École des mines de Paris
Collège doctoral

N° attribué par la bibliothèque
/__/__/__/__/__/__/__/__/__/__/

Thèse

pour obtenir le grade de

Docteur de l'École des Mines de Paris
Spécialité : Informatique Temps Réel, Robotique et Automatique

présentée et soutenue publiquement
par

Christophe CHOTTEAU

le 12 décembre 2003

CORRÉLATION SÉMANTIQUE ENTRE DOCUMENTS
Application à la recherche d'information juridique sur le Web

Jury

M.	P.	Constant	Examinateur
M.	J-J.	Girardot	Rapporteur
M.	R.	Mahl	Directeur de thèse
M.	C.	Roche	Examinateur
M.	P.	Zweigenbaum	Rapporteur

Remerciements

Je tiens tout d'abord à remercier Robert Mahl, Directeur du Centre de Recherche en Informatique de l'École des Mines de Paris, qui m'a accueilli au sein de son équipe et m'a toujours offert les moyens de poursuivre ma thèse dans les meilleures conditions. Je lui exprime une sincère gratitude pour avoir su me laisser libre de pousser jusqu'à leur accomplissement les idées qui me tenaient à cœur tout en leur insufflant sa conception de l'ingénierie des connaissances.

Je désire exprimer ma reconnaissance envers Jean-Jacques Girardot et Pierre Zweigenbaum pour avoir accepté d'être les rapporteurs de ce travail ainsi que pour les remarques détaillées qu'ils m'ont faites et qui ont ainsi contribués à l'amélioration de ce manuscrit. Mes remerciement vont également à Christophe Roche pour avoir accepté de présider le jury.

Je remercie Patrick Constant ainsi que les membres de la société Pertimm, qui ont accompagné mes travaux de recherche dans leur phase de réflexion mais également d'un point de vue matériel, pour m'avoir fait bénéficier de leurs avancées technologiques.

La réussite d'une thèse, ou tout du moins son déroulement dans une relative sérénité, tient aussi pour beaucoup à l'aide et au soutien que l'on peut trouver dans l'équipe de recherche qui vous accueille. De ce point de vue, travailler au sein du Centre de Recherche en Informatique de l'École des Mines de Paris a été un atout pour moi. Je remercie donc l'ensemble des membres du centre de m'avoir accepté en leur sein.

Enfin, je remercie mes proches, parents et amis, pour m'avoir soutenu durant ces trois années.

Table des matières

1 Introduction **11**

1.1 Contexte et objectif . . . 11

1.2 Organisation de la thèse . . . 13

I État de l'art **15**

2 La recherche d'information **17**

2.1 L'accès à l'information sur Internet . . . 18

2.1.1 Internet : un média de taille ! . . . 18

2.1.2 Rechercher et trouver . . . 21

2.1.2.1 Les annuaires . . . 21

2.1.2.2 Les moteurs de recherche . . . 22

2.1.2.3 Les méta-moteurs . . . 23

2.2 Savoir interroger . . . 24

2.2.1 Une requête inaccessible . . . 24

2.2.1.1 Savoir formuler ses requêtes . . . 25

2.2.1.2 Des résultats pertinents ? . . . 26

2.2.2 Guider les utilisateurs . . . 27

2.2.2.1 L'expansion de requêtes . . . 27

2.2.2.2 La classification interactive . . . 28

2.2.2.3 La corrélation de documents . . . 28

2.3 Prise de décision et recherche d'information . . . 29

2.3.1 Prise de décision et accès à l'information ? . . . 29

2.3.2 Le processus de prise de décision . . . 29

2.3.2.1 Comprendre pour extraire de l'information . . . 30

2.3.2.2 Comprendre pour répondre à une question . . . 31

2.4 Notre orientation . . . 32

3 Les visages de la corrélation **35**

3.1 La corrélation topologique . . . 35

3.1.1 La notion de liens hypertextes . . . 37

3.1.2 PageRank 37
3.1.3 L'algorithme HITS 38
3.1.3.1 Déterminer un sous-graphe utile du Web 39
3.1.3.2 L'analyse des liens 40
3.1.3.3 Le cadre mathématique 41
3.1.4 L'algorithme Companion 42
3.1.4.1 Détermination de S pour l'algorithme Companion 43
3.1.4.2 Simplification de S 43
3.1.4.3 Ébauche des scores de « sites » et de « feuilles » 44
3.1.4.4 Utilisation de HITS 44
3.1.5 L'algorithme de Co-citation 44
3.1.6 Autres techniques 45
3.1.7 Discussion 45
3.2 La corrélation empirique 47
3.2.1 Une étude empirique 47
3.2.2 Discussion 48
3.3 La corrélation hiérarchique 49
3.3.1 La classification 50
3.3.2 La classification de termes 50
3.3.3 La classification de documents 51
3.3.3.1 La classification non-hiérarchique 51
3.3.3.2 La classification hiérarchique 52
3.3.4 Des approches nouvelles 53
3.3.4.1 La classification interactive : exemple de Scatter/Gather 53
3.3.4.2 La classification topologique 54
3.3.4.3 Une approche hybride de classification 55
3.3.5 Discussion 56
3.4 La corrélation linguistique 57
3.4.1 Extraire les termes saillants 57
3.4.1.1 Les méthodes de pondération linguistique 57
3.4.1.2 Les méthodes de pondération statistique 58
3.4.2 L'extraction de documents corrélés 60
3.4.2.1 Calcul de distance entre signatures 60
3.4.2.2 La signature vue comme une requête 61
3.4.2.3 La corrélation vue sous la forme d'un réseau de termes 62
3.4.3 Discussion 62

II Matériel et méthodes **65**

4 Une corrélation dans l'usage **67**
4.1 Un corpus de référence . . . 67
4.1.1 Définition du corpus de référence . . . 68
4.1.2 Choix du corpus de référence . . . 70
4.1.2.1 Les codes du droit français . . . 70
4.1.2.2 Le Journal Officiel de la République française . . . 71
4.2 Quelle corrélation pour notre corpus . . . 72
4.2.1 La définition des besoins . . . 72
4.2.1.1 Une base documentaire vivante . . . 72
4.2.1.2 Utiliser la richesse du corpus . . . 73
4.2.1.3 Le document comme unité de base pour la corrélation . . . 74
4.2.2 Vers une voie de corrélation linguistique . . . 75
4.3 Définition de la méthode utilisée . . . 76
4.3.1 L'élaboration des signatures lexicales . . . 77
4.3.2 La recherche des résultats corrélés . . . 77
4.4 Les degrés de liberté . . . 78
4.4.1 L'élaboration des signatures lexicales . . . 80
4.4.2 Interrogation de la base . . . 81
4.4.2.1 La formulation des requêtes . . . 81
4.4.2.2 L'unité de recherche . . . 82
4.4.3 Le classement des résultats . . . 82

5 La quête du sens **83**
5.1 Sélectionner les candidats termes . . . 84
5.1.1 Définir les catégories de candidats termes . . . 84
5.1.1.1 Les expressions chiffrées . . . 84
5.1.1.2 Les mots composés . . . 85
5.1.1.3 Les syntagmes . . . 85
5.1.1.4 Les mots simples . . . 86
5.1.2 Extraire les candidats termes . . . 86
5.1.2.1 Méthodes d'extraction de termes . . . 87
5.1.2.2 Le choix de l'outil Sylex . . . 87
5.1.3 Filtrage des candidats termes . . . 88
5.2 Identifier les termes valides . . . 88
5.2.1 Indice T_f . . . 89
5.2.2 Indice I_{df} . . . 90
5.2.3 Indice $T_f.I_{df}$ . . . 91
5.2.4 Entropie d'un terme . . . 92
5.3 Pondération statistique avec T_{ifr} . . . 93

5.3.1 Mesure de répartition par T_{ir} . . . 94
5.3.2 Lissage par T_{if} . . . 96
5.3.3 Masque d'hapax pour T_{ifr} . . . 97
5.4 Mise en place d'une expérimentation . . . 99
5.4.1 Procédure de validation . . . 99
5.4.2 Résultats . . . 100
5.4.2.1 Nombre de termes validés . . . 101
5.4.2.2 Accord entre évaluateurs . . . 103
5.5 Comparaison des pondérations . . . 105
5.5.1 Créations de listes pondérées . . . 106
5.5.2 Protocole de comparaison . . . 106
5.5.3 Résultats . . . 107
5.5.4 Analyse . . . 107
5.6 Enseignements . . . 109

6 Définir une unité de recherche 111
6.1 Des unités de recherche pré-existantes . . . 111
6.1.1 Une justification sémantique . . . 112
6.1.2 Une justification pratique . . . 113
6.2 Une approche systématique . . . 114
6.2.1 Création d'un panel de documents . . . 115
6.2.2 Procédure d'interrogation systématique . . . 115
6.2.2.1 Parcours d'arborescence . . . 117
6.3 Expérimentation . . . 122
6.3.1 Résultats . . . 122
6.3.1.1 Le nombre de requêtes valides générées . . . 123
6.3.1.2 Le taux d'utilisation des termes . . . 124
6.3.1.3 Le rang moyen du dernier terme . . . 125
6.3.2 Enseignements . . . 127
6.3.2.1 L'unité de recherche . . . 127
6.3.2.2 Comparaisons des différentes pondérations . . . 129

III Évaluation 131

7 Optimiser les signatures lexicales 133
7.1 Un benchmark expérimental . . . 133
7.1.1 Élaboration du benchmark . . . 134
7.1.1.1 L'extraction des noyaux de documents . . . 135
7.1.1.2 L'enrichissement des noyaux . . . 136
7.2 Optimisation des signatures lexicales . . . 137

7.2.1 La taille de la signature et le facteur sémantique 137
7.2.2 Évaluation des performances 138
7.2.3 Observations 138
7.3 La recherche des résultats corrélés 141
7.3.1 Classement des résultats 142
7.3.2 Résultats 143
7.4 Enseignements 144

8 La qualité des résultats corrélés 147
8.1 Le choix du corpus de validation 147
8.1.1 TREC 148
8.1.2 AMARYLLIS 149
8.1.3 Vers une méthode de validation manuelle 150
8.2 Mise en place d'un protocole de validation 153
8.2.1 Procédure de validation 153
8.2.2 Choix et accords des évaluateurs 155
8.3 Comparaison des méthodes de corrélation 156
8.3.1 Protocole de comparaison 156
8.3.1.1 La précision des résultats 156
8.3.1.2 La qualité des résultats 158
8.3.2 Résultats 160
8.4 Enseignements 161

IV Synthèse 163

9 Description d'une méthode de corrélation 165
9.1 Bilan sur la corrélation 165
9.1.1 Entre performance et qualité des résultats 165
9.1.2 Description de la méthode de corrélation employée 167
9.2 Un outil de corrélation entre documents 169
9.3 Discussion autour d'un panel de document corrélés 172
9.3.1 Les termes du domaine juridique 172
9.3.2 Taille des signatures et taille de documents 173
9.3.3 L'aspect linguistique 174

10 Conclusions et perspectives 177
10.1 Contributions 178
10.1.1 Une corrélation sémantique 178
10.1.2 Élaboration d'un nouvel indice de pondération 178
10.1.3 Des signatures lexicales dédiées à la corrélation 179

10.1.4 Description d'une méthodologie dédiée à la corrélation 179
10.1.5 Implémentation et réalisation 179
10.2 Perspectives 180
10.2.1 Vers une meilleure intégration 180
10.2.2 Des signatures plus compactes 180
10.2.3 Une mixité mal évaluée 180
10.2.4 Utiliser les caractéristiques du corpus 181

Table des figures **186**

Liste des tableaux **188**

Bibliographie **199**

A Panel de documents juriques **201**

B Panel réduit de documents juridiques **207**

C Évaluation des résultats corrélés **211**

D Exemple de documents corrélés issus du JO **215**

Chapitre 1

Introduction

1.1 Contexte et objectif

Que ce soit dans le monde professionnel, dans le domaine politique, dans le cadre des loisirs ou encore dans celui de l'éducation, la communication est omniprésente [Mie89]. L'essor des Nouvelles Technologies de l'Information et de la Communication est aujourd'hui indéniable. L'obligation actuelle de communiquer se traduit notamment par le développement exponentiel d'Internet et du nombre de ses usagers. Si en 1993, par exemple, la toile comportait quelques milliers de pages accessibles, on en compte aujourd'hui plus de deux milliards avec un taux de croissance estimé à plus de soixante millions de pages par mois [Bou99]. De la même manière, le profil des utilisateurs a évolué : les experts (ingénieurs, chercheurs, étudiants) ne sont plus les seuls à surfer. Face à cet engouement pour ce flot d'informations, les outils de recherche dont nous disposons sont le plus souvent limités par la quantité d'information à traiter.

Il est un domaine où cette quête d'information prend toute sa signification : le domaine juridique. En effet, selon la célèbre maxime *Nul n'est censé ignorer la loi*, mais encore faut-il pouvoir y accéder. Or, à l'image de la croissance effrénée d'Internet, les bases de données juridiques électroniques se sont étoffées avec le temps, l'augmentation des capacités de stockage et la demande des internautes. L'informatique, ou plutôt le domaine de l'ingénierie des connaissances, tient, depuis quelques temps déjà, une place importante dans la problématique de l'accès au droit. Grâce à ses capacités de diffusion et ses facilités d'accès, Internet devient un acteur majeur de cette problématique. La création, en France, d'un service public d'accès au droit par l'instauration du service public de la diffusion du droit par l'Internet (Décret n° 2002-1064 du 7 août 2002) en est un exemple.

Dès lors qu'un grand nombre de documents relatifs au droit français sont gratuitement consultables sur Internet (l'essentiel des textes législatifs et réglementaires ainsi que de nombreuses jurisprudences), la question devient celle-ci : comment accéder à ces documents ? Si nul ne peut se soustraire à la loi sous prétexte qu'il l'ignore, toute personne devrait être en mesure d'y accéder.

L'accès à l'information sur Internet se ramène le plus souvent à l'utilisation d'outils tels que les moteurs de recherche. Ils représentent de loin les instruments les plus utilisés pour satisfaire le besoin informationnel croissant des internautes. Ils ont pour caractéristique de prendre une question en entrée et de renvoyer, comme réponses, des informations triées, le plus souvent, en fonction d'une pertinence propre à chaque moteur. Cette utilisation qui comble, en principe, les attentes des internautes, présente une sévère limitation qui est de savoir correctement formuler sa question. Sans bonne question, aucune réponse intéressante n'est envisageable. Les travaux exposés à présent décrivent une méthode complémentaire d'accès à l'information : la corrélation entre documents.

Cette thèse reste dans le prolongement des travaux menés par G. Lame [Lam02] dont l'aboutissement a permis de réaliser un système d'aide à la reformulation de requêtes dédié aux internautes. À leur image, ils prennent place sur un corpus juridique et tentent d'améliorer la diffusion du droit. Néanmoins, nous nous plaçons en aval du travail précédent puisque notre but est de réaliser une méthode interactive de corrélation entre documents permettant aux utilisateurs de retrouver des documents sémantiquement proches du document initial.

La composante sémantique de notre méthode doit être explicitée car cette notion revêt de multiples aspects. Pour schématiser, la sémantique peut être vue sous deux aspects distincts [Cha01] : l'un linguistique d'une *théorie visant à rendre compte des phénomènes signifiants dans le langage*, l'autre logico-symbolique d'*étude générale des relations entre les signes et leurs référents*.

L'approche linguistique peut être séparée en l'étude du sens des mots et celle du sens des énoncés. On distingue deux voies compatibles dans la détermination de l'origine du sens d'un mot :

– une voie componentielle [Gre86, Ras89] dans laquelle le sens d'un mot se construit à partir d'un ensemble d'éléments de sens ;
– une voie relationnelle [Har68] dans laquelle le sens d'un mot provient du fait qu'il apparaît dans les mêmes contextes ou des contextes similaires à ceux d'autres mots.

L'approche logico-symbolique étudie le lien entre les signes et leurs référents, autrement dit la représentation des connaissances. Le postulat est que le monde et les connaissances sur le monde peuvent être représentés d'une manière logique. Manipuler des connaissances revient alors à effectuer des opérations logiques sur des symboles [Sab00].

L'approche de la présente thèse est linguistique et relationnelle, mais à la différence d'autres travaux contemporains, notre problématique n'est pas de rechercher le contexte d'un terme dans un document ou de désambiguïser sa signification. Nous construisons, pour un document donné, une brève description terminologique ou signature qui reflète ses thématiques et permet de rechercher des documents corrélés. Les termes présents dans une même signature sont ainsi mis en relation et comparés à d'autres documents dans le but de retrouver des contextes d'utilisations similaires.

Même si nous nous éloignons quelque peu de la définition consensuelle donnée de la sémantique, ces travaux montrent notamment lors de l'élaboration des signatures ou de la recherche de documents corrélés que la notion de sémantique peut être appliquée ici.

1.2 Organisation de la thèse

La finalité de notre méthode de corrélation entre documents est de pouvoir faciliter au quotidien les recherches documentaires opérées par les internautes, tout en leur fournissant une information de meilleure qualité. La conception et la réalisation d'un tel outil s'inscrit dans la lignée des travaux réalisés dans le cadre de la recherche d'information et l'ingénierie des connaissances. Avant d'entrer pleinement dans l'univers de la corrélation, nous abordons dans le chapitre 2 la problématique de la recherche et de l'accès à l'information sur Internet et plus généralement sur des documents électroniques.

Le domaine de la corrélation entre documents est restreint par rapport à des thématiques de recherche comme la classification ou la catégorisation de documents, néanmoins cette notion de corrélation revêt de multiples aspects. Le chapitre 3 constitue un état de l'art et permet de définir les types de corrélation existants qui sont au nombre de quatre : corrélations topologique, empirique, hiérarchique et linguistique.

Le travail entrepris a pour finalité de déterminer la meilleure stratégie de corrélation possible adaptée à notre base documentaire. Dans ce contexte, la motivation du chapitre 4 est dans un premier temps de définir un corpus de référence qui sera représentatif de notre base documentaire pour, ensuite, à partir de ses caractéristiques et de nos besoins informationnels, rechercher parmi les voies de corrélation évoquées au chapitre 3, celle qui est la mieux adaptée. Ce chapitre 4 énonce, dans un deuxième temps, les étapes clefs de la méthode de corrélation retenue ainsi qu'une partie des problématiques qui lui sont propres.

Au terme du chapitre 4, nous adoptons une méthode de corrélation linguistique. L'étape clef de cette voie de corrélation est la création de signatures lexicales. Le chapitre 5 permet de définir la notion de signature lexicale et énumère les nombreux degrés de liberté qui sont à envisager lors de leur génération. L'objet de ce chapitre 5 est double : définir le concept de signature lexicale et justifier leur utilisation en tant que signature sémantique d'un document.

Le chapitre 6 répond à une question évoquée au chapitre 4 qui concerne la définition d'une unité de recherche. Cette notion d'unité de recherche est primordiale pour la recherche de documents corrélés ; de sa définition dépend, en partie, leur qualité et leur quantité. L'objectif de ce chapitre 6 est de déterminer, pour notre corpus de référence, la taille optimale de l'unité de recherche à considérer.

Les signatures lexicales peuvent avoir de multiples utilisations. Dans notre approche celles-ci sont exclusivement consacrées à la recherche de résultats corrélés. Dans cette optique, les signatures que nous générons et utilisons doivent être optimisées pour la recherche de documents corrélés. Les chapitres 7 et 8 étudient les signatures et leur capacité à retrouver des résultats corrélés suivant un axe quantitatif et un axe qualitatif. La finalité de ces deux chapitres est de choisir, parmi les différentes méthodes possibles, celle qui reflète le meilleur compromis entre quantité et qualité des documents corrélés.

Le bilan de nos investigations concernant la recherche d'une méthode de corrélation est exposé au travers du chapitre 9. Ce chapitre décrit la méthode de corrélation la mieux adaptée et son intégration au sein de notre base documentaire.

Enfin le chapitre 10 fait office de conclusion. Il permet d'envisager les perspectives de nos travaux. Il synthétise également nos contributions vis à vis du domaine de la corrélation, sans oublier d'aborder les limites de la démarche adoptée.

Première partie

État de l'art

Chapitre 2

La recherche d'information sur Internet

Depuis l'avènement d'Internet et la démocratisation des ordinateurs personnels, le contexte d'utilisation des systèmes de recherche d'information a profondément changé : jadis réservés et conçus pour être utilisés par des utilisateurs aguerris, ils sont désormais accessibles par tous. L'information désormais se *consomme* à toute heure et sous toutes ses formes et en grande quantité, que ce soit dans un cadre ludique ou professionnel, les utilisateurs accèdent sur une même machine à des journaux en ligne, lisent leur courrier, explorent le Web et rédigent leurs rapports.

Sous ses aspects communautaires, accueillant, sa grande diversité d'information et son accès planétaire sans frontière, Internet semble être une porte ouverte sur le monde idéal pour que l'internaute puisse mener sans encombre sa quête d'information. La réalité est tout autre. À défaut d'être un monde paradisiaque, Internet regorge de pièges et de limites où le voyageur virtuel tombe volontiers. Le but de ce chapitre 2 est de brosser un rapide portrait de la recherche d'information sur Internet pour comprendre au final dans quelle logique s'inscrit notre étude sur la corrélation entre documents.

Dans cette optique, la section 2.1 sera consacrée à décrire plus finement les caractéristiques d'Internet, mettre en avant les problèmes rencontrés et énumérer quelques outils classiques de recherche d'information couramment utilisés. Dans la section 2.2 nous évoquerons la principale difficulté rencontrée par les internautes pour accéder à l'information : formuler leurs besoins d'information sous forme de questions ou de requêtes. Nous verrons dans la section 2.3 que l'une des solutions envisagées pour faciliter l'accès à l'information est de pouvoir prendre des décisions en lieu et place de l'utilisateur. Finalement la section 2.4 tirera profit de nos enseignements pour décrire l'orientation de notre méthode de corrélation.

2.1 L'accès à l'information sur Internet

Le nombre et la diversité des informations disponibles sur Internet fait de ce média une source d'information quasi inépuisable à la disposition de tous. Malgré un grand nombre d'outils de recherche d'information disponibles et leurs perpétuelles évolutions, la recherche d'une information précise reste toujours délicate et ressemble, pour l'utilisateur néophyte, à un chemin de croix.

La principale raison de ce difficile accès provient de la nature même d'Internet. Les outils de recherche utilisent des algorithmes et des techniques de recherche [Sal89, FC85] maintes fois améliorés ; malheureusement ces techniques s'adressent à une information structurée, cohérente et de taille modeste comparée aux milliards de documents qui constituent le Web [ACGM+01].

2.1.1 Internet : un média de taille !

Pour bien comprendre les difficultés rencontrées pour accéder aux informations qui en sont captives, il faut s'imaginer le Web comme un espace quasi infini, un espace non structuré où les informations sont éparpillées et dispersées aux quatre coins de la galaxie virtuelle, un espace en perpétuelle évolution où les informations sont constamment mises à jour.

L'internaute est confronté à de multiples problèmes pour étancher sa soif d'information mais le premier est de pouvoir explorer et exploiter simultanément autant de données. De nombreuses études [BYBC+00, LG98b, LG99, BB99] ont été menées pour tenter d'estimer la taille du Web et ainsi mesurer l'ampleur de la tâche à accomplir. Leurs résultats sont légèrement différents mais tous s'accordent à dire que le nombre de pages disponibles sur Internet se compte désormais en milliards, l'incertitude est de savoir s'il s'agit de dizaines ou de centaines de milliards de pages. La figure 2.1 donne une idée de la progression du nombre de pages publiées sur Internet entre 1999 et 2001. Sachant que la taille moyenne d'un document publié sur Internet est compris entre 5 et 10 Ko[1], la masse représentée par le Web est d'au moins 10 To[2]. Une masse en perpétuelle expansion puisque la quantité d'information publiée double en moyenne tous les deux ans [Bou99], un rythme soutenu qui représente 2 millions de nouvelles pages par jour et qui ne semble pas près de fléchir.

Le deuxième défi imposé par Internet est lié à la nature versatile de ce média qui se renouvelle constamment par des opérations de mise à jour ou de modifications du contenu des documents existants [PP97, WM99, DFKM97, CGM00]. Un renouvellement qui est également amplifié par la création et la disparition quotidienne de millions de documents.

[1] Kilo-octets.
[2] Tera-octets.

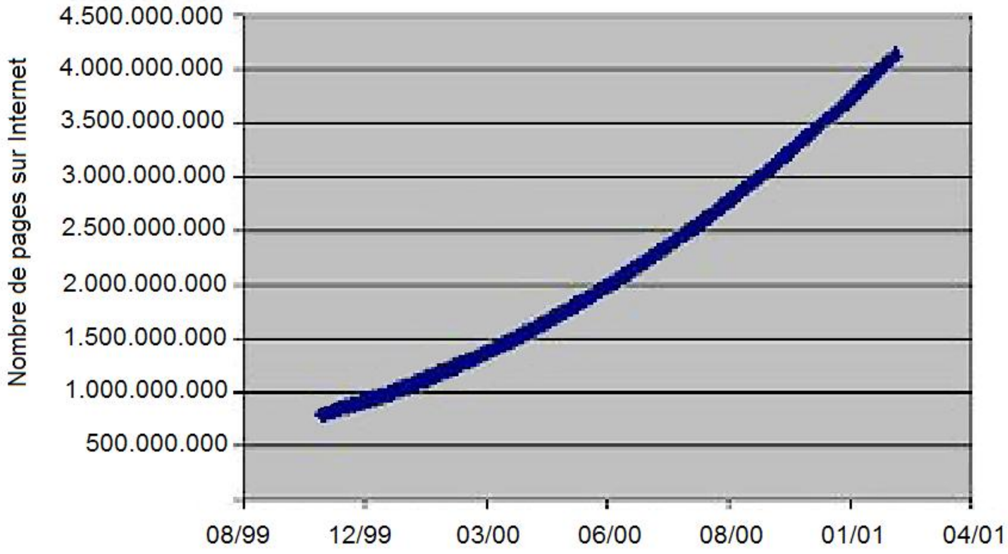


FIG. 2.1 – *La croissance du Web entre 1999 et 2000 [Cyv00]*

À titre d'exemple, une enquête menée en 2000 [CGM00] qui portait sur l'étude d'un demi million de pages Web pendant une durée de 4 mois a montré que 23% des pages avaient été quotidiennement mises à jour, c'est à dire que leur contenu avait changé de manière significative et que 40% des pages en `.com` avaient été quotidiennement modifiées. Sur cet échantillon de pages, l'étude a établi que la durée de vie moyenne des documents est de l'ordre de dix jours, période au bout de laquelle l'URL du document n'était plus valide. Tout ceci dans le but d'insister sur la grande volatilité des informations publiées et de montrer que les outils actuels ne sont pas capables pour l'instant de gérer cette masse d'information mouvante sans en omettre une grande partie.

Internet est un monde de démesure, de par la quantité d'information présente, la difficulté de mettre à jour cette base d'information ; mais les problèmes ne se limitent pas à cela : Internet est également un monde obscur, difficile à sonder, où une grande partie des informations présentes ne sont jamais retrouvées. Le Web comporte ainsi une face cachée nommée “ Web invisible[3] ” qui correspond à l'ensemble des documents non indexés par les outils de recherche traditionnels que sont les moteurs de recherche.

Ce Web invisible se compose de plusieurs types de documents qui sont :

- des documents à accès sécurisé (Accès restreint par mot de passe) ;
- des formulaires ;
- des documents déclarés non indexables (Utilisation du fichier de référencement *robots.txt*) ;
- des informations contenues dans des bases de données ;

[3]L'expression “ Invisible Web ” fut lancée en 1994 par Jill Ellsworth.

– des documents générés dynamiquement.

Cette liste non exhaustive donne un aperçu des types de documents incriminés qui sont des documents que l'on croise couramment sur Internet. La figure 2.2, reproduite d'après une étude de la société BrightPlanet consacrée au Web Invisible, illustre à cet effet la répartition de ses documents [Pla00].

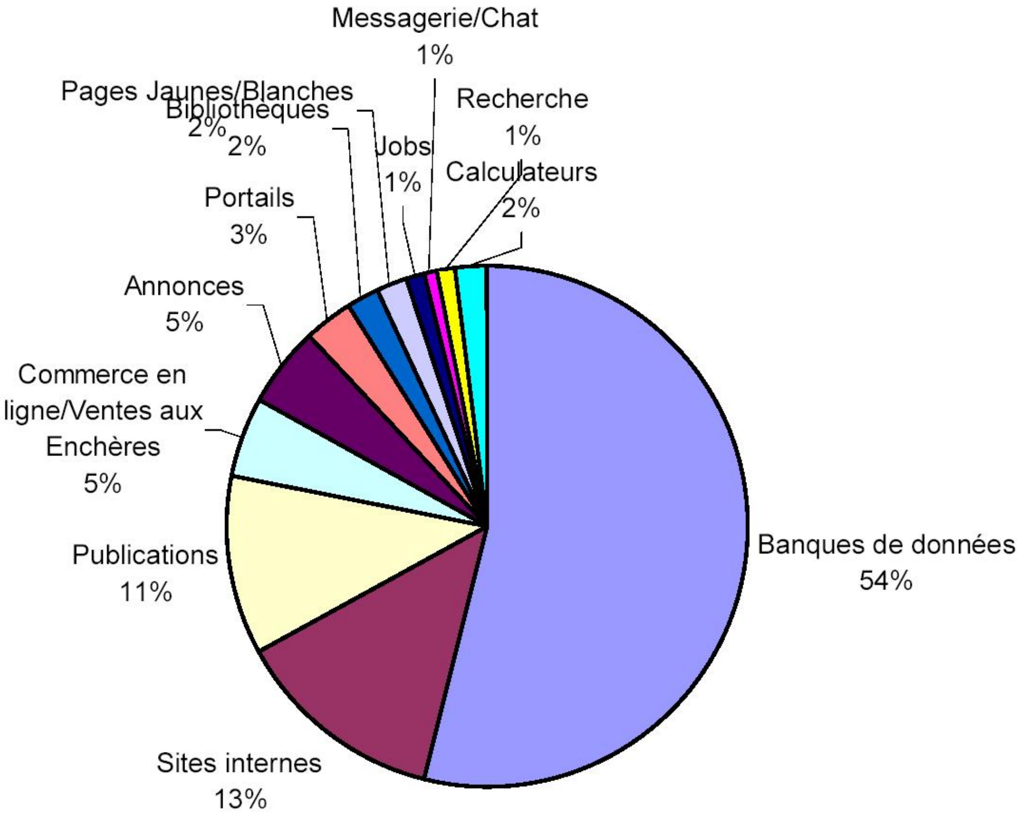


FIG. 2.2 – *Le Web invisible par type de document [Pla00]*

Le Web invisible n'est pas non plus synonyme d'information non référencée ; cette information n'est pas indexable par des moteurs traditionnels mais malgré tout elle reste accessible et indexée sur les sites concernés et à en croire certains, il s'agirait même de l'information la mieux structurée et la mieux organisée du Web [Ouf01]. Dernière caractéristique concernant ce Web invisible : sa taille. Différentes études menées pour évaluer sa taille [Ouf01] ont montré qu'elle serait environ 250 fois plus importante que la partie visible et indexée. Cette taille impressionnante a pour origine son expansion et son taux de croissance, bien supérieur à ceux de son alter ego le " Web Visible ".

La constatation qui s'impose est de dire qu'Internet est un média difficile à explorer quels que soient les outils de recherche envisagés de par sa taille et sa volatilité. De plus, Internet impose de sévères restrictions quant à son exploitation puisque la plus grande partie des informations qui y circulent appartiennent au Web Invisible.

2.1.2 Rechercher et trouver

La finalité d'une recherche d'information sur Internet est de pouvoir retrouver des documents intéressants et pertinents pour l'utilisateur final. Ceux-ci doivent le plus souvent répondre à l'expression d'un besoin que l'on matérialise sous la forme d'une question. Pour y répondre, on utilise des outils spécialisés. Il en existe une grande variété. L'objet de ce mémoire n'étant pas de tous les énumérer, nous n'évoquerons que trois types d'outils, ceux qui sont les plus fréquemment utilisés [WW01], afin de mieux appréhender leurs limites et de comprendre d'une manière générale la problématique de l'accès aux ressources qui sont publiées sur Internet.

Nous illustrons donc notre travail par l'étude succincte des outils suivants :

- les annuaires ;
- les moteurs de recherche ;
- les méta-moteurs de recherche.

2.1.2.1 Les annuaires

Première catégorie abordée : les annuaires. À l'instar des annuaires de services *papiers*, les annuaires virtuels sont des outils où les informations sont rangées et cataloguées de manière hiérarchique : chaque document trouve sa place dans une arborescence de thèmes.

La création de ces annuaires est le plus souvent confiée à des documentalistes spécialisés qui archivent manuellement une parties des informations présentes sur le Web, mais une partie seulement. Parmi les annuaires les plus connus, on peut citer entre autres Librarian's Index, Infomine, Academic Info, Yahoo ! ou encore About.com dont la richesse est illustrée par le tableau 2.1 qui restitue, pour chacun d'eux, le nombre de sites indexés.

Annuaires	Nombre de sites indexés	Adresse (URL)
Librarian's Index	> 7.000	`http://www.lii.org`
CISMeF	> 11.000	`http://http://www.chu-rouen.fr/cismef`
Infomine	> 20.000	`http://infomine.ucr.edu`
Academic Info	> 20.000	`http://www.academicinfo.net`
Yahoo !	≈ 1.000.000	`http://www.yahoo.com`
About.com	≈ 1.000.000	`http://www.about.com`

TAB. 2.1 – *Taille des principaux annuaires mondiaux [Bak00]*

L'indexation manuelle des ressources, telle qu'elle est réalisée pour créer ou enrichir un annuaire existant, apporte une grande valeur ajoutée aux informations recherchées, les informations obtenues sont souvent plus précises que celles qui pourraient être retrouvées via des méthodes automatiques. Ce procédé d'indexation manuelle de l'information se heurte toutefois à une grande restriction, la couverture des documents reste faible comparée aux dimensions titanesques imposées par le Web.

2.1.2.2 Les moteurs de recherche

Deuxième acteur incontournable de la recherche d'information sur Internet : les moteurs de recherche. Contrairement aux annuaires, la traque et l'étude des informations présentes sur le Web se fait sans intervention humaine. Les informations sont collectées et archivées automatiquement ; de ce fait, les bases de connaissances générées, aussi appelées index, couvrent pour les moteurs les plus conséquents plusieurs milliards de documents.

Les moteurs de recherche sont des outils de recherche d'information privilégiés puisque 85% des recherches faites sur Internet débutent par leur utilisation [Abo]. Ce plébiscite tient sans doute au fait qu'ils restent simples d'emploi, de prise en main rapide et qu'ils renvoient dans presque tous les cas de figure une réponse, ce qui rassure l'internaute.

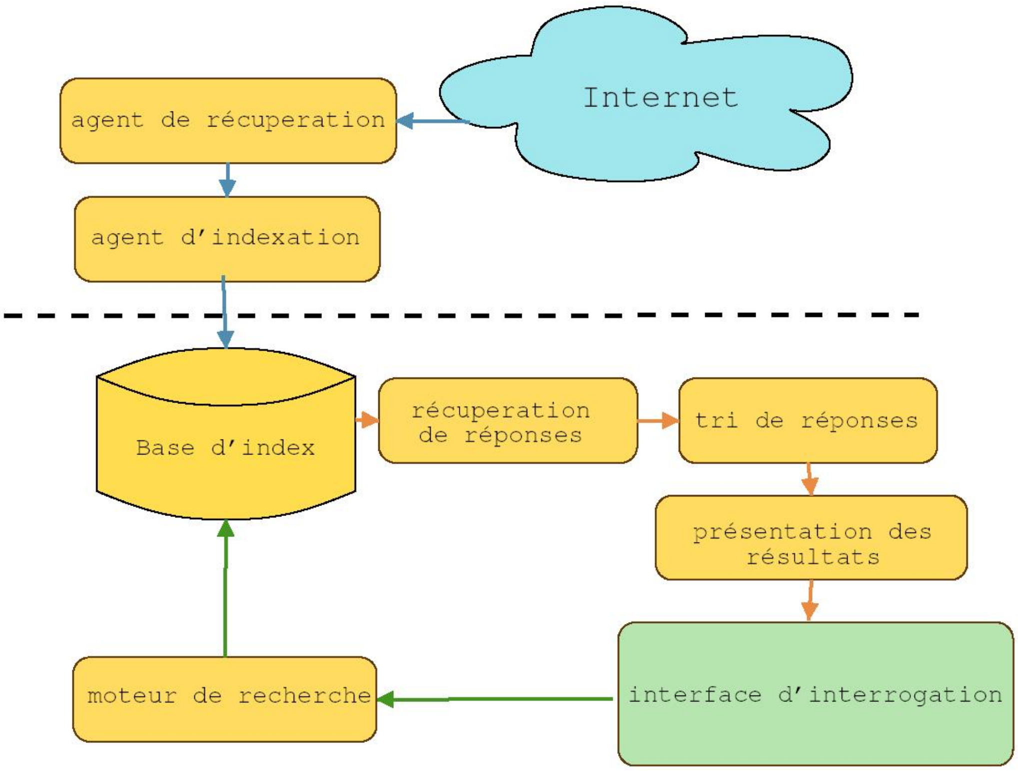


FIG. 2.3 – *Fonctionnement interne d'un moteur de recherche*

Leur grande simplicité d'emploi se retrouve aussi dans leur fonctionnement interne ;

celui-ci peut être divisé en deux étapes distinctes que l'on retrouve schématisées sur la figure 2.3 :

- une étape de récupération et d'indexation des documents collectés sur Internet ;
- une étape d'interrogation, phase durant laquelle l'utilisateur interroge le moteur pour obtenir les informations voulues.

Malheureusement cette extrême simplicité et la gestion d'une base de documents dépassant parfois le milliard de documents a son revers. Ainsi, la gestion de grandes bases documentaires reste problématique et l'information renvoyée aux utilisateurs est d'une qualité souvent décevante car trop approximative (problème du classement des résultats) ou périmée (problème de mises à jour des informations). Néanmoins, malgré ces nombreux manques et lacunes les moteurs de recherche restent plébiscités.

2.1.2.3 Les méta-moteurs

Les moteurs de recherche allient simplicité et richesse d'information. Malgré tout leur couverture, même si elle est très nettement supérieure à celle des annuaires, reste parcellaire et ne couvre pas l'intégralité du " Web visible ". Pour tenter de combler ce manque, l'idée est d'interroger simultanément plusieurs bases documentaires afin de croiser les résultats et ainsi d'augmenter leur couverture. Les méta-moteurs sont des outils de ce type permettant d'interroger simultanément plusieurs sources d'information.

L'interrogation simultanée de plusieurs sources répond également à d'autres exigences. La mise à jour des documents sur différents moteurs de recherche étant plus ou moins fréquente, une interrogation distribuée vers plusieurs cibles semble être un moyen pour prendre en compte les nouveaux documents dès l'instant où ils sont mis à jour sur l'une des sources interrogées. En créant une plateforme d'interrogation unifiée, les méta-moteurs répondent également à l'attente des utilisateurs qui espèrent ainsi maximiser leur chance d'obtenir une réponse intéressante. Dernier point, le croisement des résultats opéré par les méta-moteurs améliore, comme nous venons de le décrire, la couverture des résultats et l'accès à une information plus récente, mais il permet également de retrouver plus facilement des documents discriminants en comparant les réponses des différents moteurs interrogés et en ne gardant que les documents communs [WW01].

Néanmoins, les méta-moteurs ne sont pas des outils infaillibles et ils se heurtent aux mêmes limitations que les moteurs qu'ils exploitent. Ils améliorent la couverture des résultats mais sont cependant loin d'une couverture totale. De plus, leur mise à jour reste dépendante du temps de réactivité des moteurs pris en compte. Le constat est donc décevant puisque l'accès à l'information est en deçà des espérances et ce quel que soit l'outil de recherche évoqué.

2.2 Savoir interroger

Même si les outils de recherche d'information, à l'instar de ceux passés en revue, ont leurs caractéristiques propres, néanmoins ils possèdent tous un dénominateur commun : la phase d'interrogation.

Malgré tous les progrès accomplis en matière de recherche d'information, il reste une étape préliminaire et personnelle indispensable à toute recherche : l'utilisateur doit définir et exprimer ses besoins informationnels s'il veut obtenir une information satisfaisante en retour.

Que l'on prenne, à titre d'exemple, le cas évident des moteurs de recherche ou des méta-moteurs pour lesquels toute recherche passe par la rédaction d'une question, ou que l'on étudie le cas des annuaires où là encore l'expression d'un besoin est une étape indispensable au parcours des éléments de la classification : toute recherche d'information est consécutive à l'expression d'une question, encore faut-il savoir la formuler.

Dans la section 2.1, nous avions décrit des outils classiques de recherche d'information, parmi ceux énumérés, il en est un qui prédomine : le moteur de recherche. Couramment utilisé, il sert de brique élémentaire à toute recherche sur Internet. La formulation d'une question passe par la définition d'une requête. Une requête est le plus souvent assimilée à une suite de termes séparés par des opérateurs booléens permettant d'exprimer une interrogation de manière synthétique. Bien entendu, plusieurs types de requêtes sont envisageables dont certains seront abordés dans la section 4.4.2.1, mais l'idée sous-tendue est la même : pouvoir exprimer et transcrire un besoin au travers d'un panel restreint de termes.

Nous allons, dans cette section 2.2, analyser le problème de la formulation des questions et de la rédaction d'une requête.

2.2.1 Une requête inaccessible

Les moteurs de recherche sont des outils polymorphes aux multiples facettes : ils peuvent se présenter sous la forme de moteurs spécialisés utilisant un vocabulaire contrôlé et monolingue travaillant sur une base documentaire réduite, ou encore prendre les traits d'outils généralistes et multilingues indexant tout sur leur passage pour traiter de grandes quantités d'informations (exemple : Google [Corb]) .

Malgré le très grand nombre de moteurs existant et leur diversité, les six plus grands moteurs de recherche mondiaux (voir le tableau 2.2) totalisent plus de 75% des recherches utilisant ce type d'outils.

Moteurs de recherche	Taille du Web indexé	Fréquentation	URL
Google	> 3.000.000.000	55,1%	`http://www.google.com`
MSN Search	> 500.000.000	9,4%	`http://search.msn.com`
AOL search	> 300.000.000	3,5%	`http://search.aol.com`
Terra Lycos	> 50.000.000	3,0%	`http://www.lycos.com`
Altavista	> 350.000.000	2,4%	`http://www.altavista.com`
Ixquick	> 1.400.000.000	1,7%	`http://ixquick.com`

TAB. 2.2 – *Principaux moteurs de recherche mondiaux. Taille exprimée en nombre de pages indexées et fréquentation représentant le pourcentage de requêtes traitées par chaque moteur par rapport au nombre total de requêtes posées sur le Web. Enquête de Septembre 2002 [dN].*

Cet engouement quasi unanime pour un nombre très restreint d'outils de recherche n'est pourtant pas dû uniquement à la qualité des résultats car, globalement, les utilisateurs restent insatisfaits des produits proposés et jugent la qualité des réponses insuffisantes même si une enquête récente tend à montrer que celle-ci s'améliore [dN].

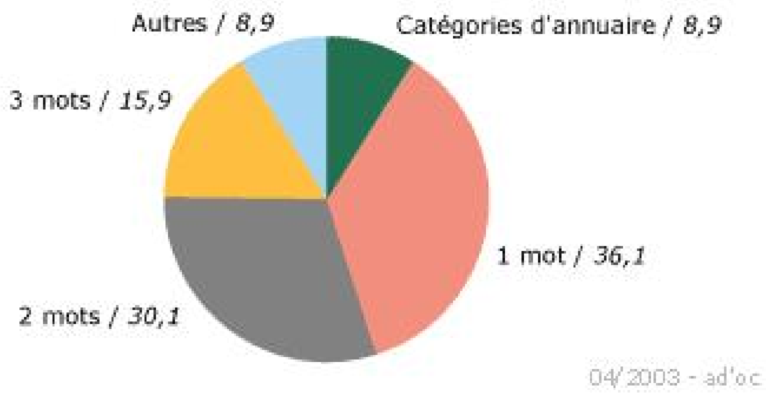


FIG. 2.4 – *Typologie des requêtes des internautes [Ad']*

2.2.1.1 Savoir formuler ses requêtes

Tous les documentalistes vous le diront : n'accède pas à l'information qui veut. Sous cette formulation légère se cache une triste réalité, si les outils de recherche d'informations ne donnent pas des résultats d'une grande précision, la faute ne leur en incombe pas systématiquement.

En effet, l'utilisateur est souvent le complice involontaire de ces mauvaises performances car il ne prend pas le temps d'apprendre à interroger un moteur et encore moins d'exprimer correctement ses besoins au travers d'une requête précise et non ambiguë.

Si l'on s'arrête sur la figure 2.4 qui montre la typologie des requêtes faites par les internautes [Ad'] on se rend compte que 66% des requêtes comportent deux mots ou moins même si l'évolution est positive et que les internautes apprennent progressivement à utiliser les outils mis à leur disposition.

