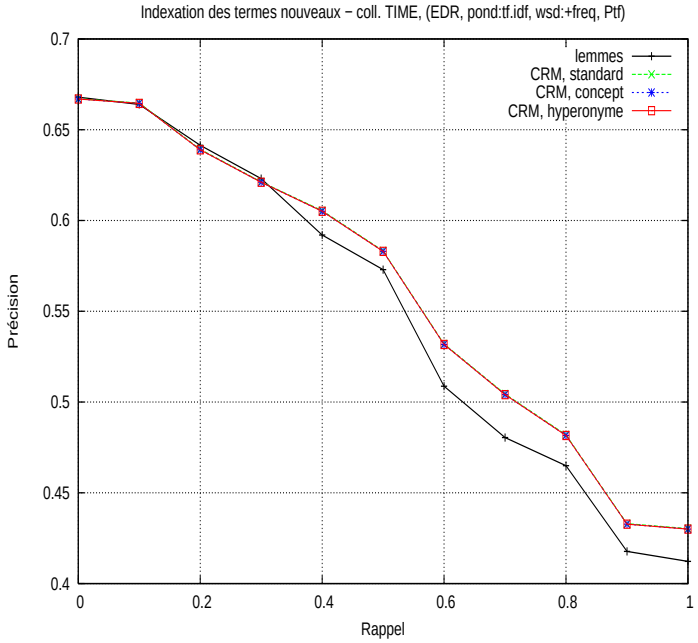


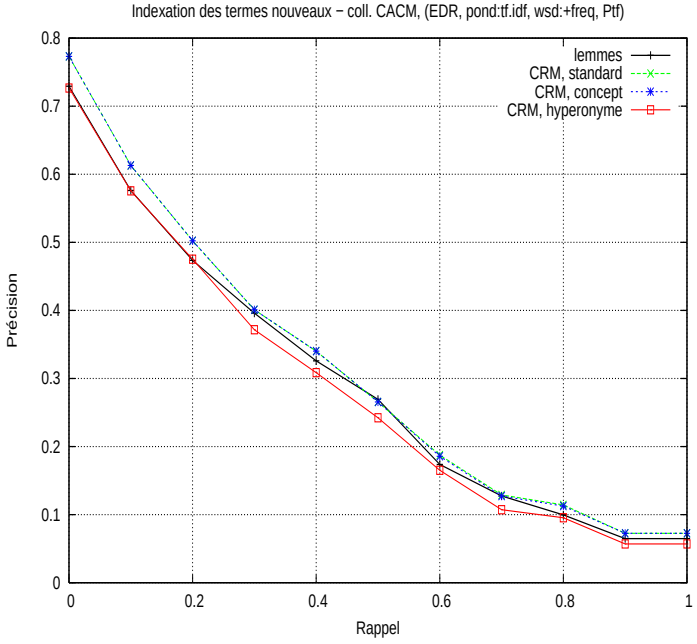


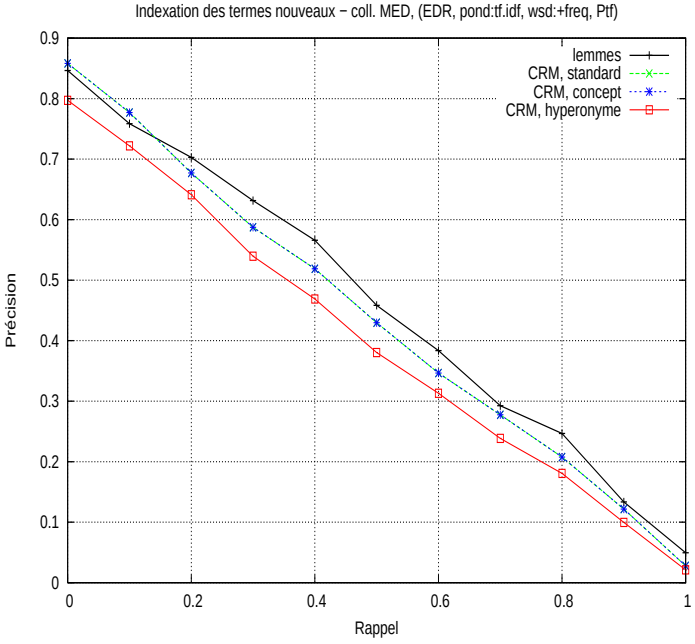
Pondération inverse en document, suppression de la polysémie





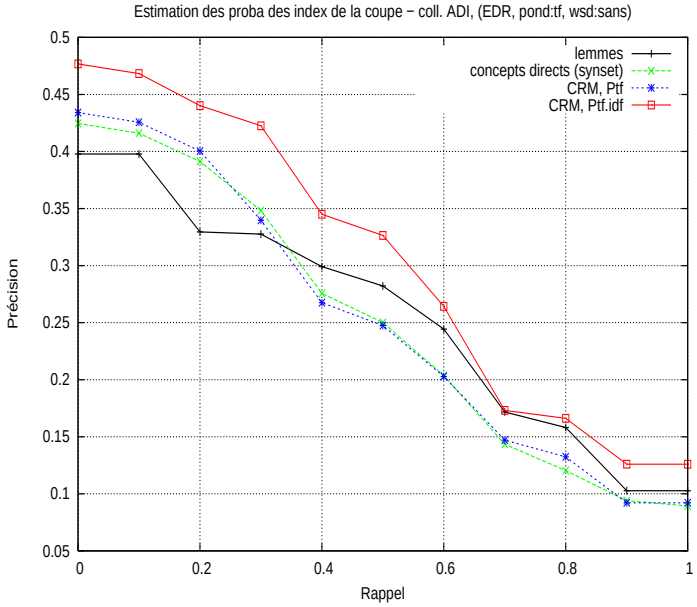


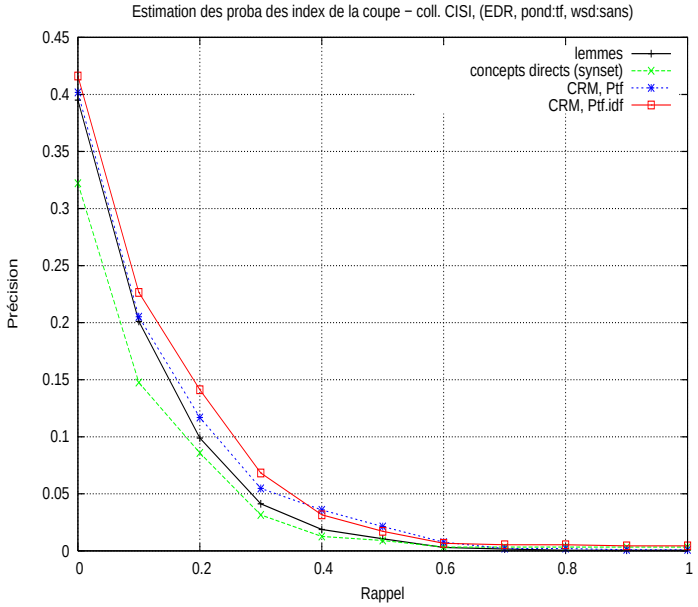


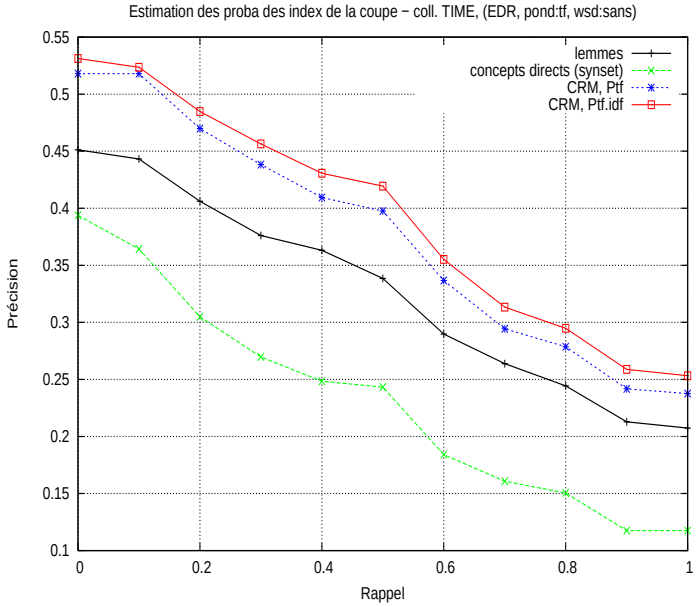


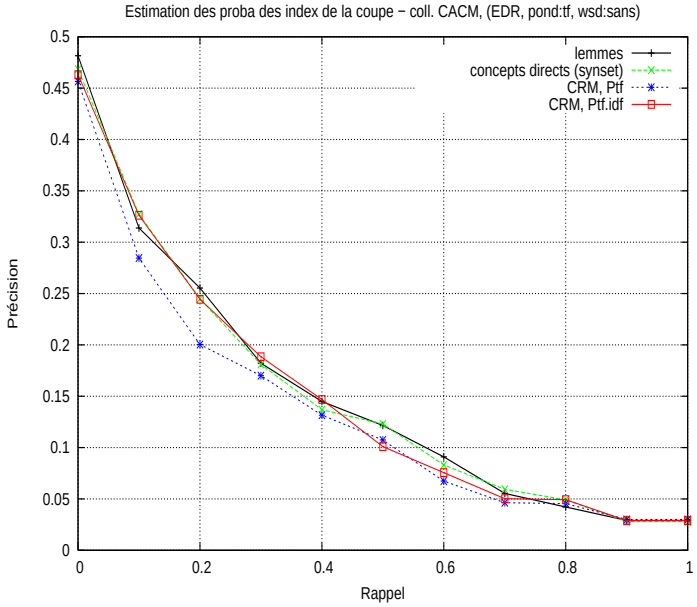
A.10 Courbes PR, estimateur $\widehat{\theta}_s$