L'enquête réalisée sur l'utilisation du moteur de recherche Altavista [SHMM98] illustre cette évolution. Ainsi, pour résumer, dans 70% des cas, l'utilisateur interroge ce moteur une unique fois, pose une requête composée d'un seul mot, se contente de regarder la première page de résultats pour finalement ne lire que le premier d'entre eux. Cette manière de procéder qui semble caricaturale est pourtant symptomatique de ce qui se passe et permet de comprendre plus aisément que bon nombre de résultats ne sont pas à la hauteur des espérances.

2.2.1.2 Des résultats pertinents ?

Le deuxième problème lié aux outils de recherche d'informations est celui du classement des résultats qui est qualifié le plus souvent de " pertinent ". Avec ce qualificatif, on court le risque de faire croire à l'utilisateur que seuls les premiers résultats sont intéressants alors que nous n'avons aucune certitude.

Parmi les différents modèles ou algorithmes de classement de résultats, on peut en dénombrer cinq principaux, qui sont :

- le modèle booléen : classement binaire des résultats en fonction de l'apparition ou non des mots de la requête ;
- le modèle statistique : classe les réponses en fonction du nombre d'apparition des mots de la requête et de leur rareté ([SJ72, SJ73, Flu77]) ;
- le modèle vectoriel : classe les résultats grâce à une mesure de similarité calculée entre les termes de la requête et ceux des documents réponses ([SM83]) ;
- le modèle probabiliste : classe les résultats en fonction de la probabilité qu'un document réponse soit pertinent ou non ([LG98a, LG98b, LG99]) ;
- le modèle topologique : classe les résultats en fonction du nombre de liens hypertextes entrants et sortant ([Kle99, BP98]).

Pour tous ces modèles à l'exception du modèle topologique, la pertinence des documents est calculée en réutilisant les éléments de la requête. Cette constatation renforce l'idée que les recherches sont compromises dès l'instant où les requêtes restent mal formulées. Nous ne remettons pas ici en cause l'intérêt des travaux effectués dans le cadre du classement des résultats, car entre des mains expertes ceux-ci donnent des résultats plus probants. Nous constatons simplement que tout le travail de recherche d'information repose exclusivement sur la requête initiale qui est le plus souvent d'une qualité moyenne.

2.2.2 Guider les utilisateurs

La section 2.2.1, en mettant l'accent sur les difficultés rencontrées par les utilisateurs pour rédiger correctement leurs requêtes et interroger les outils de manière adéquate, nous pousse à nous interroger sur la nécessité d'outils complémentaires à l'utilisation des requêtes. La requête apparaît comme le point de départ à toute recherche pour que puisse s'initier le processus de recherche d'information. Néanmoins, la question qui nous brûle les lèvres est de savoir s'il est possible d'aider l'utilisateurs à formuler correctement ses choix et ainsi à lever les ambiguïtés potentielles.

De nombreux travaux s'orientent désormais dans cette voie qui permet d'aider et de guider les utilisateurs dans leur quête d'information pour ainsi dépasser le simple cadre de la requête qui reste une source d'incompréhension pour les internautes non aguerris. Pour faciliter la recherche d'information et la navigation, on peut donner l'exemple de trois voies d'études qui tentent de pallier les insuffisances des techniques standard de recherche par mots clefs ; il s'agit de :

- l'expansion de requête ;
- la classification interactive ;
- la corrélation de documents.

2.2.2.1 L'expansion de requêtes

L'expansion de requête est une technique utilisée par certains moteurs de recherche pour préciser la question et désambiguïser les mots clefs employés dans la requête. Cette expansion de requête peut être réalisée de deux manières.

Première approche, l'approche directe, technique que l'on peut retrouver sur de nouveaux types de moteur de recherche [Exa, Kar], permet à l'utilisateur de formuler une requête, de consulter les résultats qui lui sont proposés et d'affiner progressivement sa requête en l'enrichissant avec des termes clefs qui sont proposés dynamiquement. Cette approche est qualifiée d'approche directe car à chaque étape le choix proposé est soumis à l'approbation de l'utilisateur.

Une deuxième approche, qualifiée cette fois-ci d'approche indirecte, ne laisse pas le choix à l'utilisateur. Cette voie, où l'outil est au cœur du processus de prise de décision, peut être réalisée de multiples manières. On peut utiliser des thésaurus ou des ontologies pour désambiguïser la requête en fonction des termes initialement présents ou bien employer des techniques plus originales. Par exemple, dans le projet Watson [BHB01] un agent intelligent " espionne " les internautes volontaires pour définir leurs profils en fonction des documents lus ou écrits et des sites consultés. Les informations collectées sont ensuite traitées pour déterminer le profil de l'internaute. Ce profil est alors joint à toute requête faite sur leur moteur de recherche pour ainsi étendre la question posée et l'orienter en fonction des centres d'intérêt préalablement décelés.

2.2.2.2 La classification interactive

La classification peut également devenir interactive à l'image de la méthode de Scatter/Gather dont l'objectif est de simplifier la recherche d'information. La finalité des outils de classification interactive est de pouvoir définir dynamiquement une requête en fonction des thématiques successivement choisies par l'utilisateur. Ce type de classification produit une arborescence qui évolue à mesure que l'internaute comprend la nature des documents disponibles et découvre ceux qui sont les plus intéressants pour lui [CKPT92, HP96].

Plusieurs expérimentations ont été menées autour de ce type d'approche dite de Scatter/Gather pour analyser les questions posées en entrée et améliorer l'interactivité [HKP95]. La conclusion qui s'impose est qu'une telle approche interactive aide l'internaute dans ses recherches et donne des résultats plus performants qu'avec des techniques standard de recherche par mots clefs.

2.2.2.3 La corrélation de documents

La dernière approche que nous abordons succinctement dans ce tour d'horizon des outils qui améliorent la recherche d'information est celle de la corrélation entre documents. La corrélation est une technique singulière qui permet à un utilisateur possédant un document électronique d'obtenir des documents similaires dans lesquels on aborde les mêmes thèmes pour compléter ou approfondir ses connaissances du sujet.

Cette approche se détache des précédentes pour deux raisons : tout d'abord cet outil ne permet pas, à l'évidence, d'effectuer directement une recherche, il s'agit d'un outil complémentaire qui permet, une fois un document " intéressant " identifié, de compléter ses connaissances et d'obtenir d'autres documents sémantiquement proches.

La corrélation est de ce fait une voie de recherche un peu marginale car toujours adossée à un autre outil de recherche ; le plus souvent elle est directement intégrée aux moteurs de recherche eux-mêmes. Ainsi, tous les moteurs listés dans le tableau 2.2 proposent un outil de corrélation qu'ils appellent *Pages Relatives*, *Liens similaires* ou encore *Documents corrélés* mais sous toutes ces dénominations se cache le même concept, celui de la corrélation entre documents. Nous ne détaillerons pas d'avantage cette technique puisque ce travail fera l'objet des chapitres suivants.

2.3 La prise de décision au cœur de la recherche d'information

Les voies actuelles de la recherche d'information tentent de plus en plus d'aider les utilisateurs à mieux formuler leurs besoins et simplifient progressivement les étapes de la recherche en participant à la prise de décisions. L'objectif est d'apporter à l'utilisateur le résultat souhaité le plus rapidement possible : en réduisant le nombre de résultats potentiels et en minimisant le nombre d'actions à effectuer (liens à suivre, choix à faire, etc.). Dans cette mouvance, on distingue deux courants : ceux qui suggèrent et ceux qui décident.

Les travaux proches de la classification interactive sont à classer dans cette première catégorie puisque l'utilisateur semble être au cœur de la prise de décision. Au contraire la corrélation de documents ressemble de loin à une boîte noire qui ne laisse pas grand choix à l'utilisateur. Malgré tout, dans ces deux cas de figure, la prise de décision suggestive ou imposée est au cœur même de l'accès à l'information.

2.3.1 Prise de décision et accès à l'information ?

La problématique qui est ici soulevée par ces nouvelles approches en matière d'accès automatisé à l'information est celle de la prise automatique de décision. Dans les deux précédents exemples, la prise de décision est au cœur même de la recherche d'information car même dans le cas de classification interactive, les thématiques présentées doivent être identifiées préalablement et donc choisies automatiquement même si l'utilisateur finira par guider le système documentaire. La prise de décision apparaît être un élément essentiel dans le cadre de la recherche d'information. Mais avant de nous étendre sur les diverses significations qui lui sont rattachées, il nous faut au préalable la situer dans le contexte de notre étude.

“ Décider ” est, ici, éloigné de son sens commun. Pour nous décider veut avant tout dire “ identifier ” : identifier les thèmes, les sujets de discussion d'un texte, d'un document. Il s'agit de repérer les éléments informatifs qui aident à la compréhension. Décider reflète un processus d'extraction d'information et d'interprétation des documents plutôt qu'une réelle prise de décision. La prise de décision effective intervenant pour choisir les thèmes, termes, mots clefs qui seront retenus du document initial et exploités par la suite dans les diverses applications citées précédemment.

2.3.2 Le processus de prise de décision

Si décider, c'est avant tout extraire de l'information, la question sous-jacente est donc comment choisir avec discernement les éléments clefs qu'il ne faut pas omettre lors de l'analyse d'un texte ? L'extraction de " l'information pertinente", c'est avant tout le problème de la compréhension du document.

Sans compréhension, pas moyen d'éviter les erreurs et contresens comme ce fut le cas en 1952 lors de la première conférence sur la traduction automatique organisée au MIT par Bar-Hillel. Les principaux travaux concernaient la recherche dans des dictionnaires et la traduction était alors vue comme une substitution de mots suivie d'un éventuel réordonnancement grammatical. Cette technique a produit ainsi l'exemple célèbre *The spirit is willing but the flesh is weak* (l'esprit est fort mais la chair est faible), qui, traduit en russe puis retraduit en anglais, donna *The vodka is strong but the meat is rotten...*(la vodka est forte mais la viande est pourrie) !

Comprendre un texte, comprendre son contenu est un exercice qui peut être plus ou moins exigeant en fonction des attentes. La compréhension d'un texte peut se faire suivant plusieurs niveaux de difficulté ; dans l'exemple donné nous n'en aborderons que deux pour illustrer simplement nos attentes en termes de compréhension :

– comprendre pour extraire de l'information ;
– comprendre pour répondre à une question.

2.3.2.1 Comprendre pour extraire de l'information

Dans ce premier cas, l'intérêt de " comprendre un texte" est très relatif puisque l'on cherche " simplement" à identifier et extraire les informations intéressantes sous différentes formes, formes qui dépendent de l'utilisation finale.

Ainsi comprendre et extraire pour Church & Gale [GC91] correspond à l'extraction de termes clefs et à leur désambiguïsation grâce à l'utilisation de modèles statistiques. Ces modèles, même s'ils n'apportent qu'une aide limitée dans le cadre de l'extraction d'information, peuvent encore être améliorés.

D'autres méthodes commencent par identifier les éléments clefs du texte comme des noms propres, des dates, des localités géographiques, les utilisent ensuite en les combinant avec des contraintes linguistiques et une base de connaissance du domaine étudié, pour identifier les informations spécifiques du texte original. De telles approches illustrées par Ciravegna et al. [CCC92] ou Mellish et al. [Mel95] permettent par exemple sur une base de textes d'économie industrielle de retrouver les entreprises qui participent à une même *joint venture*.

2.3.2.2 Comprendre pour répondre à une question

Comprendre, cela sous-entend également avoir la capacité de formaliser l'existant, de le " mettre en équation " pour pouvoir lui appliquer des règles de logique et en déduire des connaissances nouvelles à partir de faits et d'hypothèses de départ.

Pour illustrer cette explication, prenons comme exemple la description morphologique et anatomique des primates telle qu'elle est formulée dans l'ouvrage *Primate Adaptation and Evolution - John. Fleagle - Ed. Academic Press.* Cette description liste un certain nombre de caractères physiques propres aux primates qui sont :

- deux mains partiellement préhensiles ;
- deux pieds préhensiles (sauf homme) ;
- membres de structure primitive à 5 doigts ;
- pouce opposable ;
- ongles plats (au lieu de griffes) ;
- coussinets tactiles ridés au bout des doigts (dermatoglyphes) ;
- calcaneum allongé dans sa partie distale ;
- réduction du museau et de l'olfaction ;
- plancher du bulbe tympanique formé par l'os pétreux ;
- vision stéréoscopique (40 à 50% des fibres sont ipsilatérales ; grand recouvrement des champs visuels des deux yeux et traitement séparé par moitié au niveau cortical) ;
- face réduite ;
- orbites de grande taille ;
- orbites convergentes (vers l'avant) ;
- présence d'une barre post-orbitaire (jonction des os frontal et zygomatique).

À partir d'une telle description, l'intérêt n'est pas tant de pouvoir redistribuer l'information, mais plutôt de pouvoir la traiter et l'interpréter pour répondre à des questions simples telles que " *Combien de doigts possède un primate ?* ". La réponse nous semble évidente, les primates étant constitués de deux pieds, de deux mains avec chaque membre lui-même constitué de cinq doigts, au final on dénombre vingt doigts sans que ce renseignement soit explicitement énoncé dans le document.

Cet exemple simpliste d'arithmétique n'est pas sans rappeler des travaux propres à la sémantique. En effet, cette discipline tente de mettre en équation les connaissances pour d'une part vérifier si les hypothèses de départ sont valides et ensuite pour en déduire de nouvelles informations qui n'ont pas été explicitées dans le texte et ainsi répondre aux éventuelles questions.

Les applications pour ce genre de systèmes sont nombreuses, on peut citer les travaux de McKeown et Kukich [MKS94] qui permettent de s'approprier les connaissances contenues dans un texte de manière à les exprimer différemment. Reprenons un exemple de reformulation donné par McKeown [MKS94] qui prend comme point de départ une base de donnée textuelle contenant comme information les horaires du vol quotidien entre Paris et Moscou. On apprend dans cette base que ce vol quotidien est assuré du 1er au 14 Novembre et du 16 au 30 Novembre. Il est alors plus utile de répondre à la question " *Au mois de novembre, à quelles dates peut-on prendre un vol pour Moscou en partance de Paris ?* " par " *Tous les jours excepté le 15.* " plutôt que de lister toutes les dates effectivement envisageables. Ce travail nécessite, comme dans l'exemple précédent, une compréhension préalable du texte pour se rendre compte que ces deux représentations sont équivalentes et répondre à la question de la manière la plus simple.

2.4 Notre orientation

Avant même d'avoir présenté un état de l'art de la corrélation entre documents (voir le chapitre 3), ce chapitre a brossé un portrait rapide et non exhaustif de la recherche d'information sur Internet et traite plus généralement du cadre de la recherche d'information sur des documents au format électronique. Malgré son rôle introductif, il nous informe d'ores et déjà des limitations propres à la discipline auxquelles devra se confronter par la suite la méthode de corrélation envisagée.

La première limitation que nous avons découverte quant aux méthodes de corrélations, est qu'elles ne peuvent être considérées que comme des outils complémentaires. Incapables de fournir à l'utilisateur le document qu'il recherche, ces méthodes ne sont qu'un moyen parmi d'autres pour étendre les connaissances d'un internaute sur un document déjà retrouvé. Ces méthodes ne prétendent donc pas améliorer la recherche d'information dans un cadre général mais elles se positionnent comme des outils permettant d'obtenir un complément d'information.

Nous avons également constaté dans les sections 2.2 et 2.3 que les internautes éprouvent, dans une grande majorité, quelques difficultés quant à l'utilisation d'outils traditionnels de recherche d'information tels que les moteurs de recherche ; la première difficulté étant de formuler correctement les questions posées. Parmi les solutions précédemment évoquées permettant d'améliorer la qualité des recherches, le dénominateur commun est d'arriver à comprendre les besoins de l'utilisateur pour les formuler à sa place. L'illustration la plus appropriée était celle de la reformulation de requêtes dont le travail de prédilection se borne à cette unique tâche.

Dans le cadre de la corrélation, comprendre les besoins de l'utilisateur signifie comprendre le document d'origine pour retrouver des documents similaires. Nous avons envisagé, dans la section 2.3, deux niveau de compréhension, mais celui qui nous intéresse tout particulièrement est le premier niveau, à savoir : comprendre pour extraire l'information. En effet, nous ne souhaitons pas mettre en place une plateforme d'interrogation en langage naturel, mais seulement retrouver des documents sémantiquement corrélés. Nos travaux se limiteront donc par la suite à une compréhension primaire des documents, c'est à dire à en extraire les informations discriminantes.

Chapitre 3

Les différents visages de la corrélation

Dans le domaine de la corrélation entre documents, la définition même de corrélation s'avère être multiple et à géométrie variable. L'objet de ce chapitre 3 est de réaliser un état de l'art des différentes corrélations pour pouvoir décider ultérieurement de la voie à suivre.

Les méthodes de corrélation contemporaines peuvent se scinder suivant quatres grandes thématiques : trois thématiques traditionnelles de la recherche d'information et des méthodes de corrélation et une quatrième qui est à un stade expérimental.

Nous décrirons tout d'abord dans la section 3.1 les méthodes dites topologiques qui examinent la structure interne des documents pour tenter de comprendre les interactions qui les lient les uns aux autres. Dans la section 3.2, nous traitons ensuite du cas des méthodes dites empiriques qui étudient le comportement des internautes pour trouver des relations entre ces documents. Enfin au travers de la section 3.3, nous abordons le cas de la dernière approche classique de corrélation, l'approche hiérarchique. Dans cette partie, nous confrontons des travaux de corrélation de documents avec des travaux inspirés du domaine de la classification pour nous rendre compte que ces deux notions sont au final étroitement liées. La dernière partie, la section 3.4, expose une voie de corrélation expérimentale dite linguistique qui met en œuvre des notions d'extraction terminologique dans le but d'extraire la signature sémantique de chaque document.

3.1 La corrélation topologique

Comme évoqué dans le chapitre 2, les outils de recherche d'information éprouvent des difficultés à analyser et à maintenir à jour les données relatives à une base documentaire aussi grande que le Web. De même, le manque de descriptions du contenu des documents analysés limite les investigations à une simple recherche par mots clefs, mots qui, mal choisis, ne sont pas toujours discriminants.

FIG. 3.1 – *Recherches sur* `www.voila.fr`

Un exemple où l'utilisation de mots clefs est insuffisante est celui qui est illustré par la figure 3.1. Dans cet exemple, l'objectif est de trouver une liste de moteurs de recherche afin de s'en servir ultérieurement. Malheureusement les limites des recherches textuelles sont atteintes car aussi paradoxal que cela puisse paraître, les principaux outils de recherche d'information sur Internet ne contiennent pas l'expression " moteurs de recherche", il n'est donc pas possible de les retrouver à partir de cette requête. La preuve en est que les résultats ainsi obtenus rassemblent des sites dont la thématique traite des moteurs de recherche, de leur fonctionnement, de leurs différences, etc, sans toutefois répondre à notre attente initiale.

Inversement, il existe des techniques de désinformation comme le " spamming " dont la finalité est de mieux positionner une page ou un site Web dans le classement des résultats opérés par un ou plusieurs moteurs de recherche. Pour ce faire, on ajoute des termes savamment choisis que l'on insère dans les méta-descripteurs des documents électroniques afin de perturber volontairement les réponses renvoyées. Cette illustration met en avant la fragilité des recherches d'information menées sur le contenu car aisément manipulables et pas toujours pertinentes.

Une solution envisagée pour améliorer l'accès aux documents électroniques et contourner les difficultés est de ne plus utiliser les termes pour faire les recherches mais les liens hypertextes. Cette section 3.1 décrit des techniques de recherche basées sur l'utilisation des liens hypertextes ainsi que leurs retombées pour le domaine de la corrélation.

3.1.1 La notion de liens hypertextes

Les documents publiés sur Internet sont majoritairement au format HTML (HyperText Markup Language). Ce format comme son nom l'indique utilise la notion de lien hypertexte qui est la clef de voûte des méthodes de corrélation topologiques. Internet, que l'on nomme communément la toile, est composé d'une multitude de liens unidirectionnels : les liens hypertextes. Les méthodes topologiques étudient ces relations pour mettre en évidence les documents qui se co-référencent le plus souvent et ainsi en déduire des notions de proximité ou de popularité.

De nombreux travaux ont reconnu qu'il pouvait être enrichissant d'utiliser la structure des liens hypertextes pour localiser l'information et améliorer les recherches [BRS92, RAS94, Fri88]. Les techniques qui utilisent cette structure hypertexte pour retrouver des documents corrélés justifient leur emploi en décrivant le lien hypertexte comme un marqueur sémantique [Kle99]. En effet, la mise en relation de deux pages par un lien hypertexte nécessite que l'auteur lui-même insère ce pointeur. Son intervention explique alors la valeur sémantique plus importante que l'on accorde aux liens hypertextes.

Les travaux utilisant la notion de liens hypertextes modélisent le Web (ou un sous-ensemble) par un graphe $G = (N, A)$ où N est l'ensemble des nœuds et A l'ensemble de ses arêtes. Dans ce graphe, un noeud symbolise une page Web et une arête représentent un lien hypertexte entre deux noeuds. Les applications évoquées par la suite utilisent cette représentation du Web.

3.1.2 PageRank

PageRank est un algorithme de classement global des résultats (traduit de l'anglais *global ranking scheme*) mis au point par Brin & Page [PBMW98]. Il permet de trier les résultats d'un moteur de recherche grâce à un score de pertinence fondé sur la notion de popularité des pages Web. Ainsi la page d'accueil du portail Yahoo ! semble intuitivement plus importante que la page d'accueil d'un site personnel quelconque car elle est plus référencée par d'autre sites ou portails qu'une page quelconque. Le rang d'une page A peut être ainsi défini par le nombre de pages Web qui pointent vers elle. Malheureusement ce type de classement par citations ou popularité (de l'anglais *citation ranking*) donne de mauvais résultats à cause notamment des techniques de spamming évoquées précédemment.

PageRank reprend cette notion de popularité ainsi que l'idée qu'une page acquiert de l'importance avec une citation accrue, et l'étend en prenant en considération l'action opposée. Toujours avec le même exemple, si Yahoo ! définit un lien qui pointe vers votre site, celui-ci en retira une importance accrue par rapport à un lien émanent d'un site quelconque et méconnu. L'algorithme PageRank contraste avec les algorithmes de citations qui ne distinguent pas les deux cas. Il est à noter que l'algorithme PageRank est récursif, que l'importance d'une page est dépendante des autres, mais qu'elle les influence également.

Cet algorithme notamment utilisé par le moteur de recherche Google [BP98] combine des techniques classiques de recherche d'information par mots clefs avec un tri hérité de l'algorithme PageRank.

3.1.3 L'algorithme HITS

L'algorithme HITS (*Hypertext Induced Topic Search*) à l'instar de PageRank permet de trier les résultats avec toutefois quelques singularités.

Tout d'abord, l'algorithme HITS évoqué pour la première fois par Kleinberg [Kle99], permet de moduler le poids ou le score d'un document en fonction de la question posée et de la visibilité du document dans le corpus ou sur Internet. Il se différencie de PageRank car il n'attribue pas de poids global et immuable au fil du temps. De plus cet algorithme ne se limite pas à classer les documents suivant une pertinence qui lui est propre. Il s'agit également d'un algorithme de recherche et d'accès à l'information qui a servi de modèle et d'exemple à des travaux alternatifs comme la corrélation topologique. À ce titre, l'algorithme HITS mérite d'être détaillé afin de mieux appréhender le fonctionnement de la corrélation topologique.

Sa philosophie est de considérer qu'une page visible (i.e : souvent référencée) aura initialement un score important, mais en fonction de l'évolution des liens qui l'entourent ce score évoluera et ainsi pourra perdre ou gagner de la valeur. Ainsi, une réponse pertinente sur un forum de discussions que tout le monde va citer aura une grande visibilité et donc un score important mais dès que le phénomène de mode aura disparu, que l'information aura été véhiculée et que progressivement les liens vont disparaître, son score diminuera.

Pour imaginer un tel fonctionnement de son algorithme, Kleinberg distingue deux catégories de pages Web : les pages dites de “ sites ” (de l'anglais *Hubs*), notées par la suite P_H, et les pages dites “ feuilles ” (*Authority*), notées P_A.

Les P_A se définissent comme étant des réponses pertinentes au vu d'une question posée. Cette pertinence est définie généralement en fonction de l'absence ou de la présence des termes clefs de la question initiale dans le document considéré. La définition des P_H est tout autre puisque les P_H sont des pages qui possèdent au moins un lien hypertexte pointant vers une P_A. Dans la suite de ce chapitre 3, nous admettrons que la formulation *la page A pointe vers B* sous-entend que la page A possède un lien hypertexte dont la cible est la page B.

L'idée de base qui prévaut dans les recherches utilisant HITS est de pouvoir initialement identifier un sous-graphe du Web de taille raisonnable (Section 3.1.3.1) et d'y analyser les liens afin de localiser les P_A et P_H (Section 3.1.3.2). Cette sélection préalable d'un sous-graphe du Web (typiquement une centaine de pages) permet de travailler sur un ensemble de documents pour ainsi diminuer le temps de traitement ce qui permet d'effectuer cette tâche dynamiquement au moment de l'interrogation.

3.1.3.1 Déterminer un sous-graphe utile du Web

Comme nous venons de l'évoquer, toute analyse utilisant HITS débute préalablement par la recherche d'un sous-graphe du Web noté S, étape préliminaire destinée à diminuer la charge de travail. Pour générer ce sous-graphe on commence par définir une question notée Q, question qui dans le cadre de l'utilisation d'un moteur de recherche est directement formulée par l'utilisateur. Q sert alors à présélectionner un ensemble de documents noté R_Q qui répondent à cette question. On ne prend alors en compte que t pages choisies aléatoirement parmi R_Q pour former l'ensemble R. Le paramètre t qui sert à alléger les temps de calculs est défini expérimentalement et est dépendant du corpus d'étude (taille, liens, etc.). L'ensemble R est ensuite enrichi en y ajoutant les pages situées à son voisinage. L'algorithme mis en place est alors le suivant :

1. on construit R_Q comme étant l'ensemble des P_A issues de Q ;
2. R est un ensemble de t pages choisies aléatoirement dans R_Q ;
3. $S \longleftarrow R$;
4. pour chaque page $p \in R$;
   (a) on inclut dans S toutes les P_A issues de p ;
   (b) on inclut dans S toutes les P_H qui pointent vers p ;
5. S devient finalement le graphe d'étude.

3.1.3.2 L'analyse des liens

L'analyse des liens est une phase de l'algorithme utilisant la propriété de renforcement mutuel des liens. Cette propriété est utilisée pour identifier les P_A et P_H parmi les éléments de S, elle nécessite quelques notations préalables :

Soit n le nombre d'éléments constitutifs de S, soit i son $i^{ème}$ élément. On note par $P(i)$ l'ensemble des pages qui pointent vers la page i (P pour *parent*) et par $F(i)$ l'ensemble des pages vers lesquelles i pointe (F pour *fils*). L'analyse des liens génère alors pour chaque page i deux scores, un pour chaque aspect de sa personnalité duale : page " site " ou page " feuille " que l'on note H_i pour les P_H et A_i pour les P_A.

Au début de l'analyse, les H_i et les A_i sont initialisés avec des valeurs arbitraires. L'algorithme analysé est itératif et engendre deux types d'opérations notées I et O. Lors des opérations de type I, le A_i de chaque page est mis à jour et reçoit la somme des valeurs de H_i des pages qui pointent vers elle. Inversement durant des opérations de type O, le H_i de chaque page est mis à jour et reçoit la somme des A_i des pages liées. Cela se résume par :

$$\begin{array}{rl} \text{opération de type } I : & A_i = \sum_{j \in P(i)} H_i \\ \text{opération de type } O : & H_i = \sum_{j \in F(i)} A_i \end{array}$$

Les étapes I et O traduisent l'intuition naissante de Kleinberg pour qui une page de type " site " est intéressante à condition qu'elle soit liée à des pages de type " feuilles " elle-mêmes pertinentes et vice versa. Cette intuition se traduit par le renforcement mutuel des pondérations que nous venons d'évoquer. On peut également remarquer qu'une page peut être à la fois " site " et " feuille ". Cet algorithme se contente de calculer ces deux scores pour chaque page et répète itérativement les phases I et O jusqu'à ce que les scores convergent (la converge est assurée dans la section 3.1.3.3). La procédure utilisée est alors la suivante :

1. on calcule les A_i et H_i pour chaque élément de S (i.e : $1 \leq i \leq n$) ;
2. tant que les A_i et H_i ne convergent pas, on réitère les trois étapes suivantes ;
   (a) $\forall i \in [1 \ldots n], A_i = \sum_{j \in P(i)} H_i$ (Opération I) ;
   (b) $\forall i \in [1 \ldots n], H_i = \sum_{j \in F(i)} A_i$ (Opération O) ;
   (c) étape de normalisation : on choisit α et β tels que $\sum_i \alpha . A_i^2 = 1$ et $\sum_i \beta . H_i^2 = 1$.

Un exemple de calcul des H_i et A_i est donné par la figure 3.2. Ainsi le score de " feuille " du noeud 5 sur la figure est obtenu en additionnant le score des noeuds qui pointent vers 5 (i.e., $0,408 + 0,816 + 0,408$) et cette valeur est ensuite normalisée (i.e., $(0,408)^2 + (0,816)^2 + (0,408)^2$).

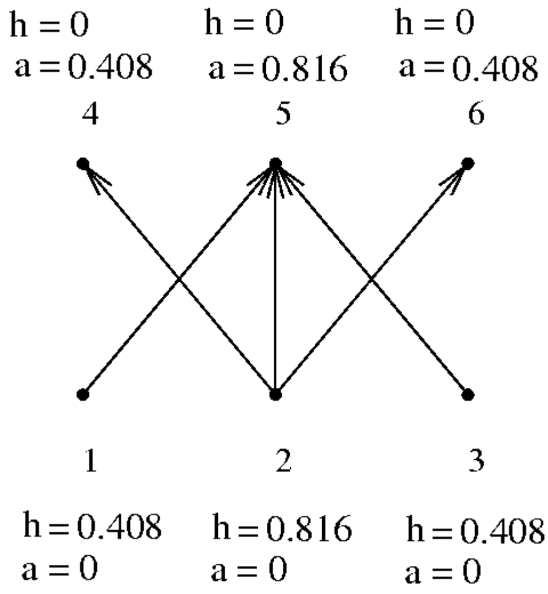


FIG. 3.2 – *Exemple de calcul des H_i et A_i*

3.1.3.3 Le cadre mathématique

Les calculs des valeurs de “ sites ” et de “ feuilles ” peuvent être vus dans un cadre plus mathématique et ont, aux même titres que les calculs menés avec PageRank, des propriétés intéressantes.

Définissons par $M_{m \times n}$ la matrice qui représente le sous-graphe S précédemment étudié. La $(k, l)^{ième}$ entrée de M vaut 1 si la page k pointe vers l et 0 autrement. On définit par $\overrightarrow{a}$ le vecteur des scores de “ feuilles ” $[A_1, A_2, \ldots, A_n]$ et $\overrightarrow{h}$ celui des scores de “ sites ” valant $[H_1, H_2, \ldots, H_n]$. Les opérations définies par I et O s'expriment alors de la manière suivante :

$$\begin{array}{rl} \text{opération de type } I : & \overrightarrow{a} = M.\overrightarrow{h} \\ \text{opération de type } O : & \overrightarrow{h} = M^T.\overrightarrow{a} \end{array}$$

Une simple substitution dans les équations montre que la convergence des deux scores satisfait les formules suivantes (avec c_1 et c_2 deux constantes ajoutées du fait de la normalisation) :

$$\begin{array}{rl} \text{opération de type } I : & \overrightarrow{a} = c_1.M.M^T.\overrightarrow{h} \\ \text{opération de type } O : & \overrightarrow{h} = c_2.M^T.M.\overrightarrow{a} \end{array}$$

Exprimés sous cette forme, les scores de “ sites ” et de “ feuilles ” sont alors obtenus en recherchant les vecteurs propres, respectivement, des matrices $M.M^T$ et $M^T.M$. Le problème de la convergence de deux scores en présence a été abordé par Kleinberg qui démontre que dans le cadre général celle-ci est toujours obtenue [Kle99].

3.1.4 L'algorithme Companion

Contrairement à PageRank, qui n'a été utilisé qu'au travers de pondérations pour les moteurs de recherche, HITS a inspiré d'autres voies de recherches et notamment en matière de corrélation. Les idées de Kleinberg ont été reprises et explorées par Dean & Henzinger qui ont ainsi mis au point deux algorithmes de corrélations : l'algorithme Companion qui dérive directement de HITS et l'algorithme de Co-citation [DH99]. Dans cette section, nous nous intéressons tout d'abord à l'algorithme Companion.

À titre de comparaison, les deux algorithme HITS et Companion ont de nombreuses similitudes, approche, conception, pondération des documents mais diffèrent sur un point : l'initialisation.

HITS recherche d'abord un ensemble de documents potentiellement corrélés noté S, avec comme point de départ de sa recherche la question posée par l'utilisateur symbolisée par les termes de la requête. Ici rien de semblable puisqu'il n'y a pas de requête, juste une page qui sert de point de départ à la recherche des documents corrélés.

Cette étape d'initialisation et de recherche des pages qui vont constituer S, le sous-graphe du Web de départ, est cruciale car elle sélectionne *in fine* tous les éléments potentiellement corrélés. La poursuite de l'algorithme sert essentiellement à classer les documents précédemment choisis via une pondération similaire à HITS pour ne garder que les plus discriminants.

Pour aborder plus en détails le déroulement de l'algorithme Companion qui se scinde en quatre parties que nous allons évoquer ci-après, nous réutiliserons les notations de la section 3.1.3 auxquelles s'ajoutent les suivantes :

On note U l'URL de la page de départ servant à l'élaboration des résultats corrélés. Pour simplifier les notations on ne fait pas de distinction entre l'URL d'une page, la page elle-même, ou encore sa représentation sous forme de nœud dans le graphe orienté qu'est le Web.

Soit p une page quelconque du Web, on admet que $P(p) = [P_p^1, \ldots, P_p^k]$ représente l'ensemble des pages " parents" de p (avec $k = card(P(p))$) et $F(p) = [F_p^1, \ldots, F_p^l]$ l'ensemble des pages " fils" (avec $l = card(F(p))$).

On admet également que l'ensemble $P \circ F(p)$ représente l'ensemble des pages " parents" des " fils" de p privé de p d'où $P \circ F(p) = P(F(p)) = P([F_p^1, \ldots, F_p^l])$. Inversement l'ensemble $F \circ P(p)$ représente les " fils" des " parents" de p privé de p.

3.1.4.1 Détermination de S pour l'algorithme Companion

Cette première étape consiste à déterminer les pages potentiellement utiles lors d'une corrélation prenant comme point de départ U. On commence par rechercher les proches voisins de U au sens généalogique, c'est à dire les pages appartenant à sa fratrie et ses ascendants et descendants immédiats (voir la figure 3.3) et pour cela on procède de la manière suivante :

1. soit U l'URL de départ ;
2. on construit les ensembles $P(U)$ et $F(U)$ privés de U (cas d'un lien récursif) ;
3. on construit les ensembles $P \circ F(U)$ et $F \circ P(U)$ privés de U ;
4. finalement $S \longleftarrow (P(U) \cup F(U) \cup P \circ F(U) \cup F \circ P(U)) - U$.

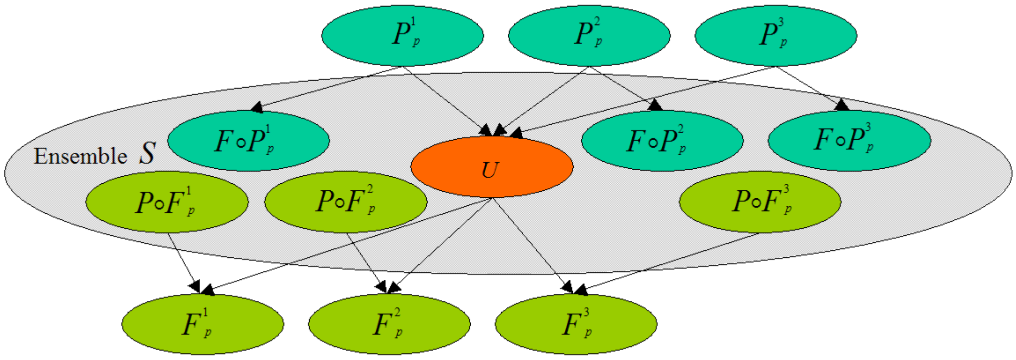


FIG. 3.3 – *Construction de S pour l'algorithme Companion*

3.1.4.2 Simplification de S

Expérimentalement on s'aperçoit que certaines URLs sont plus populaires que d'autres sur le Web. Ainsi un très grand nombre de liens entrants et sortants sont dirigés tout naturellement vers des outils de recherche d'information tels que des annuaires ou des moteurs de recherche. On admet que lors de la recherche des éléments de S, on élimine les pages y faisant référence. Pour ce faire, on construit de manière empirique une liste d'exclusion d'URLs. Cette précaution empêche les résultats de faire systématiquement référence à ces adresses qui n'apportent pas d'informations supplémentaires. Toutefois, si la page de référence est un élément de cette liste, on prend alors en compte toutes les URLs retrouvées sans distinctions.

Après avoir simplifié S, on s'attaque au problème des pages doubles. On considère que deux pages sont des doublons si elles comportent plus de 10 liens hypertextes et que 95% de ces liens sont en commun. Elles sont alors remplacées dans S par un noeud dont les liens sont l'union des liens précédents.

3.1.4.3 Ébauche des scores de “ sites ” et de “ feuilles ”

La finalité de cette partie est de pouvoir choisir les liens entre documents qui sont a priori les plus intéressants et les plus variés possible car le but recherché est de ne présenter à l'utilisateur qu'une seule page HTML par site Web. Le but visé est d'obtenir des résultats qui couvrent le plus grand nombre de sites Web possible pour obtenir un panel de choix conséquent et éviter qu'un même site n'ait trop d'influence sur les résultats corrélés.

Dans ce but, on applique une pondération à chaque élément de S pour tenter d'éliminer les liens internes aux sites Web. La pondération préconisée est celle mise au point par Bharat et Henzinger [BH98] qui permet de déterminer pour chaque liaison entre sites un poids de “ site ” et de “ feuille ”. Contrairement à HITS, ce ne sont plus les pages composant les sites qui obtiennent directement un score de “ site ” ou de “ feuille ” mais les liaisons entre ces mêmes sites.

Pour éliminer les liens internes, une liaison du graphe S reliant deux noeuds d'un même site ne reçoit aucun point. Par contre s'il existe k liens provenant de pages d'un site S_1 pointant vers un unique document d'un site S_2, chacune de ces k liaisons de ce document reçoit $1/k$ points comme score de “ feuille ”. De manière analogue, s'il existe k' liaisons provenant d'un unique document d'un site S_3 pointant vers des documents d'un site S_4, chacune de ces k' liaisons reçoit $1/k'$ points comme score de “ site ”. Pour savoir si deux éléments appartiennent au même site, la solution envisagée est de travailler sur les URLs des pages pour voir si elles partagent un radical commun.

3.1.4.4 Utilisation de HITS

La dernière étape de l'algorithme consiste à appliquer l'algorithme HITS en utilisant comme scores initiaux de “ sites ” et de “ feuilles ” les poids de liaisons précédents. La convergence de l'algorithme démontrée par Kleinberg pour HITS reste valable pour Companion même si expérimentalement Dean & Henzinger montrent que l'on peut arrêter les itérations des opérations I et O après dix passes en moyenne. Les résultats corrélés sont alors choisis comme étant les dix sites obtenant les meilleurs scores avec l'algorithme HITS.

3.1.5 L'algorithme de Co-citation

La deuxième méthode de corrélation mise au point par Dean & Henzinger, l'algorithme de Co-citation, est beaucoup plus simple que la précédente. Initialement cet algorithme, qui s'inspire des co-citations issues de la bibliométrie [Sma73], a été élaboré pour tester la validité des résultats de l'algorithme Companion.

La principale idée est d'identifier les pages parents de U et de repérer parmi les liens qu'elles comportent, ceux qui sont souvent co-cités avec U. En pratique on construit l'ensemble des pages $F \circ P(U)$ qui constitue l'ensemble de départ S (voir la figure 3.4) et pour chaque élément de $F \circ P(U)$ on calcule son score de co-citation avec U. Ce score de co-citation étant égal au nombre de parents distincts qu'ont en commun deux pages.

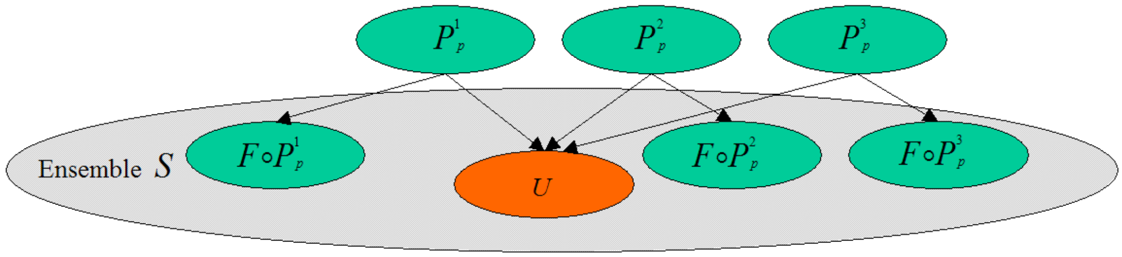


FIG. 3.4 – *Construction de S pour l'algorithme Co-citation*

Une fois ce calcul effectué, on choisit comme résultats corrélés les dix pages qui ont les scores de co-citations les plus élevés. Bien que cette méthode soit simple, elle s'avère assez performante même si elle est en léger retrait par rapport aux résultats obtenus par Companion [DH99]. Sa principale différence et son principal avantage par rapport à l'algorithme Companion est qu'elle s'intéresse directement aux pages Web et non pas aux sites.

3.1.6 Autres techniques

D'autres recherches ont été entamées pour mettre en avant l'utilisation des liens hypertexte [AMM97, BP98, BH98, CK97, Kle99, PPR96, PP97, Spe97, TH98a, TH98b] mais la plupart de ces résultats ne traitent pas des problèmes de corrélations à quelques exceptions près [CDR$^+$98] qui utilisent des techniques proches de l'algorithme Companion pour trouver des “ pages apparentées ” avec, toutefois, une pondération légèrement différente lors du tri des résultats corrélés. D'autres auteurs suggèrent de prendre en considération des notions de co-citations et d'étudier la connectivité entre les pages Web. Ainsi Spertus [Spe97] utilise le critère de co-citation entre les liens hypertextes pour faire de la corrélation. Pitkow & Pirolli [PP97], quant à eux, fondent leurs analyses sur de la clusterisation de pages Web grâce à une analyse de co-citation, et Terveen et Hill [TH98a, TH98b] étudient la connectivité de la structure du Web pour trouver des groupes de pages relatives.

3.1.7 Discussion

Concevoir les liens hypertextes comme des marqueurs sémantiques pouvant être utiles à la corrélation nous a permis de voir en détail quelques algorithmes de corrélation topologique. En théorie, cette manière de procéder assure des résultats pertinents car le lien hypertexte reflète la volonté de l'auteur de lier des documents dans un certain ordre et avec une certaine logique. Malheureusement deux auteurs différents ont peu de chance d'avoir des logiques en tous points similaires ; en pratique leurs pointeurs auront donc des sens différents ou des interprétations différentes. Toutefois, si l'on admet l'existence d'une logique consensuelle ou cartésienne, on peut penser qu'en règle générale deux auteurs emploieront les mêmes liens hypertextes avec des finalités assez proches.

Il existe, néanmoins, des limitations à l'emploi des hyperliens. Leur choix est un critère qui doit être pris en considération. En effet bon nombre de liens sont placés de manière automatique et dans un but publicitaire ; ils n'ont alors plus rien avoir avec un choix de l'auteur et n'apportent que du bruit. Même si la plupart de ces liens peuvent être éliminés grâce à de simples considérations fréquentielles, ce qui est fait succinctement dans l'algorithme Companion, cette problématique n'est quasiment pas abordée par les méthodes topologiques.

Un autre problème soulevé par les méthodes topologiques provient de l'utilisation d'un critère de popularité. En effet, le principe utilisé par ces méthodes pour donner de l'importance à un site ou à une page Web est la notion de popularité. Or il n'y a pas de lien de cause à effet entre la popularité d'un site et sa qualité informationnelle. Le risque est donc d'exclure des liens intéressants car ils ne sont pas assez visibles, donc pas assez populaires. La conséquence directe est une perte d'information qui peut être préjudiciable. Une étude menée par Savoy et Picard [SP01] a mis en avant ce problème. L'expérimentation portait sur l'étude de plusieurs méthodes topologiques de recherche d'informations et elle montrait que de tels algorithmes topologiques tendent à ignorer une grande partie des documents disponibles, et donc contribuent à une perte d'information.

La dernière critique que l'on peut évoquer concerne le choix des documents renvoyés ; les méthodes décrites ci-dessus ont une vision macroscopique du Web qui supposent que l'unité sémantique intéressante est le site Web. Or Internet n'est pas qu'un ensemble de sites mais également un ensemble de pages, et ce sont ces pages qui contiennent, a priori, les informations utiles. Malheureusement, hormis l'algorithme de co-citations, les méthodes de corrélations topologiques ne s'intéressent qu'aux sites. Cela s'explique simplement par la nécessité qu'elles ont d'obtenir un nombre suffisant de données pour relier deux éléments entre eux. Plus l'entité de base est conséquente, plus le nombre de liens et de données la concernant est importante. Sans cette préférence pour les sites Web, des méthodes proches de HITS ne peuvent pas donner de résultats satisfaisants car elles n'obtiennent initialement pas assez de liens reliant les mêmes éléments.