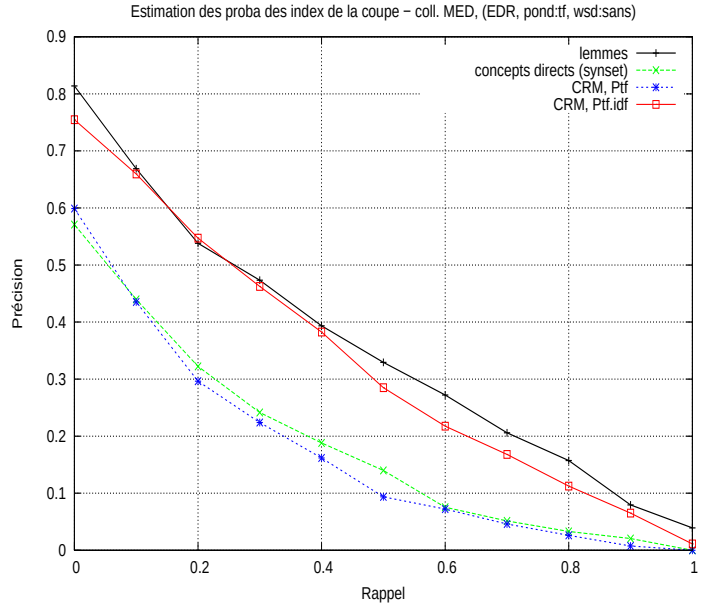
Sans pondération, conservation de la polysémie



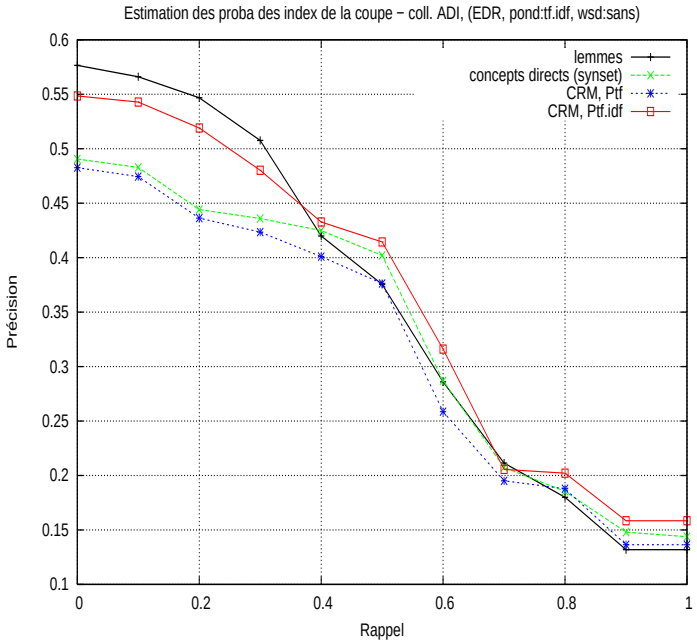


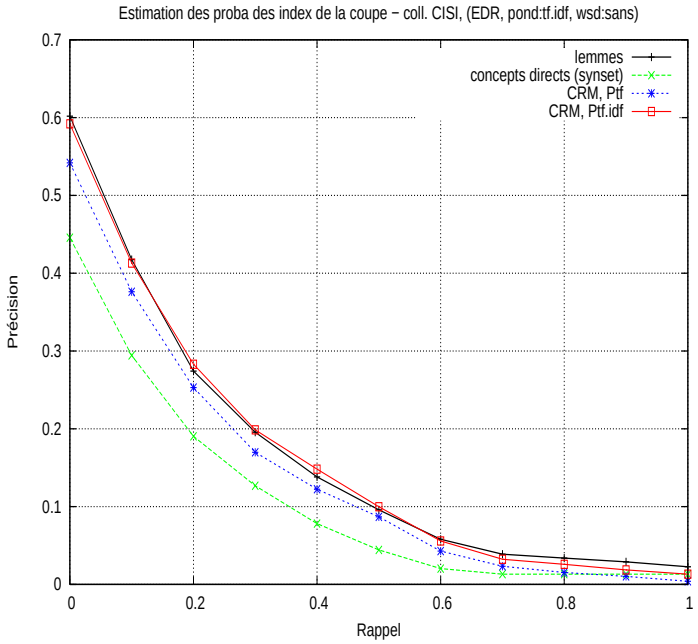


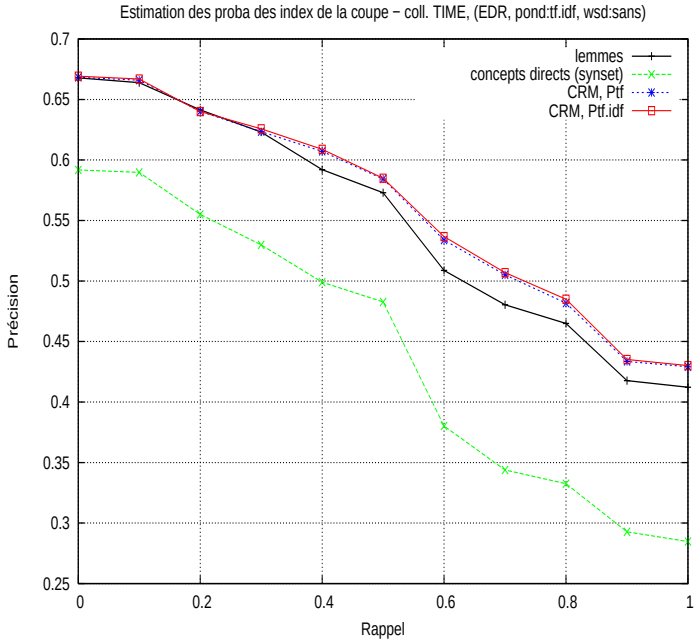


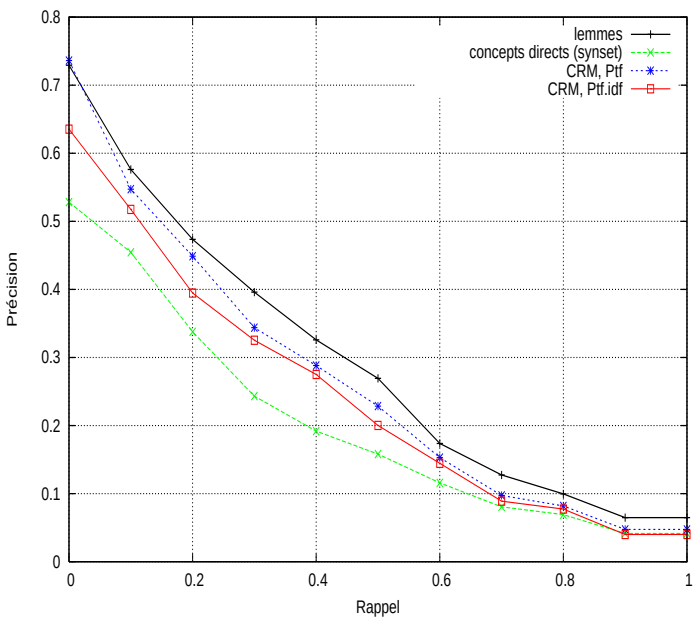


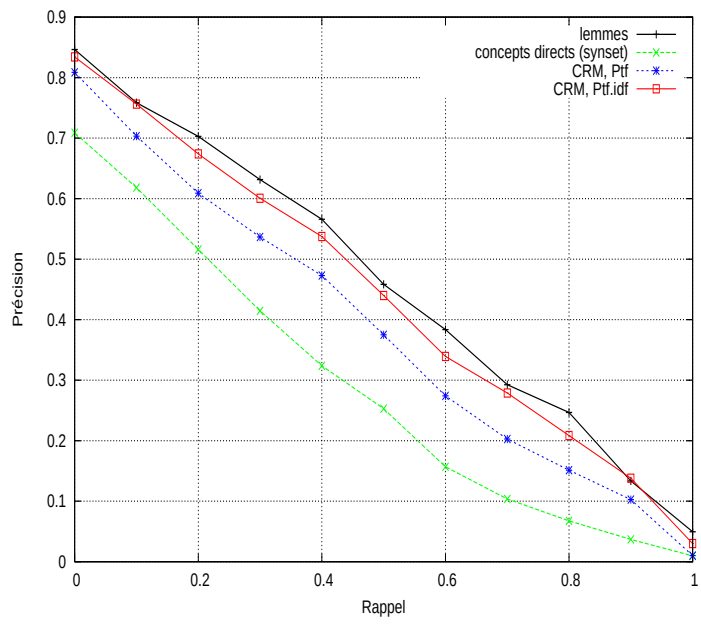
Pondération inverse en document, conservation de la polysémie



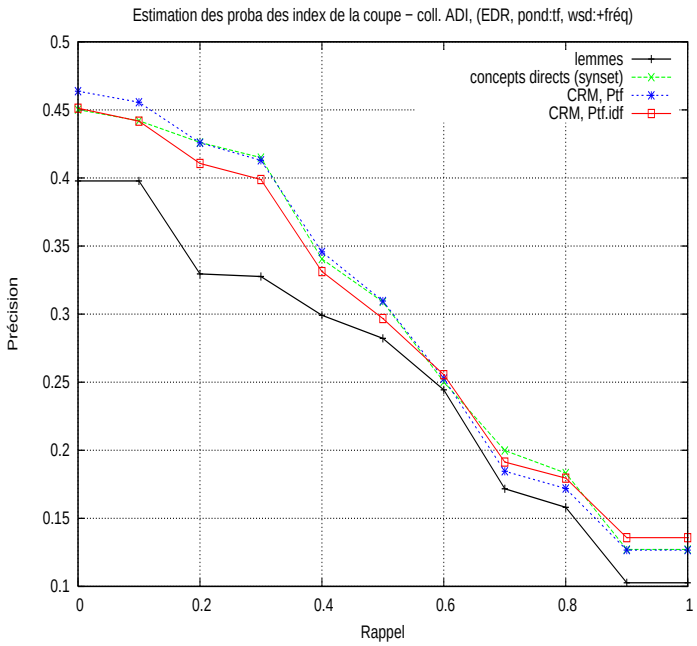


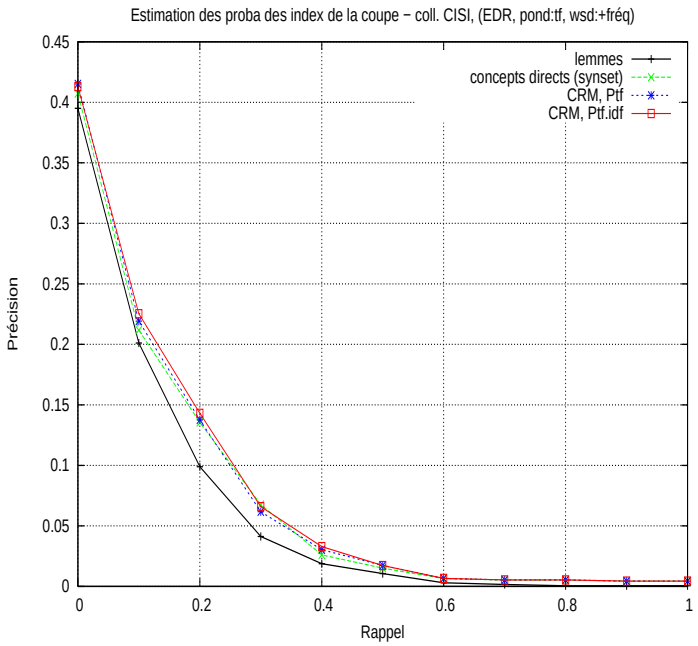


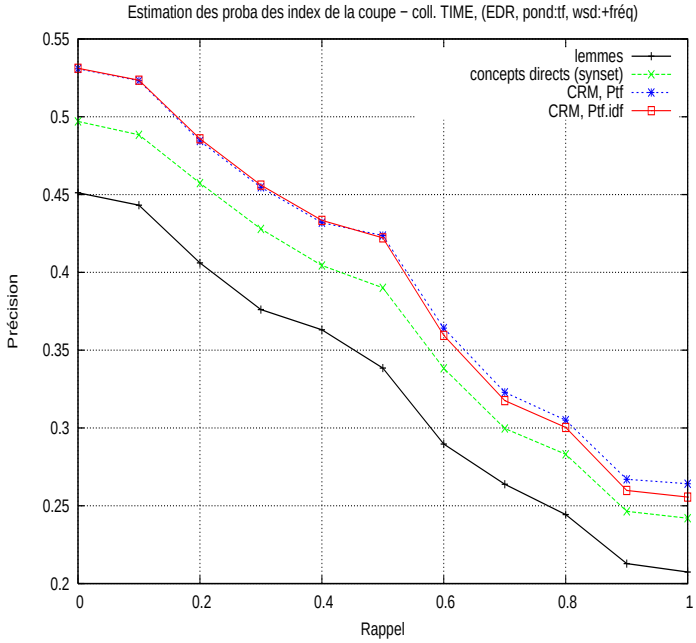


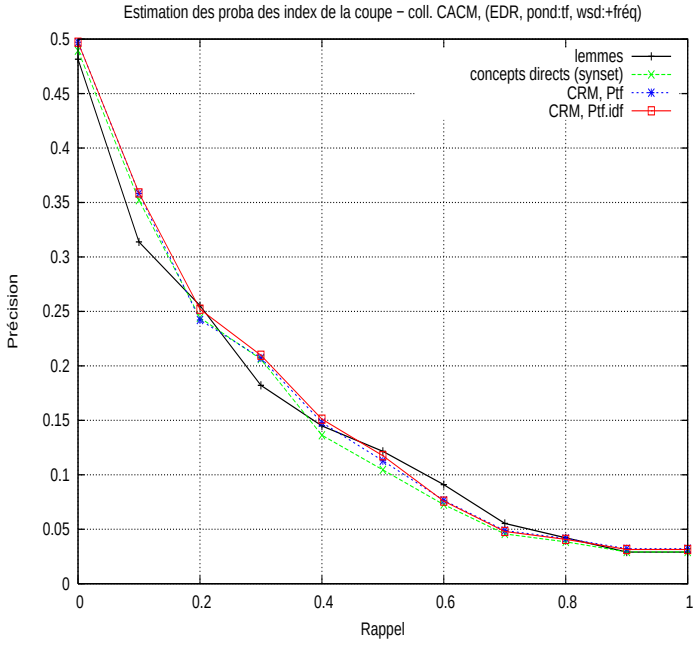


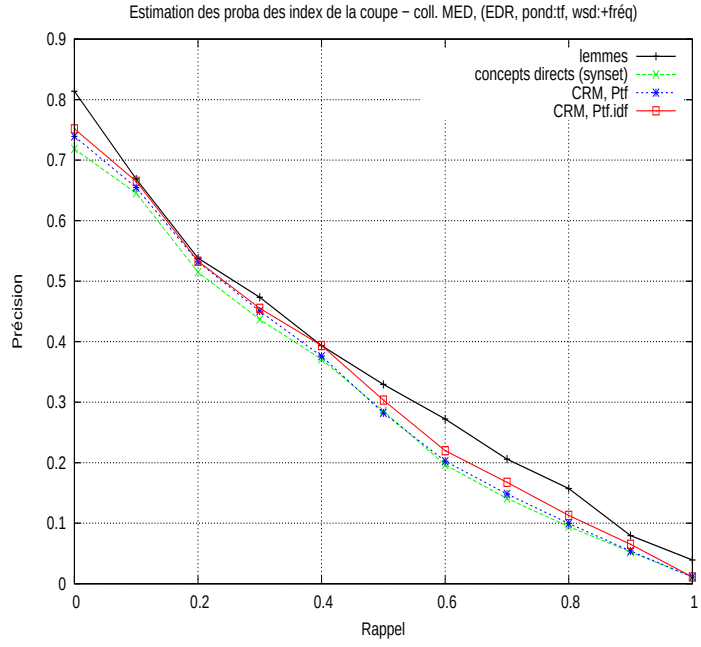
Sans pondération, suppression de la polysémie