3.2 La corrélation empirique

Après avoir passé en revue quelques méthodes topologiques dédiées à la recherche de résultats corrélés nous allons nous intéresser à une méthode tout aussi originale de corrélation qui est la corrélation empirique. L'idée est d'étudier et de retracer le parcours des internautes sur la toile pour déduire de leur faits et gestes passés les sentiers qu'ils préfèrent emprunter et tenter d'en déduire des liens de proximité entre les différents documents visités.

3.2.1 Une étude empirique

Il existe peu d'acteurs qui s'intéressent aux possibilités de la corrélation empirique ; on peut néanmoins citer deux outils de recherche d'information utilisant cette technique : le site de Netboussole [S.A] et le site d'Alexa [Cora].

Les outils développés par la société Alexa sont de loin les plus évolués. Le concept d'une corrélation empirique a été imaginé par Brewster Kahle, connu pour avoir réalisé en 1990, le premier système de publication sur Internet, le WAIS, ou Wide Area Information Server [Kah91], qui permet, par exemple, à l'Encyclopaedia Britannica ou au New York Times de porter leur contenu en ligne.

En 1996, il lance en parallèle deux sites Web : Alexa[1] et Archive.org[2]. Ces deux sites qui ont des tâches bien spécifiques jouent des rôles complémentaires dans le cadre de la corrélation. Le second site, Archive.org, représente une bibliothèque virtuelle qui se veut l'égale de la mythique bibliothèque d'Alexandrie. Une copie de chaque page Web ayant existé doit en principe y être stockée. Depuis 1996, et à raison de 120 Go indexé par jour, la base documentaire disponible s'élève à plus de 240 To.

Le site d'Alexa propose de son coté un outil en ligne de corrélation que l'on trouve également directement intégré au navigateur Web Netscape [Corc] sous la forme d'une fonction nommée " What's related ".

La technique de corrélation mise en place utilise les archives du Web détenues par Archive.org pour construire une représentation du Web sous la forme d'un graphe orienté. Comme précédemment, les noeuds représente les pages Web et les arcs, les liens hypertextes. Le travail initial consiste à appliquer à ce graphe des techniques de clusterisation pour regrouper les pages en fonction de leurs contenus. Sur ce graphe initial, on fait quotidiennement des opérations de mise à jour pour ajouter ou enlever de nouvelles pages en fonction des modifications constatées sur la toile.

---
[1] `http://www.alexa.com`
[2] `http://www.archive.org`

À partir de cette représentation du Web sous forme de graphe, on affine sans cesse les liens reliant les sites en étudiant le comportement de milliers d'internautes utilisant les produits d'Alexa. Produits qui se trouvent intégrés dans certains navigateurs Web (Netscape 4.06) et qui enregistrent les préférences et le parcours des internautes à chaque fois qu'ils sont utilisés. En suivant leur parcours et en l'analysant, on peut en déduire des relations de proximité entre différents sites Web. Par exemple, si beaucoup d'utilisateurs vont visiter le site S_1 puis directement le site S_2, on peut supposer que les deux sites sont en relation, cette relation est même orientée puisque dans le cas présent S_2 est en relation ou pointé par S_1.

Cette méthode de corrélation est basée sur une double étude : clusterisation et étude empirique. Le regroupement initial via des techniques classiques de clusterisation donne un certain cadre d'étude qui est affiné ensuite par les liens empiriques, découverts grâce à l'étude du comportement des internautes. Ce travail préliminaire est nécessaire car il évite de faire reposer toute la méthode sur une étude empirique qui pourrait donner des résultats surprenants. Cette étape assure une plus grande sécurité pour la validité de résultats corrélés. Au final, les résultats corrélés sont les pages d'un même cluster, classées en fonction des liens empiriques qui les unissent à la page initiale.

3.2.2 Discussion

Si l'on compare cette méthode de corrélation " empirique" avec la précédente, on constate que l'on travaille sur des données plus volatiles et plus difficile à évaluer, car on manipule des données recueillies via une étude comportementale. Toute la difficulté de cette méthode est donc de trouver une interprétation plus ou moins rationnelle aux parcours des internautes sur la toile. Cette difficulté explique, en partie, la phase d'initialisation du procédé pendant laquelle on crée une structure primaire de liens corrélés qui consiste à lier les documents les uns aux autres par l'intermédiaire de procédés de clusterisation. Ce travail préalable permet d'imposer un cadre lors de l'élaboration des résultats corrélés et évite ainsi que cette méthode ne donne trop de résultats incohérents.

Néanmoins, tout repose sur une interprétation du parcours de l'utilisateur. Or cela pose un problème ; pour s'en convaincre il suffit de regarder les statistiques réalisées sur le moteur de recherche Altavista [SHMM98], pour lesquelles nous avions relevé, section 2.2.1.1, que 70% des utilisateurs qui se connectent à un moteur l'interrogent une unique fois par session et se contentent de regarder la première page de résultats renvoyés. Même si les statistiques ont évolué, pour utiliser une méthode empirique de corrélation, il faudrait préalablement pouvoir vérifier et évaluer la qualité du parcours réalisé par les internautes. Dans le cas contraire, une telle méthode ne peut pas assurer que les documents corrélés ne doivent pas en fait leur rapprochement à une pondération particulière des résultats sur un ou plusieurs moteurs de recherche. De ce fait, on peut craindre que les résultats retournés ne fassent apparaître que des sites " populaires ", car de par leur statut ils briguent les premières places dans les recherches et ont plus de chance d'être liés les uns aux autres.

La deuxième absence que l'on peut trouver avec cette méthode est, comme pour les méthodes topologiques, que les travaux effectués ne permettent pas de mettre en relation des pages Web les unes avec les autres, mais uniquement des sites. L'explication est similaire : pour obtenir suffisamment d'information, de type empirique, entre les éléments à corréler on ne regarde que les éléments de taille suffisante, c'est à dire les sites.

3.3 La corrélation hiérarchique

Pour aborder les problématiques de la corrélation, on se doit également d'aborder des sujets voisins tels que la classification de documents. Classification et corrélation sont étroitement liées. On pourrait même ajouter que le processus est identique et que seules les conditions initiales changent. Pour schématiser, la classification tente de ranger des documents sémantiquement proches dans des mêmes classes, alors que la corrélation part d'un document et construit la classe à laquelle il appartient. En poursuivant dans cette voie, une méthode triviale de corrélation serait de construire les classes de documents corrélés à partir de celles de la classification. De tels rapprochements ont d'ailleurs été exploités par Weiss [WVS+96] pour mettre au point le système " hypursuit " qui est une méthode de corrélation fondée sur une classification topologique. La question qui nous anime dans cette partie est de savoir si un rapprochement entre corrélation et classification peut être envisagé pour mettre au point notre méthode de corrélation.

En effet, même si les deux techniques sont très proches d'un point de vue conceptuel, on peut s'interroger sur la faisabilité d'un tel procédé et sur la validité d'employer des outils de classification dans un but détourné. En terme de faisabilité, corrélation et classification sont assez éloignées l'une de l'autre. La motivation de la seconde étant de créer physiquement l'ensemble des classes de documents alors que la première se contente de les créer en fonction des besoins des utilisateurs. En pratique ces deux méthodes peuvent employer des réalisations techniques différentes qui risquent de les éloigner.

Les deux méthodes divergent quelque peu dans leurs réalisations et dans leur finalité, mais ont un dénominateur commun : regrouper des documents sémantiquement proches. Nous allons examiner les différents travaux de classification existants en vue de les transposer à des travaux de corrélation. Nous allons, préalablement, clarifier le vocabulaire. Nous parlerons indifféremment de groupes, de clusters ou de classes pour désigner les briques élémentaires de la classification que sont les classes. De la même manière, pour désigner le terme de classification, nous pourrons utiliser celui de clusterisation. Néanmoins, il ne faut pas non plus confondre le concept de la classification avec celui de la catégorisation, confusion due au problème d'interprétation des sources documentaires anglo-saxonnes. La classification ne part que des documents du corpus et essaye de construire ces classes en analysant le contenu, alors que dans la catégorisation, les classes sont initialement créées en fonction des résultats auxquels on veut arriver et on essaye de les remplir en fonction de leur description.

3.3.1 La classification

On peut étudier la classification au travers de plusieurs axes de recherches. Il y a tout d'abord une approche topologique de la classification qui comme son nom l'indique étudie la structure des liens hypertextes pour trouver des éventuels rapprochements entre les documents [BS91, Bot93]. On trouve ensuite des techniques de classification dites interactives [CKPT92, HP96] qui permettent à l'utilisateur d'exprimer son choix. On peut aussi entrevoir des méthodes de classification qui permettent d'identifier des communautés de documents [KRRT99a, KRRT99b, Muk00a] ou qui permettent d'extraire des informations utiles sur le Web [Lar92, PSHD96]. Mais avant d'entrevoir toutes ces possibilités nous allons revenir aux sources de la classification et aborder des méthodes fondées sur des approches terminologiques.

3.3.2 La classification de termes

L'idée première de la classification est de regrouper et d'organiser l'information en fonction des similarités que l'on peut y trouver [SS77]. Par exemple on peut agréger différentes sources d'information proches sémantiquement pour fabriquer une source d'information condensée de plus haut niveau ou bien alors tenter de rapprocher des concepts proches dans des concepts plus généraux (ainsi un aigle, une hirondelle ou un moineau peuvent être inclus dans le concept d'oiseau).

La classification textuelle de documents permet de regrouper les documents en fonction des similarités que l'on peut retrouver au sein des textes considérés, et/ou en fonction des citations éventuelles que l'on peut rencontrer. Les premières investigations faites se sont intéressées au degré de co-occurrence des termes du corpus. Le degré de co-occurrence entre deux termes reflète le nombre de fois où ces termes sont utilisés simultanément dans un même contexte (phrase, paragraphe, segments, etc.) au sein d'un même document. Plus ce degré de co-occurrence est élevé plus ces termes sont simultanément associés. De telles considérations devaient améliorer la précision des résultats (voir définition section 7.2.2) en permettant d'indexer les documents sur la base d'un vocabulaire controlé provenant d'un thésaurus automatiquement généré à partir des degrés de co-occurrence des termes initiaux qui était la traduction de leurs regroupements.

Cette méthode a obtenu des premiers résultats concluants [SJ71], mais les expériences successives n'ont apporté que des résultats mitigés et l'efficacité de cette approche a été largement discutée [Sal72, MWZ72, Sal86, PW91]. Dans des études plus récentes Voorhees [Voo93] utilise les ressources du thésaurus anglophone WordNet, grâce auquel elle explore les relations sémantiques entre les termes contenus dans cette base de connaissance. Voorhees procéda à des expérimentations sur 5 corpus différents (CACM, CISI, CRAN, MED, TIME) en utilisant WordNet pour désambiguïser le sens des termes étudiés et ensuite pour comparer son expérimentation avec des techniques classiques de désambiguïsation de termes. La conclusion de cette expérience est que les résultats obtenus sont de moins bonne qualité que les techniques traditionnelles, la principale difficulté rencontrée se manifeste notamment pour désambiguïser la signification des termes dans des requêtes courtes.

Une autre approche de la classification de termes est d'utiliser d'autres techniques propres à la recherche d'information en amont comme des techniques de pondération de termes. Malheureusement, il a vite été montré que ce choix pouvait être vu de manière critique dans le cadre de l'utilisation des méthodes de classification [RSJ76, Sal71, SB88, vR79]. Les travaux de Rasmussen [Ras92] ont par ailleurs souligné que les méthodes de pondération n'avaient que peu d'impacts positifs sur les travaux de classification. La conséquence directe est qu'une telle voie a été abandonnée au profit d'autres méthodes de classification.

3.3.3 La classification de documents

La classification documentaire peut se scinder suivant deux thématiques principales : la classification hiérarchique et la classification non-hiérarchique.

La différence entre les deux formes de classification est conceptuelle, les classifications non-hiérarchiques mettent tous les documents sur un pied d'égalité et les répartissent dans différentes classes, sans hiérarchiser les classes. La classification hiérarchique permet au contraire d'obtenir une arborescence des classes, plus au moins précise en fonction de leur niveau dans la hiérarchie.

3.3.3.1 La classification non-hiérarchique

La classification non-hiérarchique est le procédé le moins utilisé car il pose un certain nombre de problèmes. Ainsi une telle méthode pose le problème du partitionnement d'un corpus en sous-ensembles optimaux, le critère d'optimalité étant à définir en fonction des besoins, celui qui nous intéresse étant une proximité sémantique optimale entre documents. En pratique, il est bien compréhensible qu'une telle méthode soit difficilement réalisable.

Pour simplifier les traitements la classification non-hiérarchique implique habituellement une heuristique de production de divisions sous-optimales et itératives. Les ensembles traités sont préalablement scindés en parties de tailles raisonnables. L'une des approches les plus répandues est de réduire progressivement le nombre total de critères de similarité entre documents pris en compte jusqu'à obtenir une stabilisation du nombre de documents ou jusqu'au dépassement d'un seuil prédéfini [Wil88].

La création des classes par une méthode non-hiérarchique peut sembler arbitraire car la génération des classes peut dépendre de multiples paramètres tels que l'ordre d'analyse des documents, le nombre d'éléments dans la classe ou bien encore la définition d'un seuil. La conséquence directe est que ce type de classification n'est pas souvent employé malgré une rapidité supérieure aux classifications hiérarchiques qui exigent d'un point de vue quantitatif des calculs de similarités entre toutes les paires possibles de documents et de classes.

3.3.3.2 La classification hiérarchique

Les méthodes de classification hiérarchiques peuvent être de deux sortes : agglomératives ou séparatives.

Les méthodes séparatives construisent des hiérarchies de classes en partant de la racine puis en se divisant en de multiples sous-classes. A chaque niveau inférieur de l'arborescence les classes deviennent de plus en plus spécialisées. Au contraire, les méthodes agglomératives construisent dès le départ des classes très spécialisées et tentent de les regrouper en éléments plus généraux. Des classes séparatives sont monothétiques, cela signifie que les membres de la classe doivent contenir certains termes, alors que les classes agglomératives sont polythétiques, ce qui signifie que les documents appartenant à une classe doivent avoir quelques termes en commun mais qu'aucun n'est exigé pour adhérer.

Les classifications hiérarchiques agglomératives sont les plus populaires, elles sont prépondérantes dans la littérature, et quelquefois on va jusqu'à assimiler les classifications hiérarchiques aux seules classifications hiérarchiques agglomératives [EHW89, GRW84, JvR71, Voo86, Wil88]. Les deux procédés étant de surcroît similaires, nous n'étudierons que des méthodes agglomératives.

La classification hiérarchique agglomérative Le procédé inhérent aux classifications hiérarchiques agglomératives est la construction d'une matrice de similitude contenant l'intégralité des scores de similarité entre chaque paire de documents du corpus. Le principe est de faire tourner un algorithme itératif sur cette matrice pour fusionner progressivement les paires de documents entre elles. La fusion des cellules respecte des critères propres à chaque méthode ; pour toutes le but est identique : former des clusters de documents de plus en plus grands. En pratique, les paires possédant les scores de similarités les plus grands sont fusionnées et la matrice est mise à jour. Les lignes et les colonnes contenant les éléments fusionnés sont remplacées par les scores de similarités de la classe ou du cluster ainsi trouvé. Le processus est répété jusqu'à l'obtention d'un unique cluster racine. Le résultat de ce calcul est un arbre binaire dont les branches correspondent aux différents regroupements successifs. Il existe plusieurs méthodes de classification hiérarchique agglomérative, mais toutes suivent dans les grandes lignes cette procédure.

La méthode dite de lien simple (de l'anglais *simple linkage*) utilise la paire de documents ayant une similarité maximale (c'est-à-dire les plus proches voisins) entre les classes ou clusters pour déterminer la similarité des autres clusters. Cette méthode a tendance à former de grands clusters, ayant peu de cohésion interne. Malgré tout elle reste la plus utilisée en raison de sa rapidité d'exécution.

La méthode de lien complet (*complete linkage*) est antinomique de la précédente et emploie comme élément de départ la paire de documents ayant la similarité la plus faible (c'est à dire un éloignement maximum) entre les clusters pour déterminer leur similarité. Cela a pour conséquence d'obtenir des clusters beaucoup plus petits possédant de forts liens de similarité.

Il existe une autre méthode, appelée méthode du lien moyen de groupe (*group average method*), qui tente de faire un compromis entre les deux méthodes précédentes. Malheureusement, elle a tendance à engendrer des aberrations et des classes de petites tailles [Wil88].

3.3.4 Des approches nouvelles

La classification est au même titre que les autres domaines de la recherche d'information une voie de recherche en devenir qui fait preuve de créativité et expérimente de nouvelles voies. Nous allons en exposer quelques unes qui pourraient être utilisées dans le cadre de travaux de corrélation.

3.3.4.1 La classification interactive : exemple de Scatter/Gather

La classification est, un outil de recherche d'information, à l'image des moteurs de recherche. Elle peut être interactive et donner un degré de liberté supplémentaire comme le fait la méthode de Scatter/Gather. Cette méthode s'utilise de manière interactive et elle a pour but de construire une arborescence qui se modifie à mesure que l'utilisateur comprend mieux la nature des documents disponibles et découvre ceux qui sont les plus intéressants [CKPT92, HP96].

Le principe de fonctionnement est le suivant : il suffit d'imaginer qu'en posant une question à un outil de recherche d'information, tel un moteur de recherche, l'on reçoive 500 réponses. La méthode Scatter/Gather examine alors ces pages réponses et les répartit en plusieurs groupes selon leurs ressemblances. L'utilisateur peut alors examiner ces groupes et choisir ceux qui lui semblent être les plus intéressants. Bien que le comportement des utilisateurs soit variable, les expériences indiquent que les arborescences facilitent l'accès aux documents pertinents même si le groupement de documents n'est pas une panacée, car il n'explique pas comment les documents sont trouvés.

La principale différence qui existe avec d'autres méthodes est qu'un ensemble de documents valides pour une question donnée peut ne pas être valide pour une question différente. De multiples expérimentations ont été faite autour de Scatter/Gather pour étudier la création des classes, pour analyser les questions posées en entrée ou bien encore pour améliorer l'interactivité [HKP95]. Ces travaux montrent que des recherches utilisant cette méthode sont plus performantes que ces mêmes recherches confrontées à des techniques standard de recherche par mots-clefs.

3.3.4.2 La classification topologique

Les premières méthodes ayant exploité cette possibilité sont basés sur des algorithmes d'analyse et de création de graphe. Botafogo & Shneiderman [BS91] utilisaient des graphes pour identifier et concevoir les regroupements de documents grâce à un calcul de distance entre liens hypertextes et entre documents. Les regroupements qui en découlent n'étant pas suffisamment discriminants, cette approche fut améliorée [Bot93] en partant de l'idée que la qualité des relations entre les liens est proportionnelle au nombre de liens indépendants dans leur URL.

Une autre approche topologique a été fondée sur une analyse des co-citations [WM89] couplée avec une technique de réduction et de simplification de matrice nommée technique de mise à l'échelle multidimensionnelle (MDS) [Lar92]. Le point de départ de cette méthode est l'étude de la matrice de co-citation que l'on tente de réduire via MDS et pour laquelle ensuite on essaye de visualiser les regroupements de documents.

Kumar [KRRT99a] décrivit un nouvel algorithme de classification combinant notion de co-citation et analyses de graphes nommé “trawling”. Son but premier était d'identifier les différentes communautés de documents mises à disposition sur le Web et de les représenter au travers d'une classification avec différents niveaux de granularité comme le fait l'annuaire Yahoo [Cord]. Cet algorithme reprend le principe de l'algorithme de co-citations qui considère que deux pages sont à mettre en relation si elles sont souvent co-citées dans les documents. Il se réfère également au principe de renforcement mutuel du poids des liens cher à HITS. Le déroulement de l'algorithme passe par l'utilisation initiale d'un panel de classes existantes qu'il nettoie de ses doublons selon sa propre politique inspirée toutefois de Broder [BGMZ97]. Cette étape permet d'éliminer en moyenne 60% des documents qui auraient fait perdre de la précision au classement des documents. Les nouveaux documents sont ensuite ajoutés aux classes existantes et finalement ces classes sont à nouveaux épurées et redessinées. Avec une telle méthode, Kumar a obtenu un corpus contenant 100.000 classes, parmi lesquelles il a vérifié aléatoirement 400 d'entre elles, qui lui font dire que 96% des classes étaient convaincantes et que 56% des documents disposés dans ces classes n'avaient pas été pris en compte par Yahoo.

3.3.4.3 Une approche hybride de classification

L'idée sous-jacente qui vient, après avoir étudié des méthodes de classification fondées sur une approche topologique et d'autres fondées sur une approche textuelle, est de vouloir mélanger ces deux approches pour ne garder que le meilleur des deux. Ainsi, dans une approche textuelle, les relations mises en avant sont les similarités (linguistiques le plus souvent) entre les documents, alors que dans une approche topologique on prend en compte les citations mutuelles. Une approche topologique isolée souffre de la nature des liens qu'elle étudie, ce qui implique de prendre des mesures complémentaires pour valider leur choix. La même constatation peut être faite avec n'importe quelle autre approche de classification textuelle qui doit être complétée d'autres renseignements destinés à améliorer la qualité des résultats obtenus. Une approche hybride entre les deux techniques doit justement permettre cette amélioration de la qualité des résultats.

Pour illustrer, nous pouvons donner plusieurs exemples de techniques hybrides qui viennent compléter les outils propres à la classification. L'une des équipes pionnières des méthodes hybrides est celle de Pirolli [PPR96] qui a rangé les pages du Web en exploitant une représentation vectorielle des données qui sont traitées sous la forme de vecteurs hybrides combinant les deux types d'approches. La représentation des documents pour Pirolli s'appuie sur des vecteurs contenant des informations topologiques, des critères de similarité textuelle et des meta-informations (titre, taille, ...). Ils servent ensuite à construire un graphe pour représenter les documents que l'on essaye de simplifier et dans lequel on cherche des regroupements pour définir les classes de documents.

Des études plus récentes sur des techniques hybrides explorent des liens plus étroits entre les interactions des méthodes textuelles et topologiques. Modha & Spangler [MS00] et Mukherjea [Muk00a, Muk00b] utilisent dans un premier temps une méthode de classification textuelle et recherchent alors un petit noyau de documents corrélés, qu'ils tentent ensuite de faire grossir par des techniques topologiques. Mukherjea augmente, par exemple, la qualité des résultats en définissant un seuil de similarité textuelle entre les documents, une démarche qui s'inspire d'une stratégie hybride de recherche menée par Bharat et Henzinger [BH98]. Mukherjea [Muk00a, Muk00b] crée le système WTMS 40 pour regrouper et analyser les documents toujours dans le cadre d'une méthode hybride. Le but est de trouver des pages relatives à un sujet particulier et de les organiser pour améliorer la recherche d'information pour les internautes.

3.3.5 Discussion

D'un point de vue général, la comparaison qui oppose des recherches à base de classification et d'autres sans classification montre que ces premières n'apportent pas nécessairement une amélioration de la qualité des résultats, et qu'elles peuvent même dans certains cas les dégrader [GLW86, JvR71, Sal71, vRC75, Wil88]. La raison invoquée est simple : la création des classes est une affaire de point de vue et ne respecte pas toujours celui de l'utilisateur.

De plus le processus de classification est souvent considéré comme très lent et trop gourmand en ressources pour être appliqué à des corpus conséquents [Ras92, Wil88]. Confrontés à ces réalités, les chercheurs tentent de trouver des alternatives pour organiser automatiquement l'information. Les mauvaises performances des recherches utilisant la classification ont poussé à regarder celle-ci comme un outil de navigation dans lequel on permet aux utilisateurs de naviguer et ainsi de rechercher plus facilement les documents appropriés.

Ainsi, Crouch & Andreas [CCA89], ont regardé l'efficacité de l'utilisation d'une classification interactive pour retrouver des documents. Le résultat de cette étude était qu'une recherche menée par l'utilisateur au sein d'une classification donnait des résultats bien supérieurs aux techniques plus traditionnelles de recherche automatique à l'aide d'une classification. Le jugement de l'être humain n'est pas seulement souhaitable, mais il est une alternative viable aux systèmes conventionnels fondés sur des recherches à base de classification.

En ce qui concerne la qualité des résultats, les méthodes interactives semblent être de bonnes voies pour la classification, en termes de rapidité ou en termes de pertinence des résultats. Néanmoins elles nécessitent l'avis d'un utilisateur, ce qui est difficilement envisageable dans le cadre d'une méthode de corrélation fondée sur une classification. Il faut nécessairement que celle-ci soit figée et non évolutive. Ce type de classification doit donc être écarté.

Hormis cette exclusion préliminaire des méthodes interactives, les autres voies de classifications restent envisageables pour bâtir une méthode de corrélation hiérarchique. Il reste à voir ensuite si de tels procédés peuvent s'adapter aux exigences que nous nous sommes fixées.

3.4 La corrélation linguistique

Pour que notre tour d'horizon des méthodes de corrélation soit complet, nous abordons enfin une voie expérimentale de corrélation : la corrélation linguistique.

Celle-ci diffère des précédentes méthodes car elle utilise un composant de base de la recherche d'information : l'extraction terminologique. On s'intéresse ici uniquement aux attributs de base des documents, c'est à dire les phrases et les termes (mots ou expressions) qui les composent. On se rapproche des travaux de résumés automatiques par extraction de termes clefs, qui reposent sur l'hypothèse qu'il existe dans tout texte des unités textuelles saillantes [DM00]. L'idée est donc d'extraire dans un premier temps une signature lexicale pour chaque document, étape qui sera abordée dans la section 3.4.1, et de s'en servir pour retrouver les documents pertinents (Section 3.4.2).

3.4.1 Extraire les termes saillants

L'extraction de termes saillants sert à retrouver dans un ou plusieurs documents donnés les termes discriminants et utiles à la compréhension du discours. Ce travail d'identification des éléments clefs du discours est à la base de nombreux travaux répartis dans différents domaines : résumé automatique [Jin00], segmentation thématique [Her02, LH97], apprentissage de connaissances [Fer98], etc.

La particularité de ces travaux est qu'ils nécessitent pour la plupart une phase initiale de pondération des termes pour extraire ceux qui apparaissent comme les plus discriminants. La corrélation linguistique n'y fait pas exception et passe par une phase de recherche et d'identification des éléments clefs via des méthodes de pondération avant d'aborder la recherche des documents corrélés.

L'identification des termes clefs passe par une étape de pondération qui va être étudiée maintenant. Cette étape se scinde en deux grandes parties : les méthodes de pondération linguistique (Section 3.4.1.1) et statistique (Section 3.4.1.2).

3.4.1.1 Les méthodes de pondération linguistique

Les méthodes de pondération linguistique sont des méthodes qui recherchent des termes clefs en étudiant la structure du discours. Des mots peuvent acquérir plus d'importance que d'autres car, dans un corpus donné, ils sont fréquemment utilisés, par exemple, dans les titres des articles. De telles considérations ont été étudiées par Dennis & Bookstein [Den67, BKR98] qui ont utilisé les occurrences d'un terme [Luh57] en y couplant une pondération des termes en fonction de leur position dans les paragraphes et dans les documents.

D'autres travaux comme ceux menés par Edmundson [Edm69] recherchent plus spécifiquement des méta-marqueurs du langage (tels que " en conclusion " ou " cet article décrit ") pour trouver des passages intéressants et donc des termes discriminants. Edmundson fut le premier à montrer l'importance d'avoir des heuristiques pour localiser les éléments clefs du discours, en s'appuyant sur les travaux de Baxendale [Bax58]. Cette voie a ensuite été suivie par d'autres dont le but était d'identifier des phrases clefs [PJ93] ou de repérer des termes de contexte négatif [MRY73]. Une étude du discours plus poussée, incluant la prise en compte des règles de grammaire, permet dans certains cas d'identifier les noms propres d'un document [MML+93, PY94].

3.4.1.2 Les méthodes de pondération statistique

La deuxième grande voie de la pondération est statistique. L'un des premiers auteurs à s'être intéressé à ces considérations fut Luhn [Luh57] qui souhaitait retrouver le vocabulaire spécifique d'un document. Dans cette optique, il émit l'hypothèse que la distribution des mots dans le corpus pouvait renseigner sur leur qualité informationnelle, et fit la constatation que les termes les plus fréquents n'apportent que peu d'informations au discours car ils sont trop communs. Il admit également qu'une faible représentation d'un mot dans un texte conséquent n'apporte pas non plus d'information. Cette remarque, que l'on doit rapprocher de l'expérimentation de Zipf [Zip49], peut s'énoncer simplement : les termes peu pertinents sont aux deux extrémités de la courbe de distribution fréquentielle des termes du corpus.

Néanmoins, lorsque l'on regarde le nombre d'occurrences d'un terme dans un document donné, on remarque que celui-ci est proportionnel à l'intérêt de ce terme pour le document. Cela provient du fait qu'un auteur répète naturellement les termes importants à la compréhension de son article. Bien entendu, la polysémie et la synonymie modèrent cette constatation, mais de manière générale, un terme fréquent dans un document est plus important qu'un terme rare. La fréquence d'un terme est le plus souvent utilisée pour montrer l'importance de celui-ci dans la représentation du contexte [Luh57, Bax58, SYY75, SBY83, Sal89].

Le problème de l'extraction terminologique est qu'un document, même après filtrage, contient toujours des mots sans intérêt qui ne sont pas représentatifs du discours. Ces mots commun appelés " mots vides " foisonnent dans les corpus documentaires et dépendent du corpus étudié, ainsi le terme " loi ", qui peut être discriminant dans un corpus général, va être porteur de bruit dans un corpus juridique. On constate qu'un terme devient discriminant et significatif pour un corpus donné s'il n'y est pas trop fréquent. L'indice I_{df} (*inverse document frequency*) est utilisé pour mesurer cette propriété [SJ73, SB88, Lee95]. L'indice I_{df} maximise le poids du terme lorsque celui-ci n'apparaît que dans un document et le minimise lorsque celui-ci apparaît dans tous les documents. Un score comme I_{df} dépend du corpus utilisé car il est fondé sur la distribution des termes dans la collection. Quand le corpus évolue au fil du temps, les indices de pondération doivent être recalculés s'ils ne sont pas obtenus dynamiquement ; ce n'est pas rédhibitoire à leur emploi mais cela peut être décourageant lorsque le corpus est en perpétuelle évolution [SB88]. Le poids amené par I_{df} est couramment utilisé pour rechercher les termes potentiellement intéressants dans le corpus ou dans un document, mais il permet a priori de déterminer les mots vides de sens dans le corpus, ceux pour lesquels I_{df} est nul ou quasi-nul [SOK94].

L'intérêt d'utiliser des poids est qu'ils peuvent aisément se combiner ; ainsi, comme on vient de le voir, la fréquence d'un terme lui assure de l'importance lorsqu'il apparaît souvent au sein d'un même texte, et I_{df} lui donne du poids à condition qu'il ne se retrouve pas dans tous les documents du corpus. Combiner les deux notions assure qu'un terme devient important pour un document donné si celui-ci apparaît souvent dans ce texte, mais relativement peu dans d'autres documents. Cette combinaison, très répandue pour déterminer les termes clefs d'un texte [SJ73, SYY75, SB88, Har86] est à la base d'une méthode de pondération nommée $T_f.I_{df}$.

Les indices qui ont été vus jusqu'à maintenant sont relativement généraux et ne tiennent pas compte de la longueur des textes analysés. Pour compenser l'effet de la longueur des textes face aux informations qu'ils distillent, on est dans l'obligation de normaliser les pondérations utilisées. Par exemple, la fréquence d'un terme est normalisée tout simplement en divisant l'ancienne valeur par la fréquence maximum de celui-ci dans le corpus [SB88, Lee95]. Une autre possibilité est d'utiliser l'indice cosinus en divisant, cette fois-ci, la précédente valeur par la racine carrée de la somme des carrés des fréquences du terme dans le corpus. Bien entendu, cette normalisation peut s'étendre à d'autres pondérations. Pour des textes relativement longs, qui peuvent contenir différents concepts et différentes significations, la normalisation a tendance à pénaliser les termes qui s'y trouvent mais il s'agit parfois de l'effet recherché [Str94].

La décision de qualifier un terme de pertinent ou non s'opère en fonction du score qu'il a obtenu. Plus le score est élevé, plus le terme sera pertinent pour une pondération donnée. Cette méthodologie peut être appliquée à toutes sortes de termes, y compris à des entités autres que les mots, comme les syntagmes nominaux (un syntagme nominal sera noté SN par la suite) ou des phrases. Les méthodes de pondération statistique proposent des pondérations dédiées à ces différents types de termes. Parmi ces techniques on peut citer les travaux sur l'information mutuelle [CH90], le coefficient ϕ^2 [GC91] ou encore le coefficient Loglike [Dun93]. Tous ces indices servant à pondérer des SN ont été comparés au travers d'expérimentations menées par Daille [Dai96] pour au final montrer que l'indice Loglike obtient les meilleurs résultats pour rechercher les SN pertinents.

De la même manière, il existe des méthodes dédiées à la recherche de phrases clefs dans un texte. On pourrait imaginer qu'une pondération triviale serait de définir le poids d'une phrase comme étant égal à la somme des poids des termes qui la composent. Malheureusement, ce n'est pas aussi simple [Fuh92, LSJ96], l'élaboration d'un poids pour une phrase est d'ailleurs un travail qui demande beaucoup plus d'efforts que pour des mots simples [Fag89, CTL91, Buc93, SGPC+97]. Ces techniques de pondération étant par ailleurs en dehors de nos axes de recherche, nous ne nous investirons pas dans ce sujet.

Dernière étape à entrevoir dans le cadre d'une extraction terminologique, le cas des méthodes alternatives. La première méthode a été évoquée par Lafon [Laf80] qui s'intéresse aux spécificités du vocabulaire. Il propose ainsi d'appliquer une distribution hypergéométrique à la répartition des termes dans le corpus. La méthode exploite les spécificités du vocabulaire propre à chaque partie du corpus pour en déduire les termes saillants. Toutefois cette approche impose de nombreuses restrictions qui rendent cette méthode difficilement utilisable sur tout type de corpus [LL00].

Autre voie alternative, celle des modèles statistiques qui proposent de nouvelles approches utiles pour traiter de grands volumes d'information. Si l'on considère que dans un corpus suffisamment grand, les mots sont répartis de manière aléatoire, on peut alors voir le processus d'écriture des textes comme un processus stochastique où les documents sont générés à partir d'une sélection aléatoire de termes. Les distributions de Poisson sont des outils qui justement peuvent être appliqués pour modéliser des phénomènes aléatoires. Une première approche [BS74] est de dire que des mots communs que l'on a appelés précédemment des mots vides sont réellement distribués au hasard parmi le corpus ; il est donc très facile de les retrouver via des modèles de Poisson [Mar92]. De la même manière les termes pertinents dévient du modèle de Poisson, car leur répartition dans le corpus est moins soumise au hasard du fait de la volonté de l'auteur.

3.4.2 L'extraction de documents corrélés

La section 3.4.1 nous a permis de passer en revue une grande partie des méthodes de pondération existantes. Une méthode spécifique de corrélation linguistique consiste à sélectionner, à partir des pondérations retenues, une liste de termes candidats pour former une signature lexicale qui servira de point de départ à la recherche des documents corrélés. Bien entendu, le choix de la pondération à utiliser est fonction des besoins et de plusieurs paramètres comme la nature du corpus ou sa taille. La recherche de documents corrélés que nous abordons maintenant peut s'envisager de différentes manières ; on en dénombre trois, en fonction des utilisations que l'on veut faire des signatures lexicales.

3.4.2.1 Calcul de distance entre signatures

Les signatures représentant un ensemble de termes discriminants pour un document donné, l'idée qui vient à l'esprit est de vouloir comparer les signatures entre elles. En effet, on peut penser que deux signatures similaires appartiendront à des documents sémantiquement proches.

En pratique, on définit une notion de distance entre les signatures en fonction du nombre de termes communs entre elles pour mesurer leur proximité. Une fois les calculs opérés, il ne reste plus qu'à ne garder que les documents les plus proches du document initial pour obtenir un pool de documents corrélés.

Ce calcul de distance peut, par exemple, utiliser la méthode des vecteurs de contexte [JT99] qui permet de mesurer une distance entre des ensembles de termes de même longueur. Une telle utilisation des signatures lexicales n'entre pas dans le champ des outils de corrélation mais met en relation des portions de textes [JT99] (phrases, paragraphes, etc.). Même s'il ne s'agit pas d'une réelle corrélation entre documents, la finalité reste proche, et la transposition de ce procédé sur des documents entiers n'est certes pas triviale mais reste envisageable.

3.4.2.2 La signature vue comme une requête

Une autre approche est de comparer la signature retrouvée non plus avec d'autres signatures, mais directement avec le vocabulaires des autres documents. L'intérêt est de pouvoir améliorer le taux de rappel des résultats (voir définition section 7.2.2). Pour réduire le temps de calcul, les tailles des signatures utilisées sont relativement faibles : environ une dizaine de termes choisis parmi des centaines. Le cas où deux signatures sont identiques est très rare et le cas où deux signatures comportent suffisamment de termes en commun pour être proches n'est pas fréquent. On obtient donc peu de résultats corrélés.

Lorsque l'on recherche maintenant les éléments communs entre une signature et un document, on augmente nécessairement le nombre de résultats corrélés et on améliore ainsi le taux de rappel. Néanmoins on ne peut plus utiliser des calculs de distances comme précédemment car ils nécessitent que les deux éléments à comparer aient la même taille, et une signature est par définition plus petite qu'un document.

L'approche présentée ici, étudiée par Park & al [PPGK02] et Phelps & Wilensky [PW00], considère la signature lexicale une fois créée comme une requête booléenne. Cette requête, soumise à un moteur de recherche, renvoie des résultats contenant plus ou moins d'éléments de la signature initiale. Ces résultats doivent ensuite être classés en fonction d'une pertinence qui reste à définir car les résultats renvoyés sont nombreux et plus ou moins proches de la question posée. Cette approche permet de construire des signature lexicales dont la finalité est multiple, mais qui permettent notamment de rechercher et de trouver des résultats corrélés.

Le principal avantage de cette technique est la simplicité de mise en œuvre ; malheureusement la qualité des résultats découle pour une grande partie de la pertinence des termes de la signature. De plus, l'utilisation des requêtes booléennes sur un nombre de termes potentiellement grand peut engendrer des problèmes d'explosion combinatoire.

3.4.2.3 La corrélation vue sous la forme d'un réseau de termes

La dernière approche permettant d'obtenir des résultats corrélés est fondée sur l'utilisation des co-occurrences. Elles peuvent être de deux sortes : syntaxiques ou purement statistiques. Les co-occurrences syntaxiques, plus généralement appellées relations de dépendances syntaxiques, utilisent exclusivement des relations syntaxiques pour trouver des mots qui co-occurrent comme des relations de type adjectif-nom, sujet-verbe, verbe-complément, etc. Cet aspect a été utilisé dans des travaux de désambiguïsation de sens [HR91] avec des résultats probants. La seule condition imposée pour fabriquer de telles co-occurrences est de disposer d'un analyseur syntaxique efficace.

Le deuxième type de co-occurrence est constitué à partir d'un calcul de distance entre les termes. Dans le cadre de la corrélation, la recherche de co-occurrences permet, à partir des mots clefs, de tisser des liens entre les termes de la signature et les documents. Les liens corrélés sont représentés par une matrice de co-occurrences qui associe des documents à des éléments de la signature. Une telle méthode rapproche les documents en fonction des liens trouvés entre les termes du corpus.

3.4.3 Discussion

La corrélation linguistique qui, pour l'instant, n'en est qu'à ses débuts, possède des aspects intéressants : elle reste simple à mettre en œuvre mais surtout elle est facilement interprétable. Contrairement aux voies topologiques et empiriques, si deux documents sont mis en relation par une approche linguistique, cela sous-entend qu'ils partagent un vocabulaire commun, ce qui semble plus intuitif comme approche de corrélation à condition, bien sûr, de tenir compte du contexte d'utilisation des termes. Néanmoins cette approche soulève de nombreuses questions qui ne sont pas encore résolues.

La première d'entre elles concerne le choix des termes qui composent la signature ; de nombreux travaux ont suggéré que l'emploi de mots pour qualifier un texte ou un document n'est pas suffisant à cause de leur trop grande polysémie et synonymie. Comme cela a été évoqué, il existe de nombreuses méthodes de pondération pour les mots, les SN, ou encore les phrases. La question qui se pose est de savoir de quels types de termes doit être composée cette fameuse signature lexicale.

Une fois le choix opéré entre les différents types de termes envisageables, encore faut-il pouvoir choisir les termes à prendre en compte. Les méthodes de pondération permettent d'obtenir des listes de termes triés par ordre décroissant de pertinence. Mais comment opérer un choix ? Doit-on imposer un seuil au delà duquel les termes ne doivent plus être pris en compte, ou plus simplement peut-on se contenter d'un nombre fixe de termes ?

Une question d'ordre général doit également être évoquée avant de persévérer dans cette voie de recherche : peut-on utiliser des méthodes de pondération dont la finalité première est de trouver des termes pertinents dans un document ou un corpus pour également rechercher des termes clefs dédiés à la corrélation ? Cette question peu souvent abordée dans la littérature est cruciale puisque de sa réponse dépend la validité sémantique de cette approche.

Le constat sur la corrélation linguistique semble évident, cette voie de recherche nouvelle soulève quelques problèmes et un certain nombre d'interrogations qui devront être nécessairement traités par la suite avant de pouvoir conclure sur la validité d'une telle méthode de recherche de documents corrélés.

Deuxième partie

Matériel et méthodes

Chapitre 4

Élaboration d'une méthode de corrélation dans l'usage

Comme tout travail, celui que nous entreprenons ne se fait pas sans finalité. Celle qui nous anime est de déterminer la meilleure stratégie possible pour réaliser une méthode de corrélation entre documents.

Le but ultime de notre méthode de corrélation n'est pas de rechercher des documents corrélés sur un large corpus hétérogène à l'image du Web, mais au contraire de fonctionner sur un ensemble de documents juridiques, définis section 4.1. Sa réalisation doit dans un premier temps être testée sur un corpus dit de référence, que nous allons également aborder section 4.1. Le chapitre 3 nous a révélé qu'il existait plusieurs voies bien distinctes de corrélation, chacune répondant à des impératifs et des finalités différentes. L'objet de la section 4.2 est de définir nos besoins afin de mieux comprendre l'orientation prise par nos recherches avant d'énoncer plus clairement, dans la section 4.3, les étapes de notre méthode pour finalement exposer dans la section 4.4 ses degrés de liberté.

4.1 Un corpus de référence

Le travail que nous réalisons a pour but de faciliter l'accès à l'information sur un portail juridique composé de plusieurs ensembles de textes de natures différentes (des textes constitutionnels, des lois organiques, des textes de jurisprudence, des textes européens, des articles de codes, etc.). La méthode de corrélation que nous développons a vocation à faciliter la recherche d'informations sur l'intégralité des textes couverts dans notre portail juridique ; néanmoins, pour sa réalisation, nous nous devons de définir un corpus de référence dont le choix revêt une grande importance.

La section 4.1.1 s'attache à définir les différents façons de construire et d'envisager un corpus de référence. Après avoir choisi une méthode, la section 4.1.2 nous permettra de définir la base documentaire la mieux adaptée pour jouer le rôle de corpus de référence. Ce corpus servira ensuite de base à la réalisation d'une méthode de corrélation entre documents juridiques. Les considérations et les positions prises quant à la définition et la réalisation de ce corpus reposent en grande partie sur les travaux de Lame [Lam02], juriste qui avait défini un corpus de référence sur la même base documentaire.

4.1.1 Définition du corpus de référence

Le corpus de référence est un ensemble de textes rassemblés avec un objectif précis, celui d'être la base d'une démarche d'acquisition de connaissances. C'est sur ce corpus que seront utilisés des outils et techniques de traitement automatique de la langue, pour expérimenter différentes adaptations de notre méthode de corrélation pour aider à sélectionner la plus pertinente.

Nous parlons de corpus de référence tandis que d'autres préfèrent parler de *corpus dédié* ou *special purpose corpus* [MP01]. Certains distinguent le corpus de référence du corpus dédié [Pea98] arguant que la caractéristique principale du premier est la représentativité et que dans le corpus dédié, la composition est déterminée par un objet précis pour lequel il est élaboré. Dans le cas présent, le corpus dédié et le corpus de référence se confondent, tout comme dans les travaux menés par Lame [Lam02] sur un corpus similaire. Nous préférons parler de corpus de référence.