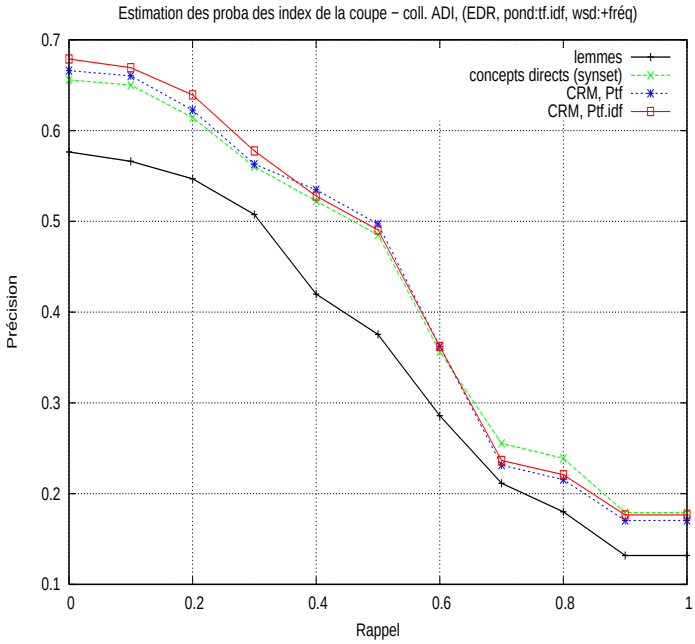


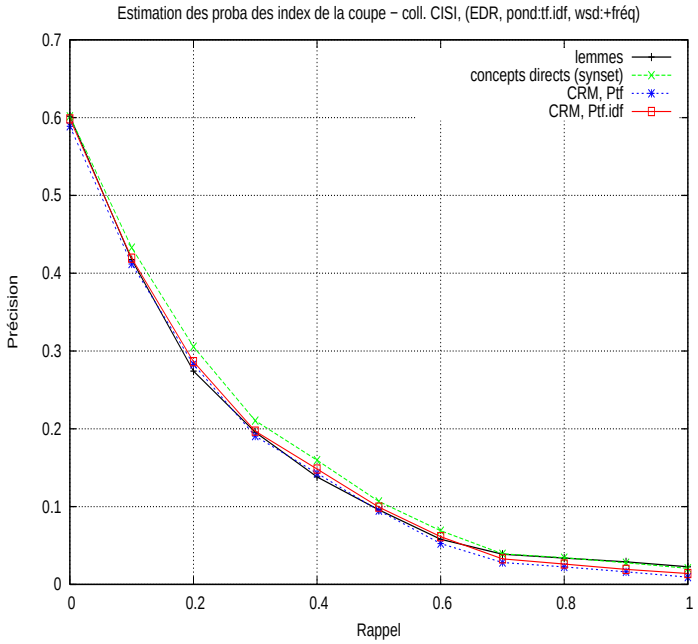


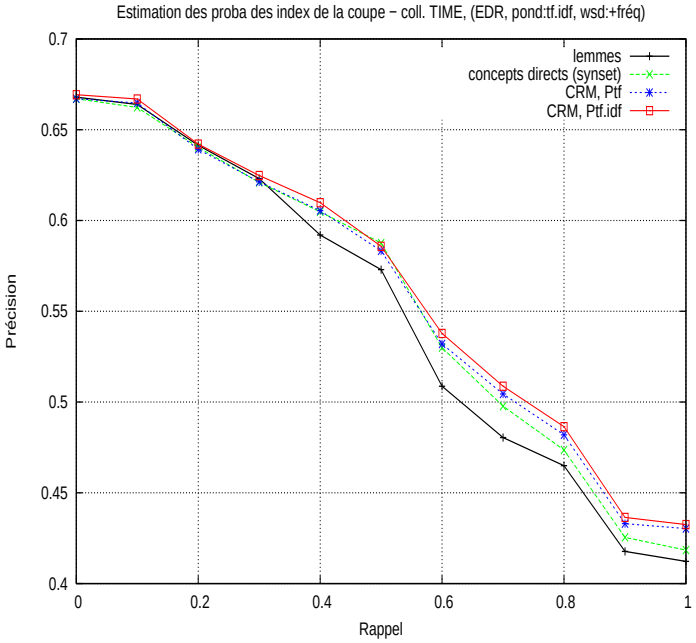


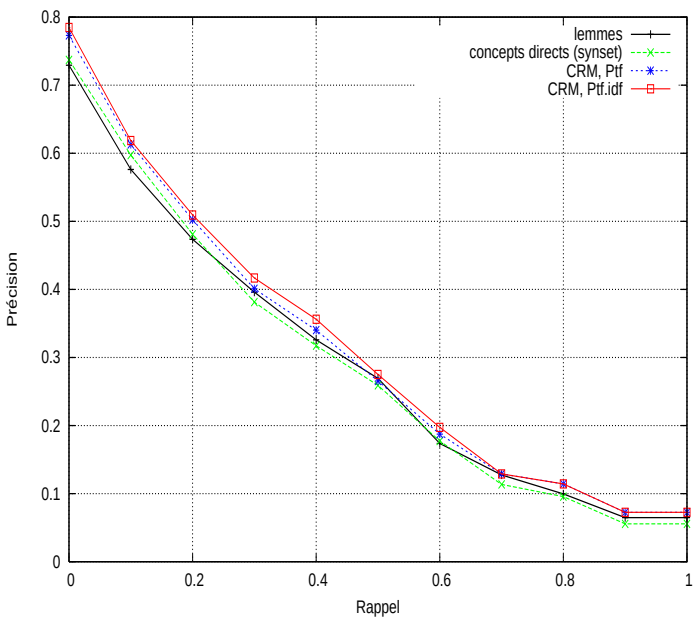


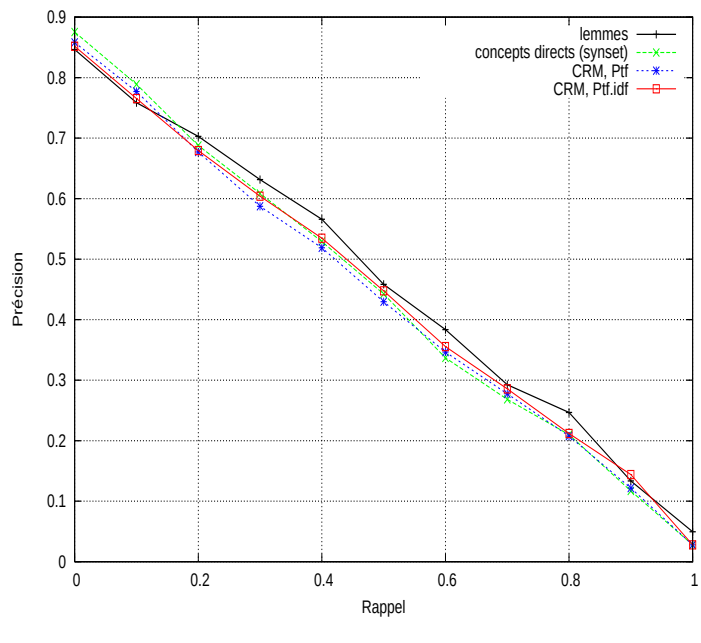
Pondération inverse en document, suppression de la polysémie











Annexe B

Non-pertinence du critère global séparable d'information mutuelle

Le critère séparable de sélection d'une coupe Γ dans un DAG (définition en section 3.5.2 [page 35]) par maximisation de l'information mutuelle entre Γ et les documents à indexer n'est pas satisfant. En effet, l'optimum de ce critère (sous réserve d'une gestion particulière de la polysémie) est toujours obtenu avec la coupe passant par les feuilles, ce qui revient à indexer les documents uniquement avec les mots qu'ils contiennent, indépendamment du thésaurus (DAG).

Avec $D = \{d_i\}$ la collection de documents à indexer et $G = [\mathcal{S}, \mathcal{R}]$ le thésaurus, ce critère est défini par :[1]

$$\begin{aligned} &\mu(s) = p(s) \\ &\lambda(\Gamma) = I(\Gamma; D) = H(D) - H(D|\Gamma) \\ &F'(\Gamma) = \delta\left(\Gamma, \underset{\Gamma_i \in \Upsilon}{\operatorname{argmax}}(\lambda(\Gamma_i))\right) \end{aligned} \tag{B.1}$$

On cherche donc Γ qui maximise :

$$I(\Gamma; D) = \underbrace{H(D)}_{\text{cte}} \underbrace{- H(D|\Gamma)}_{\text{à maximiser}} \tag{B.2}$$

L'entropie conditionnelle de D sachant Γ, $H(D|\Gamma)$ est donnée par :

$$\begin{aligned} -H(D|\Gamma) &= \sum_{d \in D} \sum_{s \in \Gamma} p(s,d) \cdot \log(p(d|s)) \\ &= \sum_{d \in D} \sum_{s \in \Gamma} p(s,d) \cdot \log\left(\frac{p(s,d)}{p(s)}\right) \end{aligned} \tag{B.3}$$

Considérons la situation de la figure B.1 [page suivante], avec deux coupes Γ et Γ', l'une passant par les feuilles $[\mathsf{a}, \mathsf{b}]$ et l'autre passant par le sommet [A]. dominant $[\mathsf{a}, \mathsf{b}]$.

Comparons $H(D|\Gamma)$ et $H(D|\Gamma')$. Pour cette dernière, les probabilités (aussi bien absolues que celles conditionnées aux documents) associées au sommet A sont obtenues en sommant les

[1] Notation et terminologie introduites au chapitre 3 [page 21].

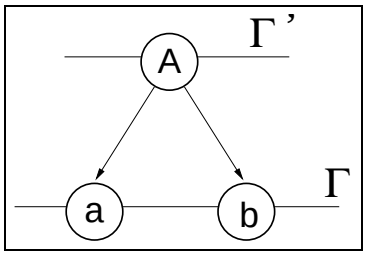


Figure B.1*:* Coupes dans une partie arborescente d'un thésaurus.

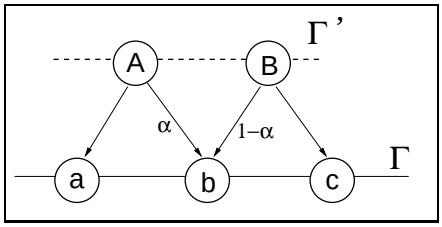


Figure B.2*:* Coupes dans une partie non arborescente d'un thésaurus.

probabilités des feuilles dominées par A :

$$p(\mathrm{A}, d) = \sum_{w \in \{\mathrm{A}^{\Downarrow} \cap \mathcal{M}\}} p(w, d) = p(\mathsf{a}, d) + p(\mathsf{b}, d)$$
$$\text{et } p(\mathrm{A}) = \sum_{d \in D} p(\mathrm{A}, d) = p(\mathsf{a}) + p(\mathsf{b})$$

avec $s^{\Downarrow} \cap \mathcal{M}$ l'ensemble des feuilles (mots) dominées par s.

Les entropies conditionées aux coupe Γ et Γ' sont alors :

$$\begin{aligned} H(D|\Gamma) &= -\sum_{d \in D} \ldots + \sum_{s \in \{\mathsf{a},\mathsf{b}\}} \left(p(s,d) \cdot \log \left(\frac{p(s,d)}{p(s)} \right) \right) + \ldots \\ H(D|\Gamma') &= -\sum_{d \in D} \ldots + \left(\sum_{s \in \{\mathsf{a},\mathsf{b}\}} p(s,d) \right) \cdot \log \left(\frac{\sum_{s \in \{\mathsf{a},\mathsf{b}\}} p(s,d)}{\sum_{s \in \{\mathsf{a},\mathsf{b}\}} p(s)} \right) + \ldots \end{aligned} \tag{B.4}$$

Comme on sait par ailleurs[2] que, pour des nombres non négatifs $x_1, \ldots x_n$ et $y_1, \ldots y_n$,

$$\left(\sum_{i=1}^{n} x_i \cdot \log \frac{x_i}{y_i} \right) \geqslant \left(\sum_{i=1}^{n} x_i \right) \cdot \log \left(\frac{\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} y_i} \right), \tag{B.5}$$

il est alors clair (avec $x_i = p(s_i, d)$ et $y_i = p(s_i)$), que :

$$I(\Gamma = \mathcal{M}; D) \geqslant I(\Gamma'; D)$$

La coupe sur les feuilles est celle qui maximise le critère, du moins sur les parties arborescentes du thésaurus.

Il reste à vérifier que cela est aussi le cas sur les parties non arborescentes du DAG (sommet partagé). Dans le cas de la figure B.2, les coupes $[\mathrm{A}, \mathsf{c}]$ et $[\mathsf{a}, \mathrm{B}]$ se ramènent trivialement au cas précédent ; ce qui n'est pas le cas de la coupe $\Gamma' = [\mathrm{A}, \mathrm{B}]$ de l'illustration.

[2] Inégalité des sommes de logarithmes, c.f Cover et Thomas [1991], page 29.

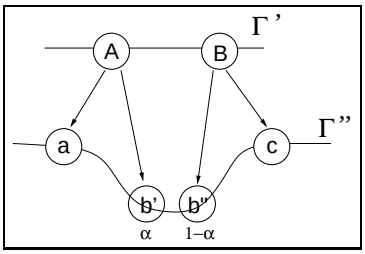


Figure B.3*:* Transformation en arborescence – dédoublement d'un sommet « polysémique ».

On fera l'hypothèse que la probabilité d'occurrence $p(\mathsf{b})$ est répartie entre les concepts A et B, selon un coefficient $\alpha \in [0,1]$:[3]

$$p(\mathrm{A}) = \sum_{d \in D} p(\mathrm{A}, d) \qquad p(\mathrm{B}) = \sum_{d \in D} p(\mathrm{B}, d)$$
$$p(\mathrm{A}, d) = p(\mathsf{a}, d) + \alpha \cdot p(\mathsf{b}, d) \qquad p(\mathrm{B}, d) = (1-\alpha) \cdot p(\mathsf{b}, d) + p(\mathsf{c}, d)$$

Mais comme par ailleurs on peut reporter le découplage entre A et B des occurrences de b au niveau des feuilles, en dédoublant ce sommet (comme indiqué sur la figure B.3)[4], ce qui n'affecte en rien l'entropie de la coupe (pas plus que l'information mutuelle avec les documents), on se ramène au cas précédent d'une arborescence, et la coupe optimale est toujours celle sur les feuilles.

$$\begin{aligned} H(D|\Gamma'') &= \ldots + \alpha \cdot p(\mathsf{b}, d) \log \left(\frac{\cancel{\alpha} \cdot p(\mathsf{b}, d)}{\cancel{\alpha} \cdot p(\mathsf{b})} \right) + (1-\alpha) \cdot p(\mathsf{b}, d) \log \left(\frac{\cancel{(1-\alpha)} \cdot p(\mathsf{b}, d)}{\cancel{(1-\alpha)} \cdot p(\mathsf{b})} \right) + \ldots \\ &= \ldots + \left(p(\mathsf{b}, d) \log \left(\frac{p(\mathsf{b}, d)}{p(\mathsf{b})} \right) \right) \cdot \underbrace{(\alpha + (1-\alpha))}_{=1} + \ldots = H(D|\Gamma) \end{aligned} \tag{B.6}$$

Donc, l'entropie conditionnelle sur Γ'' est égale à celle sur Γ ; la coupe sur les feuilles maximise donc également le critère sur les parties non arborescentes (ou du moins égale le maximum).

Ce critère (et tous ceux qui lui sont apparentés, pour lesquels on retrouve cette propriété de sommation) ne peut donc être utilisé pour déterminer une coupe dans un arbre ou un DAG.

[3] On impose une non simultanéité des occurrences de A et B lors d'occurrences de b ; cette « désambiguisation sémantique » suppose des concepts clairement distincts quant à leur sens.

[4] La « désambigüisation » est ramenée au niveau des mots.