Pour l'élaboration d'un tel corpus de référence trois principaux cas peuvent être distingués [Lam02] :

1. Le premier cas est trivial, il aborde le cas d'un corpus spécialement élaboré à cet effet, et dont les textes n'existaient pas. Ce cas regroupe des démarches d'acquisition de connaissances à partir de textes créés par la récolte préalable des avis d'experts du domaine juridique afin de pouvoir constituer une base documentaire solide et éprouvée. De tels corpus sont éventuellement créés pour des applications particulières comme par exemple la création d'une base documentaire de la mémoire d'une entreprise. Cela implique, notamment, la prise en compte de connaissances organisationnelles sur l'entreprise. Ainsi, une acquisition de connaissances peut résulter d'un recueil des savoirs des experts sous forme textuelle grâce à des interviews. Les savoirs des experts sont en effet souvent opposés aux connaissances explicites. Les connaissances des experts sont dites tacites, dans le sens où elles sont souvent inexprimées. Ces connaissances et surtout les connaissances organisationnelles, sont ainsi difficilement communicables donc accessibles. L'enjeu des systèmes de gestion des connaissances organisationnelles est alors de faire passer ces connaissances du tacite vers l'explicite.

La constitution de corpus de référence incluant des interviews d'experts est fonction du domaine de l'organisation et de l'application visée. Notre contexte d'étude ne nécessite pas pour sa part de recueillir des connaissances tacites d'experts puisque notre application se résume à corréler entre elles des connaissances préalablement exprimées et physiquement répertoriées.

2. Le deuxième cas est rencontré lorsque les informations constitutives du corpus final sont explicites mais qu'elles ne sont pas rassemblées au sein d'une même base documentaire homogène. La création d'un corpus de référence consiste donc à rassembler des documents épars préexistants dont l'objet premier n'était pas de servir de fondement à l'élaboration d'une quelconque base documentaire. Ce genre de travail s'attache à la constitution de corpus de référence en déterminant des mots clef du domaine, en cherchant des documents correspondants sur le Web et en sélectionnant parmi eux les documents les plus représentatifs [GB01]. Un corpus de référence est alors élaboré, avec un impératif : réunir les documents représentatifs du domaine. La sélection peut se faire soit en lisant ces documents, soit en ne gardant que ceux qui présentent un nombre de termes du domaine suffisamment élevé. Malgré tout, cette étape reste le plus souvent manuelle ou au mieux supervisée. Pour adopter cette méthode, il faut bien sûr que le domaine s'y prête et que les documents existent. Le domaine juridique français est typique de ce point de vue. Le droit est un domaine présidé par un ensemble de normes. Le droit français n'étant pas un droit coutumier, les normes qui le composent, les connaissances qu'elles expriment, le sont sous la forme de textes. Ainsi, l'un des fondements du droit français est le Code civil, document créé sous Napoléon. Les sources des normes du droit français sont donc bien textuelles ; l'essence même du droit français transparaît dans des textes.

3. Dernier cas de figure, le nôtre, la définition du corpus de référence consiste à identifier un corpus préexistant, et valide, pour servir de corpus de référence. Dans notre approche les sources existantes sont celles mises à disposition sur notre portail juridique. La question se pose alors quant au choix à faire parmi les différentes sources qui le composent sans avoir à toutes les sélectionner. Il faut pouvoir trouver un corpus de référence qui couvre suffisamment l'ensemble des thématiques présentes dans les documents, tout en gardant une taille raisonnable, c'est-à-dire analysable par un logiciel de traitement automatique des langues.

Au regard des documents dont nous disposions et des expériences que nous explicitons ci-dessous, nous en sommes arrivés à envisager l'ensemble des documents publiés au Journal Officiel de la République Française, édition lois et décrets (noté JO), de l'année 2000 comme corpus de référence (voir section 4.1.2).

4.1.2 Choix du corpus de référence

Nous disposons au Centre de Recherche en Informatique de l'École des Mines de Paris d'un ensemble de documents juridiques du droit français ainsi qu'un ensemble de documents de droit communautaire. Ces documents sont ceux diffusés par les sites officiels tels que Légifrance[1] ou Europa[2]. Nous disposons ainsi de documents issus du JO, des codes du droit français et des directives et règlements européens.

Tenant à nous limiter au seul droit français pour l'instant, nous avons choisi de ne pas prendre en considération les textes européens, et de nous concentrer sur les possibilités restantes à savoir : les codes ou le JO. Pour l'élaboration de notre corpus de référence, la problématique abordée par dans cette section 4.1.2 est de pouvoir faire un choix entre ces deux possibilités.

4.1.2.1 Les codes du droit français

Les codes ont la particularité d'avoir un vocabulaire relativement ramassé sur le domaine juridique. Ainsi, la fonction première des codes est de rassembler thématiquement les normes. Ils sont actuellement créés par la Commission supérieure de codification dont l'objet est de rassembler thématiquement les normes, de les rationaliser et de les organiser. La totalité du droit n'est pas codifié : les codes ne couvrent pas l'intégralité des sources normatives du droit français même s'ils en couvrent une bonne partie. Les précédentes expérimentations faites autour de ce corpus montrent que, globalement, le vocabulaire utilisé dans les codes est condensé [Lam02]. Cela résulte de l'objet même des codes : rassembler les normes de façon rationnelle.

Concrètement et physiquement, les codes dont nous disposons sont stockés sous la forme de documents HTML, un document par article de code. Divisée en 58 codes, à la date du 4 avril 2002, cette base documentaire contient 64 184 articles pour un poids de 194 Mo et un vocabulaire rassemblant un total de 6 403 216 mots.

Les codes représentent par opposition aux documents du JO un cadre applicatif idyllique car dénué de bruit. Ainsi dans les documents du JO, il est fréquent de retrouver des informations parasites telles que des indications sur les mesures d'applications d'une norme, ou les ministres chargés de son application. Dans les codes, rien de tout cela, et c'est en grande partie la raison qui nous pousse à écarter cette base documentaire, car son vocabulaire homogène, ses textes dénués de bruit et de termes parasites l'éloignent des autres documents de notre base documentaire.

Contrairement aux travaux mené par Lame [Lam02, Lam01a, Lam01b] qui préconise

[1] `http://www.legifrance.gouv.fr/jo`
[2] `http://www.europa.eu.int`

d'utiliser les codes comme corpus de référence comme base d'acquisition terminologique et d'acquisition de connaissances, nous lui préférons les documents du JO car les codes ne sont pas assez représentatifs du reste des documents qui seront par la suite à analyser.

4.1.2.2 Le Journal Officiel de la République française

Le corpus des JO dont nous disposons est celui qui est diffusé par le site officiel Légifrance. Le Journal Officiel de la République française, édition lois et décrets, rassemble un ensemble de documents propres au droit français : les lois, les décrets, les arrêtés émanant des différents ministères du gouvernement et des avis d'autorités telles l'Autorité de Régulation des Télécoms ou le Conseil Supérieur de l'Audiovisuel. Nous disposons de l'ensemble de ces documents depuis le 1er janvier 1998, et de textes relativement importants pour les dates antérieures. Ce corpus de JO avoisine (avril 2003) 130 000 documents, chaque loi, décret, arrêté ou avis constituant un seul et unique document.

Ce corpus représente un ensemble d'environ 600 000 mots différents et 4 000 000 de SN différents. Ces chiffres élevés sont expliqués, ou du moins explicables, par le fait que de nombreux noms propres apparaissent dans ces documents, des noms de personnes dans le cas des décrets de nomination par exemple, ou des noms de localités.

La richesse de cette base documentaire peut également être un frein pour les expérimentations vu le nombre de documents à analyser (voir à titre d'exemple les expérimentations de la section 6.1.2). Raisonnablement, nous ne pouvions pas envisager de traiter l'ensemble des 130 000 documents du JO. Une sélection a donc été opérée pour ne garder que les documents de l'année 2000.

Ce corpus de l'année 2000 qui constitue désormais notre corpus de référence est composé de 24 178 documents, soit un total de 139 410 mots distincts pour 447 324 SN différents [3]. Ce sous-ensemble du JO représente 172 Mo de données au format HTML, soit 132 Mo de données textuelles.

La principale crainte à avoir lorsque l'on sélectionne un sous-ensemble de documents comme c'est le cas ici, est que l'échantillon ne soit plus représentatif de l'ensemble initial. Il est ainsi envisageable de ne plus avoir des documents de natures similaires dans les deux entités en terme de taille moyenne de document, de richesse du vocabulaire ou de distribution fréquentielle des termes. Pour tous ces paramètres, le choix entre l'un ou l'autre ensemble ne change pas les caractéristiques énoncées car nous avons pris en compte un échantillon suffisamment conséquent pour qu'en moyenne il reste comparable au corpus initial. L'unique différence que l'on peut relever se situe au niveau de la richesse du vocabulaire plus grande dans l'ensemble des JO que dans la seule année 2000. Cette différence

[3]Données fournies par l'analyseur morpho-syntaxique Sylex [Con95].

est imputable au nombre d'hapax qui est différent dans les deux ensembles. Un hapax est défini comme un terme qui n'apparait que dans un unique document ; ces termes sont légion dans le JO et sont la plupart du temps des noms propres. Cette différence reste marginale car sans grande conséquence pour la réalisation d'un outil de corrélation.

4.2 Quelle corrélation pour notre corpus

Toutes les méthodes de corrélation énoncées dans le chapitre 3 ne peuvent être utilisées sur notre corpus de référence ; elles imposent certaines restrictions quant à leur utilisation. Nous aborderons dans la section 4.2.2 ces restrictions pour finalement dévoiler la voie envisagée mais d'abord nous devons définir, dans la section 4.2.1, les attentes auxquelles doit répondre la méthode retenue.

4.2.1 La définition des besoins

Notre méthode de corrélation, en s'attachant à l'étude d'un corpus de référence particulier, doit naturellement s'adapter à ses contraintes et/ou à ses singularités. La définition des besoins inhérents à notre corpus juridique est traitée section 4.2.1.1 et section 4.2.1.2. Nous y aborderons les problématiques de mise à jour de la base documentaire et de richesse de son vocabulaire.

La section 4.2.1.3 expose un objectif que nous nous sommes fixé, celui de vouloir réaliser une corrélation entre documents et non une corrélation entre sites Web, qui est le type de corrélation le plus populaire (Voir chapitre 3).

4.2.1.1 Une base documentaire vivante

Notre méthode de corrélation s'applique à un portail juridique [4] qui regroupe de nombreux documents juridiques tels que :

- les textes du JO qui représentent 130 000 documents ;
- les codes, soit un peu moins de 65 000 articles ;
- les textes législatifs européens, soit environ 17 000 documents.

Cette base documentaire qui regroupe à ce jour [5] un peu plus de 210 000 documents n'est pas figée dans le temps et subit plusieurs contraintes, à commencer par son évolution perpétuelle signalée au chapitre 2.

[4] `http://www.admi.net`
[5] Juin 2003.

Notre base subit principalement deux types de modifications :

1. des ajouts de documents ;
2. des modifications de documents existants.

Le premier type de modification est de loin le plus fréquent, il est quotidien et voit chaque jour apparaître de nouveaux documents qui proviennent des nouveaux textes publiés au JO (environ une soixantaine). Ce mécanisme d'ajout de documents correspond également à l'apparition de nouvelles rubriques, comme par exemple ce fut le cas avec la mise à disposition des textes européens ou encore la future mise à disposition des textes de jurisprudence.

Le second type, qui est plus rare concerne la modification de documents existants. L'exemple le plus parlant étant celui des textes de lois consolidées dont l'essence même est d'évoluer en permanence.

Que l'on parle d'ajout de nouveaux documents ou de modifications de documents existants, la conséquence est que la voie de corrélation envisagée doit satisfaire en premier lieu à deux contraintes : savoir gérer une grande base documentaire et pouvoir prendre en compte les modifications apportées aux données existantes.

4.2.1.2 Utiliser la richesse du corpus

Notre corpus de référence, comme il a pu être défini au travers de la section 4.1, est un corpus spécifique dont la thématique centrale est le domaine juridique. Comme tout corpus spécifique, celui-ci permet de travailler sur une base documentaire maîtrisée et homogène en terme de thématique ce qui permet de réduire les ambiguïtés sémantiques car un terme fréquemment utilisé dans un même contexte admet généralement une signification unique [MC91].

“ Le langage du droit est un langage de spécialité au sein de la langue commune. Simplement, son entendement nécessite plus que la maîtrise de la langue commune. Ainsi, la communication du droit se heurte à ce que l'auteur appelle un *écran linguistique* ”. C'est en ces termes que Cornu [Cor00] exprime le fait que les textes juridiques sont des textes qui s'adressent à des spécialistes du domaine mais que néanmoins le vocabulaire utilisé est faussement porteur d'une double signification : celle adressée aux experts et celle que les non-initiés ont cru percevoir.

Ainsi certains termes de la langue française n'ont de sens que dans un contexte juridique à l'image de *antichrèse*, *nantissement* ou *synallagmatique*. La plupart sont cependant polysémiques : le même terme ayant un sens commun, voir plusieurs, compréhensible par tout un chacun et un autre, voire plusieurs, dans le contexte juridique. Le vocabulaire juridique français est ainsi défini comme *l'ensemble des termes de la langue française qui reçoivent du droit un ou plusieurs sens.* Selon Cornu [Cor00], ce vocabulaire serait composé d'environ 10.000 termes. Ces termes, qui représentent dès lors les termes du domaine juridique, sont autant de pistes à exploiter pour rapprocher les termes d'un point de vue sémantique car au sein même de notre corpus d'étude, leur emploi est précis et surtout non-ambigu.

Nous souhaitons réaliser une méthode de corrélation qui tire profit de la richesse du vocabulaire jurique et de ses particularités pour faciliter la recherche de résultats corrélés ou améliorer leur qualité. Ainsi, à titre d'exemple, des documents comportant des termes spécifiques à la matière juridique tels que *licéité*, *stellionataire*, *colon partiaire*, *répétition de l'indu* ou *tontine* sont autant de pointeurs discriminants à utiliser pour comparer des textes potentiellement corrélés.

4.2.1.3 Le document comme unité de base pour la corrélation

La finalité d'une méthode de corrélation entre documents est de faciliter l'accès à l'information vis à vis des internautes et de lui amener des informations complémentaires sur un document existant. Pour apporter ce complément d'information il n'existe pas de voie prédéterminée et l'on se doit alors de s'interroger sur les différents types de corrélations envisageables.

Il convient d'abord de décider si l'on apporte à l'utilisateur un complément d'information local, c'est à dire en provenance d'une base documentaire bien délimitée, ou au contraire si celle-ci doit être globale, c'est à dire que l'information complémentaire est située quelque part sur Internet. Entre approche locale et approche globale, nous avons déjà pris position section 4.1. La base documentaire visée est représentée par l'ensemble des documents juridiques publiés sur notre portail du droit. La corrélation fournira aux utilisateurs un complément d'information, composé de documents corrélés, recherchés dans les documents de cette même base.

On peut ensuite distinguer dans la corrélation trois grands axes que sont :

- la corrélation entre documents ;
- la corrélation documents-sites ;
- la corrélation documents-communautés.

La principale différence concerne la portée des résultats corrélés. Le point de départ est un document et l'on tente de lui associer des résultats corrélés, mais pour le premier cas de figure ces résultats sont des éléments de même nature, c'est à dire des documents. Dans le second cas, on a associé à un document un ensemble de sites corrélés (voir section 3.1.4) pour permettre à l'utilisateur d'aller, non pas directement consulter des documents, mais visiter des sites dont la thématique est proche de son document initial, et donc susceptibles de lui fournir des informations complémentaires. Enfin, le troisième cas se trouve être proche du second et retourne, cette fois-ci, un ensemble de communautés de sites Web. À une communauté donnée correspond une thématique en relation avec le document initial qui rassemble plusieurs sites sémantiquement proches [TK01], l'idée étant de classer les sites préalablement au travers d'une thématique pour faciliter la navigation.

4.2.2 Vers une voie de corrélation linguistique

Dans la partie précédente (Section 4.2.1), nous avons vu les impératifs auxquels devaient répondre notre méthode de corrélation. Ceux-ci peuvent être résumés en quatre points.

La méthode envisagée doit :

1. pouvoir être appliquée sur un large corpus ;
2. tenir compte des mises à jour de la base documentaire ;
3. exploiter les particularités linguistiques d'un corpus juridique ;
4. permettre de corréler les documents entre eux.

Nous allons voir maintenant comment les quatre grandes voies de corrélation étudiées au chapitre 3 s'adaptent à ces exigences.

En ce qui concerne la gestion d'un large corpus, toutes les approches, à l'exception des voies hiérarchiques, permettent de traiter un grand nombre de documents. Toutefois les méthodes empiriques qui, certes, n'ont pas de réelles limitations quant aux nombres de documents analysés, ne fonctionnent que si la majeure partie des documents ont été consultés pour obtenir des statistiques en nombre suffisant pour exploiter le parcours des internautes.

Au niveau de la mise à jour des informations et de la prise en compte des modifications subies par le corpus de référence, les choses se compliquent pour les méthodes topologiques et hiérarchiques qui font le plus souvent appel à des calculs matriciels sur un grand nombre de données. La modification d'un document ou son apparition dans le corpus nécessite alors de recalculer les matrices, une fois les données mises à jour, opération qui reste coûteuse en temps. L'approche empirique, pour schématiser, ne nécessite que de " compter " le nombre d'internautes allant d'une page A vers une page B ; tout nouveau document ou toute modification est alors facilement prise en compte. Dans le cadre d'une corrélation linguistique, le problème de la mise à jour disparait car la génération des signatures lexicales nécessaires à la recherche des documents corrélés peut être opérée dynamiquement. Les modifications ou les ajouts sont alors pris en compte presque instantanément.

Concernant l'exploitation des ressources linguistiques présentes dans notre corpus de référence, seules les approches hiérarchiques et linguistiques travaillent concrètement sur des termes extraits. Elles apparaissent donc comme les seules méthodes qui répondent à nos critères de sélection. Pour les méthodes topologiques qui se concentrent a priori sur l'utilisation d'hyperliens, on peut élargir la notion d'hyperliens pour y inclure les termes du domaines juridique. A cette condition, une approche topologique exploite les données spécifiques du corpus ; néanmoins ce cadre d'étude n'ayant pas été expérimenté, nous considérons qu'à ce jour les méthodes topologiques n'exploitent pas les spécificités du vocabulaire.

Dernière étape, vérifier que la voie choisie permet de corréler des documents entre eux. Au chapitre 3 et dans la section 4.2.1.3, nous avons vu de nombreuses méthodes qui font correspondre à un document des résultats corrélés de nature différente ; c'est le cas des méthodes topologique ou empirique. Dans ces conditions, seules les voies hiérarchiques et linguistiques sont conformes à nos attentes.

TAB. 4.1 – *Comparaisons entre les différentes voies de corrélation*

	Large Corpus	Mises à jour	Exploiter le corpus	Corrélation de documents
Voie Topologique	Oui	Non	Non	Non
Voie Empirique	Oui	Oui	Non	Non
Voie Hiérarchique	Non	Non	Oui	Oui
Voie Linguistique	Oui	Oui	Oui	Oui

Les résultats de cette investigation sont synthétisés dans le tableau 4.1 qui résume les avantages et inconvénients de chaque voie. On y voit apparaître que l'approche de la corrélation linguistique semble la plus appropriée à la recherche de documents corrélés sur notre corpus de référence.

4.3 Définition de la méthode utilisée

Pour développer une méthode de corrélation nous retenons la voie linguistique. Au sein même de cette voie linguistique, il existe plusieurs orientations possibles. La finalité de cette section 4.3 étant de pouvoir définir avec précision les temps forts de la méthode retenue.

Nous avons vu au chapitre 3, lors de l'état de l'art des méthodes de corrélations linguistiques, qu'une telle approche se scinde en deux phases principales symbolisées par :

- l'élaboration des signatures lexicales ;
- la recherche des résultats corrélés.

Tout naturellement, le plan de cette section 4.3 va dans une première partie (Section 4.3.1) aborder la problématique de la construction des signatures lexicales avant de se pencher dans une deuxième partie (Section 4.3.2) sur le problème de la recherche des résultats corrélés.

4.3.1 L'élaboration des signatures lexicales

Les signatures lexicales sont vues comme de brèves descriptions textuelles faites à partir d'un document initial. Cette description est représentée par un ensemble non ordonné de termes dont le nombre et la nature (mot, SN, etc.) sont prédéfinies. Leur finalité est d'être exploitées pour interroger un moteur de recherche dont les premières réponses sont assimilées à des documents corrélés. Cette approche de la corrélation propre à Park et al. [PPGK02] est celle que nous abordons et souhaitons étendre.

La principale difficulté est de pouvoir composer des signatures lexicales, c'est à dire pouvoir fabriquer des listes de termes discriminants propres à chaque document du corpus. Comme évoqué au chapitre 3, la voie privilégié pour retrouver les termes discriminants est de les pondérer pour ne garder *in fine* que les plus intéressants, mais pour cela nous devons préalablement opérer un choix. En effet, en terme de pondérations, deux courants s'affrontent : l'un linguistique et l'autre statistique. Le choix que nous avons retenu est d'utiliser des méthodes de pondération statistiques pour élaborer des signatures lexicales, et ce pour plusieurs raisons.

Les méthodes statistiques restent dans leur ensemble plus simple à utiliser que des méthodes linguistiques qui supposent l'emploi d'analyses syntaxiques préalables et/ou l'étiquetage de certains marqueurs spécifiques du langage. Même si ce travail préalable peut grandement améliorer la pertinence des analyses et le choix des termes clefs, il reste en grande partie manuel et nécessite des experts. Cette voie a été écartée, et nous nous replions vers des méthodes statistiques qui ont comme principal avantage de pouvoir traiter rapidement les grands flots d'information qui caractérisent notre corpus de référence.

4.3.2 La recherche des résultats corrélés

La recherche des résultats corrélés peut être réalisée de plusieurs façons (Section 3.4.2). En s'orientant vers une voie de corrélation linguistique, le terme, au travers des signatures lexicales, est au cœur de toutes nos préoccupations. Ce terme ou ces termes, plus exactement, iront, dans notre approche, jusqu'à servir de requêtes pour interroger un moteur de recherche et en extraire les résultats corrélés. Or ces termes occupent des rôles bien spécifiques dans notre corpus de référence.

Le domaine juridique est un domaine bien délimité dans lequel les termes peuvent avoir un sens commun mais ont surtout un sens propre à la matière. Ces richesses propres à notre corpus ont été étudiées pour montrer qu'il est possible à partir de termes du domaine d'aider les utilisateurs dans leurs recherches [Lam00, Lam01a, Lam01b]. Concrètement, ils ont été utilisés comme point de départ à la création de requêtes qui, une fois traitées par un moteur de recherche, ont amélioré l'accès à l'information dans notre base documentaire [Lam02] (voir la figure 4.1, illustrant des techniques d'expansion de requêtes développées sur notre corpus d'étude). Fort de cette constatation, nous préconisons également d'utiliser les signatures lexicales générées comme des requêtes directement utilisables par un moteur de recherche pour retrouver des résultats corrélés.

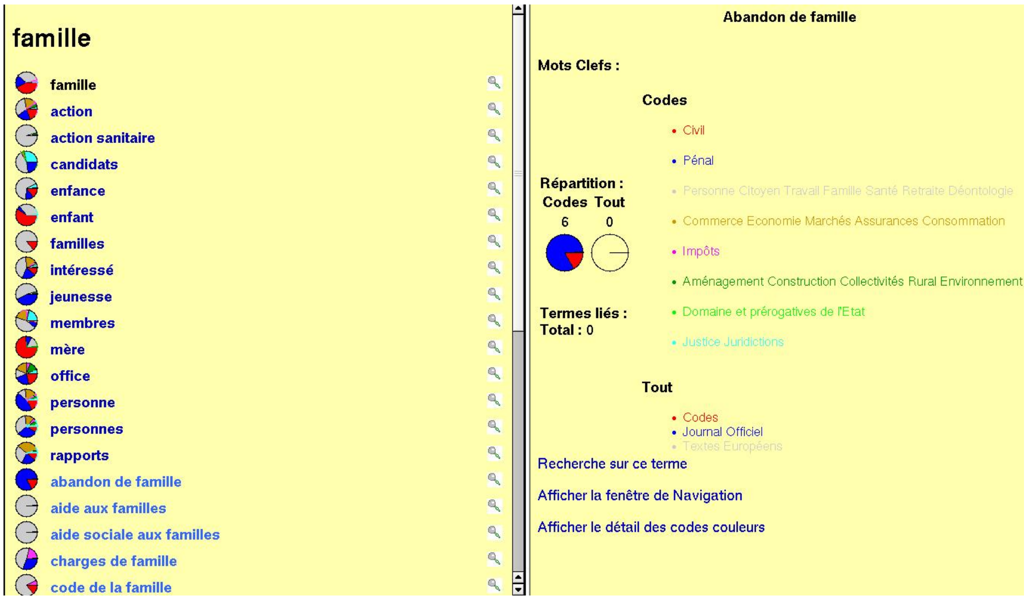


FIG. 4.1 – *Illustration des travaux de Lame : typologie de termes juridiques [Lam02].*

Dans notre travail, contrairement à Lame [Lam02], les signatures lexicales ne sont pas systématiquement constituées de termes du domaine mais sont un savant mélange de termes ordinaires et de termes juridiques. Néanmoins, nous estimons que ceci n'est pas un handicap majeur car l'ajout de termes non-discriminants du point de vue juridique peut d'un point de vue de la corrélation enrichir les résultats sans toutefois générer trop de bruit. Le bruit est par nature inévitable et doit être minimisé par le classement des résultats corrélés, opéré a posteriori.

4.4 Les degrés de liberté

Après avoir défini nos choix et nos orientations en termes de corrélation, nous allons définir plus finement les choix et les étapes à mettre en oeuvre pour réaliser notre méthode de corrélation. Pour illustrer nos choix nous pouvons nous référer à la figure 4.2 qui rassemble ces étapes clefs.

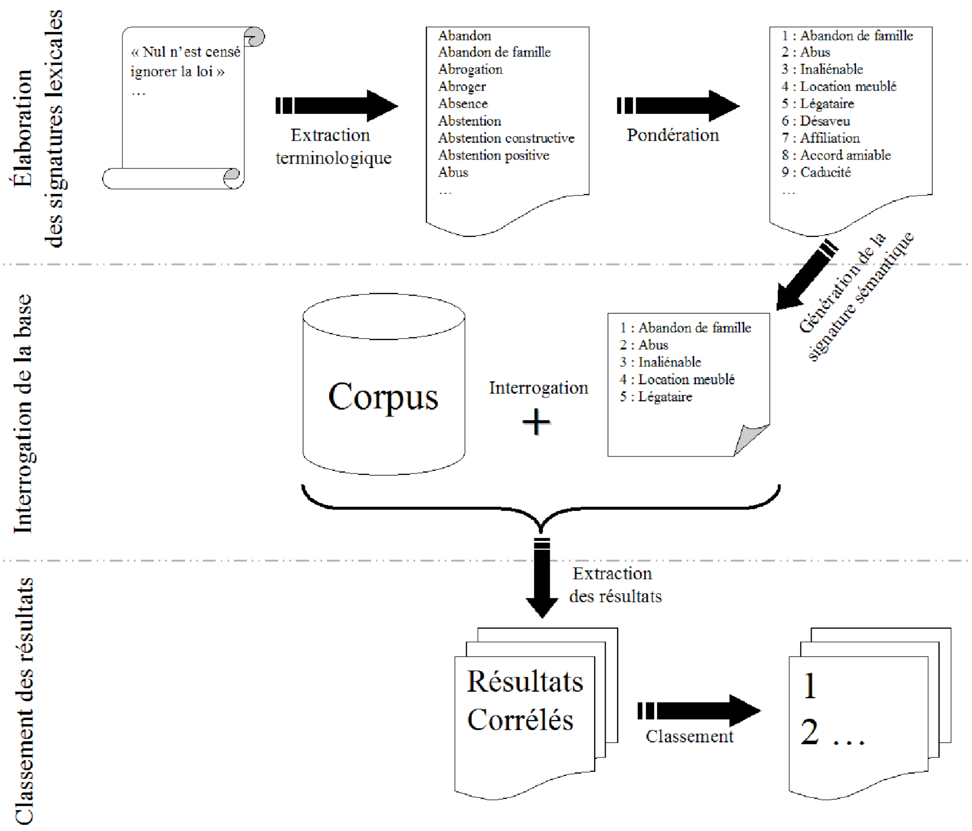


FIG. 4.2 – *Les différentes étapes de la corrélation*

Les différentes étapes s'articulent autour de trois temps forts :

- l'élaboration des signatures lexicales ;
- l'interrogation de la base documentaire ;
- la présentation et le classement des résultats corrélés.

Ces étapes laissent entrevoir un très grand nombre d'inconnues quant à la réalisation d'une méthode de corrélation. En effet, beaucoup de questions restent pour l'instant en suspens comme, notamment, les types de termes (mots, SN, etc.) qui doivent être utilisés pour élaborer les signatures lexicales, ou bien encore les méthodes de pondération que l'on doit considérer pour extraire les termes les plus discriminants. Autant de questions schématisées sur la figure 4.3 et qui représentent, en quelle sorte, les degrés de liberté que nous offre cette méthode de corrélation.

La finalité de cette section 4.4 est de décrire l'ensemble des degrés de liberté laissés vacants par notre approche de la corrélation. Les sections qui suivent aborderont les problématiques liées et, à défaut de toutes les résoudre, renverront vers d'autres chapitres ou sections détaillant plus finement le problème. La motivation première de la présente section 4.4 est dans un premier temps de susciter l'interrogation et la réflexion.

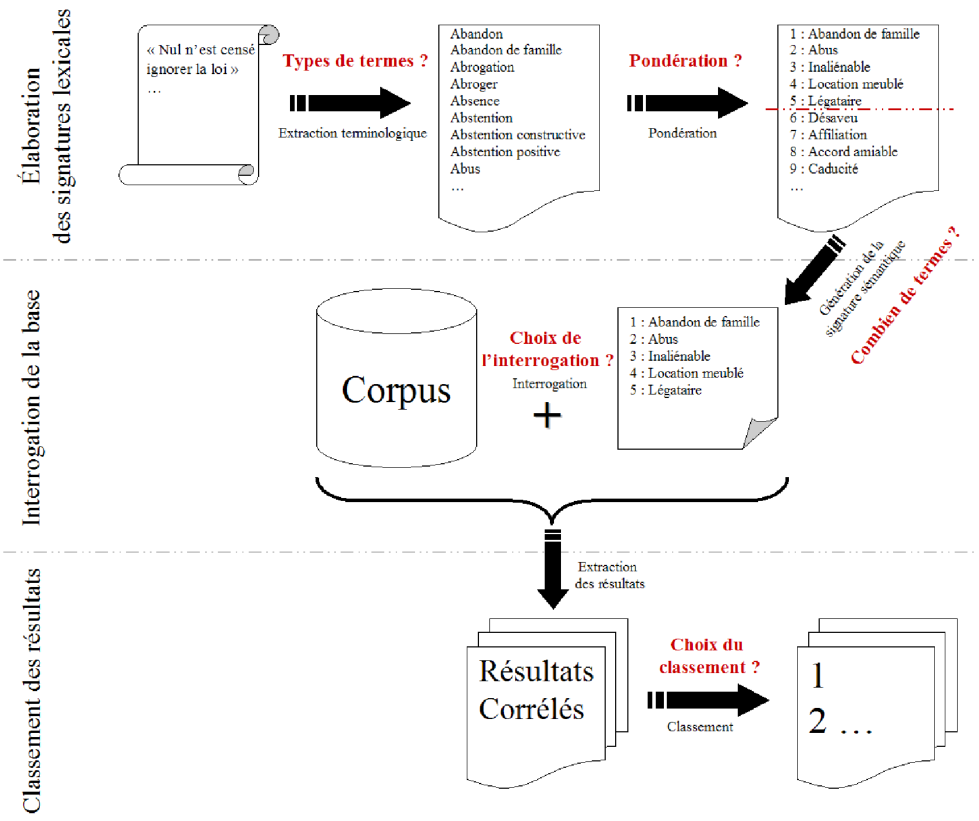


FIG. 4.3 – *Les différents degrés de liberté*

4.4.1 L'élaboration des signatures lexicales

La première phase d'une méthode de corrélation linguistique passe par la recherche des termes qui vont composer les signatures lexicales. La question est de savoir : de quels termes parle-t-on au juste ?

Les termes peuvent être de natures différentes comme : des mots, des mots composés, des syntagmes nominaux, des expressions numériques ou encore alphanumériques. La première interrogation concerne donc le choix de la nature des éléments constitutifs des signatures lexicales. Ensuite, une fois ce choix opéré, faut-il encore savoir si les signatures envisagées doivent être constituées de termes de nature homogène ou hétérogène, question qui sera abordée au cours du chapitre 5.

Les signatures ne sont générées qu'après avoir choisi un panel de termes considérés comme discriminant pour le document. Nous avons envisagé de faire ce choix grâce à l'utilisation de méthodes de pondération statistiques (Section 4.3.1) sans pour autant définir concrètement celle qui sera utilisée, plusieurs ont été évoquées (Section 3.4.1) mais leur choix dépend de nombreux paramètres comme notamment celui du type de termes employé ; certaines permettent de pondérer toutes sortes de termes alors que d'autres sont par exemple réservées à la pondération des SN [CH90, GC91, Dun93].

Le dernier degré de liberté propre aux signatures lexicales concerne le choix des termes qui participent finalement à l'élaboration de la signature. La figure 4.3 montre que la pondération permet de " classer par pertinence décroissante " les différents termes qui composent un document mais le choix des termes n'est pas encore effectué. Doit-on considérer qu'il faut choisir les termes en fonction d'un seuil relatif à une méthode de pondération ou bien encore en prendre un nombre fixe ? Là encore le chapitre 5 permettra de répondre à cette question par l'expérimentation.

4.4.2 Interrogation de la base

L'interrogation de la base est, au même titre que la création des signatures, une étape primordiale pour la recherche des documents potentiellement corrélés car le type de recherche à opérer influe sur la qualité des résultats. Une question floue va générer trop de bruit alors qu'au contraire une question trop restrictive va perdre en rappel (voir définition section 7.2.2). Le choix du mode d'interrogation doit apporter un juste équilibre entre le bruit généré et la précision des documents (voir définition section 7.2.2).

Pour interroger la base on peut distinguer deux degrés de liberté supplémentaires : la formulation de la requête ainsi que l'unité de recherche choisie.

4.4.2.1 La formulation des requêtes

La formulation de la requête correspond au mode de recherche envisagé ; une requête peut être [Lef00] :

- en langage naturel ;
- booléenne ;
- conceptuelle.

La première catégorie ne nous est pas d'un grand secours car des requêtes en langage naturel se présentent sous la forme de questions telle que " Dans quel(s) document(s) fait-on référence à la loi Aubry ? ", qui ne peuvent pas être obtenues à partir des descriptions synthétiques qu'expriment les signatures lexicales.

Nous devons donc choisir entre requêtes booléennes et requêtes conceptuelles. Les premières sont représentées par une suite de mots orchestrés autour d'opérateur booléens qui définissent les règles d'acceptation des documents alors que les secondes sont constituées d'une liste de mots et d'un seuil qui détermine le nombre minimum de mots qui doivent être présents pour que celui-ci soit accepté.

Dans notre cas, l'utilisation des requêtes booléennes est difficilement envisageable car elles nécessitent indirectement de définir des règles d'acceptation telle que " (Terme1 **ET** Terme2) **SAUF** Terme3", règles difficilement compatibles avec le contenu épuré des signatures lexicales. Notre choix en matière d'interrogation se porte tout naturellement vers les requêtes conceptuelles qui permettent de définir une liste de termes ainsi qu'un seuil d'acceptation. Là encore ce seuil sera étudié en détail dans les chapitres 6 et 7.

4.4.2.2 L'unité de recherche

Traditionnellement les outils de recherche de type moteur de recherche ne laissent que peu de liberté quant à la définition d'une unité de recherche. Nous définissons par unité de recherche *le segment de texte (nombre de mots, phrases, paragraphes, etc) à prendre en considération pour rechercher les éléments d'une même requête.*

L'unité de recherche qui est couramment utilisée est le document quelle que soit sa taille ; cela implique que pour qu'un document soit valide, il faut que les éléments de la requête soient retrouvés au sein de ce document. Si on change d'unité de recherche et si on prend la phrase, la recherche est toute autre ; cette fois-ci pour qu'un document soit accepté il faut qu'au moins une des phrases qui le compose valide la requête précédemment définie. L'unité de recherche joue le rôle d'un curseur qui permet de jouer sur le nombre de résultats retrouvés ainsi que leur valeur sémantique.

Entre le document lui-même, la phrase ou le groupe nominal, toutes les tailles pour déterminer l'unité de recherche sont envisageables. Le chapitre 6 abordera cette question.

4.4.3 Le classement des résultats

Dernier élément constitutif de la méthode de pondération, le classement des résultats. Sous la dénomination anglaise " ranking ", il désigne le plus souvent l'algorithme qui permet de classer les résultats en sortie d'un outil comme un moteur de recherche. Son rôle est, en quelque sorte, de masquer les résultats médiocres en les enterrant dans les profondeurs du classement.

En effet, nul ne prétend pouvoir, à partir d'une simple requête et des considérations statistiques, obtenir des résultats parfaits, sans bruit, ni silence. Nous ne dérogeons pas à cette règle. Le classement des résultats permet en théorie de rejeter dans les profondeurs du classement les résultats insuffisants voir médiocres. Sa détermination est le plus souvent empirique et ne répond pas à de véritables justifications mais elle s'appuie principalement sur le type d'interrogation utilisé pour rechercher les résultats pertinents. À cet effet, le chapitre 7 au travers de la section 7.3 présente une expérimentation qui étudie les effets d'un algorithme de classement des résultats dédié à la corrélation.

Chapitre 5

Des signatures lexicales en quête de sens

Notre vision linguistique de la corrélation passe par la création de signatures lexicales, étape qui nous laisse de nombreux degrés de liberté en matière de génération ou d'utilisation des signatures lexicales. L'objet de ce chapitre est double ; il lève certaines incertitudes concernant la création des signatures et leur composition, et il met en avant l'aspect sémantique des signatures lexicales générées.

En effet, la principale difficulté éprouvée lors de la génération de signatures lexicales est de trouver des groupes de termes sémantiquement proches. La clef de voûte de l'édifice est d'arriver à agglomérer des termes via des considérations statistiques tout en espérant qu'ils auront un rapport les uns avec les autres. Cet objectif va être étudié dans ce chapitre 5 et c'est pourquoi nous comparerons des signatures lexicales générées manuellement avec des signatures générées automatiquement via différentes méthodes de pondération.

L'articulation du présent chapitre est la suivante : la section 5.1 est consacrée aux choix des types de termes que nous allons privilégier pour notre méthode de corrélation. La section 5.2 apporte un éclairage nouveau sur les méthodes classiques de pondération qui sont utilisées pour la génération des signatures. La section 5.3 met en avant les insuffisances de ces méthodes de pondération, nous incitant à élaborer un indice de pondération expérimental. Les sections 5.4 et 5.5 détaillent une expérience qui vise à confronter les signatures générées manuellement avec celles obtenues automatiquement. Finalement, la section 5.6 compare les résultats obtenus et fait le point sur les enseignements recueillis.

5.1 Sélectionner les candidats termes

Pour retrouver les termes qui participent à l'élaboration des signatures lexicales, on passe par une étape préliminaire de sélection terminologique qui consiste à ne prendre en compte que des termes potentiellement intéressants. Parmi les termes, nous effectuons une distinction : il y a tout d'abord des termes qualifiés de *candidats termes* qui représentent, pour un document donné, l'ensemble des termes susceptibles de participer à sa signature lexicale. Ensuite nous distinguons les *termes valides* qui, pour un document donné, constitueront sa signature après pondération (Section 5.1.2). L'objet de la présente section est de comprendre comment sont choisis les *candidats termes* parmi les différentes catégories ou types de termes existants (Section 5.1.1) et de voir comment ils sont extraits (Section 5.1.2).

5.1.1 Définir les catégories de candidats termes

La recherche des candidats termes est une étape de pré-sélection où l'on choisit parmi les catégories de termes celles qui pourront participer à l'élaboration des signatures lexicales. Pour les documents présents dans notre corpus de référence, nous distinguons quatre grandes catégories de termes :

- des expressions chiffrées ;
- des mots composés ;
- des syntagmes ;
- des mots simples.

5.1.1.1 Les expressions chiffrées

Les expressions chiffrées constituent une catégorie à part entière car elles ne suivent aucune règle simple et revêtent différentes fonctions (pourcentage, formule chimique, etc.). Dans notre corpus, ces expressions revêtent deux formes principales : elles apparaissent préférentiellement dans des formules chimiques, ou correspondent à des références internes entre articles ou codes.

Leur interprétation dans notre corpus de référence montre qu'il s'agit souvent d'hapax, qu'elles peuvent aisément être ignorées car elles ne sont pas caractéristiques des termes juridiques [Lam02] et n'apportent pas d'indices supplémentaires pour faciliter la corrélation entre documents.

5.1.1.2 Les mots composés

Le principal problème lors de la recherche de termes pertinents provient de la polysémie des unitermes ; à cause de cette polysémie les unités de sens sont rarement réductibles à des mots simples [Lef00]. De cette constatation découle l'engouement de certains travaux de recherches pour l'étude d'unités syntaxiques plus complexe comme les mots composés. Grâce aux investigations menées, on peut, à ce jour, dénombrer un peu plus de 100.000 noms composés dans la langue français, mots composés extraits à partir d'articles du journal *Le Monde* [Sil93].

Mots composés et syntagmes ne diffèrent pas d'un point de vue syntaxique, ils correspondent tous deux à des regroupements de mots, dont les premiers sont un sous-ensemble des seconds. Néanmoins, les mots composés forment des entités sémantiques plus cohérentes car n'est pas mot composé tout syntagme qui le désire ; pour cela il doit posséder les trois propriétés suivantes [Lef00] :

- posséder une atomicité sémantique ;
- être institutionnalisé dans le langage ;
- avoir des composantes sémantiquement inséparables.

Grâce à cette définition, les mots composés sont sémantiquement plus précis que leurs alter ego les syntagmes ; néanmoins ces mots composés ne comblent pas toutes nos attentes. Leur faible nombre apparait comme une première limitation. De plus dans un contexte juridique ; nous aurions aimé disposer de termes du domaine et donc de mots composés juridiques, or ceux-ci ne sont pas nombreux. Pour étoffer leur nombre et en identifier davantage, le travail doit être effectué manuellement. Pour ces raisons, l'emploi de mots composés est hypothéqué et nous finissons par leur préférer les syntagmes, certes moins précis mais d'une utilisation plus aisée.

5.1.1.3 Les syntagmes

Les syntagmes sont des séquences de mots qui peuvent être classés suivant plusieures catégories, le tableau 5.1 illustre certaines d'entre elles à l'aide d'exemples pris dans notre corpus. La catégorie prédominante est celle des syntagmes nominaux, unité composée, comme son nom l'indique, exclusivement de noms. Parmi les différents types de syntagmes possibles, nous ne nous intéresserons qu'aux syntagmes nominaux car il s'agit de la sous-catégorie qui, d'une manière générale, est la plus porteuse de sens et la mieux adaptée à notre corpus d'étude [Cho02, Lam01a, Lam01b, Lam02].

Type	Exemple
Syntagme adjectival	*accessible aux personnes handicapées*
Syntagme nominal	*accueil thérapeutique*
Syntagme participial	*notifié à l'intéressé*
Syntagme verbal	*arriver à échéance*

TAB. 5.1 – *Différentes catégories de syntagmes*

5.1.1.4 Les mots simples

Dernière catégorie envisagée : celle des mots simples, qui correspondent aux éléments unitaires d'un texte à savoir les mots et que là encore, comme le montre le tableau 5.2, nous pouvons scinder en plusieurs sous-catégories.

Type	Exemple
Adjectif	*budgétaire*
Adverbe	*éventuellement*
Nom propre	*Hauts-de-Seine*
Nom	*avis*
Participe passé	*élaboré*
Verbe	*rétracter*

TAB. 5.2 – *Les différents catégories d'unitermes*

Notre choix, quant à la sélection des différentes catégories à utiliser est paradoxalement de ne pas en faire. Au premier abord, certaines catégories telles que les adverbes, les verbes ou encore les participes passés peuvent apparaître comme inutiles. Cependant nous ne souhaitons pas opérer de choix car la sémantique d'une phrase n'est pas uniquement portée par les noms qui la composent, il serait donc préjudiciable d'éliminer arbitrairement certains types de termes. D'autre part, la sélection des candidats termes ne constitue qu'une étape de filtrage préalable, il n'est donc pas utile d'effectuer un tri à cet instant, l'étape de pondération et de sélection des termes valides étant prévue à cet effet.

5.1.2 Extraire les candidats termes

Nous avons indiqué que notre choix de termes s'orientait vers deux types particuliers, les mots simples sans distinction entre les différentes catégories syntaxiques et les syntagmes nominaux. Si l'extraction des mots simples est une opération aisée, il n'en n'est pas de même pour la seconde. Pour cette étape de recherche des SN nous avons eu recours à un extracteur terminologique.

Divers outils permettent d'identifier les catégories syntaxiques des syntagmes qui leur sont soumis. Ils opèrent leur identification en utilisant deux types de méthodes appliquées séparément ou alternativement : les méthodes statistiques et les méthodes syntaxiques.

5.1.2.1 Méthodes d'extraction de termes

Les méthodes statistiques identifient les syntagmes en repérant des suites de mots qui apparaissent suffisamment souvent dans les textes et qui possèdent une cohérence lexicale. La détermination des SN nécessite des procédures itératives gourmandes en temps et en ressources. De plus, de telles approches ne sont pas exploitables sur tous les types de corpus, car ceux-ci doivent posséder certaines caractéristiques quelque peu restrictives en termes de fréquences d'apparition des mots, de nombre de répétition, etc. Ces restrictions sont contraires aux caractéristiques propres de notre corpus de référence.

La deuxième catégorie d'outils repose sur une analyse syntaxique préalable pour identifier les syntagmes dans les textes étudiés. Cette approche passe par une phase préalable d'étiquetage des mots sur lesquels on applique des règles de grammaire pour finalement identifier les syntagmes. Les règles sont souvent utilisées à plusieurs reprises et permettent ainsi de lever progressivement les ambiguïtés pour faire émerger les différents types de syntagmes : syntagme nominal, syntagme adverbial, etc.

5.1.2.2 Le choix de l'outil Sylex

Le logiciel Sylex [Con91, Con95] est présenté comme un analyseur linguistique qui cherche à *comprendre et analyser un texte quelconque.* L'outil effectue une analyse syntaxique partielle des textes et s'attelle à la sémantique en déplaçant son étude vers un *axe paradigmatique*, étude qui s'enrichit des analyses syntaxiques précédentes.

L'outil Sylex incorpore des règles de grammaire de la langue française et exploite également la théorie de Tesnière pour minimiser la création de concepts linguistiques, ce qui réduit d'autant le nombre d'arbitraires potentiels. Ainsi, toute ambiguïté lexicale (par exemple, des mots qui sont à la fois des noms et des verbes) est potentiellement résolue par le choix arbitraire d'une des interprétations possibles, l'avancement de l'analyse permettant de revenir sur ces choix arbitraires. L'analyse syntaxique du texte se fait donc par *couches*, la structuration syntaxique du texte évoluant au fur et à mesure du passage des différents modules grammaticaux. Grâce à ces différentes analyses successives nous obtenons en sortie des syntagmes étiquetés répondant à nos besoin. Concrètement, pour rechercher les termes, nous utilisons une version révisée de Sylex qui se nomme Genet et qui est incorporée au moteur de recherche Pertimm [1].

[1] Développée par la société Pertimm [Per].

5.1.3 Filtrage des candidats termes

Avant de passer à l'étape suivante qui consiste à rechercher les termes valides qui constitueront au final les signatures lexicales, nous devons nous pencher une dernière fois sur la recherche et le filtrage des candidats termes.

Nous avons vu qu'ils pouvaient être de plusieurs types et que parmi les différents choix possibles, nous en avions gardé deux : les mots simples et les SN. Néanmoins tous ces termes ne doivent pas être pris en compte, non pas à cause de leur nature syntaxique mais à cause de certaines de leurs caractéristiques fréquentielles. Nous souhaitons éliminer les termes trop fréquents ou les hapax.

La première catégorie couramment appelé *mots-stop*, *mots-vides* regroupe des termes très fréquents dans notre corpus et qui, en termes de corrélation et de génération de signatures lexicales, sont porteur de bruit. Contrairement à d'autres travaux où l'on sélectionne manuellement les termes participant à cette liste, nous préférons définir un seuil de rejet au-delà duquel les termes ne sont plus pris en compte. Ce seuil est fonction du taux d'apparition du terme dans le corpus. Il a été arbitrairement fixé à 10% ce qui dans notre cas représente déjà plus de 2.000 documents et signifie qu'un terme présent dans plus de 10% des documents du corpus sera banni.

La deuxième catégorie de terme visée par ce rejet préventif est celle des hapax. Ce choix se justifie de lui-même par la finalité de la création des signatures lexicales : rapprocher des documents en fonction de leur vocabulaire commun. Or un hapax n'est, par définition, présent que dans un unique document, il n'apporte donc aucune information utile.

5.2 Identifier les termes valides

La sélection des termes saillants parmi un ensemble de candidats termes a déjà été évoquée section 4.3.1. Pour opérer ce choix, la voie des méthodes de pondération statistiques a été choisie sans toutefois définir quelles techniques de pondérations allaient être utilisées.

Les approches statistiques proposent de multiples techniques de pondérations pour quantifier la valeur informationnelle portée par un terme. On peut évoquer des pondérations telles que le concept d'information mutuelle [CH90], le coefficient Φ^2 [GC91] ou encore le coefficient Loglike [Dun93], des indices utilisés par Daille [Dai96] pour étudier les mesures d'associations entre mots au sein de SN.

L'extraction des termes valides a été également abordée par P. Lafon [Laf80] qui propose d'appliquer une distribution hypergéométrique à la répartition des termes du corpus. Toutefois, cette approche impose de nombreuses restrictions [LL00]. Ainsi, il faut, à titre d'exemple, étudier un corpus découpé en parties de tailles sensiblement égales ou encore comparer des mots de même nature (noms, verbes, pronoms,etc.). Toutes ces restrictions nous empêchent d'utiliser cette approche car les conditions initiales ne peuvent être respectées de par la nature même de notre corpus de référence.

Pour retrouver les termes valides parmi les candidats termes, nous préférons dans un premier temps utiliser des méthodes classiques de pondération qui satisfont aux caractéristiques de notre corpus. Notre démarche est d'utiliser quatre méthodes statistiques, exposées ci-après, pour extraire les termes valides et comparer leurs résultats avec une méthode de pondération expérimentale abordée dans la section 5.3.

5.2.1 Indice T_f

Le premier indice de pondération que nous allons étudier est la fréquence d'un terme $\mathcal{T}_j$ relatif à un document quelconque D_i pris dans le corpus de référence noté Γ. L'idée sous-jacente est de supposer que l'importance d'un terme va de paire avec son nombre d'occurrences dans D_i par rapport à son utilisation dans le reste de Γ [Luh57]. L'indice noté T_f mesure cette fréquence :

$$T_f(D_i, \mathcal{T}_j) = \frac{N(D_i, \mathcal{T}_j)}{N(\mathcal{T}_j)} \tag{5.1}$$

avec $N(D_i, \mathcal{T}_j)$ le nombre d'occurrence de $\mathcal{T}_j$ dans D_i et $N(\mathcal{T}_j)$ le nombre d'occurrences de $\mathcal{T}_j$ dans Γ.

Pour chaque pondération que nous allons étudier par la suite, T_f inclus, nous allons nous pencher sur leurs caractéristiques en étudiant successivement un même exemple pour permettre de mieux comparer les indices.

Par exemple, considérons un terme quelconque apparaissant 30 fois dans Γ. Ce terme peut être réparti de multiples façons : 30 fois dans un même document, 1 fois dans 30 documents et bien entendu il reste toutes les distributions intermédiaires. L'intérêt est de pouvoir envisager toutes les valeurs prises par T_f pour les différentes répartitions possibles d'un terme T_j présent 30 fois dans le corpus.

Envisager toutes les répartitions possibles de T_j dans Γ revient à rechercher toutes les décompositions d'un entier positif, en somme de nombres entiers strictement positifs. Cette notion, nommée le partionnement simple d'un entier, a été traitée par Euler et permet de trouver et de dénombrer toutes les décompositions possibles. Dans l'exemple, il y a ainsi 5.604 décompositions différentes du nombre 30 en somme d'entiers positifs non nuls.

La figure 5.1 représente l'ensemble des valeurs normalisées qu'est susceptible de prendre T_f. Nous pouvons apercevoir en abscisse le nombre d'entiers qui participe à la décomposi-

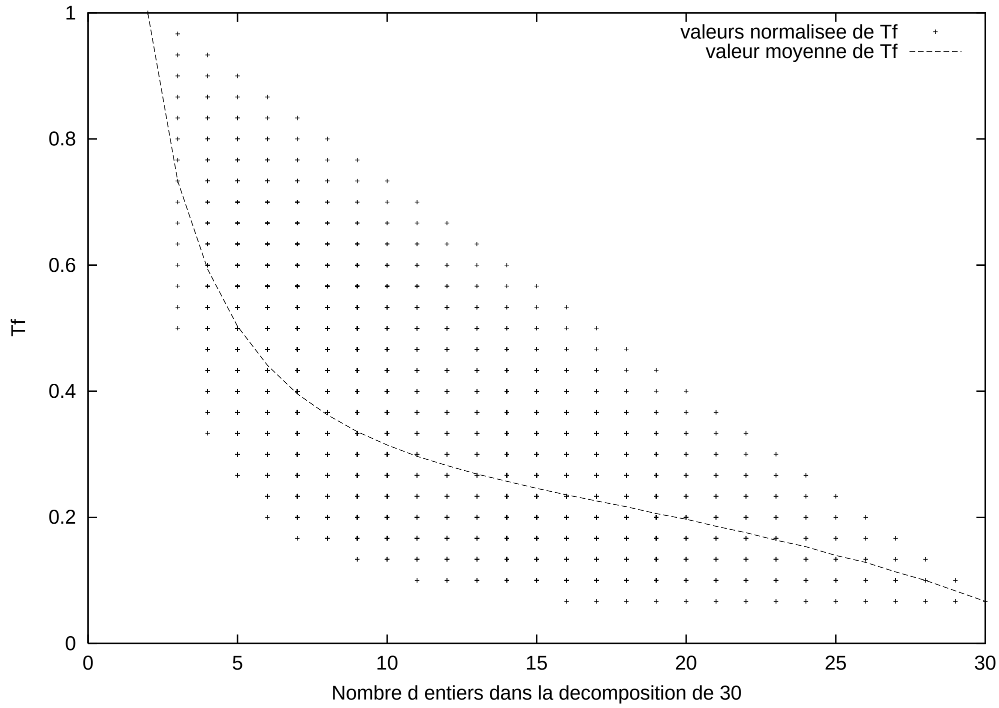


FIG. 5.1 – *Valeur moyenne de T_f pour un terme présent 30 fois dans le corpus (moyenne sur 5604 décompositions possibles)*

tion de 30, c'est à dire le nombre de documents distincts dans lesquels le terme est distribué et en ordonnée la valeur normalisée de T_f. La courbe moyenne de T_f montre qu'un terme sera alors considéré comme pertinent s'il n'apparaît que dans un ensemble très réduit de documents où il est fortement présent.

5.2.2 Indice I_{df}

Cet indice est nommé I_{df} pour *Inverse Document Frequency*. Il caractérise la répartition d'un terme dans le corpus, en partant du principe que l'importance d'un terme est inversement proportionnelle au nombre de documents du corpus dans lequel il apparait. Introduit par Spark-Jones [SJ72], cet indice permet de retrouver les termes caractéristiques d'un corpus donné.

$$I_{df}(\mathcal{T}_j) = \log(\frac{N}{n_\Gamma(\mathcal{T}_j)}) \quad (5.2)$$

avec N le nombre de documents dans Γ et $n_\Gamma(\mathcal{T}_j)$ le nombre de documents distincts de Γ contenant $\mathcal{T}_j$

La figure 5.2 représente l'ensemble des valeurs normalisées qu'est susceptible de prendre I_{df}, elle nous montre que cet indice décroît lentement et ne donne des pondérations discriminantes que pour un terme présent dans un nombre restreint de documents (moins de 5). De plus, le graphique se résume à une courbe et non un nuage de points car la valeur de I_{df} est valable pour un terme quelque soit sa T_f dans le document considéré, seul compte le nombre de documents dans lequel est réparti le terme.

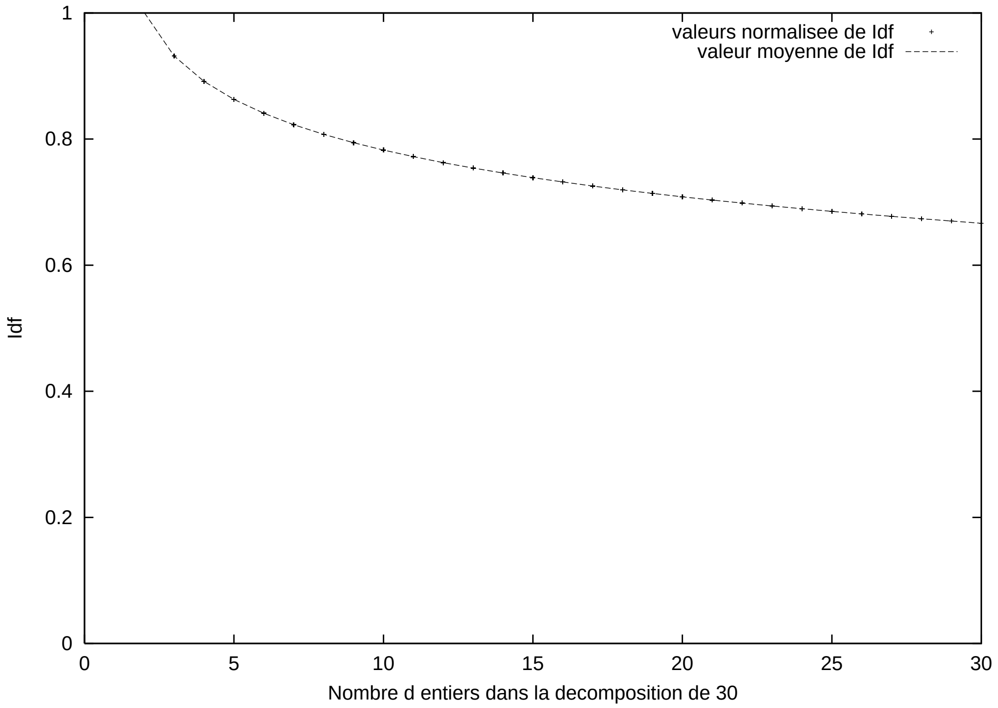


FIG. 5.2 – *Valeur moyenne de I_{df} pour un terme présent 30 fois dans le corpus (moyenne sur 5604 décompositions possibles)*

5.2.3 Indice $T_f.I_{df}$

Pour élaborer une méthode de corrélation, le critère de sélection doit finalement être un juste compromis entre termes fréquents et termes bien répartis dans le corpus. L'expérience montre que des termes sélectionnés uniquement à partir des indices T_f ou I_{df} ne sont pas nécessairement de bons descripteurs ; par contre ils servent de référence pour pouvoir comparer avec d'autres méthodes de pondération.

Un moyen simple de retrouver des termes fréquents mais présents dans peu de documents de Γ est finalement de combiner les deux critères de pondération que sont T_f [Luh57] et I_{df} [SJ72], on obtient alors l'indice $T_f.I_{df}$ [SJ73] :

$$T_f I_{df}(\mathcal{T}_j) = \frac{N(D_i, \mathcal{T}_j)}{N(\mathcal{T}_j)}.\log(\frac{N}{n_\Gamma(\mathcal{T}_j)}) \tag{5.3}$$

La figure 5.3 qui représente l'ensemble des valeurs normalisées prises par $T_f.I_{df}$ pour un terme présent 30 fois dans le corpus, résulte de la combinaison des figures 5.1 et 5.2 et accentue les caractéristiques propres de ces deux dernières. La décroissance de la courbe est, au début, plus rapide que pour les deux autres ce qui permet de dire que les termes discriminants qui seront retrouvés doivent être présents dans un pool de documents encore plus réduit que pour les deux indices précédents.

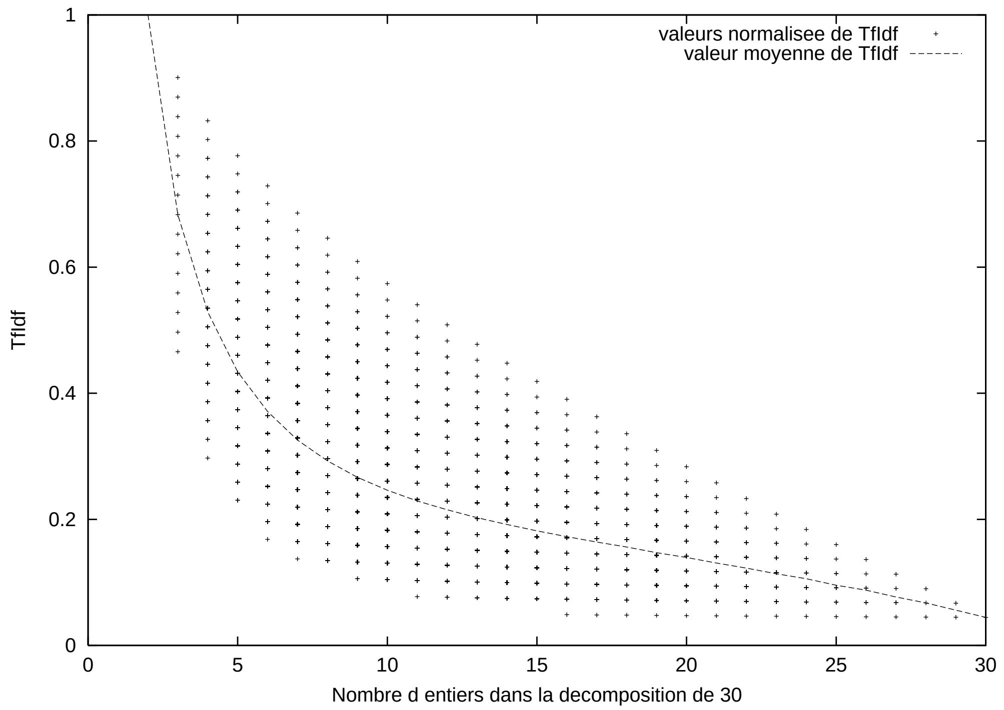


FIG. 5.3 – *Valeur moyenne de $T_f.I_{df}$ pour un terme présent 30 fois dans le corpus (moyenne sur 5604 décompositions possibles)*

5.2.4 Entropie d'un terme

Pour mesurer l'importance d'un terme dans un texte, on peut également regarder la quantité d'information qu'il apporte. Ce renseignement peut être obtenu en se basant sur le modèle de la théorie de l'information, qui nous permet de calculer la valeur informationnelle de $\mathcal{T}_j$ dans D_i [BH80]. L'entropie, notée $\mathcal{H}_\Gamma(\mathcal{T}_j)$, représente la valeur informationnelle moyenne de $\mathcal{T}_j$ dans Γ. Son calcul est donné par l'équation (5.4) ; elle est aussi appelée ratio d'indication de bruit par Salton [SY73], qui l'utilise pour sélectionner des termes valides dans un document :

$$\mathcal{H}_\Gamma(\mathcal{T}_j) = - \sum_{\substack{d \epsilon \Gamma \\ N(d,\mathcal{T}_j)>0}} P(d,\mathcal{T}_j).\log_2(P(d,\mathcal{T}_j)) \qquad (5.4)$$

avec $P(D_i, \mathcal{T}_j)$ la probabilité d'occurrence de $\mathcal{T}_j \epsilon D_i$

En pratique, $P(d, \mathcal{T}_j)$ est égale à la probabilité d'apparition d'un terme dans le document considéré, sa valeur est donnée par l'équation 5.5. Le calcul d'entropie s'obtient alors grâce à la formule donnée par l'équation 5.6.

$$P(D_i, \mathcal{T}_j) = \frac{N(D_i, \mathcal{T}_j)}{N(\mathcal{T}_j)} \tag{5.5}$$

$$\mathcal{H}_\Gamma(\mathcal{T}_j) = - \sum_{\substack{d \epsilon \Gamma \\ N(D_i, \mathcal{T}_j) > 0}} \frac{N(d, \mathcal{T}_j)}{N(\mathcal{T}_j)} . \log(\frac{N(d, \mathcal{T}_j)}{N(\mathcal{T}_j)}) \tag{5.6}$$

La figure 5.4 permet de mieux appréhender cette notion de valeur informationnelle et nous montre là encore que les termes répartis dans un petit nombre de documents sont mis en avant. Notons tout d'abord que contrairement aux autres indices, de petites valeurs pour l'entropie sont synonymes de termes discriminants et inversement. Ensuite on peut constater une différence entre les indices comme le nombre de points présents sur la figure, on s'aperçoit que la figure 5.4 est beaucoup plus dense que les autres. Cela s'explique simplement par le fait qu'à chaque distribution particulière d'un terme dans le corpus correspond une entropie différente alors que pour T_f ou pour $T_f.I_{df}$, à plusieurs répartitions différentes correspondent un même poids pour le terme. Ces indices sont composés de moins de points car moins influencés par la distribution des termes, ce qui peut être pénalisant pour différencier les termes discriminants par la suite.

5.3 Pondération statistique avec T_{ifr}

Les méthodes statistiques de pondérations existantes permettent de retrouver des termes intéressants pour le domaine considéré, mais elles ne couvrent pas tous nos besoins : comme nous venons de le voir la plupart d'entre elles ne tiennent pas assez compte de la répartition des termes dans le corpus ou, lorsqu'elles en tiennent compte, elles attribuent un poids global pour le terme dans un corpus donné sans relation avec le document considéré : c'est le cas de l'entropie.

L'idée qui nous anime est de pouvoir retrouver dans les documents des termes discriminants en s'appuyant sur leurs distributions. Nous souhaitons rechercher avant tout des termes porteurs d'informations au sein d'un petit groupe de documents. Rien n'empêche à ce type de termes d'être fréquent dans le corpus ; il faut simplement que dans un sous-ensemble de documents plus restreints, celui-ci ait une distribution singulière et qu'il soit plus présent qu'à son habitude, un peu à l'image de la spécificité positive évoquée par P.

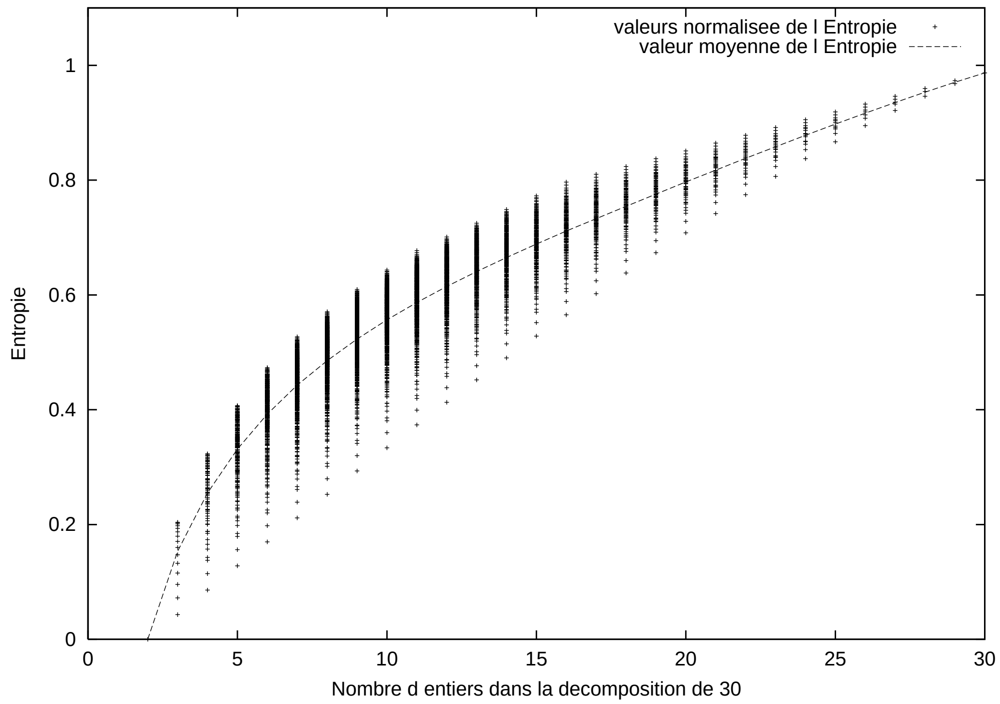


FIG. 5.4 – *Valeur moyenne de l'entropie d'un terme présent 30 fois dans le corpus (moyenne sur 5604 décompositions possibles)*

Lafon [Laf80]. Nous considérons alors ce terme comme pertinent dans ce regroupement de documents.

Pour palier ce manque, nous expérimentons un nouvel indice statistique noté T_{ifr} qui fait lui-même interagir deux autres indices nommés T_{if} et T_{ir}. L'indice T_{if}, à l'image de T_f tente de mettre en valeur un terme occurrant dans le document analysé alors que T_{ir} met en avant des termes discriminants en étudiant leur répartition dans le corpus.

5.3.1 Mesure de répartition par T_{ir}

Pour définir T_{ir}, nous allons tout d'abord étudier un indice intermédiaire noté iR et montrer qu'il permet de quantifier la répartition uniforme d'un terme $\mathcal{T}_j$ dans Γ :

$$iR(\mathcal{T}_j) = \sum_{\substack{d \epsilon \Gamma \\ N(d,\mathcal{T}_j)>0}} \frac{N(\mathcal{T}_j)}{N(d,\mathcal{T}_j)} \tag{5.7}$$

D'après l'équation (5.7), iR, admet un maximum lorsque le terme est uniformément réparti dans le corpus, et inversement, il admet un minimum pour un terme présent dans un unique document, un hapax.

Pour mieux évaluer l'intérêt de iR, nous poursuivons nos expérimentations dans le cadre d'un terme quelconque apparaissant 30 fois dans Γ, la figure 5.5 illustre toutes les valeurs prises par cet indice iR.

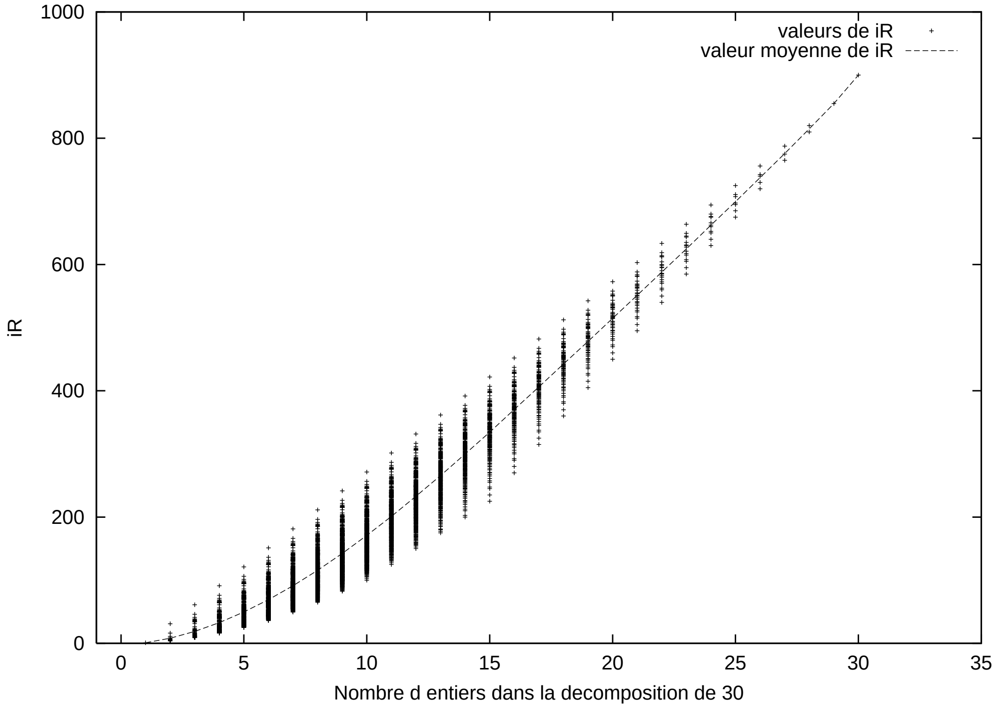


FIG. 5.5 – *Valeur moyenne de iR pour un terme présent 30 fois dans le corpus (moyenne sur 5604 décompositions possibles)*

De faibles valeurs en ordonnée correspondent à des décompositions comportant peu d'éléments telles que : $29 + 1, 28 + 1 + 1, 28 + 2, \ldots$. Les valeurs de iR qui en découlent sont faibles car elles représentent des termes présents dans peu de documents donc a priori discriminants.

À l'autre bout de l'axe des ordonnées, on retrouve des décompositions comportant beaucoup d'entiers, ces termes ayant des répartitions quasi uniformes dans les documents où ils apparaissent : les valeurs de iR sont donc importantes.

Les barres de valeurs verticales correspondent à des décompositions ayant le même nombre d'éléments telles que par exemple : $29+1, 28+2, 27+3, \ldots, 15+15$. N'ayant pas la même répartition, on obtient donc des plages de valeurs correspondant aux différentes valeurs possibles.

A partir de cet indice iR qui pondère les termes en fonction de leur distribution, nous allons construire, pas à pas, un poids donnant toute son importance aux termes porteurs d'informations.

En partant de notre définition de iR, nous obtenons l'inéquation (5.8) en l'encadrant par ses bornes atteintes dans le cas où $\mathcal{T}_j$ est un hapax et où il est présent dans tous les documents.

$$1 \leq iR(\mathcal{T}_j) \leq N(\mathcal{T}_j)^2 \tag{5.8}$$

La deuxième étape consiste à normaliser la formule (5.8) pour obtenir l'inéquation (5.9) :

$$\frac{1}{N(\mathcal{T}_j)^2} \leq \frac{iR(\mathcal{T}_j)}{N(\mathcal{T}_j)^2} \leq 1 \tag{5.9}$$

On retranche ensuite à l'inéquation (5.9) sa borne inférieure pour obtenir la formule (5.10). Cela permet de dilater les écarts de poids obtenus par les termes et de faciliter la comparaison entre des termes provenant de documents différents.

$$0 \leq \frac{iR(\mathcal{T}_j)}{N(\mathcal{T}_j)^2} - \frac{1}{N(\mathcal{T}_j)^2} \leq 1 - \frac{1}{N(\mathcal{T}_j)^2} \tag{5.10}$$

Dernière étape, obtenir un indice qui mesure le désordre amené par un terme. On prend alors l'opposé de l'équation (5.10) pour obtenir une formule qui mesure l'information apportée par chaque terme.

Nous appelons alors T_{ir} l'indice construit à partir de iR donné par l'équation (5.11) que nous allons utiliser pour rechercher des termes porteurs d'informations et donc a priori discriminants :

$$T_{ir}(\mathcal{T}_j) = 1 - (\frac{iR(\mathcal{T}_j)}{N(\mathcal{T}_j)^2} - \frac{1}{N(\mathcal{T}_j)^2}) \tag{5.11}$$

5.3.2 Lissage par T_{if}

Pour faire le parallèle avec le concept énoncé par Luhn [Luh57] pour qui un terme acquiert d'autant plus d'importance qu'il est présent dans le document, nous nous inspirons de T_f pour définir notre indice T_{ir}.

Néanmoins nous ne souhaitons pas masquer les poids obtenus par T_{ir} : nous souhaitons tout au plus les lisser et accroître le poids d'un terme occurrant dans un document. Nous avons alors retenu l'équation (5.12) pour T_{if}.

$$T_{if}(D_i, \mathcal{T}_j) = 1 + (\frac{N(D_i, \mathcal{T}_j)}{N(\mathcal{T}_j)}) \tag{5.12}$$

Elle permet alors dans le pire des cas[2] de se rapprocher de T_{ir} et dans le meilleur des cas[3] de multiplier son score.

5.3.3 Masque d'hapax pour T_{ifr}

L'indice final noté T_{ifr}, produit des deux précédents à savoir T_{if} et T_{ir}, est celui que nous allons étudier en ajoutant toutefois un Dirac de façon à masquer temporairement l'action des hapax.

En effet, l'hapax est le terme qui par excellence permet de différencier des documents ; or comme nous l'avons précédemment vu, il n'apporte pas d'information supplémentaire pour la corrélation. Notre corpus étant en perpétuelle évolution par l'apport journalier de nouveaux documents, nous devons donc prendre en compte tout terme dès qu'il apparaît dans plus d'un document.

C'est le rôle joué par le Dirac introduit dans l'équation (5.13) qui représente la formule de T_{ifr} valable pour un terme $\mathcal{T}_j$ dans un document D_i.

$$T_{ifr}(D_i, \mathcal{T}_j) = (1-\delta_1(n_\Gamma(\mathcal{T}_j)))(1+(\frac{N(D_i,\mathcal{T}_j)}{N(\mathcal{T}_j)}))(1-(\frac{iR(\mathcal{T}_j)}{N(\mathcal{T}_j)^2}-\frac{1}{N(\mathcal{T}_j)^2})) \tag{5.13}$$

avec $1 - \delta_1(n_\Gamma(\mathcal{T}_j)) = 0$ si $n_\Gamma(\mathcal{T}_j) = 1$ donc si $\mathcal{T}_j$ est un hapax

La figure 5.6 représente toutes les valeurs possibles prises par T_{ifr} pour notre exemple précédent, à savoir celui d'un terme présent 30 fois dans le corpus. En abscisse on retrouve toutes les valeurs de T_{ifr} et en ordonnée le nombre de documents dans lesquels le terme est réparti. Les plages verticales de valeurs signifient comme dans l'étude de iR qu'il existe plusieurs façons de répartir un même terme entre n documents. À noter qu'il y a cette fois-ci plus d'éléments que dans la figure 5.5, car à un point particulier de cette figure correspondent plusieurs valeurs possibles de T_{if} et donc plusieurs valeurs de T_{ifr}.

Sur la figure 5.6, il est possible de voir que, conformément aux intentions de départ, les hapax sont " temporairement " éliminés et que les termes les mieux notés sont ceux qui apparaissent dans un ensemble restreint de documents. D'autre part, les larges plages de

[2]Terme uniformément réparti dans le corpus.
[3]Hapax.

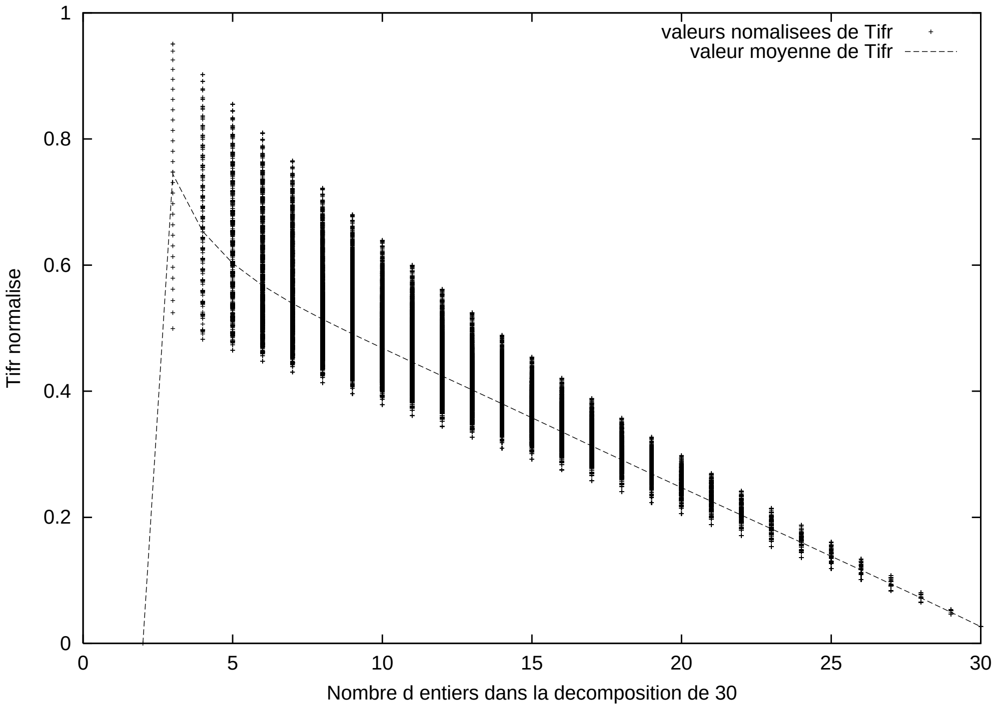


FIG. 5.6 – *Valeur moyenne T_{ifr} pour un terme présent 30 fois dans le corpus (moyenne sur 5604 décompositions possibles)*

valeurs prises par T_{ifr} permettent à des termes fréquents d'avoir de meilleurs scores que des termes moins présents ; il leur suffit pour cela d'être répartis dans un nombre plus resserré de documents, un peu à l'image de ce que fait $T_f.I_{df}$, mais en utilisant comme renseignement la distribution d'un terme dans Γ.

Concrètement, prenons l'exemple de deux termes $\mathcal{T}_1$ et $\mathcal{T}_2$ apparaissant respectivement 8 et 14 fois dans 4 documents du corpus. Le tableau 5.3 nous donne les valeurs de T_{ifr} pour les deux termes dans chaque document.

Occurrences de $\mathcal{T}_1$ dans un document	2	2	2	2
T_{ifr}	0,9967	0,9967	0,9967	0,9967
Occurrences de $\mathcal{T}_2$ dans un document	6	4	3	1
T_{ifr}	1,1313	1,0684	1,0370	0,9742

TAB. 5.3 – *Exemple de valeurs de T_{ifr}*

Grâce à ce tableau, on s'aperçoit que le terme $\mathcal{T}_2$ est d'après T_{ifr} plus intéressant que $\mathcal{T}_1$ dans 3 cas sur 4, notamment grâce à l'action de T_{ir}. En effet $T_{ir}(\mathcal{T}_1) = 0,9344$ alors que

$T_{ir}(\mathcal{T}_2) = 0,9427$. Néanmoins la combinaison avec T_{if} permet de déterminer que $\mathcal{T}_2$ joue un rôle moins important que $\mathcal{T}_1$ dans le quatrième document car, au vu de la répartition de $\mathcal{T}_2$ dans les trois autres, il apporte une information moindre dans le dernier document.

L'important est de retrouver des termes de Γ porteurs d'informations, pour cela ils doivent être discriminants dans une partie des documents où ils apparaissent. Cette formule permet à des termes occurrents d'être qualifiés de pertinents et inversement à un terme peu fréquent de ne pas être pris en compte : tout va alors dépendre de leurs distributions.

5.4 Mise en place d'une expérimentation

Cette section 5.4 s'intéresse à la mise en place d'une expérimentation permettant de comparer les cinq méthodes statistiques de pondérations décrites précédemment, à savoir les indices T_f, I_{df}, $T_f.I_{df}$, le calcul d'entropie donné par l'équation (5.6) et notre indice expérimental T_{ifr}.

Le but de cette expérimentation n'est pas uniquement de comparer les caractéristiques des indices statistiques entre eux. En effet, les méthodes statistiques de pondérations sont utilisées dans notre méthode de corrélation pour extraire les termes valides des documents parmi les candidats termes trouvés. Il faut donc pouvoir quantifier leurs pertinences dans le choix des termes valides et comparer les résultats avec une méthode d'extraction de référence, à savoir une extraction terminologique manuelle. L'expérimentation qui va être menée est bien entendu effectuée sur notre corpus de référence.

5.4.1 Procédure de validation

L'évaluation, réalisée sous la forme d'une enquête Web, consiste à faire extraire par des évaluateurs les termes valides contenus dans 50 textes différents ; cette liste de documents d'expérimentation est notée $[D_{exp}(1) \dots D_{exp}(50)]$. Les documents ont été sélectionnés aléatoirement dans notre corpus de référence (Voir Annexe A).

À partir de ces textes sont créés des formulaires d'évaluation, formulaires illustrés par la figure 5.7 et composés de deux parties. Une partie permet de lire le document d'expérimentation provenant de la liste $[D_{exp}(1) \dots D_{exp}(50)]$ et la seconde permet de valider la liste des candidats termes extraits de ce document donné par ordre lexicographique.

Pour chaque candidat terme, l'évaluateur peut répondre par Oui ou Non à la question suivante : “ Le terme vous paraît-il important pour la compréhension du document ? ”. Les évaluateurs doivent alors lire chaque document et sélectionner les termes qui leur semblent être importants.

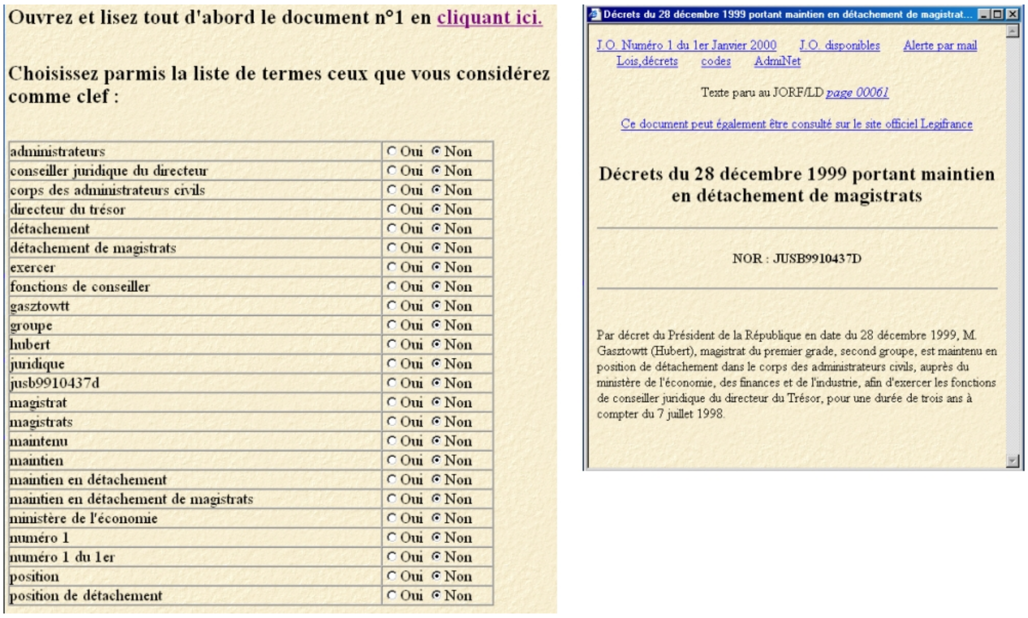


FIG. 5.7 – *Illustration d'un questionnaire d'évaluation*

Chaque évaluateur pré-sélectionné qui souhaite effectuer la validation doit fournir quelques informations nous permettant de le catégoriser comme candidat *candide* ou *expert* dans le domaine juridique. Ces informations sont collectées par l'intermédiaire d'un formulaire qui est illustré par la figure 5.8. Nous vérifions aussi par la même occasion que sa contribution est digne de confiance et peut être acceptée. La confrontation des deux types d'évaluateurs, les candides et les experts, doit permettre également de détecter les éventuelles différences existant entre les deux catégories d'utilisateurs et ainsi pouvoir améliorer les recherches juridiques sur notre site qui s'adressent en priorité à un public averti.

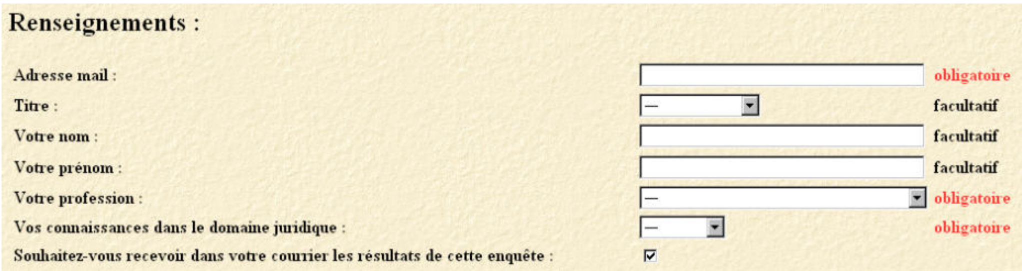


FIG. 5.8 – *Renseignements demandés aux évaluateurs*

5.4.2 Résultats

La mise en place de l'évaluation à débuté en février 2002 par l'ouverture d'une enquête menée sur le Web[4]. Dans ce cadre, nous avons pré-sélectionné nos évaluateurs en diffusant

[4] `http://evaluation.w3sites.net`

plusieurs annonces, par envoi de mail, à plusieurs instituts universitaires, écoles d'ingénieurs, centres de recherches et listes de diffusion juridique. Malheureusement le retour a été très faible : 11 inscriptions sur les centaines de candidatures potentielles. Sur ces 11 candidatures, 6 d'entre elles proviennent d'évaluateurs candides et 2 d'experts ; les 3 restantes ne peuvent pas être prises en compte car les validations n'ont pas été achevées ou les résultats ne sont pas dignes de confiance. Les résultats présentés ici sont ceux des évaluateurs candides, les experts n'étant pas suffisamment nombreux pour que les résultats puissent être significatifs.

5.4.2.1 Nombre de termes validés

Nous commençons notre analyse des résultats par l'étude de la figure 5.9 qui nous permet de comparer la taille des documents en nombre de termes distincts (Courbe $T1$) avec le nombre de termes validés par au moins un évaluateur pour un document donné (Courbe $T2$).

Premier constat, le panel de documents observé est hétérogène, à l'image de notre corpus, les tailles des documents allant de 7 termes à plus de 1100 pour le document 45. Deuxième point, on peut s'inquiéter de l'apparition de certains pics corrélés entre les deux courbes qui pourraient nous laisser penser que le nombre de termes extraits par au moins un utilisateur est proportionnel à la taille des documents.

Pour vérifier cette hypothèse, on étudie la courbe $T3$ qui prend en compte le nombre de termes cochés par au moins deux évaluateurs pour un même document. Le but est de ne prendre en considération que des termes pour lesquels il existe un consensus minimum entre les évaluateurs. Ce nombre de 2 a été retenu car il exprime un nombre minimal dans l'idée d'un consensus. Il est aussi maximal car au-delà certains documents se retrouvent dépourvus de termes consensuels[5]. Il va de soi que pour un panel plus important d'évaluateurs, ce poids de 2 est à redéfinir.

Pour prouver ou infirmer une éventuelle corrélation entre $T1$ et $T3$, on peut prendre en compte la figure 5.10 qui met en avant le nombres de termes validés en fonction de la taille des documents. On s'aperçoit alors que la courbe concernant les termes validés suit l'évolution de la taille des documents alors que la courbe propre aux termes consensuels reste quasi-horizontale.