Bibliographie

Agirre, Eneko et Rigau, German [1996] : *Word sense disambiguation using conceptual density.* Dans *Proceedings of COLING'96.* — cité en page(s) 110

Albesano, Dario, Baggia, Paolo, Danieli, Morena, Gemello, Roberto, Gerbino, Elisabetta et Rullent, Claudio [1997] : *Dialogos : A robust system for human-machine spoken dialogue on the telephone.* Dans *Proceedings of ICASSP '97*, (p. 1147–1150) (Munich, Germany). — cité en page(s) 78

Allan, James [1996] : *Incremental relevance feedback for information filtering.* Dans *Proceedings of ACM/SIGIR* (Zürich). — cité en page(s) 18

Allen, James, Byron, Donna, Dzikovska, Myroslava, Ferguson, George, Galescu, Lucian et Stent, Amanda [2000] : *An architecture for a generic dialogue shell.* Journal of Natural Language Engineering, special issue on Best Practices in Spoken Language Dialogue Systems Engineering, *tome 6*(3) :p. 1–16. — cité en page(s) 83

Allen, James, Byron, Donna, Dzikovska, Myroslava, Ferguson, George, Galescu, Lucian et Stent, Amanda [2001] : *Towards conversational human-computer interaction.* AI Magazine. — cité en page(s) 78, 82, 83

Armstrong, Susan, andGiovanni Coray andMaria Georgescul, Alexander Clark, Pallotta, Vincenzo, Popescu-Behs, Andrei, Portabella, David, Rajman, Martin et Starlander, Marianne [2003] : *Natural language queries on natural language data : a database of meeting dialogues.* Dans *8th International Conference on Applications of Natural Language to Information Systems (NLDB 2003)* (NLDB, Burg/Cottbus, Germany). — cité en page(s) 78

Baayen, R. Harald [2001] : *Word Frequency Distributions*, tome 18 de *Text, Speech and Language Technology* (Kluwer Academic Publishers, Dordrecht, The Netherlands), Nancy Ide et Jean Véronis édition. — cité en page(s) 11

Baeza-Yates, Ricardo et Ribeiro-Neto, Berthier [1999] : *Modern Information Retrieval* (Addison-Wesley). — cité en page(s) 60, 109

Bellman, Richard Ernest [1957] : *Dynamic Programming* (Princeton, New Jersey). — cité en page(s) 30, 38

Bellman, Richard Ernest [1961] : *Adaptive Control Processes : A Guided Tour.* — cité en page(s) 29

Bellman, Richard Ernest [2003] : *Dynamic Programming* (Dover Publications, Incorporated). ISBN 0486428095. — cité en page(s) 30, 38

Benzécri, Jean-Paul [1992] : *Correspondence Analysis Handbook* (Marcel Dekker, New York). — cité en page(s) 92

Besançon, Romaric [2001] : *Intégration de connaissances syntaxiques et sémantiques dans les représentations vectorielles de textes.* Thèse de doctorat N°2508, École Polytechnique Fédérale de Lausanne. — cité en page(s) 9, 15

Besançon, Romaric, Rozenknop, Antoine, Chappelier, Jean-Cédric et Rajman, Martin [2001] : *Intégration probabiliste de sens dans la représentation de textes.* Dans *Actes de la 8ème conférence sur*

le Traitement Automatique des Langues Naturelles (TALN'2001), tome 1, (p. 83–91). — cité en page(s) 10, 11

Beyer, Kevin S., Goldstein, Jonathan, Ramakrishnan, Raghu et Shaft, Uri [1999] : *When is "nearest neighbor" meaningful ?* Dans *Proceedings of the 7th International Conference on Database Theory (ICDT)*. — cité en page(s) 29

Bienkowski, Marcin, Korzeniowski, Miroslaw et Räcke, Harald [2003] : *A practical algorithm for constructing oblivious routing schemes*. Dans *Proceedings of the fifteenth annual ACM symposium on Parallel algorithms and architectures (SPAA '03)*, (p. 24–33) (ACM Press, New York, NY, USA). ISBN 1-58113-661-7. — cité en page(s) 39

Bourlard, Hervé et Morgan, Nelson [1995] : *Continuous speech recognition : An introduction to the Hybrid HMM/Connectionist Approach*. IEEE Signal Processing Magazine, *tome 12*(3) :p. 25–42. — cité en page(s) 80

Boyce, S. et Gorin, A. [1996] : *User Interface Issues for Natural Spoken Dialog Systems*. Dans *Proceedings of the International Symposium on Spoken Dialogue (ISSD)*, (p. 65–68) (Philadelphia, USA). — cité en page(s) 82

Brockmann, Carsten et Lapata, Mirella [2003] : *Evaluating and combining approaches to selectional preference acquisition*. Dans *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics (EACL '03)*, (p. 27–34) (Association for Computational Linguistics, Morristown, NJ, USA). ISBN 1-333-56789-0. — cité en page(s) 36

Buckley, Chris, Salton, Gerard et Allan, James [1992] : *Automatic retrieval with locality information using smart*. Dans *Gaithersburg*, (p. 59–72). — cité en page(s) 14

Buckley, Chris et Voorhees, Ellen M. [2000] : *Evaluating evaluation measure stability*. Dans *Proceedings of the 23rd annual international ACM/SIGIR conference on Research and development in information retrieval*, (p. 33–40) (ACM Press, Athens, Greece). Isbn = 1-58113-226-3. — cité en page(s) 57

Chevallet, J.P. et Haddad, H. [2001] : *Proposition d'un modèle relationnel d'indexation syntagmatique : mise en œuvre dans le système iota*. Dans *Actes du XIX^{e} congrès INFORSID*, (p. 465–483) (Genève). — cité en page(s) 11

Clark, Stephen et Weir, David [2002] : *Class-based probability estimation using a semantic hierarchy*. Dans *NAACL'01 : Second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies 2001*, (p. 1–8) (Association for Computational Linguistics, Morristown, NJ, USA). — cité en page(s) 36

Comanor, William S. et Scherer, Frederic M. [1969] : *Patent statistics as a measure of technical change*. Journal of Political Economy, *tome 77*(3) :p. 392–398. — cité en page(s) 92

Constant, Patrick [1995] : *Manuel de développement SYLEX-BASE*. Decan Ingénia-LN, Paris, France. — cité en page(s) 62

Cover, Thomas M. et Thomas, Joy A. [1991] : *Elements of Information Theory* (John Wiley & Soons, Inc.). — cité en page(s) 17, 132

Daly-Jones, Owen, Bevan, Nigel et Thomas, Cathy [1999] : *Wizard-of-Oz prototyping*. Dans *Handbook of User-Centred Design*. — cité en page(s) 82

Danieli, Morena [1997] : *On the use of expectations for detecting and repairing human-machine miscommunication*. Dans *Proceedings of AAAI-96 Workshop on Detecting, Preventing, and Repairing Human-Machine Miscommunications*, (p. 87–93) (Portland, OR). — cité en page(s) 88

Danieli, Morena, Gerbino, Elissbetta et M.Moisa, Loretta [1997] : *Dialogue strategies for improving the usability of telephone human-machine communication*. Dans *Interactive Spoken Dialog Systems : Bridging Speech and NLP Together in Real Applications* (edité par Julia Hirschberg, Candace Kamm et Marilyn Walker), (p. 114–120) (Association for Computational Linguistics, New Brunswick, New Jersey). — cité en page(s) 88

DAVISON, A. C. et HINKLEY, D. V. [1997] : *Bootstrap Methods and their Application* (Cambridge University Press). — cité en page(s) 102

DEERWESTER, S. C., DUMAIS, S. T., LANDAUER, T. K., FURNAS, G. W. et HARSHMAN, R. A. [1990] : *Indexing by latent semantic analysis.* Journal of the American Society of Information Science, *tome 41*(6) :p. 391–407. — cité en page(s) 13

DOU, Henri [1995] : *Veille technologique et compétitivité – L'intelligence économique au service du developpement industriel* (Dunod, Paris). — cité en page(s) 92

DUMAIS, Susan [1994] : *Latent semantic indexing (LSI) : TREC-3 report.* Dans *TREC-3 Proceedings*, (p. 219–230) (Gaithersburg, Maryland). — cité en page(s) 13

DUMAIS, Susan, LANDAUER, Thomas as et LITTMAN, Michael [1996] : *Automatic cross-linguistic information retrieval using latent semantic indexing.* Dans *SIGIR'96 - Workshop on Cross-Linguistic Information Retrieval*, (p. 16–23). — cité en page(s) 13

EFRON, Bradley [1982] : *The jacknife, the bootstrap, and other resampling plans.* Dans *CBMS-NSF Regional Conference Series in Applied Mathematics*, tome 38 (Society for Industrial and Applied Mathematics, Philadelphia). — cité en page(s) 101

EFRON, Bradley et TIBSHIRANI, Robert J. [1994] : *An Introduction to the Bootstrap*, tome Monographs on Statistics and Applied Probability (Chapman & Hall/CRC). — cité en page(s) 102

ESCOFIER, Brigitte et PAGÈS, Jérôme [1998] : *Analyses factorielles simples et multiples : Objectifs, méthodes et interprétation* (Paris), 3e édition. — cité en page(s) 95

ESCOFIER, Brigitte et ROUX, Brigitte Le [1972] : *Etude de trois problèmes de stabilité en analyse factorielle.* Publication de l'Institut Statistique de l'Université de Paris, *tome 11* :p. 1–48. — cité en page(s) 100

FAGAN, Joel L. [1987*a*] : *Automatic phrase indexing for document retrieval - syntactic andnon syntactic methods.* — cité en page(s) 11

FAGAN, Joel L. [1987*b*] : *Experiments in automated phrase indexing for document retrieval : a comparison of syntactic and non-syntactic methods.* Thèse de doctorat, Dpt of Computer Science, Cornell University. — cité en page(s) 11

FOLTZ, Peter et DUMAIS, Susan [1992] : *Personalized information delivery : An analysis of information filteringmethods.* Communications of the ACM, *tome 35*(12) :p. 51–60. — cité en page(s) 13

FOX, Edward A. [1983] : *Characterization of two new experimental collections in computer and information science containing textual and bibliographic concepts.* Rapport technique, Cornell University. — cité en page(s) 60, 109

FRASER, N. et GILBERT, N. [1991] : *Simulating Speech Systems.* Computer Speech and Language, *tome 5* :p. 81–89. — cité en page(s) 81

GAUSSIER, Eric et STÉFANINI, Marie-Hélène [2003] : *Assistance intelligente à la recherche d'informations.* Traité des sciences et techniques de l'information (Hermes Science, 11 Rue Lavoisier, 75008 Paris), lavoisier édition. — cité en page(s) 8, 81