Pour lever définitivement tout doute de corrélation entre le nombre de termes consensuels et la taille des documents on effectue un test de χ^2 entre ces courbes $T1$, $T2$ et $T3$ de la figure 5.9. Pour un degré de liberté valant ici 49 et un seuil de signification placé à 0.01, la valeur seuil du χ^2 est de $76,2$. Le calcul du χ^2 entre $T1$ et $T3$ donne une valeur de $400,26$,

[5]Termes validés par au moins deux évaluateurs.

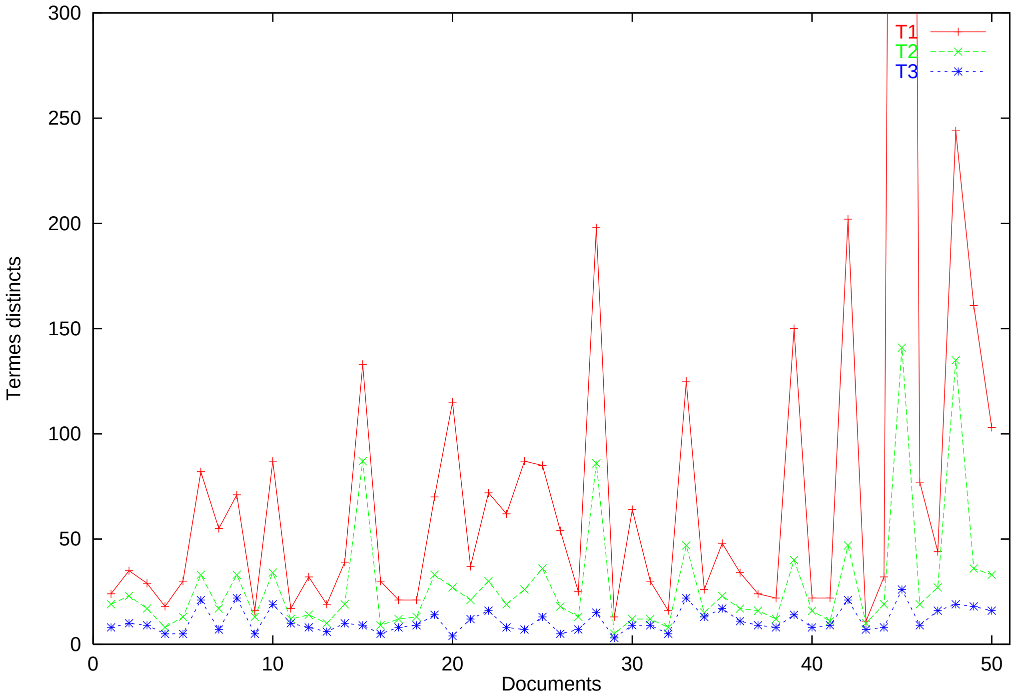


FIG. 5.9 – *Caractéristiques du corpus expérimental au regard de la taille des documents ($T1$), le nombre de termes validés ($T2$) et consensuels ($T3$)*

largement supérieure à la valeur du seuil, ce qui infirme toute possibilité de corrélation. On peut alors poser le concept de l'existence d'un consensus minimal entre les réponses, car il permet de prendre en compte un nombre de termes qui n'est pas proportionnel à la taille des documents. Cela nous permet d'envisager la possibilité d'extraire un nombre borné de termes valides pour chaque document pour lesquels la taille du document n'intervient pas dans la définition de la borne supérieure.

Le tableau 5.4 est une synthèse des résultats de la courbe 5.9, mis à part qu'il représente des nombres moyens de termes et qu'il fait apparaître de manière séparée les deux types de termes rencontrés : les mots et les SN.

	$T1$	$T2$	$T3$
Nb de mots	48.3	14.56	3.88
Nb de SN	41.46	13.36	7.2
Nb de Termes	89.76	27.92	11.08

TAB. 5.4 – *Taille moyenne des documents*

Le but est de vérifier qu'il n'y a pas de différence fondamentale entre la taille des textes pour les mots et les SN, colonne $T1$. On constate que les deux catégories ont à peu près

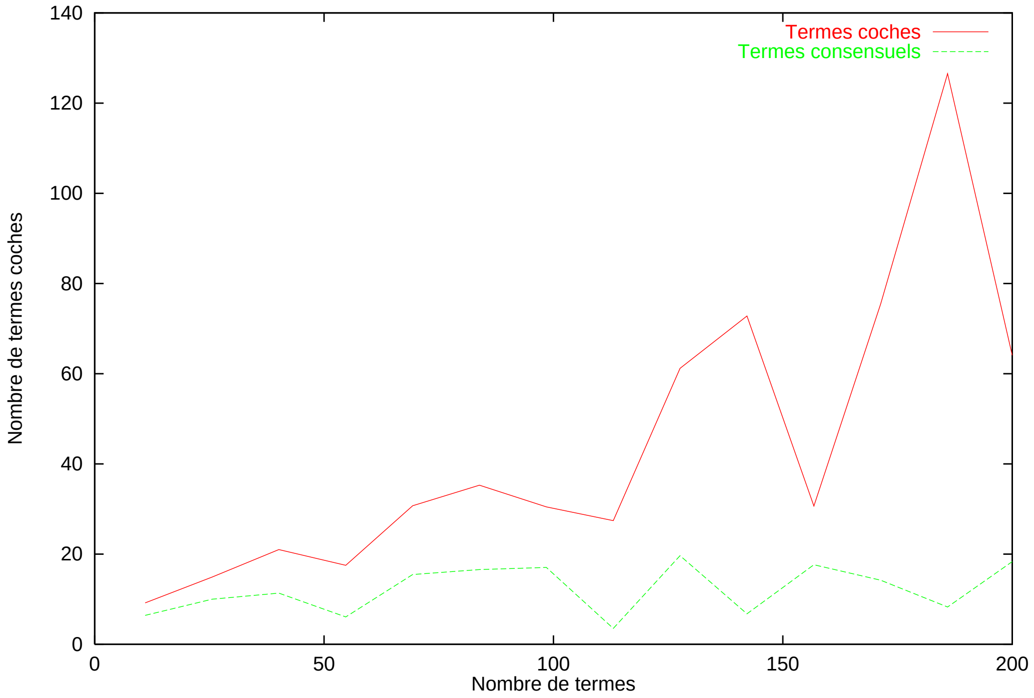


FIG. 5.10 – *Corrélation entre taille des documents et nombre de termes validés. En abscisse, les documents sont triés par taille croissante.*

le même profil car elles ont quasiment le même nombre d'éléments. La colonne $T2$ qui représente le nombre de termes cochés par au moins un utilisateur est identique pour les mots ou les SN. La seule différence visible, colonne $T3$, est constatée entre le nombre de mots et de SN consensuels. Cette différence provient du fait que les SNs sont par nature moins ambigus que les mots, lorsqu'ils sont choisis par les évaluateurs ; on assiste donc le plus souvent à l'établissement d'un consensus.

5.4.2.2 Accord entre évaluateurs

La figure 5.11 montre le taux d'accord existant entre les évaluateurs. Le but est de savoir, si en moyenne, les personnes interrogées sont d'accord entre elles. Un désaccord prononcé ne nous permettant pas d'utiliser les résultats, nous cherchons à vérifier qu'il existe une concordance suffisante entre les réponses.

La courbe $A1$ que l'on peut tracer en utilisant la formule (5.14) utilise une première définition du concept d'accord. Dans le cas présent, l'accord se caractérise par le fait que tous les évaluateurs réagissent de la même manière pour un terme et un document donné. Deux cas sont envisageables : personne ne choisit un terme pour un document considéré, ou au contraire tout le monde le choisit.

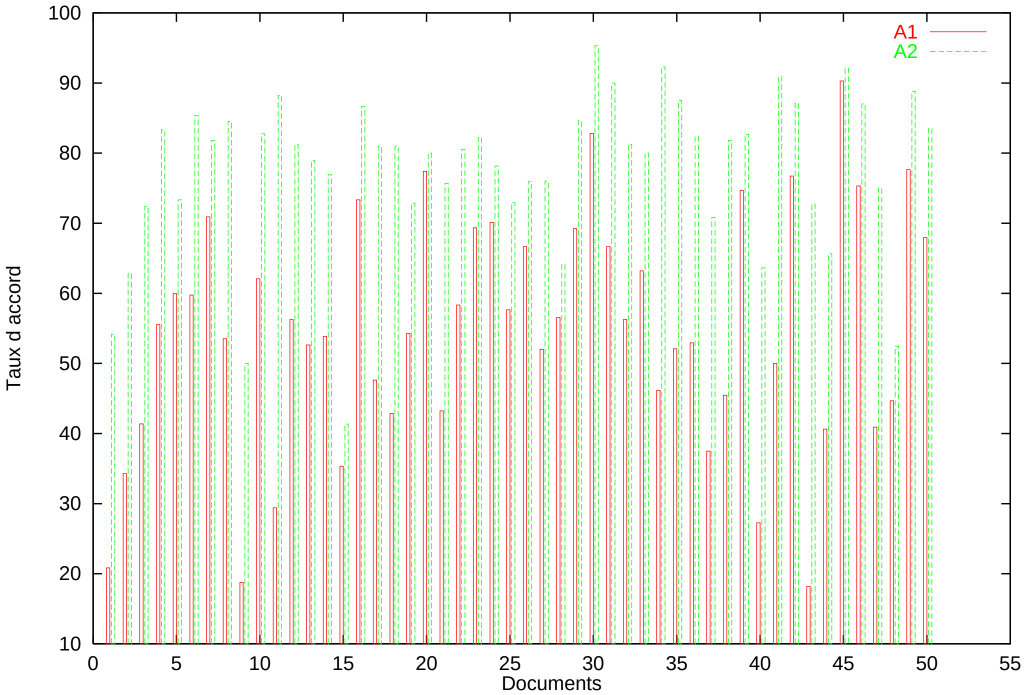


FIG. 5.11 – *Taux d'accord entre évaluateurs*

La courbe $A2$ qui peut se tracer grâce à la formule (5.15), exprime le désaccord d'une autre manière ; cette fois-ci le désaccord est représenté par les termes non-consensuels. Ainsi il y a désaccord pour un terme et un document donné si le terme est validé par un unique évaluateur et il y a accord dans tous les autres cas de figure.

$$f_1(d) = \frac{\mathcal{N}_{tnc}(d) + \mathcal{N}_{tvt}(d)}{\mathcal{N}_t(d)} \tag{5.14}$$

$$f_2(d) = \frac{\mathcal{N}_t(d) - \mathcal{N}_{tu}(d)}{\mathcal{N}_t(d)} \tag{5.15}$$

avec $\mathcal{N}_{tnc}(d)$: $\|$termes cochés par aucun évaluateur$\|$
$\mathcal{N}_{tvt}(d)$: $\|$termes cochés par tous les évaluateurs$\|$
$\mathcal{N}_{tu}(d)$: $\|$termes cochés une unique fois$\|$
$\mathcal{N}_t(d)$: $\|$ensemble des termes de $d\|$
et $d \,\epsilon [D_{exp}(1) \ldots D_{exp}(50)]$

Qu'apprend-t-on de ces deux courbes ? On constate de grandes irrégularités sur la courbe $A1$ car il y a peu ou pas de termes cochés pour lesquels tous les évaluateurs sont en accord. Le taux moyen d'accord présenté dans le tableau 5.5 n'est pas significatif, car il représente un taux d'accord moyen de 54% qui masque des irrégularités profondes sur la figure 5.11.

Par contre les valeurs de la courbe $A2$ sont nettement plus régulières, quasiment tous les taux sont au-dessus de 65%. On peut noter une valeur moyenne d'accord de près de 78%. Cela confirme l'importance des termes consensuels pris en compte sur la courbe $A2$. La comparaison de ces deux courbes montre qu'a priori le désaccord provient des termes cochés de manière unitaire, ce qui nous pousse à vouloir ne pas en tenir compte dans le reste de nos expérimentations.

	$A1$	$A2$
Mots	58%	75.03%
SN	51.51%	82.20%
Termes	54.61%	77.83%

TAB. 5.5 – *Accord moyen entre les évaluateurs.*

Le reste des valeurs du tableau 5.5 est là pour témoigner des différences minimes qui existent pour les évaluateurs entre les taux d'accords concernant les mots et les SN ; la colonne $A1$ représente les taux d'accords moyens trouvés à l'aide de l'équation (5.14). La colonne $A2$ représente ceux trouvés à partir de la formule (5.15).

Cette partie de vérification nous permet de confirmer notre hypothèse, à savoir que seuls les termes consensuels sont à prendre en compte. Cette hypothèse se justifie en étudiant les courbes 5.9 et 5.11. Elles montrent que si l'on se borne à n'utiliser que les seuls termes consensuels, le nombre de termes validés par les utilisateurs n'est plus proportionnel à la taille des documents et que le taux d'accord entre évaluateurs est élevé. Deux conséquences qui nous poussent à ne travailler par la suite, exclusivement, qu'avec des termes consensuels dans le cadre des comparaisons entre les méthodes d'extractions statistiques et l'extraction manuelle.

Un autre élément intéressant que l'on retire de cette expérimentation est le taux d'accord élevé existant entre les évaluateurs ; cela permet de considérer que ces résultats expérimentaux sont homogènes et sont une bonne base d'étude pour la suite de notre expérience.

5.5 Comparaison des pondérations

La finalité de cette section est de comparer les listes de termes résultant du choix des évaluateurs avec des listes de termes pondérés et choisis automatiquement par les trois méthodes de pondérations précédemment étudiées. Pour effectuer cette comparaison, nous allons étudier et relever les similitudes existantes entre les listes des évaluateurs et nos listes expérimentales.

5.5.1 Créations de listes pondérées

Dans un premier temps, on doit formaliser les résultats des évaluateurs sous forme de listes de termes pondérés pour chaque document de $[D_{exp}(1) \ldots D_{exp}(50)]$; ces listes seront appelées par la suite listes de références. On crée, alors, deux listes par document, une première liste qui regroupe les mots et une autre les SN. Ces deux listes sont ordonnées en fonction des choix de termes faits par les évaluateurs. Les termes placés au sommet des listes sont ceux qui ont été cochés par le plus grand nombre, puis, en descendant dans le classement, ils l'ont été de moins en moins. Dans ces listes, conformément à l'hypothèse formulée dans la section 5.4, on ne garde pour la comparaison finale que les termes consensuels.

Ensuite, pour chaque méthode de pondération statistique, on crée comme précédemment deux listes pondérées par document, une pour les mots et une autre pour les SN. Pour chaque liste, le poids associé à un terme est celui qu'il obtient via l'une des méthodes de pondération. Une fois que chaque terme des différentes listes a reçu un poids, elles sont triées par ordre de pertinence décroissante en fonction de la pondération reçue et de la méthode employée.

Pour pouvoir comparer le plus justement possible des méthodes de pondération comme l'entropie, $T_f.Idf$ d'un coté et T_{ifr} de l'autre qui réagissent de manière opposée aux problèmes des hapax, ceux-ci ont été ôtés des listes des évaluateurs et des listes de termes pondérés.

5.5.2 Protocole de comparaison

La comparaison des résultats entre les listes de termes de référence de chaque document et celles obtenues grâce aux méthodes statistiques de pondération se résume à une comparaison deux à deux de listes de termes pondérés $\mathcal{L}_1$ et $\mathcal{L}_2$. Pour comparer deux listes de termes, nous calculons leur précision (formule (5.16)) à chaque rang n, en allant jusqu'à épuisement des termes de la plus grande des deux listes. Nous définissons le rang n comme étant la position d'un terme dans une liste triée par pondération décroissante.

$$P(\mathcal{L}_1, \mathcal{L}_2, n) = \frac{\mathcal{N}_{tc}(\mathcal{L}_1, \mathcal{L}_2, n)}{n} \tag{5.16}$$

avec $\mathcal{N}_{tc}(\mathcal{L}_1, \mathcal{L}_2, n) = \|\text{termes communs entre } \mathcal{L}_1 \text{ et } \mathcal{L}_2 \text{ au rang } n\|$

Pour chaque document, nous calculons, en fonction du rang, la précision des listes de termes pondérés par les différentes méthodes statistiques avec les deux listes de références. Les résultats sont regroupés sur les figures 5.12 et 5.13 qui expriment la précision moyenne obtenue par les cinq méthodes de pondération mise en concurrences : l'entropie, $T_f I_{df}$, T_f, I_{df} et T_{ifr}. La figure 5.12 exprime les résultats pour les mots et la figure 5.13 pour les SN.

En complément du taux de précision, le taux de rappel a été étudié ; malheureusement il n'apporte pas d'information supplémentaire. Il tend à être maximal lorsque l'on se rapproche d'un rang proche de la taille du document. Cela nous enseigne, simplement, qu'il manque toujours des termes par rapport à la liste idéale des évaluateurs, et ce quelle que soit la méthode de pondération choisie.

5.5.3 Résultats

Les résultats proposés par la courbe 5.12 montrent que les méthodes obtiennent des résultats sensiblement différents. En effet, les deux méthodes de pondération reposant sur des pondérations telles que T_f ou I_{df} obtiennent de très mauvais scores et sont loin en dessous des trois autres.

Inversement des pondération telles que l'entropie, $T_f I_{df}$ ou T_{ifr} obtiennent des taux similaires proches en ce qui concerne les mots. La précision maximale est obtenue pour un rang de 3 et vaut 17%. Ce faible rang et cette précision médiocre indiquent que les résultats ne sont pas assez concluants pour permettre de montrer la qualité d'une méthode par rapport à une autre. Ils soulignent plutôt leurs insuffisances. Malgré tout on peut conserver une note d'espoir concernant notre méthode expérimentale en remarquant que la courbe T_{ifr} croît légèrement plus rapidement que ses concurrentes pour atteindre, un rang plus tôt, sa précision maximale.

Les résultats des SN illustrés par la figure 5.13 montrent un rapprochement des différentes courbes, il n'y a plus de cassure nette entre l'entropie, $T_f I_{df}$ et T_{ifr} d'un coté et T_f et I_{df} de l'autre mais plutôt une lente dégradation de la précision. La précision maximale des SN est obtenue pour un rang supérieur valant 5 avec une précision de 30%, ceci pour la courbe T_{ifr}. En acceptant une dégradation de la précision, on peut aller jusqu'au rang 10 et obtenir une précision respectable de 25%. Bien entendu les résultats nous montrent, aussi, que l'on est assez loin de retrouver tous les termes d'une extraction manuelle, mais ils permettent déjà d'obtenir quelques termes clefs choisis par des évaluateurs. On remarque aussi que notre méthode de pondération expérimentale se détache légèrement de ses deux concurrentes en atteignant, comme pour le cas des mots, plus rapidement sa précision maximale et en la maintenant quelque temps.

5.5.4 Analyse

Globalement les résultats évoqués ci-dessus montrent qu'il est difficile de statuer sur l'efficacité des méthodes, car elles fournissent des résultats mitigés. Malgré tout ils nous

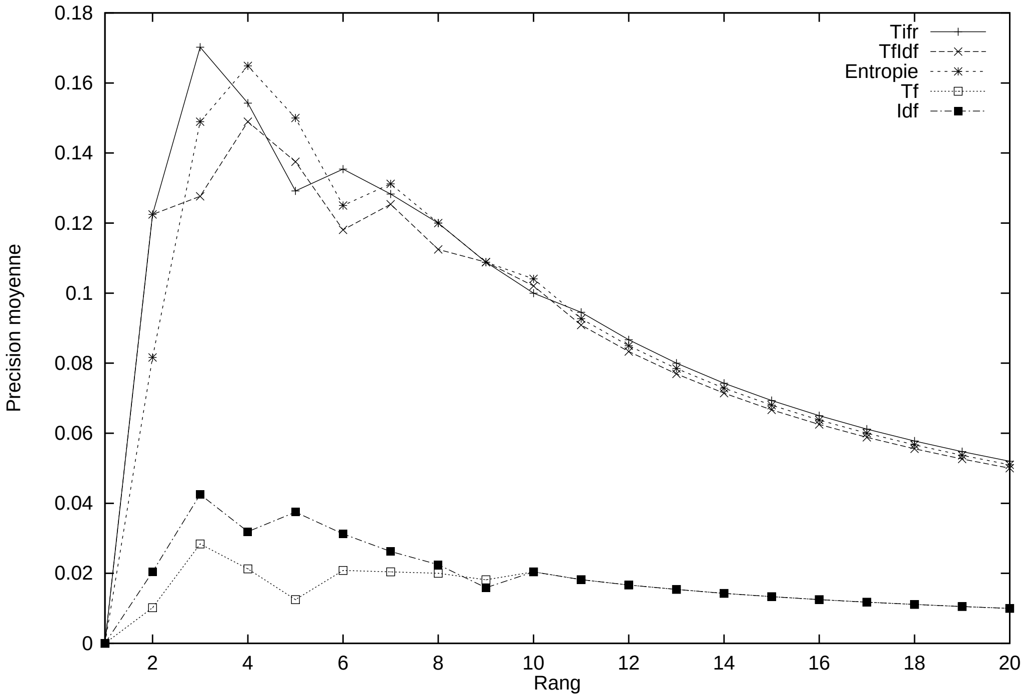


FIG. 5.12 – *Comparaison des méthodes de pondération pour les mots. En abscisse, le nombre de termes comparés.*

montrent que l'hypothèse formulée à l'égard de T_{ifr} est validée, cet indice permet d'extraire des termes discriminants en nombre comparable avec les autres méthodes étudiées.

Néanmoins ces résultats sont globalement décevants concernant les mots car la précision maximale est atteinte pour un rang très faible, sans que celle-ci soit significative. Ces résultats étaient pourtant prévisibles en étudiant les chiffres du tableau 5.4. Il nous montre, en effet, qu'en moyenne, $3,88$ mots sont validés sur un texte d'une longueur moyenne de $48,3$ mots, soit 8% des mots. Les méthodes expérimentées traitent et ordonnent ces $48,3$ mots que l'on compare avec une liste d'environ 4 mots, ce qui laisse peu de chance d'obtenir des résultats satisfaisants.

D'autre part, les résultats sur les SN sont meilleurs. En effet, globalement la précision est plus grande pour un rang plus élevé et ceci pour deux raisons. Premièrement, le nombre de termes choisis par les évaluateurs est plus important, $7,2$ SN d'après le tableau 5.4 pour une longueur moyenne des textes de $41,46$ SN, ce qui nous permet statistiquement d'espérer de meilleurs résultats.

Deuxièmement, les SN sont pertinents sémantiquement pour un document si leur signification est précise et est proche des thématiques qui y sont abordées. Or plus un SN est pertinent pour un document, plus il doit être considéré comme tel par les évaluateurs et donc plus il va, en théorie, être coché. Mais simultanément, plus il est pertinent, plus il est

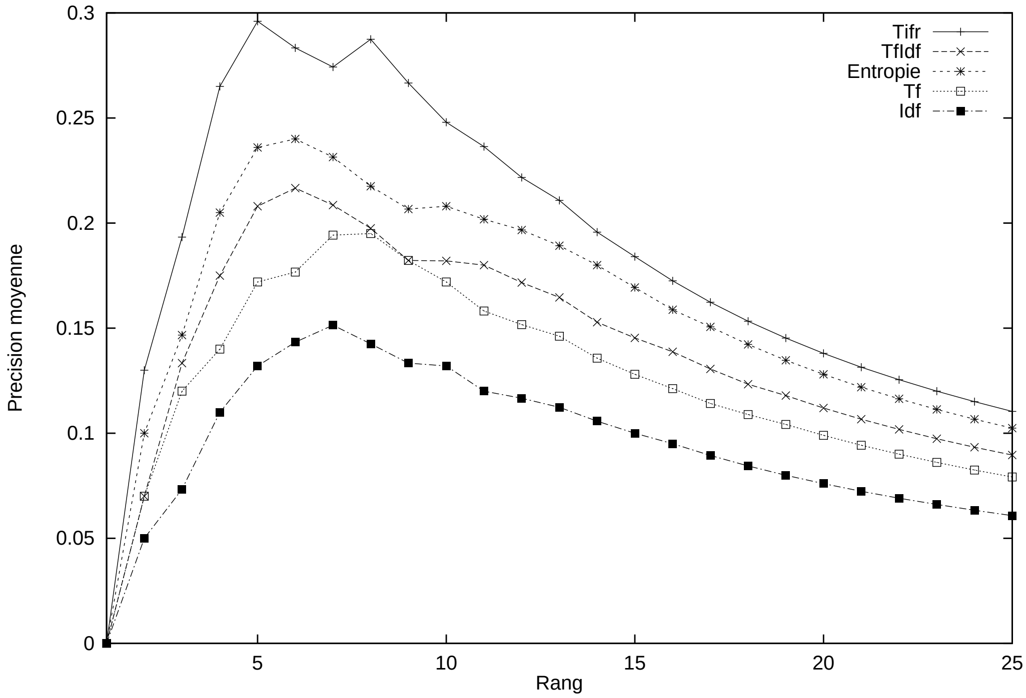


FIG. 5.13 – *Comparaison des méthodes de pondération pour les SN. En abscisse, le nombre de termes comparés.*

caractéristique du vocabulaire de ce document et donc moins il a de chance d'être présent dans le corpus. La conséquence directe est que moins ce terme est fréquent, meilleur sera le poids obtenu par une méthode statistique de pondération comme celles expérimentées ici.

5.6 Enseignements

Dans ce chapitre, nous avons expérimenté des méthodes statistiques de pondération pour comparer leurs résultats avec une sélection terminologique manuelle. Le premier enseignement à retirer de ce travail est qu'une comparaison entre des listes de termes pondérés est un travail difficile à orchestrer et à interpréter, notamment lorsque les écarts de taille entre les listes sont trop importants (cas de l'expérimentation sur les mots). Dans ce cas, il est difficile de conclure.

Néanmoins notre travail préliminaire sur l'étude des résultats des évaluateurs nous a permis de prouver qu'il peut exister entre eux un réel consensus. L'idée de vouloir calquer, via des méthodes statistiques de pondération, une extraction terminologique manuelle n'est donc pas insensée. Les termes provenant de cette extraction manuelle sont a priori sémantiquement les plus proches des thématiques des documents étudiés ; ils peuvent être considérés comme de bons indicateurs pour la recherche de documents corrélés.

Cette première étude nous a également permis de montrer que le nombre de termes consensuels dans un document n'est pas proportionnel à sa taille. Cela conforte notre volonté de créer pour chaque document une signature sémantique dont la taille peut être définie à l'aide d'un seuil, que l'on peut par exemple trouver en recherchant un maximum sur l'équation (5.16).

La dernière partie de l'expérimentation nous montre qu'il existe une différence significative entre les comparaisons de listes pondérées de mots et de SN. De manière intuitive, un SN est plus porteur de sens qu'un mot simple, car il n'est presque pas polysémique et sa fréquence moyenne d'apparition dans le corpus est beaucoup plus faible. Cette remarque nous incite à penser qu'il est judicieux de concentrer les efforts sur l'extraction et la pondération des SN pour leur donner un rôle prépondérant.

La comparaison des autres méthodes statistiques de pondération avec notre méthode empirique T_{ifr} montre que l'on obtient des résultats comparables avec des méthodes classiques (figures 5.12 et 5.13) avec, toutefois, la particularité de retrouver à rang égal plus de termes consensuels que ses rivales. Cette croissance plus rapide peut être un atout majeur pour élaborer une signature et minimiser le bruit parmi les résultats corrélés. En effet le choix des termes étant primordial pour la qualité des résultats, obtenir une précision élevée pour un rang faible est une nécessité.

L'utilisation des méthodes de pondération est une étape intermédiaire de notre méthode de corrélation. Le but premier de la mise au point de l'indice T_{ifr} est de retrouver des termes discriminants pour la corrélation. Les termes extraits ne collent peut être pas trait pour trait à l'extraction manuelle étudiée mais le fait de s'en rapprocher signifie que l'on se dirige vers une extraction sémantique.

Chapitre 6

Définir une unité de recherche

Au chapitre 5 nous avons comparé des listes de termes extraites automatiquement avec des listes composées manuellement. Les résultats de cette expérimentation ont mis en évidence la possible utilisation des méthodes de pondérations statistiques pour rechercher des termes discriminants. Malheureusement cette expérience met également en évidence son insuffisance quant à la déduction de critères de sélection des termes.

Avant de rechercher ces critères de sélection terminologique, nous devons tout d'abord répondre à une autre question énoncée au chapitre 4 concernant la définition de l'unité de recherche. L'unité de recherche, à titre de rappel, est *le segment de texte (nombre de mots, phrases, paragraphes, etc.) à prendre en considération pour rechercher les éléments d'une même requête.* Cette notion d'unité de recherche est primordiale ; de sa définition dépend la qualité des résultats retrouvés. Trop grande, ils risquent d'être imprécis ; trop restrictive, la méthode engendrera du silence. L'objectif de ce chapitre 6 est de déterminer, pour notre corpus de référence, la taille optimale de l'unité de recherche.

La section 4.4.2.2 a montré que l'unité de recherche pouvait prendre de multiples formes (nombre de mots, phrases, paragraphes, etc.), suffisamment pour qu'il ne soit pas possible de toutes les expérimenter. La finalité de la section 6.1 est, dans un premier temps, de réduire le nombre d'unités de recherche potentiellement envisageables. La section 6.2 décrit le principe et le but de notre expérimentation, alors que la section 6.3 s'attache à présenter les résultats obtenus et les enseignements à en retirer.

6.1 Des unités de recherche pré-existantes

Définir une taille optimale de l'unité de recherche est une étape qui peut devenir très rapidement insurmontable si l'on ne réduit pas au plus vite le nombre de cas envisageables. Afin de le réduire, nous postulons qu'une unité de recherche doit répondre à deux critères distincts :

- avoir une justification sémantique ;
- répondre à des impératifs pratiques.

6.1.1 Une justification sémantique

La taille choisie pour définir une unité de recherche ne peut pas se réduire à un choix arbitraire dénué de sens et sans lien avec les textes ou documents étudiés ; la définition d'une unité de recherche doit au minimum posséder une justification sémantique. Dans le cas contraire, celle-ci pourrait prendre des valeurs singulières telles qu'une portion de texte de 37 mots consécutifs ou bien encore un ensemble de 3 phrases consécutives, des valeurs prises arbitrairement dans le seul but d'optimiser les performances du système en terme de rappel ou de précision.

Si l'on ne veut pas obtenir de tels résultats ou si l'on ne souhaite pas que l'unité trouvée se justifie uniquement au travers de considérations statistiques, nous devons rechercher les unités de recherche, en priorité, parmi des segments qui peuvent se justifier sémantiquement. Cette réflexion nous amène à retrouver dans un premier temps les sous-ensembles sémantiquement homogènes de notre corpus de référence pour y piocher, ensuite, les unités adéquates.

Nous définissons deux types de sous-ensembles sémantiquement homogènes : des sous-ensembles statiques et dynamiques.

Les éléments *statiques* sont en fait des unités qui pré-existent dans un document et ont été insérées au moment de sa création par l'auteur lui-même. Ces parties qui servent à découper le texte en unités homogènes sont les *phrases, paragraphes, sections et autres chapitres* qui orchestrent le déroulement logique du récit. Leur utilisation en tant qu'unité de recherche dépend alors du corpus d'étude, car il est nécessaire de vérifier si ce découpage logique a été correctement effectué et est utilisable.

Les éléments *dynamiques* sont quant à eux inexistants dans le texte initial ; ils doivent donc être fabriqués pour l'occasion. Ils sont constitués à partir d'une réorganisation des éléments textuels initiaux pour obtenir des segments thématiquement homogènes. La recherche de ces parties dynamiques est l'objet d'étude des travaux de segmentation thématique qui permettent de découper les documents en sous-ensembles sémantiquement homogènes [Fer98, Her02] par recombinaison d'éléments existants. La finalité est d'améliorer, par exemple, la recherche de candidats termes qui participeront à la construction de signatures lexicales ou thématiques [Fer98].

Qu'ils soient statiques ou dynamiques, ces deux types de sous-ensembles thématiquement homogènes sont des candidats sérieux qui doivent être pris en considération pour définir l'unité de recherche de notre corpus.

6.1.2 Une justification pratique

Le choix entre ces différentes unités d'information est désormais fonction de notre corpus de référence et doit exploiter ses particularités. En effet, inutile de définir comme unité de recherche le *chapitre* si les documents analysés en sont dépourvus.

Notre corpus de référence, décrit section 4.1, est composé d'un grand nombre de documents (un peu plus de 24.000) de longueurs variables et de types différents : des articles, des lois, des décrets, etc. Le traitement de ce corpus révèle un problème particulier : les textes étudiés sont de nature hétérogène et ne partagent pas la même structure. Néanmoins, on peut y retrouver un élément commun : le paragraphe. En effet, dans les textes juridiques, le principe qui prévaut est que le paragraphe définit l'unité sémantique de base et que son achèvement termine le traitement d'un point particulier du document. La structure des textes que nous étudions est donc à deux étages ; nous avons à la fois le document lui-même et les paragraphes qui le composent. À l'exception des phrases qui sont à la base de tout document mais qui sont des unités trop petites, les seuls sous-ensembles statiques communs à l'ensemble des documents sont les documents eux-mêmes ou les paragraphes.

Dans le travail que nous présentons, la recherche des sous-ensembles dynamiques et l'utilisation des travaux relatifs à la segmentation thématique n'ont pas été expérimentés en raison de leur complexité. La recherche de segments thématiques cohérents sur notre corpus de référence est un problème important, qui ne doit pas être considéré hâtivement comme un sous-problème de notre méthode de corrélation entre documents. Ensuite nous avons déjà fait remarquer que les documents juridiques sont naturellement segmentés en paragraphes homogènes sémantiquement. Nous supposons dès lors que ces paragraphes constituent une bonne approximation des segments thématiques qui auraient pu être retrouvés après analyse.

Ce raccourci permet d'écourter les expérimentations menées dans le cadre de la sélection de l'unité de recherche optimale. Dans l'optique d'une amélioration de notre méthode, il va de soi que cette recherche devrait inclure des sous-ensembles dynamiques. Pour l'instant, cette voie dynamique n'étant pas prise en compte, l'unité de recherche ne peut être choisie que parmi deux éléments : le document ou le paragraphe.

6.2 Une approche systématique

Le but de la section 6.1 était de réduire le nombre de possibilités inhérentes au choix de l'unité de recherche. Un doute persiste quant à la taille du segment à considérer : le document ou le paragraphe. La finalité de la présente section est de lever l'incertitude restante par le biais de l'expérimentation. Pour déterminer expérimentalement le meilleur segment, l'idée est d'interroger notre corpus de référence par l'intermédiaire d'un panel de requêtes pour lesquelles on fera varier la taille de l'unité de recherche, celle-ci prenant successivement les deux valeurs évoquées.

Cette expérimentation est évidemment liée à l'obtention de résultats corrélés et donc en relation avec la création des signatures lexicales. Nous pouvons donc utiliser directement les acquis du chapitre 5 pour créer à partir d'un panel de textes sélectionnés les différentes signatures lexicales, une pour chaque pondération envisagée. Nous utiliserons ensuite ces signatures comme des requêtes conceptuelles afin d'interroger notre base documentaire pour y analyser les résultats obtenus.

Malgré tout, plusieurs interrogations subsistent concernant la génération des signatures lexicales. Nous avons évoqué au chapitre 4 que plusieurs degrés de liberté restaient indéfinis ; cela concernait notamment la taille de la signature ainsi que le seuil d'acceptation des requêtes conceptuelles. Seule avancée en la matière, le chapitre 5 a apporté quelques éléments de réponse en montrant que pour notre corpus de référence le nombre de termes de la signature peut être défini par une constante ; sa valeur n'a malheureusement pu être déterminée.

Afin d'alléger le discours, nous utiliserons par la suite les notations suivantes :

- F_{lon} : pour facteur de longueur, désigne le nombre de termes d'une signature lexicale ;
- F_{sem} : pour facteur sémantique, désigne le seuil d'acceptation d'une requête conceptuelle, c'est à dire le nombre de termes devant être simultanément présents dans la même unité de recherche pour que le document soit accepté.

La problématique est que les paramètres manquants, à savoir l'unité de recherche, F_{lon} et F_{sem}, sont interdépendants. Le problème est donc circulaire. Pour déterminer la valeur de l'unité de recherche, nous avons besoin de connaître F_{lon} et F_{sem} et inversement. Dans ces conditions, nous décidons d'adopter une attitude algorithmique quant à la création des signatures lexicales et à l'interrogation de la base pour limiter les effets de F_{lon} et F_{sem} sur la valeur de l'unité de recherche.

Cette section 6.2 s'articule comme suit. Dans un premier temps, section 6.2.1, nous définissons le panel de documents qui sera utilisé par la suite dans la section 6.2.2, dans laquelle nous décrivons le principe de l'expérimentation.

6.2.1 Création d'un panel de documents

Le panel de documents considéré réutilise en partie les documents ayant servi à l'expérimentation de la section 5.4.1. Tous ces documents n'ont cependant pas été retenus : seuls ceux comportant au minimum deux paragraphes ont été pris en considération. La motivation de cette partie étant de comparer des recherches menées avec des unités de recherche différentes, il fallait pouvoir distinguer les résultats et donc disposer de documents constitués de plusieurs paragraphes.

Les documents sont composés de plusieurs paragraphes. Le paragraphe ne peut pas être défini au travers d'une unité de longueur traditionnelle, car il n'a pas de taille prédéfinie. La fin d'un paragraphe coïncide avec la fin du traitement d'un point particulier du document ; cette transition sémantique ne permet hélas pas de les différencier automatiquement.

On ne peut que s'appuyer sur quelques règles typographiques simples pour faire cette distinction. Ces règles ont été étendues sous la forme d'heuristiques nous permettant d'identifier ces paragraphes. Ces heuristiques, conçues pour des documents HTML, tirent parti des indices donnés par le balisage propre au langage.

En pratique, deux paragraphes sont séparés par des marques typographiques telles que :

- la définition d'un paragraphe (utilisation de la balise <P>) ;
- la fin d'une phrase (via un point) et le retour volontaire à la ligne (utilisation des tags
,<LI>, etc) ;
- la définition d'un titre (utilisation des balises <H1>, <H2>, etc) ;
- l'utilisation d'un mot-clef de début de section (Article, Chapitre, etc).

La liste de documents servant, finalement, à notre expérimentation est un sous-ensemble des documents décrit en Annexe A et a été reportée en Annexe B.

6.2.2 Procédure d'interrogation systématique

Le principe est de créer à partir du panel de documents défini en Annexe B un ensemble de signatures lexicales, une pour chaque document et chaque pondération étudiés.

Pour chaque signature, on génère l'ensemble des requêtes qui lui sont associées. Il s'agit de l'ensemble des combinaisons réalisables en prenant P termes parmi N sans tenir compte de l'ordre des termes de la requête, N étant égal au nombre de termes de la signature considérée et P tel que $1 \leq P \leq N$. Une fois les requêtes construites, on interroge le corpus de référence pour obtenir des résultats et les analyser.

L'intérêt de l'expérimentation n'est pas de savoir si les résultats obtenus sont en adéquation avec le texte initial, mais de définir l'unité de recherche grâce à l'étude de critères choisis. De tels critères sont : le taux d'utilisation des mots dans les requêtes, le nombre de requêtes valides[1] ou encore le choix du dernier terme de la requête. La principale difficulté est de déterminer les valeurs de F_{lon} et F_{sem} pour qu'ils n'influencent pas le choix de l'unité de recherche.

Parmi les degrés de liberté évoqués au chapitre 4, il en est un qui n'a pas encore été abordé : les types des signatures lexicales. Les signatures peuvent être de type homogène et composées uniquement de mots ou de SN, ou bien de type mixte, composées d'un savant mélange des deux. Les chapitres à venir n'étudient que des signatures de type homogène.

Pour F_{lon}, le parti a été de prendre comme valeur un nombre suffisamment grand pour générer une grande quantité de requêtes. Ce nombre doit également être inférieur à la taille (en nombre de termes) du plus petit document de la sélection pour qu'il y ait des différences entre les listes de termes pondérés. De cette manière, on peut vérifier que les différences observées sont dues au choix de l'unité de recherche et non aux pondérations étudiées. Le texte le plus petit comportant 28 termes, une valeur de 20 pour F_{lon} a été retenue.

Enfin, dernier point à aborder concernant l'interrogation de la base : le temps de traitement de l'expérimentation. Jusqu'à présent, nous n'avions pas décrit les aspects algorithmiques de nos différents traitements car ils ne présentaient pas de difficultés particulières. Or dans le cas présent nous sommes confrontés à un impératif qu'il nous faut aborder. En effet, nous avons décidé de prendre en compte des signatures composées de 20 termes ; l'ensemble des requêtes uniques qu'il est possible de générer à partir de ces 20 termes représente $2^{20} - 1$ requêtes. Le moteur de recherche que nous utilisons, Pertimm [Per], est l'un des plus rapides du marché mais prend, tout de même, entre 30 milli-secondes (ms) et 300 ms pour répondre à une question avec un temps de réponse moyen proche de 200 ms [2]. Avec de tel résultats, il nous faut un peu plus de 58 heures pour générer toutes les réponses possibles d'une signature lexicale composée de 20 termes. Il devient alors difficile de mener à bien notre expérimentation sachant que pour chaque document (au nombre de 20, voir Annexe B), il faudrait refaire l'interrogation pour chaque pondération envisagée (au nombre de 5), pour chaque unité de recherche à tester (au nombre de 2) ainsi que pour chaque type de terme (au nombre de 2).

Au final, on estime le temps de calcul à 1000 jours ce qui n'est pas envisageable ; on doit donc opérer quelques choix et optimisations pour mener à bien cette expérimentation.

Tout d'abord, nous ne prendrons pas en compte toutes les pondérations évoquées au chapitre 5. Les pondérations primaires telles que T_f et I_{df} ne seront pas abordées car elles

[1] Requêtes admettant au moins deux réponses.

[2] Les temps de calculs énoncés ont été mesurés en interrogeant notre corpus de référence à l'aide d'une station de travail équipée d'un Pentium IV à 1,7 GHz, dotée de 1 Go de RAM à 133 MHz et d'un disque dur de 40 Go, 7200 rpm IDE ATA 100.

ne donnent pas a priori de résultats satisfaisants ; de plus leurs comportements peuvent se déduire de celui de $T_f.I_{df}$. Ces deux pondérations sont donc temporairement mises de coté.

Ensuite nous divisons le temps de calcul par deux en n'étudiant que des signatures lexicales homogènes d'un seul type. Les travaux menés au chapitre 5 ont montré que la recherche et la génération des signatures lexicales sont délicates lorsque celles-ci sont composées de mots. Pour éviter de nous placer dans des conditions idéales éloignées de la réalité, l'expérimentation menée dans ce chapitre 6 ne portera que sur des signatures homogènes composées exclusivement de mots.

Une fois ces deux choix opérés, l'expérimentation menée nécessite encore plus de 150 jours de temps de calcul. Pour éviter d'interroger systématiquement la base documentaire avec des requêtes inutiles (dont la réponse peut se déduire des précédentes), nous employons un algorithme de parcours d'arbre en ordre préfixe que nous modifions afin d'accélérer le traitement.

6.2.2.1 Parcours d'arborescence en ordre préfixe

Cette section présente un algorithme dont le but est de parcourir une arborescence suivant un ordre préfixe le plus rapidement possible. Dans notre cas, l'arborescence représente les questions ou requêtes qu'il est possible de générer à partir d'une signature lexicale, soit virtuellement un arbre de $2^{20} - 1$ questions. La figure 6.1 illustre le parcours de l'arbre des questions que l'on peut générer à partir de trois termes.

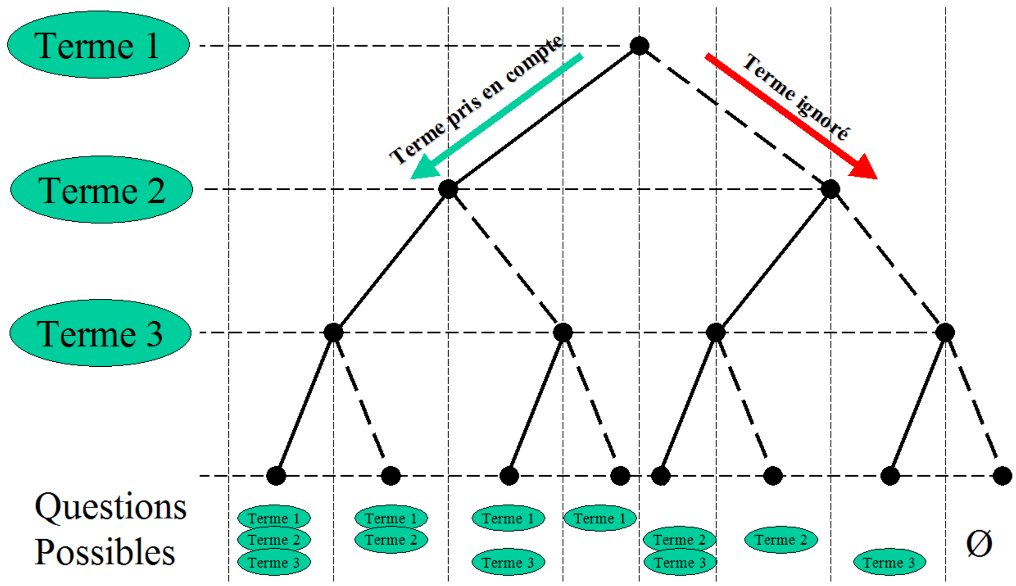


FIG. 6.1 – *Parcours de l'ensemble des questions possibles grâce à un arbre binaire*

Nous cherchons à optimiser le parcours de l'arbre pour ne sélectionner que des questions utiles. Nous imposons une condition d'arrêt pour ne pas développer une branche si les conditions ne sont plus réunies ; cette condition d'arrêt dépend de l'heuristique décrite ci-après, notée ζ_A. Le parcours d'arbre ne développe ainsi que les branches utiles et de ne garde que des requêtes valides. Un tel algorithme de parcours est illustré par la figure 6.2.

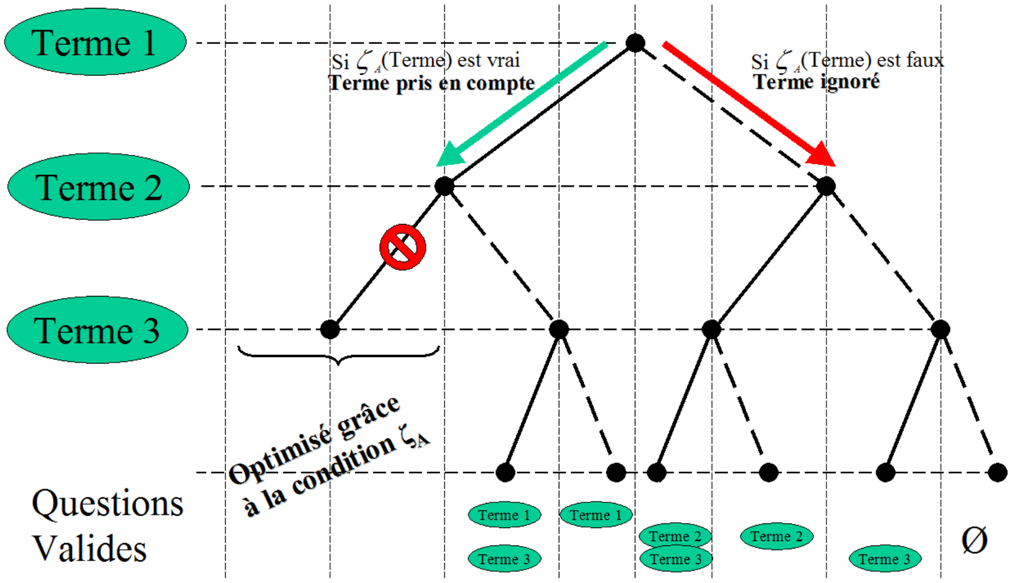


FIG. 6.2 – *Parcours de l'arbre des questions après optimisation*

La condition d'arrêt ζ_A permet de n'étudier que des requêtes utiles. Pour être utile, une requête doit nécessairement obtenir au moins deux réponses. En effet, une seule réponse n'est pas suffisante puisque celle-ci peut provenir du document d'origine de la signature. De telles requêtes seront par la suite définies sous le terme de *requêtes valides*, notion qui désigne *toutes requêtes admettant un minimum de deux réponses distinctes.*

Tous les termes de la requête doivent être présents dans l'unité de recherche réponse. Autrement dit, F_{sem} permet d'avoir une valeur égale au nombre de termes de la requête considérée. Nous n'avons pas pris en compte une valeur de F_{sem} inférieure au nombre de termes de la signature pour deux raisons :

- pour accélérer le déroulement de l'algorithme, puisque l'on réduit ainsi le nombre de branches à explorer ;
- pour éviter de traiter des requêtes inutiles.

Pour illustrer cette deuxième raison, voici un exemple simple. Considérons une requête composée de quatres termes $R = \{T_1, T_2, T_3, T_4\}$ avec $F_{sem} = 3$. On constate que T_1, T_2, T_3, T_4 sont répartis dans plusieurs documents et toujours distribués dans les mêmes unités de recherche. R est donc une requête valide d'après notre définition.

On considère maintenant la requête R' qui s'écrit de la manière suivante $R' = \{T_1, T_2, T_3, T_4, T_5\}$

où T_5 est un terme qui n'est jamais présent dans les unités de recherche où sont présent les autres termes T_1, T_2, T_3, T_4, si l'on excepte le document d'origine de la signature. Toujours d'après notre définition, R' est également une requête valide que nous devrions prendre en compte alors qu'elle n'apporte rien par rapport à R. Pour éviter ce désagrément nous imposons une valeur de F_{sem} égale à la taille de la requête étudiée.

En résumé, pour créer nos signatures lexicales et interroger la base, nous imposons que la requête admette au moins deux réponses et que pour chacune de ces réponses, tous les termes de la signature soient présents au minimum une fois dans l'unité de recherche considérée.

L'algorithme de parcours d'arbre décrit dans le tableau 6.1 permet de générer l'ensemble des questions valides à partir d'une signature lexicale de longueur F_{lon}.

Cette utilisation de l'algorithme permet de gagner du temps pour construire l'ensemble E des requêtes valides pour chaque document, chaque pondération et surtout pour chaque unité de recherche considérée. Le gain de temps n'est toutefois pas suffisant, car on estime que cette première version de l'algorithme demande encore 50 jours de calcul pour venir à bout des calculs.

Nous améliorons encore les performances de notre algorithme en réutilisant les données déjà enregistrées dans E et en constatant que toute requête incluse dans une requête valide est elle-même valide sans qu'il soit nécessaire de procéder à une interrogation de la base documentaire. On obtient alors une deuxième version de l'algorithme qui sera cette fois-ci la version définitive. Cette version est décrite dans le tableau 6.2.

Avec cette deuxième version de l'algorithme, nous arrivons à générer l'ensemble des résultats pour les 20 documents considérés en un peu moins d'une semaine. Le processus s'avère encore long, mais cette fois-ci il devient acceptable et démontre un gain de temps considérable par rapport à une recherche systématique non optimisée estimée à 152 jours.

Soit un ensemble E qui contiendra au final toutes les requêtes valides. Cet ensemble est initialement vide.
Soit ζ_A la condition d'arrêt qui teste la validité d'une requête ; elle retourne *vrai* si la requête est valide et *faux* dans le cas contraire.
Soit P_R une procédure récursive qui prend en paramètre deux listes de termes ordonnées par pertinence décroissante. On initialise P_R avec les listes suivantes :
- $S = \{T_1, T_2, \ldots, T_{20}\}$, la signature initiale ;
- $R = \emptyset$ une requête initialement vide.

$P_R(S, R)$
Début
  Si $S \neq \emptyset$
  Alors
    Soit T le premier terme de S et $S' \leftarrow \emptyset$ une liste de termes
    $S' \leftarrow S - \{T\}$
    Soit R' une requête telle que $R' \leftarrow R + \{T\}$
    $F_{sem} \leftarrow \|R'\|$
    Si $\zeta_A(R')$ est *vrai*
    Alors
      R' est une requête valide
      $E \leftarrow E + R'$
      On réitère la procédure avec $P_R(S', R')$
    Fin Si
    R' n'est pas une requête valide
    On réitère la procédure avec $P_R(S', R)$
  Fin Si
Fin

TAB. 6.1 – *Première version de notre algorithme de parcours d'arbre*

Soit un ensemble E qui contiendra au final toutes les requêtes valides. Cet ensemble est initialement vide.
Soit ζ_A la condition d'arrêt qui teste la validité d'une requête ; elle retourne *vrai* si la requête est valide et *faux* dans le cas contraire.
Soit P'_R une procédure récursive qui prend en paramètre deux listes de termes ordonnées par pertinence décroissante. On initialise P_R avec les listes suivante :
- $S = \{T_1, T_2, \ldots, T_{20}\}$ la signature initiale ;
- $R = \emptyset$ une requête initialement vide.

$P'_R(S, R)$
Début
  Si $S \neq \emptyset$
  Alors
    Soit T le premier terme de S et $S' \leftarrow \emptyset$ une liste de termes
    $S' \leftarrow S - \{T\}$
    Soit R' une requête telle que $R' \leftarrow R + \{T\}$
    $F_{sem} = \|R'\|$
    Si R' est une sous-requête non nulle de E **OU** $\zeta_A(R')$ est *vrai*
    Alors
      R' est une requête valide
      $E \leftarrow E + R'$
      On réitère la procédure avec $P'_R(S', R')$
    Sinon
      R' n'est pas une requête valide
      On réitère la procédure avec $P'_R(S', R)$
    Fin Si
  Fin Si
Fin

TAB. 6.2 – *Version améliorée de notre algorithme de parcours d'arbre*

6.3 Expérimentation

L'algorithme défini, la phase d'étude des résultats peut débuter, phase qui est l'objet de la section 6.3.1. La section 6.3.2 conclura ce chapitre 6, en revenant sur les enseignements que l'on peut tirer de cette expérimentation pour choisir de l'unité de recherche mais également pour comparer les pondérations étudiées.

6.3.1 Résultats

Pour analyser les résultats de l'expérimentation, nous allons nous intéresser à plusieurs critères tels que :

- le nombre de questions générées ;
- le taux d'utilisation des termes ;
- le taux d'utilisation du dernier terme.

Chacun de ces critères apporte de précieux renseignements. Ainsi la connaissance du nombre de questions générées donne un indice de leur qualité. Trop de questions valides peut prouver deux choses : une unité de recherche trop grande ou des signatures toujours bien choisies. À l'opposé, un nombre insuffisant de questions montre que les unités sont trop petites ou que les signatures ne sont pas pertinentes.

Le taux d'utilisation des termes complète la précédente étude, avec des signatures de longueur constante ($F_{lon} = 20$) ; les termes les plus significatifs devraient être normalement classés parmi les premiers éléments de la liste. Un taux d'utilisation des derniers termes anormalement élevé peut être un signe précurseur pour nous indiquer que le segment de recherche est trop grand. Le recoupement des différentes sources d'information comme la comparaison entre les différentes pondérations et l'utilisation de résultats moyennés sur l'ensemble des documents listés en Annexe B permet de lever les incertitudes qui pèsent sur l'interprétation de ces même résultats.

Nous traitons des requêtes composées de termes de rangs différents ; de ce fait nous ne pouvons pas comparer toutes les requêtes valides trouvées sans prendre quelques précautions au préalable. Pour illustrer le type de précautions à prendre en considération, étudions l'exemple, voir tableau 6.3, d'une signature quelconque S composée de 5 termes ainsi que R_1 et R_2 deux requêtes valides issues de S.

Dans le cas présent, les deux requêtes sont valides et pourtant elles n'ont aucun terme en commun, même si elles sont issues de la même signature. Pour comparer des requêtes valides provenant d'une même signature, il nous semble plus judicieux de les regrouper en fonction de points communs et ensuite seulement de les comparer.

$S \quad = \{ T_1, T_2\ , T_3, T_4, T_5\}$
$R_1 = \{ T_1, T_2\ \}$
$R_2 = \qquad\qquad \{T_3, T_4, T_5\}$

TAB. 6.3 – *Exemple de requêtes sans terme pivot commun issue d'une signature S composée de cinq termes*

Dans ce but, nous introduisons la notion de *terme pivot* que nous définissons comme étant *le premier terme de la requête.* Ainsi dans l'exemple précédent, T_1 était le terme pivot de R_1 et T_3 celui de R_2, l'intérêt étant de pouvoir regrouper les requêtes en fonction de leur pivot respectif.

A notre avis, il est ainsi plus aisé de comparer des requêtes valides partageant le même pivot comme dans l'exemple décrit tableau 6.4 plutôt que dans celui du tableau 6.3, surtout lorsque l'on tente de comparer des taux d'utilisation de termes dans des requêtes.

$S \quad = \{ T_1, T_2, T_3, T_4, T_5\}$
$R'_1 = \{ T_1\}$
$R'_2 = \{ T_1, T_2\}$
$R'_3 = \{ T_1, T_2, T_3\}$
$R'_4 = \{ T_1, T_2, T_3, T_4\}$
$R'_5 = \{ T_1, \qquad T_3\}$
$R'_6 = \{ T_1, \qquad\qquad T_4\}$
...

TAB. 6.4 – *Exemple de requêtes avec un terme pivot commun issue d'une signature S composée de cinq termes*

Dans les résultats donnés par la suite, nous comparons des requêtes ayant le même pivot. Celui qui a été retenu comme pivot de référence est le premier terme de chaque liste pondérée. C'est en prenant ce pivot particulier que nous devrions observer les différences les plus marquantes entre les deux unités de recherche considérées.

6.3.1.1 Le nombre de requêtes valides générées

Le premier critère que nous étudions est le nombre de requêtes valides que nous générons en moyenne pour l'ensemble des documents listés en Annexe B, suivant les trois pondérations envisagée. La figure 6.3 présente ainsi les résultats obtenus pour une unité de recherche égale au document, alors que la figure 6.4 prend en compte le paragraphe.

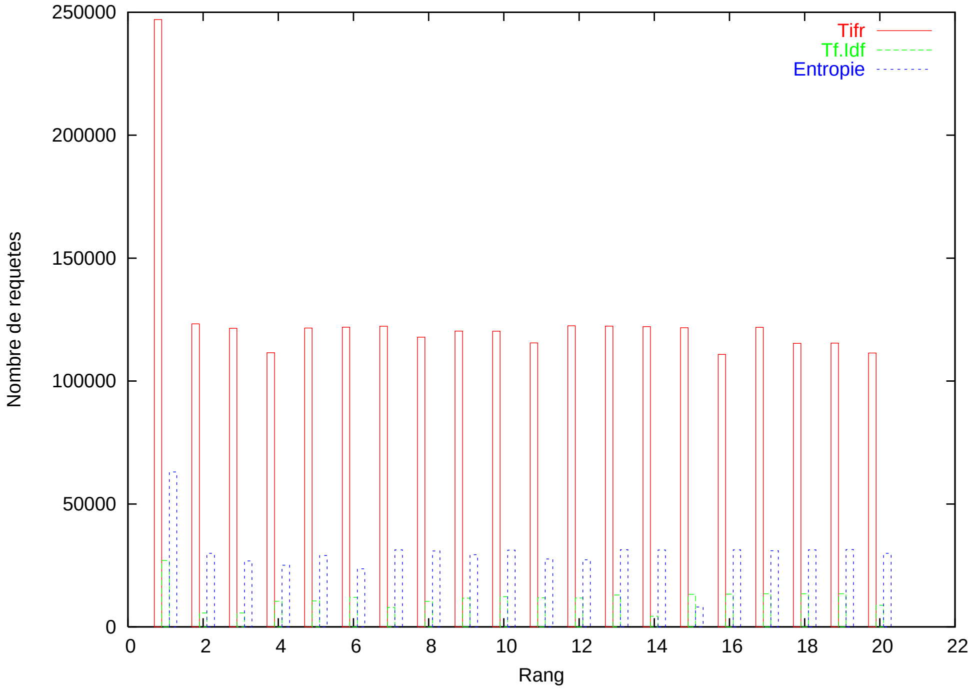


FIG. 6.3 – *Nombre de requêtes valides contenant le mot de rang N avec le document comme unité de recherche. Les mots sont ordonnés selon la pondération décroissante de l'un des 3 indices.*

Entre ces deux approches, le nombre de requêtes valides retrouvées est différent. En effet, avec le paragraphe, le nombre de requêtes valides ne dépasse pas 700 réponses alors qu'en considérant le document, celui-ci dépasse allègrement 200.000 réponses.

La différence est suffisamment importante pour montrer d'une part, l'enjeu du choix de l'unité de recherche car celle-ci joue a priori un rôle clef, d'autre part, que le document semble être une unité de trop grande taille.

Un nombre de 200.000 réponses valides n'est pas très éloigné du nombre maximum de requêtes possibles de $2^{19} - 1 = 524287$ (un terme pivot et 19 termes indépendants). Ce constat est un autre élément négatif pour le document, car cela signifie que 40% des requêtes potentiellement envisageables ont été validées. Par expérience, nous savons pertinemment que dans des listes de mots générées automatiquement, un certain nombre de termes parasites apparaissent. Il est donc très peu probable qu'une requête sur deux soit valide d'un point de vue sémantique.

6.3.1.2 Le taux d'utilisation des termes

Dans un deuxième temps, nous nous intéressons au taux moyen d'utilisation des termes. La figure 6.5 représente ce taux pour les différentes pondérations étudiées pour une unité

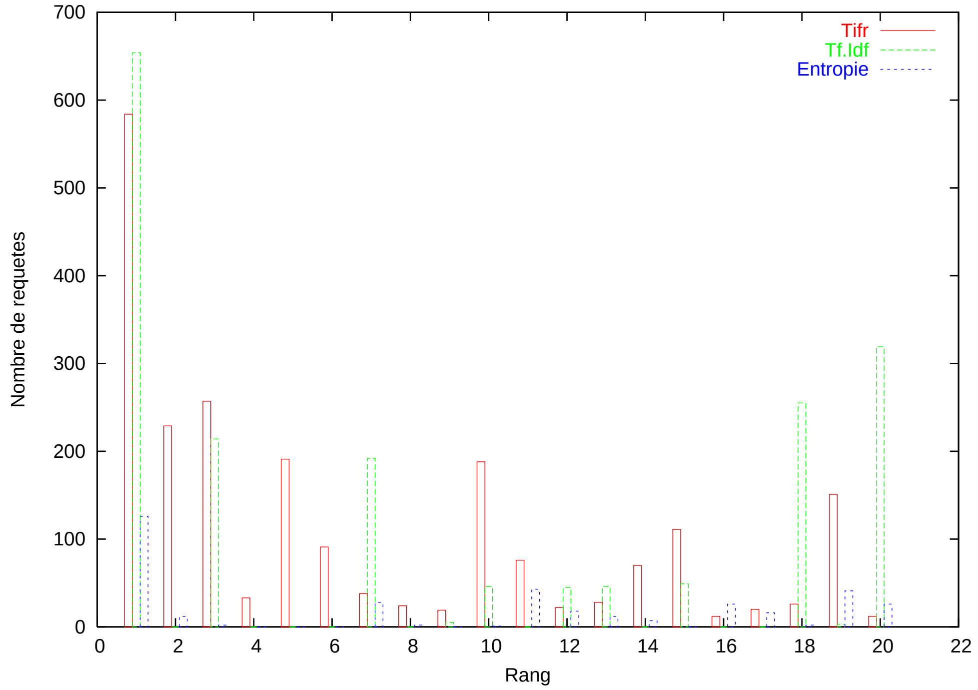


FIG. 6.4 – *Nombre de requêtes valides contenant le mot de rang N avec le paragraphe comme unité de recherche. Les mots sont ordonnés selon la pondération décroissante de l'un des 3 indices.*

équivalente au document, alors que la figure 6.6 traite le cas du paragraphe.

Force est de constater, comme évoqué précédemment, que cela n'a rien d'anormal puisque l'on étudie les mêmes chiffres, mais sous des angles différents. Les résultats sont présentés sous la forme de pourcentages pour s'attacher aux variations du taux d'utilisation des termes plutôt qu'au nombre de fois où ils ont été utilisés.

Les deux figures sont singulièrement différentes ; la première nous montre une évolution quasi-nulle de la variation du taux d'utilisation, et ce quelle que soit la pondération considérée, alors qu'au contraire la figure 6.6 traitant du cas des paragraphes présente une décroissance plus marquée pour des termes de rang plus élevé. En toute logique, les termes les plus intéressants doivent être trouvés parmi les premiers éléments de la liste. L'absence de variation est un indice inquiétant qui trahit là encore une unité de recherche mal adaptée.

6.3.1.3 Le rang moyen du dernier terme

Le dernier indice analysé dans cette expérience est le rang moyen du dernier terme. Le principe de cette expertise est de vérifier l'idée pressentie que les documents sont des unités de recherche trop grandes, avec l'étude du taux d'utilisation des termes de rang élevé.

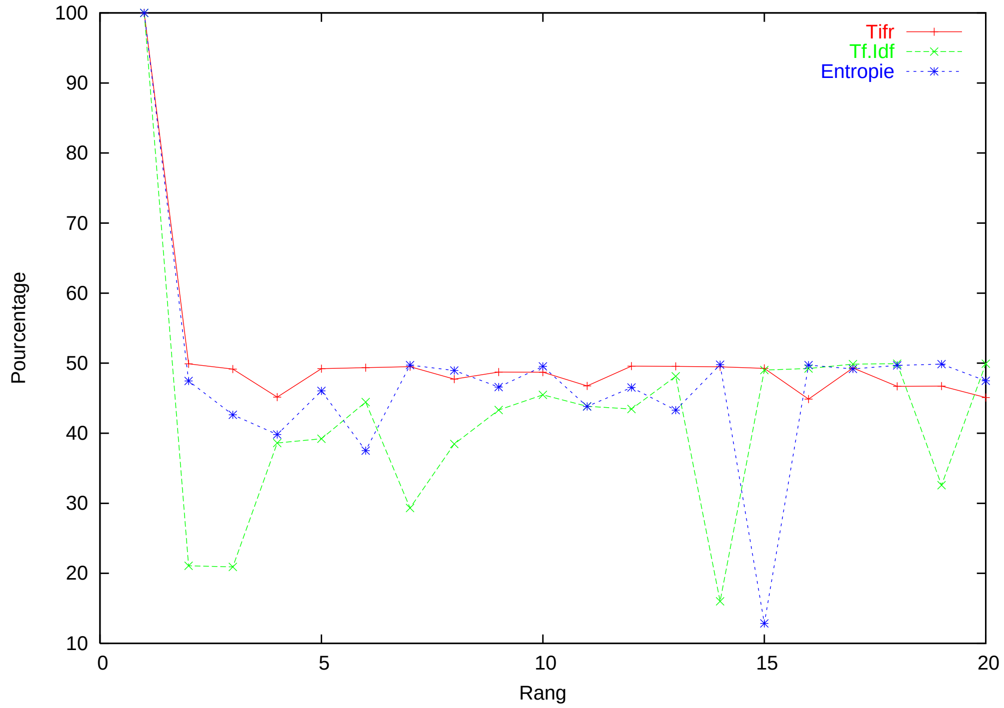


FIG. 6.5 – *Taux d'utilisation des mots dans les questions en prenant le document comme unité de recherche. Les mots sont ordonnés selon la pondération décroissante de l'un des* 3 *indices.*

Cette intuition est confirmée par l'étude des deux graphiques qui traduisent, pour la figure 6.7, le rang moyen du dernier terme lorsque l'unité considérée est le document, et le paragraphe pour la figure 6.8.

La différence est encore en défaveur du document. En effet, on constate sur la figure 6.7 que le dernier terme le plus couramment utilisé est, quelle que soit la méthode de pondération utilisée, le terme de rang le plus élevé ; cela signifie que la majeure partie des requêtes utilisant comme pivot le terme de rang le plus faible associe ce terme à celui de rang le plus élevé, qui est normalement le terme le moins pertinent de la liste.

La figure 6.8 trahit également une augmentation du taux d'utilisation du dernier terme pour un rang élevé, mais dans des proportions moindre. Cette augmentation est normale car, au vu des pondérations utilisées, plus le terme est de rang élevé, plus la fréquence de celui-ci augmente dans le corpus. Il est donc normal de constater cette légère augmentation ; le tout est de trouver une unité qui n'amplifie pas ce défaut propre aux pondérations statistiques que nous utilisons. À titre de remarque, la définition d'un nombre maximum de termes par signatures (F_{lon}) est également un moyen pour ne pas prendre en compte les termes les plus fréquents du document. Cette parenthèse refermée, on constate que le document pris comme unité de recherche amplifie ce désagrément.

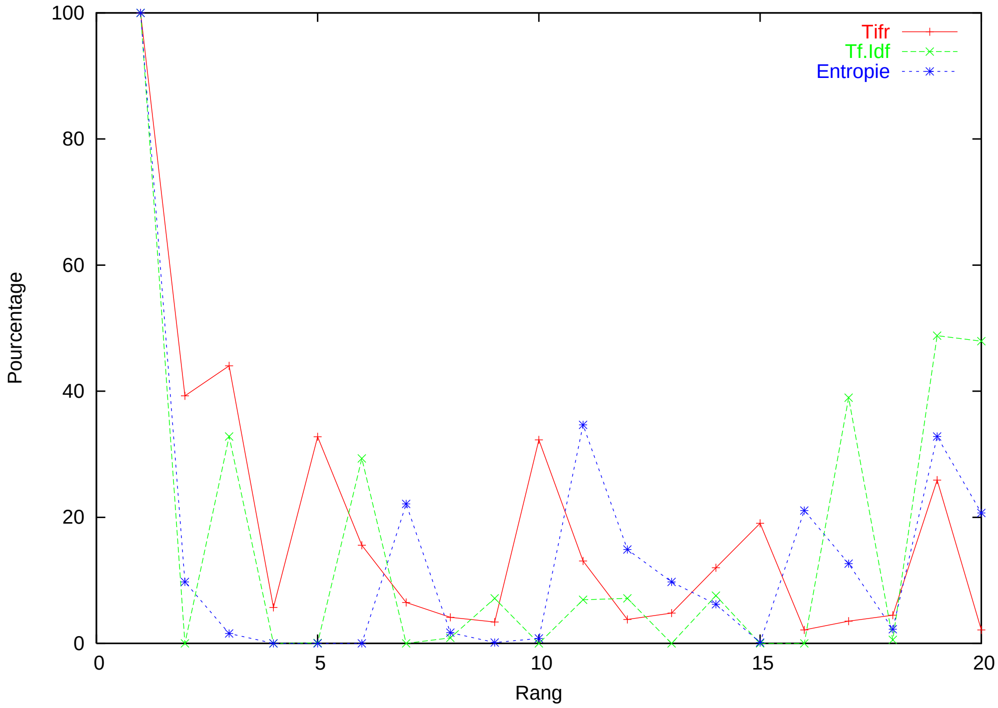


FIG. 6.6 – *Taux d'utilisation des mots dans les questions en prenant le paragraphe comme unité de recherche. Les mots sont ordonnés selon la pondération décroissante de l'un des 3 indices.*

6.3.2 Enseignements

Les enseignements que l'on peut tirer de ces résultats sont doubles. Il y a tout d'abord des conclusions évidentes qui s'imposent en terme de choix de l'unité de recherche. Mais il y a également des conclusions à apporter quant aux caractéristiques des pondérations évoquées qui en trahissent certains aspects négatifs.

6.3.2.1 L'unité de recherche

Les différents résultats précédemment obtenus nous ont apporté quelques précisions sur l'unité de recherche à utiliser. Tout d'abord il faut remarquer que les résultats visualisés sont relativement homogènes et surtout dépendent peu des pondérations étudiées. Cette remarque a son importance car, pour pouvoir conclure sur l'unité à choisir, il faut d'abord que notre expérimentation soit pertinente. Cette invariance des résultats vis-à-vis de la pondération utilisée nous permet à ce titre de juger de la validité de celle-ci. Ainsi nous pouvons exploiter en toute quiétude les résultats afin de de choisir, entre document et paragraphe, l'unité de recherche la plus adaptée à notre corpus de référence.

Les différents tableaux et figures exploités dans la section 6.3.1 nous ont donné de nombreux indices permettant de définir notre choix :

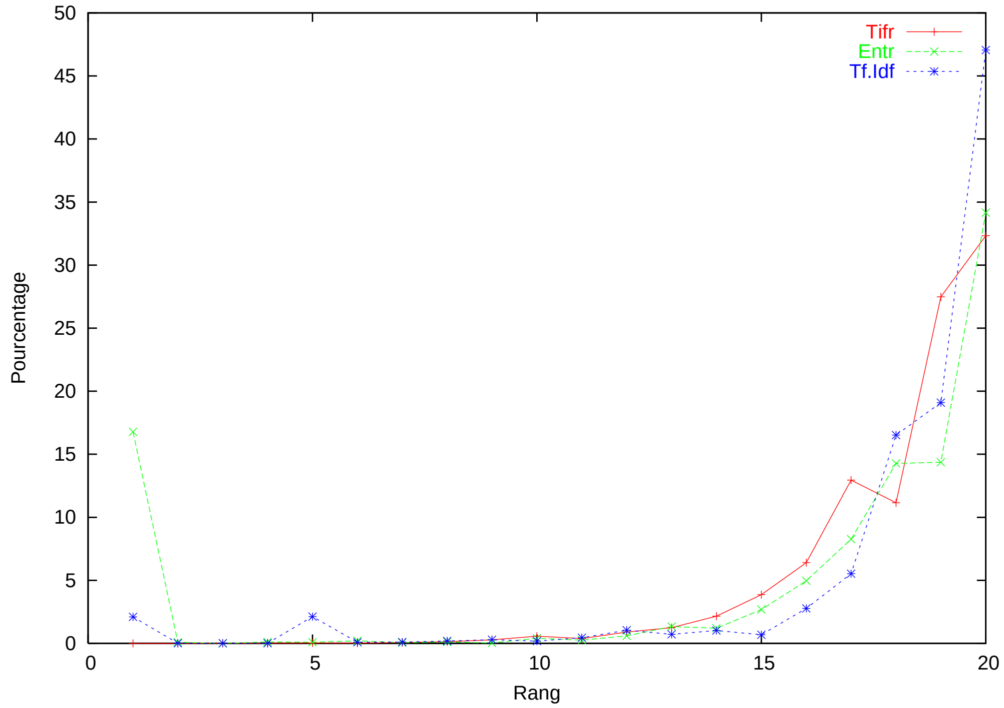


FIG. 6.7 – *Rang moyen du dernier mot constituant une question valide en prenant le document comme unité de recherche. Les mots sont ordonnés selon la pondération décroissante de l'un des* 3 *indices.*

- l'importance du nombre de requêtes valides moyennes générées montre que le document est un segment trop large, qui accepte malheureusement tous les termes sans distinction ;
- le taux d'utilisation moyen des termes, quasi-constant dans le cas du document, renforce l'idée précédente que ce segment de recherche est trop large et ne permet pas d'opérer de choix entre les différents termes présents. Du second au dernier terme, tous se retrouvent dans un nombre équivalent de requêtes, quelle que soit leur pertinence ;
- dans le même esprit, le rang moyen élevé du dernier terme, toujours dans le cas du document, laisse penser que cette unité n'est pas le choix à retenir.

Pour ces raisons, notre choix d'unité de recherche se tourne naturellement vers le paragraphe. Nous n'irons pas jusqu'à dire que cette unité est le meilleur choix envisageable mais plutôt qu'il s'agit du choix qui minimise les défauts précédemment évoqués. Ainsi, si l'on considère le rang moyen du terme des requêtes valides pour une unité de recherche égale au paragraphe (figure 6.8), on constate que ce taux présente également une légère croissance pour des rangs élevés mais que, contrairement au cas du document, cette croissance reste raisonnable.

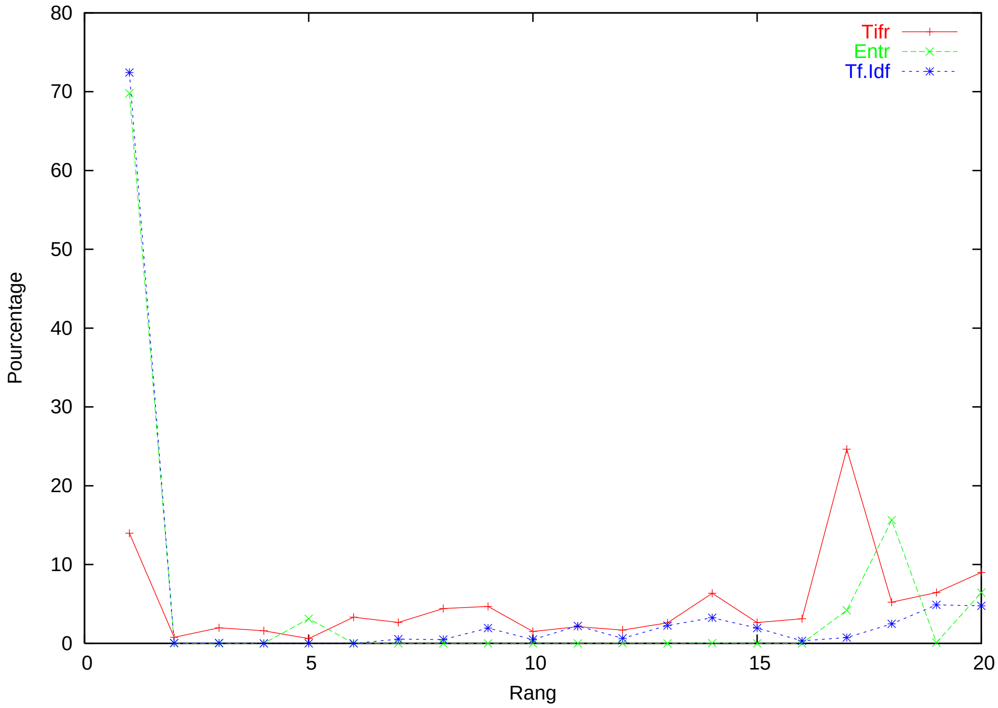


FIG. 6.8 – *Rang moyen du dernier mot constituant une question valide en prenant le paragraphe comme unité de recherche. Les mots sont ordonnés selon la pondération décroissante de l'un des* 3 *indices.*

6.3.2.2 Comparaisons des différentes pondérations

Cette expérience, dont le but initial était de répondre la question suivante : *Quelle unité de recherche pour notre corpus de référence ?*, permet également de comparer indirectement les pondérations présentes et d'émettre quelques jugements quant à leurs qualités respectives.

Au premier abord, les pondérations ont pour cette expérimentation des comportements assez proches. Néanmoins, en s'y intéressant de plus près, on note quelques différences.

La première qui apparait oppose notre pondération ad hoc T_{ifr} aux autres pondérations étudiées ($T_f.I_{df}$ et entropie). Si on prend en considération les figures 6.4 et 6.6, on constate certaines annulations de valeurs pour les courbes $T_f.I_{df}$ et d'entropie et ce pour des valeurs de rang faible.

Cela signifie que certains termes apparemment bien classés par ces méthodes de pondérations classiques ne se retrouvent pas associés au terme pivot dans un même paragraphe. On pourrait supposer que cela est dû au choix particulier de l'unité de recherche : le paragraphe. Mais l'examen des figures 6.3 et 6.5 montre que ce phénomène existe également avec l'autre unité de recherche, et que le pararaphe ne fait que l'amplifier.

D'un point de vue pratique, cette caractéristique peut être choquante, notamment si l'on tente de constituer des signatures lexicales de petites tailles (moins de cinq termes choisis parmi les premiers éléments pondérés). À un niveau sémantique cette constatation est toute aussi gênante, car elle montre que les premiers termes retrouvés par ces deux méthodes classiques ne partagent pas les même unités de recherche et donc vraisemblablement pas les mêmes contextes d'utilisation.

Troisième partie

Évaluation

Chapitre 7

Optimiser la construction des signatures lexicales pour la recherche des résultats corrélés

Le chapitre 4 nous a montré qu'il existe de nombreux degrés de liberté pour réaliser une méthode de corrélation telle que nous l'envisageons. Les paramètres optimaux de certains de ces degrés de liberté ont pu être trouvés ou déduits expérimentalement. Néanmoins, il en reste deux, précédemment évoqués, non encore déterminés : F_{lon} et F_{sem}.

L'objectif de ce chapitre est de déterminer les valeurs optimales de ces deux paramètres afin de générer automatiquement des signatures lexicales dédiées à la corrélation. Les valeurs qui seront trouvées dépendent évidemment du corpus de référence envisagé.

L'articulation de ce chapitre 7 est la suivante. La section 7.1 aborde les problèmes liés à la définition d'un benchmark dédié à la corrélation, seul moyen permettant de mesurer la qualité des documents retrouvés. La section 7.2 présente une expérimentation qui affine les valeurs de F_{lon} et F_{sem} afin de maximiser le nombre de résultats corrélés retrouvés. La section 7.3 est une étape complémentaire qui introduit un algorithme de classement des documents corrélés et discute de ses effets. Finalement la section 7.4 résume les enseignements apportés.

7.1 Un benchmark expérimental dédié à la corrélation

Le problème majeur est de pouvoir évaluer simplement la pertinence de nos résultats corrélés en comparant des critères tels que la précision ou le rappel. Jusqu'à présent, les précédentes évaluations ont étudié les résultats corrélés et vérifié leur intégrité en utilisant principalement deux approches :

- en s'aidant d'indices de similarité tel que l'indice cosine [PPGK02] qui permet de vérifier que les documents en présence sont proches en terme de distance ;
- en faisant appel à une validation manuelle de la pertinence des résultats obtenus [DH99].

Ces deux types d'évaluation permettent sans aucun doute d'améliorer la précision des premiers documents retrouvés et assurent qu'ils sont pertinents et utiles pour l'utilisateur final, mais ces évaluations restent insuffisantes. Par exemple, elles ne permettent pas de prendre en compte le nombre de documents pertinents non retrouvés. De ce fait, nous essayons de résoudre ce problème en définissant et en construisant un corpus dédié à la corrélation pour ainsi mieux étudier les effets de F_{lon} and F_{sem} sur les performances globales du système.

Ce benchmark doit au final améliorer la recherche d'information sur notre portail juridique [1] ; il est donc constitué d'éléments de notre corpus de référence, à savoir le *Journal Officiel de la République Française* de l'année 2000.

7.1.1 Élaboration du benchmark

Pour évaluer la pertinence des différentes méthodes de sélection terminologique, à savoir les cinq méthodes de pondération étudiées, nous construisons un *benchmark* composé de plusieurs groupes ou *clusters* de documents corrélés extraits du corpus de référence.

Le JO a été l'objet de nombreuses études depuis nombre d'années. Les documents qui le composent ont été dès son origine classés suivant différentes catégories permettant d'accéder à l'information plus facilement. Malheureusement cette classification ne nous est pas d'un grand secours car les différentes catégories existantes ne regroupent pas les documents en fonction de leur signification. En effet, la classification en place permet de retrouver les documents en fonction de critères fonctionnels vis-à-vis des juristes les utilisant mais sans lien direct avec la sémantique des documents.

Notre travail a donc été de sélectionner un panel de thématiques non-ambiguës extraites de notre corpus de référence et d'y rattacher les documents traitant de la même thématique pour former ainsi des regroupements de documents corrélés. Comme ce travail comporte une sélection manuelle des documents, il doit être limité dans le temps et par conséquent ne pas être trop ambitieux. Ainsi, nous nous sommes limités dans un premier temps à construire un *benchmark* regroupant huit thématiques qui sont répertoriées dans le tableau 7.1.

[1] `http://www.admi.net`

Ces thématiques représentent des thèmes clairement différenciés car leur signification est la même que l'on s'adresse ou non à des experts juridiques. Elles doivent aussi répondre à un impératif de représentativité des thèmes dans le corpus. On a ainsi dans le *benchmark* le thème portant sur la *Nomination* ; ce thème, très répandu dans le corpus de référence, couvre $\frac{1}{5}$ des documents qui le composent alors qu'à son opposé le thème portant sur les *O.G.M.* (Organismes génétiquement modifiés) n'est représenté que par trois documents.

TAB. 7.1 – *Liste des thèmes et nombre de documents*

Thèmes	Descriptions	N_{doc}	L_{doc}
Nomination	Nominations dans l'administration Française	5303	478
Télécommunication	Régulation dans les télécommunications	833	162
Médical	Rembousement médicaux	108	75
Aviation	Régulation du traffic aérien	90	68
Vin	Contrôle de la qualité du vin	88	45
Permis de conduire	Législation sur le permis de conduire	17	8
Immigration	Regroupement familial	5	1
O.G.M.	Protection contre les O.G.M.	3	1

où N_{doc} est le nombre de documents par thème et L_{doc} le nombre de documents de grande taille (composés de plus de 30 mots)

La création du *benchmark* est un processus composé de deux phases :

- une phase d'extraction pour chaque thématique d'un noyau de documents corrélés ;
- une phase d'enrichissement des noyaux de documents.

7.1.1.1 L'extraction des noyaux de documents

Lors de cette première étape on cherche un noyau de documents pour chaque élément du tableau 7.1. La construction de ces différents *clusters* de documents suit trois étapes :

1. la rédaction manuelle d'une signature lexicale ;
2. l'extraction de documents en utilisant cette signature comme une requête conceptuelle (avec $F_{sem} = 2$) ;
3. le filtrage manuel des documents non pertinents.

La première étape concerne la rédaction d'une signature lexicale ; il s'agit à cette étape de choisir les termes clefs qui vont être employés pour rechercher les documents corrélés qui appartiendront à une même thématique. Cette étape est importante car de la pertinence de la question posée dépend la qualité des réponses.

Dans la deuxième étape, on interroge la base documentaire pour obtenir un panel de documents ; on notera que le facteur sémantique F_{sem} vaut deux et non pas un, qui aurait en toute logique conduit à une approche plus exhaustive. Cette valeur a été choisie pour faciliter le filtrage manuel opéré par la suite ; en effet, une valeur trop faible de F_{sem} implique un filtrage manuel de plusieurs milliers de réponses.

Dernière étape : le filtrage manuel. Cette étape assure de manière supervisée que les documents qui sont inclus dans un *cluster* y ont bien leur place. On s'aperçoit que cette construction laisse une place prépondérante à l'appréciation humaine, seule apte à juger de la valeur fournie ou non par un document.

Une fois les noyaux de documents construits, nous avons une certitude : les documents constitutifs des *clusters* sont des documents corrélés. Il est concevable que tous ne soient pas fortement liés les uns aux autres, mais néanmoins ils participent à la même thématique et répondent à la même requête conceptuelle initiale, ce qui en fait des *clusters* de documents corrélés. Cependant, cette manière d'opérer ne nous assure pas d'avoir retrouvé la totalité des éléments potentiellement corrélés ; c'est pour cette raison que l'étape suivante tend à enrichir les noyaux de documents corrélés pour obtenir des *clusters* de documents les plus complets possibles.

7.1.1.2 L'enrichissement des noyaux

Pour enrichir les noyaux de documents existants et détecter de nouveaux éléments corrélés, nous allons constituer des requêtes conceptuelles en utilisant les signatures lexicales des documents déjà identifiés. Le principe est de sélectionner, dans chacun des huit noyaux de documents précédement ébauchés, dix documents choisis aléatoirement (ou moins si le *cluster* comporte moins de dix documents). Pour chacun de ces documents on calcule alors les différentes signatures lexicales possibles, une par méthode de pondération envisagée. Pour chacune de ces signatures lexicales, on recherche les résultats potentiellement corrélés que l'on classe suivant les valeurs décroissantes de F_{sem}. Pour un document donné, on fusionne alors l'ensemble des documents corrélés généré par les cinq signatures lexicales possibles en prenant soin d'éliminer les doublons. À partir de cette liste de résultats, on ne considére que les 200 premiers, qui sont ensuite filtrés manuellement pour en extraire les documents discriminants qui serviront à enrichir le noyau de documents constitué lors de la première étape.

Cette deuxième étape nous a permis de trouver en moyenne 6% de nouveaux documents. Cette faible valeur est encourageante car elle semble indiquer un taux d'enrichissement des *clusters* relativement faible dès la première passe. Cela nous permet d'arrêter là cette phase d'enrichissement qui fait principalement appel à un travail de validation manuelle, travail nécessaire mais qu'il n'est pas utile de réitérer car les gains seraient peu importants comparés à l'investissement humain requis. Cette faible valeur nous permet également de penser qu'un nombre peu important de documents sont portés manquants et que leur absence peut être considérée comme négligeable. Le tableau 7.1 donne pour chaque thème constitutif du *benchmark* le nombre de documents (noté N_{doc}). On peut consulter les documents participant à la constitution des différents *clusters* à l'adresse suivante : `http://cri.ensmp.fr/~chotteau/benchmark.html` .

7.2 Optimisation des signatures lexicales

Cette section décrit maintenant l'expérience menée pour construire et optimiser des signatures lexicales dans le cadre de la recherche de documents corrélés. Cette optimisation des signatures passe par une étape clef qui est de pouvoir déterminer les deux seuils précédemment étudiés, à savoir F_{sem} et F_{lon}, et de comprendre leur rôle lors de la recherche des documents réponses.

7.2.1 La taille de la signature et le facteur sémantique

Le facteur sémantique F_{sem} est défini comme *le nombre minimum de termes simultanément présents dans une unité de recherche pour que le document correspondant soit considéré comme discriminant.* On peut imaginer que cette valeur est influencée par de nombreux facteurs, y compris F_{lon}. Pour pouvoir déterminer ces valeurs, l'idée est de faire varier simultanément les deux variables pour trouver un couple de valeurs qui maximise les performances du système, et ce pour chaque méthode de pondération considérée.

Dans notre corpus de référence, la génération des signatures lexicales est une opération d'autant plus délicate que les documents sont imposants. Un document de *petite taille* se limite souvent dans notre corpus à un unique paragraphe qui est sémantiquement homogène et ne présente pas de difficulté majeure pour rechercher les termes clefs. Au contraire, un document plus imposant constitué de nombreux paragraphes et couvrant différentes thématiques impose de prendre un nombre de termes limités qui risque de n'être représentatif que d'un sous-ensemble du texte initial. Notre intention est de n'étudier que des textes de *grande taille* où le vocabulaire utilisé est varié et peut engendrer des confusions. Pour ces raisons nous ne prendrons comme documents d'étude pour la génération des signatures lexicales que des documents de grande taille (plus de 30 mots et 20 SN, ce qui représente 10% des documents du corpus). Le nombre de documents de ce type dans le *benchmark* est indiqué dans le tableau 7.1 sous l'intitulé L_{doc}.