GLOVER, F., TAILLARD, E. et DE WERRA, D. [1993] : *A user's guide to Tabu Search.* Annals of Operations Research, *tome 41* :p. 3–28. — cité en page(s) 44

GONDRAN, Michel et MINOUX, Michel [1995] : *Graphes et algorithmes* (Paris), 3e édition. — cité en page(s) 38, 44

GONZALO, J., VERDEJO, M.F., PETERS, C. et CALZOLARI, N. [1998*a*] : *Applying EuroWordNet to multilingual text retrieval.* Journal of Computers and the Humanities, Special Issue on EuroWordNet Reprinted in Vossen, P. (ed). EuroWordNet : a multilingual database with lexical semantic networks. Kluwer Academic Publishers, Dordrecht, 1998, *tome 32*(2-3) :p. 185–207(23). — cité en page(s) 19

Gonzalo, Julio, Chugur, Irina et Verdejo, Felisa [2000] : *Sense clusters for information retrieval : Evidence from semcor and the EuroWordNet interlingual index.* Universidad Nacional de Educacidn a Distancia, Departamento de Lenguajes y Sistemas Informáticos. Publié en 2000 ou apres. — cité en page(s) 19

Gonzalo, Julio, Verdejo, Felisa, Chugur, Irina et Cigarran, Juan [1998*b*] : *Indexing with WordNet synsets can improve text retrieval.* Dans *Proceedings of the COLING/ACL 1998 Workshop on Usage of WordNet for Natural Language Processing, Montreal*, (p. 38–44) (Montreal, Canada). — cité en page(s) 19, 64

Goutte, Cyril et Gaussier, Éric [2005] : *A probabilistic interpretation of precision, recall and F-Score, with implication for evaluation.* Dans *Advances in Information Retrieval, 27th European Conference on IR Research (ECIR2005)* (edité par David E. Losada et Juan M. Fernández-Luna), tome 3408 de *Lecture Notes in Computer Science*, (p. 345–359) (Springer, Santiago de Compostela, Spain). — cité en page(s) 72

Guellec, D. et van Pottelsberghe, B [1998] : *New indicators from patent data.* Dans *Proceedings of Joint NEST/TIP/GSS Workshop.* — cité en page(s) 92

Gull, C.D. [1956] : *Seven years of work on the organization of materials in the special library.* American Documentation, *tome 4* (7) :p. 320–329. — cité en page(s) 56

Harman, Donna [1988] : *Towards interactive query expansion.* Dans *SIGIR '88 : Proceedings of the 11th annual international ACM SIGIR conference on Research and development in information retrieval*, (p. 321–331) (ACM Press). ISBN 2-7061-0309-4. — cité en page(s) 18

Harrelson, Chris, Hildrum, Kirsten et Rao, Satish [2003] : *A polynomial-time tree decomposition to minimize congestion.* Dans *Proceedings of the fifteenth annual ACM symposium on Parallel algorithms and architectures (SPAA '03)*, (p. 34–43) (ACM Press, New York, NY, USA). ISBN 1-58113-661-7. — cité en page(s) 39

Hirst, G., McRoy, S., Heeman, P., Edmonds, P. et Horton, D. [1994] : *Repairing conversational misunderstandings and non-understandings.* Speech Communication, *tome 15* :p. 213–230. — cité en page(s) 83

Hofmann, Thomas [1999] : *Probabilistic Latent Semantic Indexing.* Dans *proc. of the 22th International Conference on Research and Development in Information Retrieval (SIGIR)*, (p. 50–57). — cité en page(s) 13

Hotho, Andreas, Staab, Steffen et Stumme, Gerd [2003] : *WordNet improves text document clustering.* Dans *Proceedings of the SIGIR2003 Semantic Web Workshop.* None. — cité en page(s) 19, 110

Jacquemin, Christian [1998] : *Improving automatic indexing through concept combination and term enrichment.* Dans *Proceedings of the 17th International Conference on Computational Linguistics (COLING'98)*, (p. 595–599) (Montr ?al). — cité en page(s) 11

Jacquemin, Christian [2001] : *Fastr – a tool for automatic indexing (version 2.03).* — cité en page(s) 11

Jacquemin, Christian, Klavans, Judith et Tzoukermann, Evelyne [1997] : *Expansion of multi-word terms for indexing and retrieval using morphology and syntax.* Dans *PROC "the 35th Annual Meeting of the Association for Computational Linguistics ((E)ACL'97)*, (p. 24– 31). — cité en page(s) 11

Johnson, R.A. et Wichern, D.A. [1998] : *Applied multivariate statistical analysis* (Prentice-Hall, Inc.). — cité en page(s) 92

Johnson, W. B. et Lindenstrauss, J. [1984] : *Extensions of lipshitz mapping into hilbert space.* Contemp. Math., *tome 26* :p. 189–206. — cité en page(s) 13

Jurafsky, D., Wooters, C., Tajchman, G., Segal, J., Fosler, E. et Morgan, N. [1994] : *The Berkeley Restaurant Project.* Dans *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, (p. 2139–2142) (Yokohama, Japan). — cité en page(s) 78

KANG, Bo-Yeong [2003] : *A novel approach to semantic indexing based on concept.* Dans *ACL '03 : Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, (p. 44–49) (Association for Computational Linguistics, Morristown, NJ, USA). ISBN 0-111-456789. — cité en page(s) 18

KERNIGHAN, B. W. et LIN, S. [1970] : *An Efficient Heuristic Procedure for Partitioning Graphs.* Bell Syst. Techn., *tome 49* :p. 291–307. — cité en page(s) 44

KIRKPATRICK, S., GELLAT, C. D. et VECCHI, M. P. [1983] : *Optimization by Simulated Annealing.* Science, *tome 220* :p. 671–680. — cité en page(s) 44

KIRYAKOV, Atanas K. et SIMOV, Kiril Iv. [1999] : *Ontologically Supported Semantic Matching.* Dans *Proceedings of NODALIDA'99 : Nordic Conference on Computational Linguistics, Trondheim.* — cité en page(s) 19, 29, 64

KOKIOPOULOU, E. et SAAD, Y. [2004] : *Polynomial filtering in latent semantic indexing for information retrieval.* Dans *Proceedings of the 27th annual international ACM/SIGIR conference on Research and development in information retrieval (SIGIR '04)*, (p. 104–111) (ACM Press, New York, NY, USA). ISBN 1-58113-881-4. — cité en page(s) 13, 33

LAND, A. H. et DOIG, A. G. [1960] : *An automatic method for solving discrete programming problems.* Econometrica, *tome 28* :p. 497–520. — cité en page(s) 38

LANDAUER, Thomas, FOLTZ, Peter et LAHAM, D. [1998] : *Introduction to latent semantic analysis.* Discourse Process, *tome 25* :p. 259–284. — cité en page(s) 13

LARSEN, Lars Bo [1997] : *A strategy for mixed-initiative dialogue control.* Dans *Proceedings of EUROSPEECH*, (p. 1331–1334) (EUROSPEECH, Rhodes). — cité en page(s) 84

LEBART, Ludovic, MORINEAU, Alain et PIRON, Marie [2000] : *Statistique exploratoire multidimensionnelle* (Paris), 3e édition. — cité en page(s) 95, 100

LEBART, Ludovic, SALEM, André et BERRY, Lisette [1998] : *Exploring Textual Data*, tome 4 (Kluwer Academic Publishers). — cité en page(s) 95

LEE, John Ho [1995] : *Combining multiple evidence from different properties of weighting schemes.* Dans *EIGHTEENTH ACMSIGIR* (edité par Ewdard A. Fox), (p. 180–188) (Seattle, Washington). — cité en page(s) 13

LI, Hang [1998] : *A probabilistic approach to lexical semantic knowledge acquisition and structural disambiguation.* Thèse de maître, Graduate School of Science, University of Tokyo. — cité en page(s) 34, 36

LI, Hang et ABE, Naoki [1998] : *Generalizing case frames using a thesaurus and the MDL principle.* Computational Linguistic, *tome 24* (2) :p. 217–244. ISSN 0891-2017. — cité en page(s) 35, 36, 38

DE LOUPY, Claude [2000] : *Évaluation de l'apport de connaissances linguistiques en désambiguïsation sémantique et recherche documentaire.* Thèse de doctorat, Université d'Avignon et des Pays de Vaucluse, Académie D'Aix-Marseille, Laboratoire d'Informatique d'Avignon. — cité en page(s) 7, 9, 57

LUHN, Hans Peter [1957] : *A statistical approach to mechanized encoding and searching of literary information.* IBM Journal of Research and Development, *tome 1* (4) :p. 309–317. — cité en page(s) 7, 11

LUHN, Hans Peter [1958] : *The automatic creation of litterature abstracts.* IBM Journal of Research and Development, *tome 2* :p. 157–165. — cité en page(s) 11

LUHN, Hans Peter [1959] : *Keyword-in-context index for technical literature (KWIC index).* Yorktown heights, IBM, N.Y. — cité en page(s) 11

MANDALA, Rila, TOKUNAGA, Takenobu et TANAKA, Hozumi [1998] : *The use of WordNet in information retrieval.* Dans *Proceedings of the COLING-ACL workshop on Usage of Wordnet in Natural Language Processing*, (p. 31 – 37). — cité en page(s) 109

MATVEEVA, Irina, LEVOW, Gina-Anne, FARAHAT, Ayman et ROYER, Christiaan [2005] : *Term Representation with Generalized Latent Semantic Analysis.* Dans *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP'05)* (edité par Galia Angelova, Kalina Bontcheva, Ruslan Mitkov, Nicolas Nicolov et Nikolai Nikolov) (European Commission as a Marie Curie Large Conference, MLCF-CT-2004-013233, Borovets, Bulgaria). — cité en page(s) 13, 33

MIHALCEA, Rada et MOLDOVAN, Dan [2000] : *Semantic indexing using WordNet senses.* Dans *Proceedings of ACL Workshop on IR & NLP* (Hong Kong). — cité en page(s) 18, 19

MILLER, Gorges A., BECKWITH, R., FELLBAUM, C., GROSS, D. et MILLER, K. [1990] : *Introduction to wordnet : an online lexical database.* International Journal of Lexicography, *tome 3*(4) :p. 235–312. — cité en page(s) 62

MIYOSHI, Hideo, SUGIYAMA, Kenji, KOBAYASHI, Masahiro et OGINO, Takano [1996] : *An overview of the EDR electronic dictionary and the current status of its utilization.* Dans *Proceedings of COLING*, (p. 1090–1093). — cité en page(s) 60

MOLDOVAN, Dan I. et MIHALCEA, Rada [2000] : *Using WordNet and Lexical Operators to Improve Internet Searches.* IEEE Internet Computing, *tome 4*(1) :p. 34–43. ISSN 1089-7801. — cité en page(s) 18

MOOERS, Calvin N. [1960] : *Mooers' law, or why some retrieval systems are used and others are not.* American Documentation, *tome 11*(3). — cité en page(s) 3

MORGAN, Nelson et BOURLARD, Hervé [1995] : *Neural networks for statistical recognition of continuous speech.* Dans *Proceedings of the Institute of Electrical and Electronics Engineers (IEEE)* (edité par Institute of Electrical et Electronics Engineers), tome 83, (p. 742–770) (New York, USA). — cité en page(s) 80

MÖLLER, Sebastian et SKOWRONEK, Janto [2003] : *Quantifying the impact of system characteristics on perceived quality dimensions of a spoken dialogue service.* Dans *EuroSpeech2003.* — cité en page(s) 78

NARIN, F. [1995] : *Patents as indicators for the evaluation of industrial research output.* Scientometrics, *tome 34*(4) :p. 489–496. — cité en page(s) 92

NG, Hwee Tou, GOH, Wei B. et LOW, Kok L. [1997] : *Feature selection, perceptron learning, and a usability case study for text categorization.* Dans *Proceedings of SIGIR-97, 20th ACM International Conference on Research and Development in Information Retrieval* (edité par Nicholas J. Belkin, A. Desai Narasimhalu et Peter Willett), (p. 67–73) (ACM Press, New York, US, Philadelphia, US). — cité en page(s) 12

PAPADIMITRIOU, Christos H., TAMAKI, Hisao, RAGHAVAN, Prabhakar et VEMPALA, Santosh [1998] : *Latent semantic indexing : A probabilistic analysis.* Dans *Proceedings of the seventeenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*, (p. 159–168). — cité en page(s) 13

PESTOV, V. [1999] : *A geometric framework for modelling similarity search.* Dans *Proceedings of the International Workshop on Similarity Search (IWOSS99)* (Florence, Italy). — cité en page(s) 29

QIU, Yonggang [1994] : *Improving retrieval effectiveness by similarity thesaurus.* — cité en page(s) 18

QIU, Yonggang et FREI, Hans-Peter [1993] : *Concept-based query expansion.* Dans *Proceedings of SIGIR-93, 16th ACM International Conference on Research and Development in Information Retrieval*, (p. 160–169) (Pittsburgh, US). — cité en page(s) 18

QUINLAN, John Ross [1986] : *Induction of decision trees.* Dans *Machine Learning*, tome 1, (p. 81–106). — cité en page(s) 86

QUINLAN, John Ross [1993] : *C4.5 : Programs for Machine Learning* (Morgan Kaufmann, San Mateo, California). — cité en page(s) 86

RÄCKE, Harald [2002] : *Minimizing congestion in general networks.* Dans *Proceedings of the 43rd Symposium on Foundations of Computer Science (FOCS '02)*, (p. 43–52) (IEEE Computer Society, Washington, DC, USA). ISBN 0-7695-1822-2. — cité en page(s) 39

Rajman, Martin, Andrews, Pierre, del Mar Pérez Almenta, María et Seydoux, Florian [2005] : *Using the EDR large scale semantic dictionary : application to conceptual document indexing.* Dans *Proceedings of the 11th International Symposium on Applied Stochastic Models and Data Analysis, AMSDA 2005* (edité par Jacques Jansen et Philippe Lenca), (p. 98–105) (ENST Bretagne, Brest, France). ISBN 2-908849-15-1. — cité en page(s) 50

Rajman, Martin, Bui, Trung H., Rajman, Andréa, Seydoux, Florian, Trutnev, Alex et Quateroni, Silvia [2004] : *Assessing the usability of a dialogue management system designed in the framework of a rapid dialogue prototyping methodology.* ACTA ACUSTICA united with ACUSTICA, the Journal of the European Acoustics Association (EAA) : International Journal on Acoustics, *tome 90*(6) :ISSN 1610-1928. S. Hirzel Verlag - Stuttgart, Germany. — cité en page(s) 82

Rajman, Martin, Rajman, Andréa, Seydoux, Florian et Trutnev, Alex [2003] : *Assessing the usability of a dialogue management system designed in the framework of a rapid dialogue prototyping methodology.* Dans *First ISCA Tutorial & Research Workshop on Auditory Quality of Systems* (Akademie Mont-Cenis). — cité en page(s) 82

Ripley, Brian D. [1987] : *Stochastic Simulation.* Wiley series in probability and mathematical statistics (John Wiley & Soons, Inc.). — cité en page(s) 103

Rissanen, Jorma [1989] : *Stochastic Complexity in Statistical Inquiry* (Singapore). — cité en page(s) 36

Roeck, Anne De, Sarkar, Avik et Garthwaite, Paul H. [2005] : *Even very frequent function words do not distribute homogeneously.* Dans *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP'05)* (edité par Galia Angelova, Kalina Bontcheva, Ruslan Mitkov, Nicolas Nicolov et Nikolai Nikolov) (European Commission as a Marie Curie Large Conference, MLCF-CT-2004-013233, Borovets, Bulgaria). — cité en page(s) 11

Sahlgren, Magnus [2005] : *An Introduction to Random Indexing.* Dans *Proceedings of the Methods and Applications of Semantic Indexing Workshop at the 7th International Conference on Terminology and Knowledge Engineering (TKE 2005)* (Copenhagen, Denmark). — cité en page(s) 13, 33

Salton, Gerard [1968] : *Automatic Information Organization and Retrieval.* McGraw-Hill computer science series. (McGraw-Hill, New York.). — cité en page(s) 6, 8, 18

Salton, Gerard [1971] : *The SMART Retrieval System : Experiments in Automatic Document Processing* (Prentice Hall). — cité en page(s) 9

Salton, Gerard et Buckley, Chris [1988] : *Term weighting approaches in automatic text retrieval.* IPM, *tome 24* :p. 513–523. — cité en page(s) 13, 14

Salton, Gerard et Buckley, Chris [1990] : *Improving retrieval performance by relevance feedback.* JASIS, *tome 41* :p. 288–297. — cité en page(s) 18

Salton, Gerard, Fox, Edward A. et Wu, Harry [1982] : *Extended boolean information retrieval.* Rapport technique, Cornell University. — cité en page(s) 8

Salton, Gerard et McGrill, Michael J. [1983] : *Introduction to Modern Information Retrieval* (McGraw-Hill, Inc., New York, NY, USA). — cité en page(s) 9

Salton, Gerard, Yang, C. S. et Yu, Clement T. [1975] : *A theory of term importance in automatic text analysis.* JASIS, *tome 26*(1) :p. 33–44. — cité en page(s) 11, 14

Savoy, Jacques [2006] : *Un regard statistique sur l'évaluation de performance : L'exemple de clef 2005.* Dans *Actes de la Troisième Conférence en Recherche d'Information et Applications (CORIA2006)*, (p. 73–84). — cité en page(s) 72

Schultz, C. K. (rédacteur) [1968] : *H. P. Luhn : Pioneer of Information Science ; selected works* (New York), spartan édition. — cité en page(s) 11

Schütze, Hinrich, Hull, David et Pedersen, Jan [1995] : *A comparison of classifiers and documents representations for the routing problem.* Dans *Proceedings of the Eighteenth Annual International ACM*

SIGIR conference on Research and development in information retrieval, (p. 229–237) (Seattle, WA.). — cité en page(s) 11, 13

Sebastiani, Fabrizio [2002] : *Machine learning in automated text categorization.* ACM Computing Surveys, *tome 34* (1) :p. 1–47. — cité en page(s) 12

Seydoux, Florian et Chappelier, Jean-Cédric [2005*a*] : *Hypernyms Ontologies for Semantic Indexing.* Dans *Proceedings of the Workshop on Methodologies and Evaluation of Lexical Cohesion Techniques in Real-world Applications (ELECTRA'2005), the 28th Annual International ACM SIGIR Conference*, (p. 49–55) (Salvador, Brazil). — cité en page(s) 40

Seydoux, Florian et Chappelier, Jean-Cédric [2005*b*] : *Indexation sémantique au moyen de coupes de redondance minimale dans une onthologie.* Dans *Actes de la 12ème conférence sur le Traitement Automatique des Langues Naturelles (TALN'2005)*, tome 1, (p. 33–42) (Dourdan, France). — cité en page(s) 40

Seydoux, Florian et Chappelier, Jean-Cédric [2005*c*] : *Minimum Redundancy Cut in Ontologies for Semantic Indexing.* Dans *Progress in Artificial Intelligence : Proc of the 12th Portuguese Conference on Artificial Intelligence, EPIA 2005 (TeMA Workshop on Text Mining and Applications)* (edité par Carlos Bento, Amílcar Cardoso et Gaël Dias), tome 3808 / 2005 de *Lecture Notes in Computer Science*, (p. 658–668) (Springer-Verlag GmbH, Covilhã, Portugal). ISBN : 3-540-30737-0. — cité en page(s) 40

Seydoux, Florian et Chappelier, Jean-Cédric [2005*d*] : *Semantic Indexing using Minimum Redundancy Cut in Ontologies.* Dans *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP'05)* (edité par Galia Angelova, Kalina Bontcheva, Ruslan Mitkov, Nicolas Nicolov et Nikolai Nikolov), (p. 486–492) (European Commission as a Marie Curie Large Conference, MLCF-CT-2004-013233, Borovets, Bulgaria). — cité en page(s) 40

Shannon, C. E. [1948] : *A mathematical theory of communication.* The Bell System Technical Journal, *tome 27* :p. 379–423. — cité en page(s) 40

Singhal, Amit [1997] : *Term weighting revisited.* Thèse de doctorat, Department of Computer Science, Cornell University. — cité en page(s) 15

Singhal, Amit, Salton, Gerard, Mitra, Mandar et Buckley, Chris [1995] : *Document length normalization.* Rapport technique, Department of Computer Science, Cornell University. — cité en page(s) 13

Stoer, Mechthild et Wagner, Frank [1997] : *A simple Min-Cut algorithm.* Journal of the ACM (JACM), *tome 44* (4) :p. 585–591. ISSN 0004-5411. — cité en page(s) 39

Stolcke, Andrea, Ries, Klaus, Coccaro, Noah, Shriberg, Elizabeth, Bates, Rebecca, Jurafsky, Dan, Taylor, Paul, Martin, Rachel, Ess-Dykema, Carol Van et Meteer, Marie [2000] : *Dialogue act modeling for automatic tagging and recognition of conversational speech.* Computational Linguistics, *tome 26* (3) :p. 339–373. — cité en page(s) 87

Sullivan, Danny [2000] : *Invisible web gets deeper.* Rapport technique 45, Search Engine Watch. — cité en page(s) 1

Tanaka, J.S. et Huba, G.J. [1984] : *Confirmatory hierarchical factor analysis of psychological distress measures.* Journal of Personality and Social Psychology, *tome 46* :p. 621–635. — cité en page(s) 100

Tomuro, Noriko [2001] : *Tree-cut and a lexicon based on systematic polysemy.* — cité en page(s) 36

Van Kommer, Robert, Rajman, Martin et Bourlard, Hervé [2000] : *Heading towards virtual-commerce portals.* Comtec, *tome 9* :p. 10–13. — cité en page(s) 77

Voorhees, Ellen M. [1993] : *Using Word Net To Disambiguate Word Senses For Text Retrieval.* Dans *Sixteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, (p. 171–80) (Association for Computing Machinery, Pittsburgh, Pennsylvania). — cité en page(s) 64

VOORHEES, Ellen M. [1994] : *Query expansion using lexical-semantic relations.* Dans *Proceedings of the 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval* (edité par W. B. Croft et C. J. van Rijsbergen), (p. 61–69) (ACM/Springer, Dublin, Ireland). — cité en page(s) 18

VOORHEES, Ellen M. [1998] : *Using WordNet for text retrieval.* Dans *WordNet : An Electronic Lexical Database* (edité par C. Fellbaum), chapitre 12, (p. 285–303) (MIT Press). — cité en page(s) 19

VOORHEES, Ellen M. [2001] : *The philosophy of information retrieval evaluation.* Dans *CLEF '01 : Revised Papers from the Second Workshop of the Cross-Language Evaluation Forum on Evaluation of Cross-Language Information Retrieval Systems*, tome 2406, (p. 355–370) (Springer-Verlag, London, UK). — cité en page(s) 57, 64

WHALEY, Jason M. [1999] : *An application of word sense disambiguation to information retrieval.* Rapport technique PCS-TR99-352, Dartmouth College, Computer Science, Hanover, NH. — cité en page(s) 19

WOODS, William A. [1997] : *Conceptual indexing : A better way to organize knowledge.* Rapport technique TR-97-61, Sun Microsystems Laboratories. — cité en page(s) 18

YANG, Yiming et PEDERSEN, Jan O. [1997] : *A comparative study on feature selection in text categorization.* Dans *Proceedings of the 14th International Conference on Machine Learning*, (p. 412–420) (Morgan Kaufmann). — cité en page(s) 12

Florian Seydoux
Avenue Préfaully 25C
CH-1022 Chavannes-Près-Renens (VD)
Tél. fixe : +41 (0)21 635 18 83
Cellulaire : +41 (0)76 495 01 02
E-mail : florianseydoux@gmail.com
né en 1973, nationalité Suisse

Ingénieur Informaticien (Software)

Formation

dès 2003 **Doctorat en informatique** – intelligence artificielle,
École Polytechnique Fédérale de Lausanne (EPFL), Suisse.
Exploitation de connaissances sémantiques externes dans les représentations vectorielles en recherche documentaire,
direction : Drs M. Rajman et J.-C. Chappelier, Laboratoire d'Intelligence Artificielle.

2002 **Spécialisation postgrade** – «Language and Speech Engineering»,
École Polytechnique Fédérale de Lausanne (EPFL), Suisse.
Mémoire : *Gestionnaire de dialogues pour serveur vocal interactif*,
direction : Dr. M. Rajman, Laboratoire d'Intelligence Artificielle.

1995 – 1999 **(M. Sc.) Ingénieur EPF en Informatique** – orientation «logiciel d'applications»,
École Polytechnique Fédérale de Lausanne (EPFL), Suisse.
Travail de diplôme : *Analyse syntaxique probabiliste du langage naturel*,
direction : Dr. J.-C. Chappelier, Laboratoire d'Intelligence Artificielle.

1989 – 1995 **Ingénieur HES en Génie Électrique** – orientation télécommunication,
École d'Ingénieurs de Genève (EIG), Suisse.
Travail de diplôme (avec mention) : *Outils de gestion pour RNIS*,
direction : G. Litzistorf, Laboratoire de Transmission de Données.

Expérience professionnelle

dès 2006 **Ingénieur de développement logiciel**,
SpinX-technologies SA – Genève, Suisse.
Optimisation de processus, planning, langage et compilation, etc.

1999 – 2005 **Assistant de recherche et d'enseignement**,
Laboratoire d'Intelligence Artificielle – École Polytechnique Fédérale de Lausanne, Suisse.

Activité de recherche :

(2003–2005)
- EXKNOWTIC [projets Fond National Suisse 2100-066901 & 200020-103529, EPFL]
  «*Intégration de sources de connaissances pour l'amélioration des modèles à base de sémantique distributionnelle*»

(2000–2002)
- Sting [projet européen EU IST99-20847, Computer Technology Institute (Patras)]
  «*Evaluation of Scientific & Technological Innovation and Progress in Europe through Patents*»

(1999–2001)
- InfoVox [projet CTI 4247.1, EPFL, IDIAP (Martigny), Swisscom]
  «*Interactive Voice Servers for Advanced Computer Telephony Application*»

(1999)
- ELSE [projet européen EU LE4-8340, LIMSI/CNRS (Paris)]
  «*Evaluation in Language and Speech Engineering*»

Activité d'enseignement :

(2000–2005)
- Cours d'informatique I et II (annuel, sections math et physique, audience $\sim$ 200 étudiants)
  Programmation orientée objet en C++, introduction à l'algorithmique, bases du système Unix ;
  préparation et encadrement des travaux pratiques, projets et examens, gestion des étudiants-assistants.
- Réalisation et conduite de tutoriaux en gestion de dialogues
  dans le cadre de workshops (conférences TALN 2000 et IJCAI 2001), École d'été (Brno 2001), module de cours postgrade EPFL (2001 et 2002).
- Encadrement de divers projets d'étudiants (semestre et master EPFL).

1997 – 2000 **Assistant de laboratoire** (temps partiel),
Laboratoire d'Informatique Industrielle – École d'Ingénieurs de Genève, Suisse.
Expérimentation et préparation de protocoles de laboratoire pour les étudiants :
(entre autres) applications embarquées (cibles MVME) et développement croisé Ada-C/C++,
validation d'algorithmes via système CSAO/CD-Lustre, analyse non-intrusive d'exécutifs temps réels.

Peer reviewed publications :

Books and Chapters

2004 Jean-Cédric Chappelier et Florian Seydoux,
(1re éd) « *C++ par la pratique. Recueil d'exercices corrigés et aide-mémoire* »
Presses Polytechniques et Universitaires Romandes (PPUR), octobre 2005 (2e éd.), Lausanne, Suisse.

Journal Papers

2004 Martin Rajman, Huu-Trung Bui, Andréa Rajman, Florian Seydoux, Alex Trutnev and Silvia Quarteroni,
« *Assessing the usability of a dialogue management system designed in the framework of a rapid dialogue prototyping methodology* »
ACTA ACUSTICA united with ACUSTICA, the Journal of the European Acoustics Association (EAA) : International Journal on Acoustics, n°6, nov-dec 2004, pp 1096-1111, ISSN 1610-1928 S. Hirzel Verlag–Stuttgart, Germany.

Conference and Workshop Papers

2005 Florian Seydoux and Jean-Cédric Chappelier,
« *Minimum Redundancy Cut in Ontologies for Semantic Indexing* »
Progress in Artificial Intelligence : Proc of the 12th Portuguese Conference on Artificial Intelligence, EPIA 2005 (TeMA Workshop on Text Mining and Applications), december 2005, Lecture Notes in Computer Science (Springer-Verlag GmbH), vol. 3808/2005, pp 658–668, ISBN : 3-540-30737-0, Covilhã, Portugal.

2005 Florian Seydoux and Jean-Cédric Chappelier,
« *Semantic Indexing using Minimum Redundancy Cut in Ontologies* »
Proc. of the International Conference on Recent Advances in Natural Language Processing (RANLP'05), pp 486–492, september 2005, Borovets, Bulgaria.

2005 Florian Seydoux and Jean-Cédric Chappelier,
« *Hypernyms Ontologies for Semantic Indexing* »
Proc. of the Workshop on Methodologies and Evaluation of Lexical Cohesion Techniques in Real-world Applications (ELECTRA'2005), the 28th Annual International ACM SIGIR Conference, pp 49–55, august 2005, Salvador, Brazil.

2005 Florian Seydoux and Jean-Cédric Chappelier,
« *Indexation sémantique au moyen de coupes de redondance minimale dans une onthologie* »
Actes de la 12ème conférence sur le Traitement Automatique des Langues Naturelles (TALN'2005) vol. 1, pp 33-42, juin 2005, Dourdan, France.

2005 Martin Rajman, Pierre Andrews, María del Mar Pérez Almenta and Florian Seydoux,
« *Using the EDR large scale semantic dictionary : application to conceptual document indexing* »
Proc. of the 11th International Symposium on Applied Stochastic Models and Data Analysis, AMSDA'05, pp 98–195, may 2005, Brest, France.

2003 Florian Seydoux, Alex Trutnev and Martin Rajman,
« *Dialogue Management with weak speech recognition : a pragmatic approach* »
ISCA workshop on Error Handling in Dialogue Systems, august 2003, Chateau-d'Oex, Switzerland.

2003 Martin Rajman, Andréa Rajman, Florian Seydoux et Alex Trutnev,
« *Prototypage rapide et évaluation de modèles de dialogue finalisés* »
Traitement Automatique des Langues Naturelles (TALN), juin 2003, Batz-sur-Mer, France.

2003 Martin Rajman, Andréa Rajman, Florian Seydoux and Alex Trutnev,
« *Assessing the usability of a dialogue management system designed in the framework of a rapid dialogue prototyping methodology* »
First ISCA Tutorial & Research Workshop on Auditory Quality of Systems, april 2003, Akademie Mont-Cenis, Germany.

2002 Martin Rajman, Vivi Peristera, Jean-Cédric Chappelier, Florian Seydoux and Antonis Spinakis,
« *Evaluation of Scientific and Technological Innovations using Statistical Analysis of Patents* »
6th Int. Conf. on the Statistical Analysis of Textual Data (JADT), pp 641–652, march 2002, Saint-Malo, France.