7.2.2 Évaluation des performances

L'évaluation des résultats corrélés se fait naturellement en utilisant le *benchmark* décrit au cours de la section 7.1. La procédure de comparaison est la suivante : pour chaque thème du *benchmark*, on sélectionne aléatoirement dix " grands " documents (ou le maximum qu'il est possible de sélectionner si le *cluster* comporte moins de dix éléments, voir tableau 7.1). Pour chacun de ces documents, dont l'ensemble est appelé θ, on construit les signatures lexicales issues des différentes pondérations utilisées. Nous imposons ensuite une valeur maximale de F_{lon} de 25 pour les mots et 15 pour les SN pour avoir des changements significatifs entre les listes pondérées étudiées.

La finalité de l'expérimentation est de rechercher, pour un document donné, les différents ensembles corrélés en utilisant les différentes valeurs possibles de F_{lon} and F_{sem}. Les résultats sont ensuite analysés en mesurant le rappel (R) et la précision (P) par l'intermédiaire d'un indice de performance noté F et décrit ci-après.

Considérons un thème particulier du *benchmark* utilisé et son ensemble de documents associés. Soient D_{ben} et D_{auto} deux *clusters* de documents qui représentent respectivement le *cluster* de document du thème considéré et celui obtenu automatiquement après interrogation de la base via une signature lexicale.

La définition des indices R, P et F est alors donnée par les équations suivantes :

$$R = \frac{|D_{ben} \cap D_{auto}|}{|D_{ben}|} \quad P = \frac{|D_{ben} \cap D_{auto}|}{|D_{auto}|} \quad F = \frac{2RP}{R+P} \tag{7.1}$$

7.2.3 Observations

La première expérimentation que nous menons ne s'intéresse qu'au facteur de performance F car il combine en une seule et même mesure les propriétés du rappel et la précision ; et correspondant au meilleur couple (F_{lon}, F_{sem}) envisageable pour retrouver le plus grand nombre de résultats corrélés possible sans toutefois générer trop de bruit.

Les figures 7.1 et 7.2 représentent la moyenne des meilleures performances (F) obtenues suivant les différentes méthodes de pondération utilisées et respectivement pour des signatures composées de mots et de SN. Les valeurs ainsi transcrites sur les figures ont été obtenues en calculant la moyenne des valeurs de F pour chaque couple de valeurs (F_{lon},F_{sem}) possible. Pour chaque méthode, nous avons reporté dans les tableaux 7.2 et 7.3 les valeurs maximales obtenues pour F (valeur notée F_{max}) ainsi que les valeurs correspondantes de F_{lon} et F_{sem}.

L'analyse des résultats donnés par les figures 7.1 et 7.2 tend à montrer que, pour la

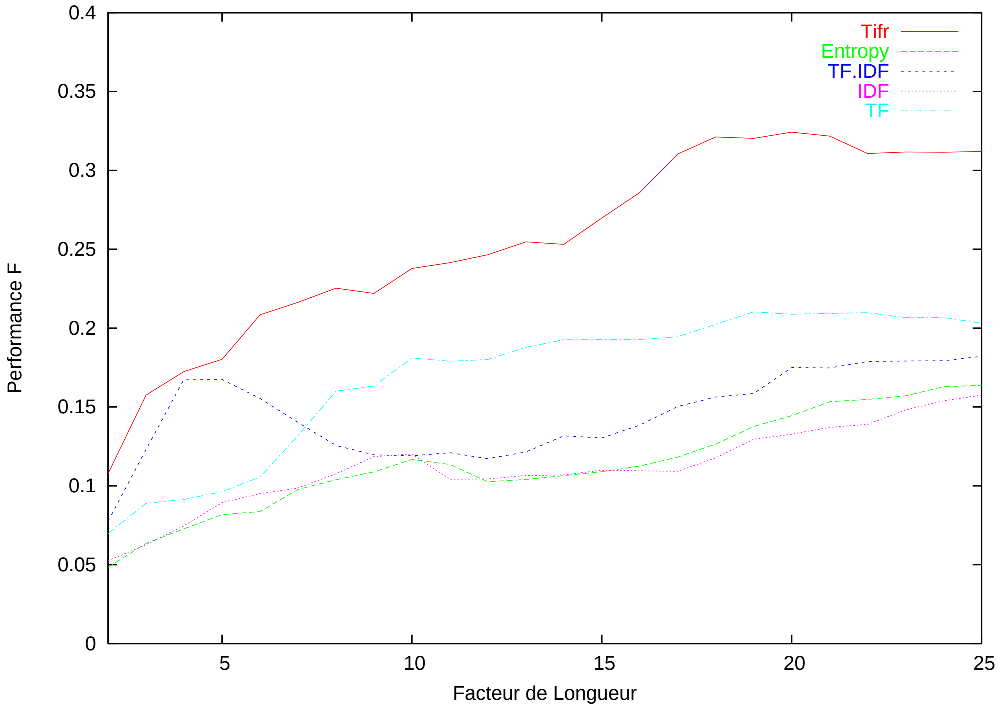


FIG. 7.1 – *Performance moyenne en utilisant des mots, en fonction du nombre de termes dans la requête*

plupart des méthodes considérées, plus la valeur de F_{lon} est importante, meilleure est la performance F. Cette corrélation entre les valeurs prises par F et celles de F_{lon} est décevante car elle ne nous permet pas de trouver de valeur optimale de F. Cette constation doit être nuancée car il y a lieu de distinguer deux catégories de pondérations. La première est représentée par $T_f.I_{df}$, I_{df} et l'entropie. Ces méthodes ont été regroupées car elles n'ont pas permis d'atteindre une valeur maximale de F. La seconde catégorie est représentée par les poids T_{ifr} and T_f ; ces poids permettent de retrouver une valeur maximale de F, donc un couple (F_{lon}, F_{sem}) permettant d'optimiser la recherche de documents corrélés. On s'aperçoit également que lorsqu'il est possible de déterminer une valeur maximale pour F, notée F_{max}, aux abords de cette valeur, F varie lentement. Pour cette raison nous définissons F_{90}, une valeur particulière de F qui équivaut à 90% de F_{max} et qui nous permettra de suivre l'impact d'une légère dégradation des performances sur les valeurs du couple (F_{lon}, F_{sem}). Cette valeur est reportée dans les tableaux 7.2 et 7.3 afin d'être utilisée ultérieurement.

La première catégorie de pondération définie précédemment montre des résultats négatifs : la performance moyenne F tourne autour de 0.13 pour les mots et 0.07 pour les SN, deux valeurs qui sont loin de celles atteintes par les deux autres méthodes de pondération, T_f ou T_{ifr}. Des poids tels que I_{df} ou une mesure d'entropie sont d'excellents choix pour collecter un nombre limité de documents [PPGK02, PW00], mais lorsque l'on analyse leur

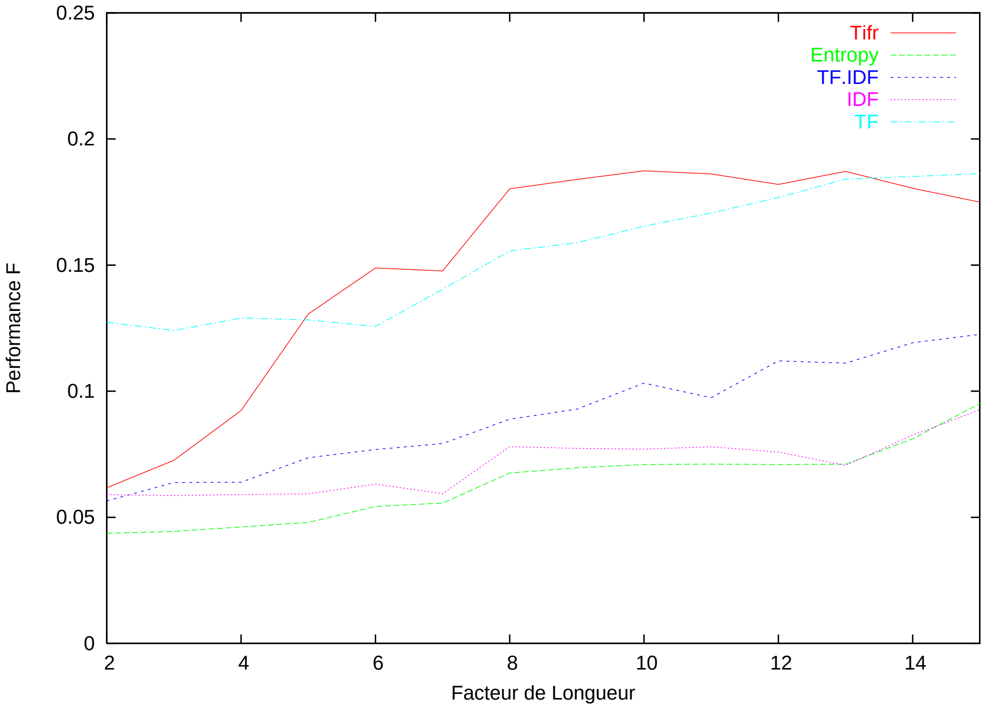


FIG. 7.2 – *Performance moyenne en utilisant des SN, en fonction du nombre de termes dans la requête*

performance F, trop de documents potentiellement corrélés ont été omis pour obtenir un taux de rappel satisfaisant et donc une performance F intéressante.

Qu'il s'agisse de signatures lexicales composées de mots ou de SN, les deux types de signatures induisent des résultats similaires. La différence fondamentale est que la valeur de la performance moyenne est plus faible pour des signatures composées de SN. Les SN, à la différence des mots, ne sont pas distribués dans tout le corpus de la même manière ; ils permettent d'accéder à un nombre plus petit de documents. La conséquence est que le taux de rappel est inférieur et de ce fait la performance est en retrait par rapport aux mots.

Les tableaux 7.2 et 7.3 nous donnent de plus amples renseignements sur les différences entre pondérations. Une remarque importante est de constater que plus la valeur de F_{sem} est faible pour une valeur élevée de F_{lon}, moins les termes sont cohérents dans une signature lexicale. Une grande signature avec une valeur faible de F_{sem} indique, en effet, que les termes présents sont rarement retrouvés simultanément dans les mêmes unités de recherche. Mais l'inverse n'est pas vrai : le meilleur exemple est celui de T_f qui obtient sa plus petite valeur de F_{lon} pour un maximum de F_{sem}. L'explication reste simple : une pondération comme T_f recherche principalement des termes fréquents dans le corpus de référence ; la probabilité qu'ils soient associés dans les mêmes unités de recherche est donc supérieure. Les tableaux 7.2 et 7.3 montrent alors que les pondérations telles que I_{df}, $T_f.I_{df}$ ou une

	T_{ifr}	Entropy	$T_f.I_{df}$	I_{df}	T_f
F_{max}	0.324	0.163	0.182	0.157	0.210
F_{lon}	20	25	25	25	19
F_{sem}	3	2	2	2	4
F_{90}	0.291	0.147	0.163	0.141	0.189
F_{lon}	17	20	20	23	14
F_{sem}	3	2	2	2	4

TAB. 7.2 – *Valeurs des paramètre en utilisant des mots*

	T_{ifr}	Entropy	$T_f.I_{df}$	I_{df}	T_f
F_{max}	0.187	0.095	0.122	0.092	0.186
F_{lon}	10	15	15	15	15
F_{sem}	3	1	3	1	6
F_{90}	0.168	0.085	0.110	0.083	0.167
F_{lon}	8	15	12	15	11
F_{sem}	3	1	3	1	6

TAB. 7.3 – *Valeurs des paramètre en utilisant des SN*

mesure d'entropie produisent des signatures lexicales incohérentes, mais cela ne signifie pas pour autant que les pondérations restantes, à savoir T_f et T_{ifr}, sont des pondérations idéales et dédiées à la corrélation.

En résumé, cette expérimentation a montré que les méthodes de pondération réagissent différement : T_{ifr} et T_f peuvent être considérées, a priori, comme des poids intéressants pour la corrélation car une valeur F_{max} a pu être trouvée. Nous avons également pu remarquer que les mots semblent être plus performants que les SN pour améliorer les performances du sytème de recherche de documents corrélés. Finalement, l'indice T_{ifr} donne les meilleurs résultats et améliore la performance F moyenne de ces résultats de 9% pour les mots et de plus de 7% pour les SN.

7.3 La recherche des résultats corrélés

Comme toutes les méthodes de recherche d'information utilisant des signatures lexicales, la méthode de corrélation envisagée dépend du choix du moteur de recherche utilisé car les documents corrélés sont influencés par le classement des résultats [PPGK02]. Pour l'utilisateur final, l'utilité d'une méthode de corrélation est notamment de pouvoir classer les documents en fonction de l'intérêt qu'ils représentent par rapport au document initial ; la finalité est de présenter des documents fortement corrélés parmi les premiers résultats obtenus. Le but de cette section est d'étudier l'influence d'un algorithme simple de classement des résultats corrélés en fonction des pondérations utilisées.

7.3.1 Classement des résultats

L'algorithme retenu pour classer les résultats corrélés a été conçu pour pouvoir tirer parti des pondérations retenues et comparer efficacement leurs performances respectives.

Le processus de classement des documents est un mécanisme scindé en quatre étapes :

1. les documents sont classés par valeurs décroissantes de F_{sem} ;
2. si les valeurs de F_{sem} sont identiques, les documents sont classés par valeurs décroissantes de F^{ϕ}_{sem} [2] ;
3. si les valeurs de F^{ϕ}_{sem} sont identiques, les documents sont triés par nombre décroissant de termes communs avec la signature lexicale du document initial ;
4. finalement, pour départager les derniers documents, on les classe par somme décroissante des T_f locales de chaque terme commun avec la signature lexicale initiale.

Pour chaque document de l'ensemble θ, nous contruisons alors la liste des documents corrélés qui sont ensuite classés grâce à l'algorithme précédemment défini. Pour construire l'ensemble des documents corrélés, nous nous mettons dans les conditions optimales d'utilisation des différentes pondérations ; les valeurs prises pour F_{sem} et F_{lon} sont alors celles qui ont été reportées dans les tableaux 7.2 et 7.3.

Dans la section 7.2, nous avions étudié comme critère de comparaison la performance moyenne du système donnée par F car notre but était de déterminer le meilleur compromis entre rappel et précision ; dans le cas présent nous cherchons à savoir si les documents les mieux classés sont significatifs ; le meilleur critère d'étude est cette fois-ci la précision des résultats en fonction du rang occupé dans le classement.

Les figures 7.3 et 7.4 représentent la précision moyenne des résultats lorsque l'on considère dans le cas d'une performance optimale déterminée dans la section 7.2 (i.e. lorsque $F = F_{max}$). Les tableaux 7.4 et 7.5 donnent une information synthétique concernant la précision moyenne des documents en fonction de leur rang et ce pour deux valeurs particulières de F, à savoir : $F = F_{max}$ et $F = F_{90}$.

[2] Pour un document, F^{ϕ}_{sem} représente le nombre d'unités de recherche où au moins F_{sem} termes ont été retrouvés.

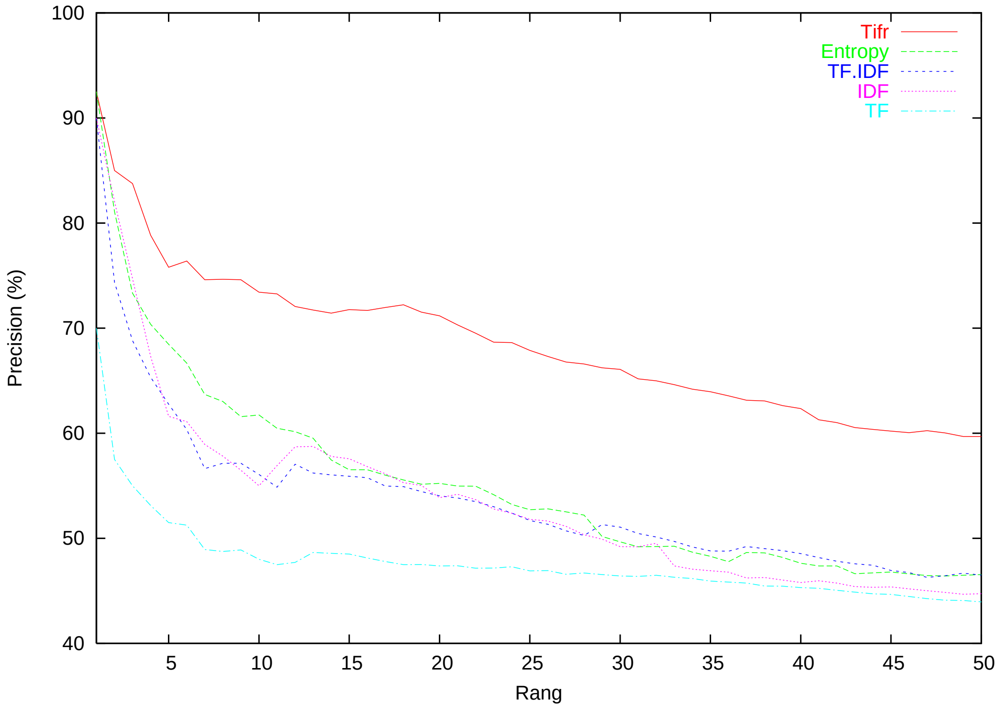


FIG. 7.3 – *Précision moyenne des résultats de la corrélation en utilisant des mots, en fonction du rang du document dans la liste des réponses*

7.3.2 Résultats

L'étude des figures 7.3 et 7.4 confirme les conclusions précédentes à propos de l'efficacité des signatures lexicales composées de mots. En effet, la comparaison des deux figures montre qu'en moyenne leur utilisation améliore de 7% la précison des résultats par rapport à l'utilisation de signatures composées de SN. Ces résultats qui semblent pessimistes quant à l'utilisation des SN pour fabriquer des signatures lexicales ne doivent en aucun cas nous dissuader de poursuivre les expérimentations avec des SN, car ces résultats sont quantitatifs et non qualitatifs. En effet, le benchmark tel qu'il a été construit apporte une information binaire entre les documents et les thèmes : corrélés ou non. L'hypothèse de départ était de dire que deux documents appartenant au même *cluster* sont corrélés, mais qu'il n'est pas possible de dire avec quel degré de corrélation. De ce fait, cette expérience montre quantitativement que l'utilisation de mots est plus profitable que l'emploi de SN mais elle ne peut rien affirmer qualitativement.

Les figures 7.3 et 7.4 montrent également que les tout premiers documents retrouvés ont, quelle que soit la méthode envisagée et ce à l'exception de T_f, des précisions comparables. En considérant des rangs plus élevés, les choses évoluent rapidement et l'on constate une soudaine dégradation de la précision moyenne à l'exception de T_{ifr} qui l'améliore de l'ordre de 10% pour les mots et de 7% pour les SN. La décroissance des autres méthodes semble

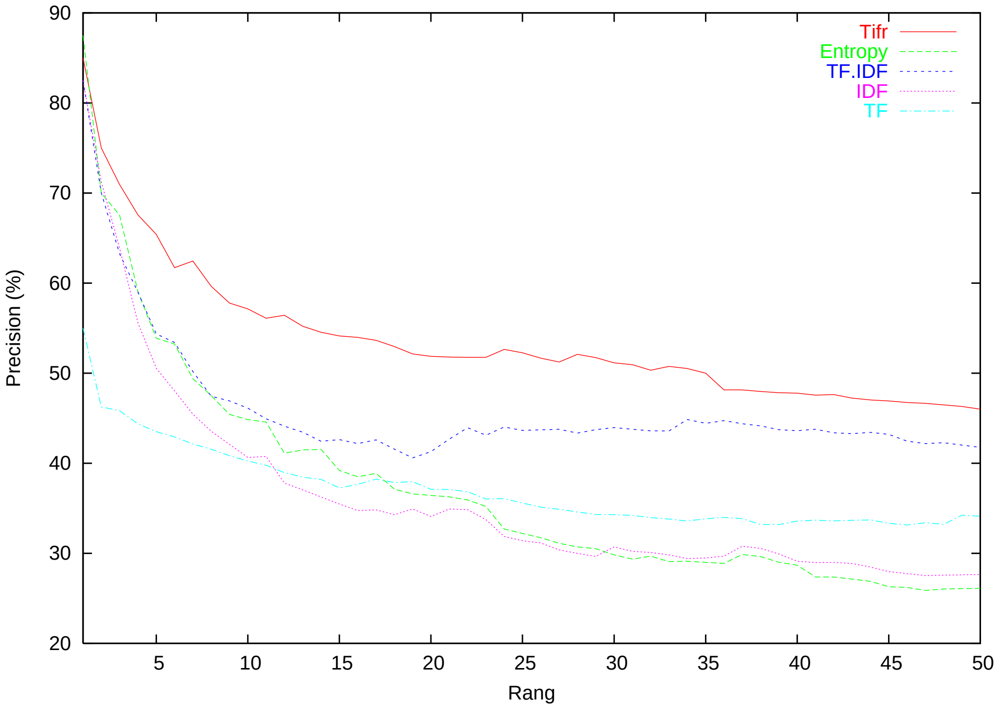


FIG. 7.4 – *Précision moyenne des résultats de la corrélation en utilisant des SN, en fonction du rang du document dans la liste des réponses*

rapide car on passe rapidement sous la barre des 60% de précision atteinte dès le $10^{ème}$ rang alors que pour T_{ifr} cette même valeur est atteinte au $50^{ème}$ rang.

Les tableaux 7.4 et 7.5 permettent de comparer les résultats lorsque l'on tolère une légère dégradation des conditions initiales en remplaçant le critère optimal $F = F_{max}$ par un critère approché $F = F_{90}$. Le résultat est que la précision reste stable. Cette constatation est très importante car, si l'on considère à nouveau les tableaux 7.2 et 7.3 qui nous donnent les différentes valeurs prises par le couple (F_{lon}, F_{sem}), force est de constater que seules les valeurs de F_{lon} changent et sont inférieures dans le cadre $F = F_{90}$. L'enseignement que l'on peut en retirer est qu'une signature de taille plus petite n'influence pas fondamentalement la précision des 50 premiers résultats. En continuant dans cette logique, on peut supposer que si nous souhaitions n'avoir qu'un panel restreint de documents corrélés, de l'ordre de 10, une diminution de la taille des signatures plus conséquente pourrait être envisagée.

7.4 Enseignements

Dans ce chapitre, nous avons comparé différentes méthodes de pondération par l'intermédiaire des signatures lexicales constituées, et au travers des documents corrélés générés. Nous avons recherché les meilleures conditions possibles pour constituer des signatures les

	Rang	T_{ifr}	Entropie	$T_f.I_{df}$	I_{df}	T_f
F_{max}	1	92.5	92.5	90.0	90.0	70.0
	10	79.2	72.0	66.1	66.8	53.2
	50	68.9	56.8	54.6	55.1	47.7
F_{90}	1	95.0	87.5	87.5	85.0	55.0
	10	78.3	68.3	67.1	69.2	46.8
	50	68.0	58.2	53.8	56.9	42.6

TAB. 7.4 – *Précision moyenne (%) en utilisant des mots*

	Rang	T_{ifr}	Entropie	$T_f.I_{df}$	I_{df}	T_f
F_{max}	1	85.0	87.5	82.5	82.5	55.0
	10	66.6	58.6	57.5	55.3	44.2
	50	54.6	40.5	47.4	39.0	37.1
F_{90}	1	80.0	87.5	82.5	82.5	47.5
	10	62.3	58.6	55.3	55.8	39.3
	50	50.8	40.5	46.5	39.9	32.1

TAB. 7.5 – *Précision moyenne (%) en utilisant des SN*

plus efficaces possibles. Cette recherche de conditions est matérialisée par la détermination du couple idéal (F_{lon}, F_{sem}) pour chaque méthode utilisée. La recherche de ce couple a montré que, pour un certain nombre de pondérations telles que $T_f.I_{df}$, I_{df} ou encore l'entropie, il n'était pas possible d'en trouver un, ce qui nous laisse penser que ces pondérations ne sont pas pertinentes pour la recherche de documents corrélés. D'un autre coté, une pondération telle que T_f permet de trouver un couple (F_{lon}, F_{sem}) idéal, mais les résultats obtenus manquent cruellement de précision.

L'expérimentation a également montré que notre méthode ad hoc de pondération, T_{ifr}, donne des résultats plus précis que les autres méthodes. Pourtant, nous ne pensons pas encore avoir amélioré de manière sûre et définitive la recherche de résultats corrélés.

L'explication tient d'abord à ce que nous sommes placés dans des conditions particulières. En effet nous avons montré que T_{ifr} donnait de meilleurs résultats que les méthodes classiques de pondération mais pour un nombre de documents réponses peu important : 50 tout au plus. Cette taille doit être confrontée avec la taille moyenne des *clusters* de documents corrélés, de l'ordre du millier de documents en moyenne.

Ensuite, une validation qualitative des documents corrélés par des méthodes quantitatives est nécessaire pour départager les pondérations en présence et s'assurer de nos avancées en termes de corrélations.

Chapitre 8

Évaluer la qualité des résultats corrélés

Jusqu'à présent, les expérimentations décrites dans les précédentes parties (Chapitre 5, chapitre 6 ou chapitre 7) ont été validées sur des critères quantitatifs. L'objectif du présent chapitre est de mettre en place une expérimentation qui permette de valider notre approche d'un point de vue qualitatif.

La section 8.1 présente les différentes orientations possibles en terme d'expérimentation et permet de choisir la voie qui sera étudiée. La section 8.2 décrit plus finement l'expérimentation réalisée ; la section 8.3 s'attache à comparer les différentes méthodes de corrélation. Finalement, la section 8.4 présente les enseignements tirés de cette expérimentation.

8.1 Le choix du corpus de validation

Le paradigme de l'évaluation n'est pas nouveau. Dans le domaine des sciences de l'information, il a trouvé une formalisation dans les années 50, avec les travaux de Berry, Kent, Luehrs et Perry [ea55]. Mais les initiatives se sont multipliées depuis une vingtaine d'années, dans la lignée des projets de l'agence fédérale de recherche technologique de la défense américaine, la DARPA, dont la plus récente et la plus connue est sans doute, depuis 1992, le cycle de compétitions TREC (Text REtrieval Conference). Les deux initiatives que nous présentons, à savoir TREC (Section 8.1.1), compétition internationale américaine, et AMARYLLIS (Section 8.1.2), cycle francophone d'évaluations partagent une même méthode d'évaluation, même si leurs objectifs et les motivations de leurs commanditaires (agence fédérale de défense et réseau de recherche francophone) sont différents. À travers cette présentation et cette comparaison nous montrerons dans la section 8.1.3 l'orientation prise pour évaluer nos résultats.

8.1.1 TREC

L'objectif de TREC est l'évaluation de méthodologies de recherche d'informations textuelles. TREC est organisée aux USA par le NIST et sponsorisée par le NIST et par la DARPA. La première opération, TREC-1, a eu lieu en 1992, la conférence finale s'étant tenue en novembre 1992 dans les locaux du NIST, à Gaithersburg, Maryland (USA). Depuis cette date, il y a une campagne TREC organisée chaque année.

TREC fournit des corpus de documents d'entraînements (brevets US, articles du Wall Street Journal, dépêches Reuters,...) et des thèmes de recherche, avec les réponses attendues permettant aux participants d'optimiser leurs systèmes pour les corpus étudiés. De nouveaux corpus de test composés de documents et de thèmes sont ensuite fournis sans les réponses pour tester les différents systèmes de recherche concurrents. Les participants à TREC ont des tâches précises à effectuer, et les résultats obtenus sont ensuite comparés selon la même méthode en utilisant le logiciel TrecEval, puis discutés lors d'une conférence commune.

Dans TREC trois types de corpus sont utilisés :

- corpus de documents ;
- corpus de thèmes de recherche (*topics*) ;
- corpus de réponses.

Dans TREC, les corpus de thèmes de recherche et de réponses sont élaborés à partir de documents d'origines diverses. TREC utilise, pour les phases d'entraînement, des documents qui représentent un volume d'information de l'ordre de 1 Go ou plus. Les documents forment un ensemble de textes primaires issus de journaux du type Financial Times et Los Angeles Times, du Federal Register, du Foreign Broadcast Information Service ou de brevets.

Chaque *topic* (thème de recherche) contient un certain nombre d'éléments : titre, question, concepts, etc. Ce nombre a varié au fil des campagnes d'évaluation en fonction des réactions des participants. Par exemple, l'opération TREC-6 n'a retenu comme éléments que le **titre** qui donne le champ de connaissance, la partie **description** qui est la question proprement dite et la partie **narration** qui donne les limites de pertinence des réponses à fournir. Un exemple de document issu de la collection TREC-6 est représenté dans le tableau 8.1.

Ces thèmes sont créés par des professionnels de l'information (*assesseurs*). Dans le cas des thèmes ad hoc de TREC-6, une nouvelle méthode d'élaboration a été introduite. Une question est posée et les 25 premiers documents retrouvés sont jugés par des assesseurs. Si parmi ceux-ci aucun document pertinent n'est ramené ou plus de 20 documents sont pertinents, le thème est rejeté. Si 1 à 5 documents sont pertinents, la question est affinée

```
< top >
< num > Number: 324
< title > Argentine/British Relations
< desc > Description: De_ne Argentine and British international
relations
< narr > Narrative:
It has been 15 years since the war between Argentina and the
United Kingdom in 1982 over sovereignty in the Falkland Islands.
A relevant report will describe their relations after that period.
Any kind of international contact between the two countries is
relevant, to include commercial, economic, cultural, diplomatic,
or military exchanges. Negative reports on the absence of such
exchanges are also desirable. Reports containing information on
direct exchanges between Argentina and the Falkland Islands are
also relevant.
< /top >
```

TAB. 8.1 – *Exemple de document extrait de la campagne TREC-6*

et reposée. Si 6 à 20 documents sont pertinents, la question est retenue. Pour les thèmes retenus, les documents sont ensuite jugés jusqu'au 100ème afin d'avoir une idée très précise des réponses possibles (et de mettre au point la partie narration). Sur 120 thèmes testés de cette façon, 91 répondent aux critères de sélection ci-dessus. 50 ont été choisis parmi les 91 possibles pour les tâches ad hoc. Pour les tâches de routage, 47 thèmes ont été choisis parmi les 300 utilisés à l'entraînement.

A l'issue des tests, chaque participant renvoie pour chaque thème de recherche ses 1000 premières réponses réponses. Les corpus de réponses sont alors construits selon une méthode dite d'échantillonage. Les 100 premières de chaque participant sont prises en compte, rassemblées, dédoublonnées pour ensuite y faire effectuer des jugements de pertinence par les *assesseurs*.

8.1.2 AMARYLLIS

L'appel d'offres de l'AUPELF-UREF[1] à l'origine du projet AMARYLLIS précisait que son action devait être proche de celle menée dans TREC. Un cycle AMARYLLIS dure un an à partir du lancement de l'appel d'offres, avec un laps de temps entre la fin d'un cycle et le début du suivant. Au moment de la réunion finale, les différents participants à l'opération se rencontrent, comme pendant la conférence TREC, pour présenter leurs

[1]Agence francophone pour l'enseignement supérieur et la recherche.

méthodes, leurs outils et les difficultés rencontrées.

D'un point de vue méthodologique, les deux types de campagne d'évaluation se ressemblent trait pour trait. Les principales différences entre TREC et AMARYLLIS portent essentiellement sur les types de documents utilisés et sur la taille des corpus. AMARYLLIS utilise, pour le cycle exploratoire, 139 Mo pour l'entraînement et 119 Mo pour les tests à comparer au giga-octets de TREC. L'ensemble des documents utilisé par AMARYLLIS est découpé suivant plusieurs thématiques et cet ensemble est très hétérogène en terme de contenu, de taille ou encore de vocabulaire, qui est très spécifique.

Chaque thème comprend 5 éléments tirés de la création des *topics* de versions précédentes de TREC : le **domaine** qui permet de situer le champ de connaissance, le **sujet** qui est un titre définissant le thème, la **question** proprement dite, les **compléments d'information** qui donnent des précisions sur les limites de pertinence des réponses à fournir et les **concepts** qui sont des descripteurs permettant de délimiter le champ de recherche.

8.1.3 Vers une méthode de validation manuelle

Le but visé par cette description des campagnes d'évaluation TREC et AMARILLYS est avant tout de déterminer si leur utilisation peut nous être utile pour évaluer la qualité de nos travaux. À ce titre nous pouvons faire plusieurs remarques.

Tout d'abord, il faut rappeler que l'utilisation des *benchmarks* TREC et AMARYLLIS sert normalement à évaluer la qualité d'un système de recherche d'information par l'intermédiaire d'un jeu de questions-réponses conçues par des professionnels (*assesseurs*). L'utilisation que nous souhaitons en faire est donc nécessairement une utilisation biaisée. Pour nous, l'intérêt d'utiliser de tels *benchmarks* est avant tout d'obtenir un corpus de documents scindés suivant différentes thématiques. Ces thématiques sont ensuite assimilées à des *clusters* de documents corrélés pour, à partir d'un document et d'une thématique donnés, trouver d'autres documents de cette thématique. La question que nous sommes en droit de nous poser est de savoir si cette pratique est viable.

Pour répondre à cette question, nous avons pris comme référence le *benchmark* TREC de la campagne 2001. Ce corpus composé de dépêches de 1996-1997 de l'agence Reuters représente un total d'environ 550.000 dépêches pour une taille proche de 3 Go de données. Parmi cet ensemble de documents, 330.000 environ sont regroupés dans 84 thématiques (thématiques listées dans le tableau 8.2), les 220.000 documents restants étant considérés comme non pertinents pour ces thématiques.

Contrairement aux classes définies section 7.1, qui avaient participé à l'élaboration de notre *benchmark* expérimental, celles présentes dans le tableau 8.2 sont pour la plupart étroitement liées. Ainsi va-t-on pouvoir trouver des classes telles que *MANAGEMENT* et

MANAGEMENT MOVES ou encore *ANNUAL RESULTS*, *INSOLVENCY/LIQUIDITY*, *SHARE CAPITAL*, *BONDS/DEBT ISSUES*, *LOANS/CREDITS* et *CREDIT RATINGS* qui sont difficiles à différencier. Cette proximité sémantique des classes en présence laisse penser que de tels regroupements de documents ne sont pas de bons candidats pour jouer le rôle de classes de documents corrélés.

De plus, les méthodes d'évaluation actuellement utilisées par TREC ou AMARYLLIS sont souvent basées sur une évaluation plus quantitative que qualitative [LKSS99]. En effet, les résultats produits se limitent à calculer une différence (rappel/précision) entre les réponses d'un système quelconque et les réponses attendues. Ce type d'évaluation est orienté vers une approche comparative de plusieurs systèmes reposant sur le principe des bancs d'essai. Cependant, elle ne fournit que peu d'information sur la qualité globale de l'outil. Or d'un point de vue pratique, nous avons déjà réalisé des expérimentations quantitatives pour évaluer la performance de notre système (Chapitre 7).

Dernier point soulevé, celui de la justesse de faire une évaluation qualitative sur un corpus autre que notre corpus de référence. Dans les chapitres précédents, nous avons essayé de définir une méthodologie de recherche des documents corrélés sur notre corpus de référence, et à ce titre nous nous sommes appuyés sur ses particularités. La méthode que nous avons conçue et réalisée est peut être générique dans son approche mais spécialisée dans sa réalisation et dédiée à un corpus juridique en langue française. Notre conviction est donc qu'une évaluation qualitative des résultats corrélés réalisée directement sur notre corpus de référence sera plus judicieuse qu'une validation plus générique utilisant des benchmarcks de type TREC ou AMARYLLIS.

STRATEGY/PLANS	LEGAL/JUDICIAL
REGULATION/POLICY	SHARE LISTINGS
ANNUAL RESULTS	INSOLVENCY/LIQUIDITY
SHARE CAPITAL	BONDS/DEBT ISSUES
LOANS/CREDITS	CREDIT RATINGS
ASSET TRANSFERS	PRIVATISATIONS
PRODUCTION/SERVICES	NEW PRODUCTS/SERVICES
RESEARCH/DEVELOPMENT	CAPACITY/FACILITIES
MARKETS/MARKETING	DOMESTIC MARKETS
EXTERNAL MARKETS	MARKET SHARE
ADVERTISING/PROMOTION	CONTRACTS/ORDERS
DEFENCE CONTRACTS	MONOPOLIES/COMPETITION
MANAGEMENT	MANAGEMENT MOVES
LABOUR	ECONOMIC PERFORMANCE
MONETARY/ECONOMIC	MONEY SUPPLY
INFLATION/PRICES	CONSUMER PRICES
WHOLESALE PRICES	CONSUMER FINANCE
PERSONAL INCOME	CONSUMER CREDIT
RETAIL SALES	EXPENDITURE/REVENUE
GOVERNMENT BORROWING	OUTPUT/CAPACITY
INDUSTRIAL PRODUCTION	INVENTORIES
EMPLOYMENT/LABOUR	UNEMPLOYMENT
TRADE/RESERVES	BALANCE OF PAYMENTS
MERCHANDISE TRADE	RESERVES
HOUSING STARTS	LEADING INDICATORS
EUROPEAN COMMUNITY	EC INTERNAL MARKET
EC CORPORATE POLICY	EC AGRICULTURE POLICY
EC MONETARY/ECONOMIC	EC INSTITUTIONS
EC ENVIRONMENT ISSUES	EC COMPETITION/SUBSIDY
EC EXTERNAL RELATIONS	EC GENERAL
DEFENCE	INTERNATIONAL RELATIONS
DISASTERS AND ACCIDENTS	ARTS, CULTURE, ENTERTAINMENT
ENVIRONMENT AND NATURAL WORLD	FASHION
HEALTH	LABOUR ISSUES
OBITUARIES	HUMAN INTEREST
BIOGRAPHIES, PERSONALITIES, PEOPLE	RELIGION
SCIENCE AND TECHNOLOGY	SPORTS
TRAVEL AND TOURISM	WAR, CIVIL WAR
ELECTIONS	WEATHER
WELFARE, SOCIAL SERVICES	BOND MARKETS
INTERBANK MARKETS	FOREX MARKETS
METALS TRADING	ENERGY MARKETS

TAB. 8.2 – *Thématiques du corpus TREC 2001*

8.2 Mise en place d'un protocole de validation

Au travers de la section 8.1, nous nous sommes positionnés quant à l'utilisation de protocoles de validation tels que TREC ou AMARYLLIS et nous leur préférons une évaluation ad'hoc de nos résultats sur notre corpus de référence. La finalité de cette section est de décrire ce protocole de validation (Section 8.2.1) et de vérifier que les résultats sont exploitables (Section 8.2.2).

8.2.1 Procédure de validation

Notre but ultime, en terme de corrélation, est de présenter à l'utilisateur une série de résultats corrélés en qualité et en quantité suffisante pour étoffer ses connaissances sur les thèmes développés par son document de référence. Néanmoins nous ne prétendons pas pouvoir retrouver tous les résultats potentiellement corrélés.

La voie ici explorée pour évaluer la qualité des différentes méthodes utilisée est de présenter à un panel d'évaluateurs des documents sources et pour chacun d'eux de lui faire noter la qualité des meilleurs résultats corrélés. Comme pour l'évaluation réalisée dans la section 5.4, celle-ci se déroule sous la forme d'une enquête Web.

Pour cette enquête, nous avons réutilisé les documents pré-sélectionnés dans la section 6.2.1 qui sont listés en Annexe B. La liste contenant cet ensemble de documents est appelée Δ. Pour chaque document $D \in \Delta$, nous extrayons ses différentes signatures lexicales comme dans le chapitre 7 en utilisant les valeur de F_{lon} et F_{sem} reportées dans les tableaux 7.2 et 7.3.

Pour ce document noté $D \in \Delta$ et un algorithme de pondération donné nous générons ainsi quatre signatures lexicales distinctes :

- une signature composée de mots générée dans les conditions $F = F_{max}$;
- une signature composée de mots générée dans les conditions $F = F_{90}$;
- une signature composée de SN générée dans les conditions $F = F_{max}$;
- une signature composée de SN générée dans les conditions $F = F_{90}$.

Pour chaque signature nous recherchons les documents corrélés. Ils sont ensuite classés par pertinence décroissante en utilisant l'algorithme décrit section 7.3.1.

À partir de ces résultats, nous construisons une liste de documents corrélés composée des cinq premiers résultats obtenus pour chaque signature lexicale, et ce pour chaque méthode de pondération utilisée.

Potentiellement, nous pouvons obtenir jusqu'à 100 documents corrélés (5 méthodes de pondération, 4 signatures lexicales, prise en compte des 5 premiers documents corrélés).

Dans la pratique, bon nombre de résultats sont en double ou en triple, et une fois les doublons ôtés, nous obtenons des listes d'une longueur moyenne de 31 documents. Nous y rajoutons 5 documents choisis aléatoirement à partir de notre corpus de référence et c'est à partir de ces listes que les évaluateurs vont réaliser leurs notations. Cet ensemble de documents est alors noté $C_D = [c_1, c_2, \ldots, c_r]$ avec r un entier positif.

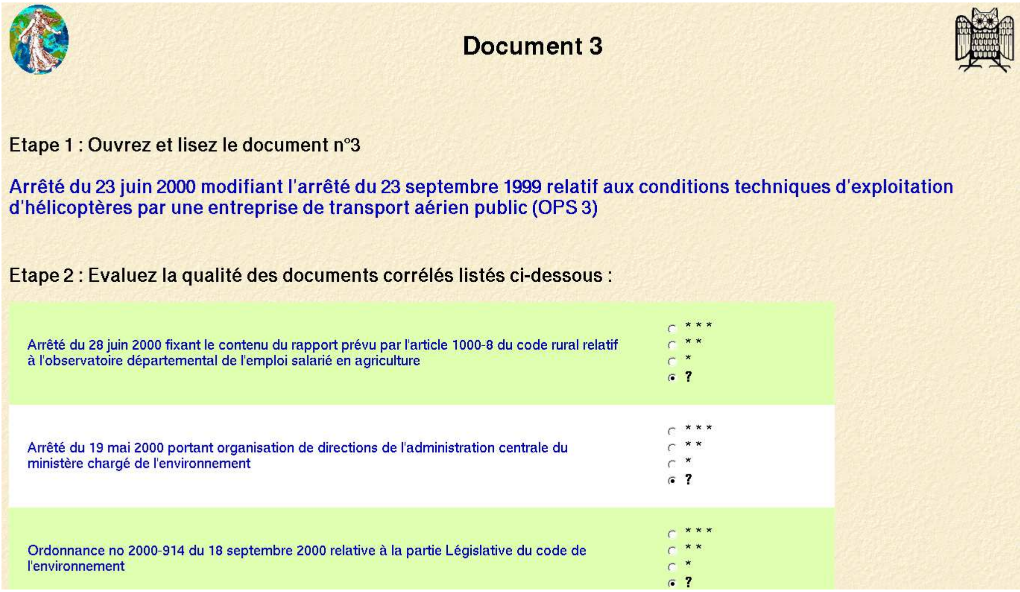


FIG. 8.1 – *Illustration d'un formulaire d'interrogation*

À partir de ces textes sont créés des formulaires d'évaluation, formulaires qui se trouvent illustrés par la figure 8.1 et qui comportent deux parties. Une première partie permet de lire le document de référence issu de la liste Δ. Une seconde partie permet de valider la liste de documents corrélés classés par ordre lexicographique de leur numéro NOR.

Pour chaque document potentiellement corrélé, l'évaluateur doit estimer la pertinence du document par rapport au document initial. Pour cela il peut noter le résultat suivant quatre critères :

- les deux textes sont en parfaite adéquation : il reçoit la note " *** " ;
- les deux textes sont assez proches : il reçoit la note " ** " ;
- les deux textes ont quelques points communs : il reçoit la note " * " ;
- si les deux textes n'ont aucun point commun : il reçoit la note " ? ".

Chaque évaluateur pré-sélectionné qui souhaite effectuer la validation doit, comme pour la première expérience décrite section 6.2.1, fournir quelques informations nous permettant de valider ses résultats et de le catégoriser comme candidat *candide* ou *expert* dans le domaine juridique.

8.2.2 Choix et accords des évaluateurs

La mise en place de cette deuxième évaluation a débuté en Mars 2003 par l'ouverture d'une enquête menée sur le Web[2]. Comme précédemment nous avons pré-sélectionné nos évaluateurs en diffusant plusieurs annonces, par envoi de mail, à plusieurs instituts universitaires, écoles d'ingénieurs, centres de recherches et listes de diffusion juridique.

Le retour n'a pas été plus important que précédemment : 14 inscriptions sur les centaines de candidatures potentielles. Sur ces 14 candidatures, 11 proviennent d'évaluateurs candides et aucune d'expert ; les 3 restantes ne peuvent pas être prises en compte car les validations n'ont pas été achevées. Les résultats présentés ici sont donc ceux des évaluateurs candides.

Avant d'analyser les réponses données par les évaluateurs, nous devons vérifier la validité des résultats. La figure 8.2 présente le taux d'accord existant entre les évaluateurs. Le but est de savoir si, en moyenne, les personnes interrogées s'accordent entre elles. Un désaccord prononcé ne nous permettant pas d'utiliser les résultats, nous cherchons alors à vérifier qu'il existe un consensus entre les évaluateurs.

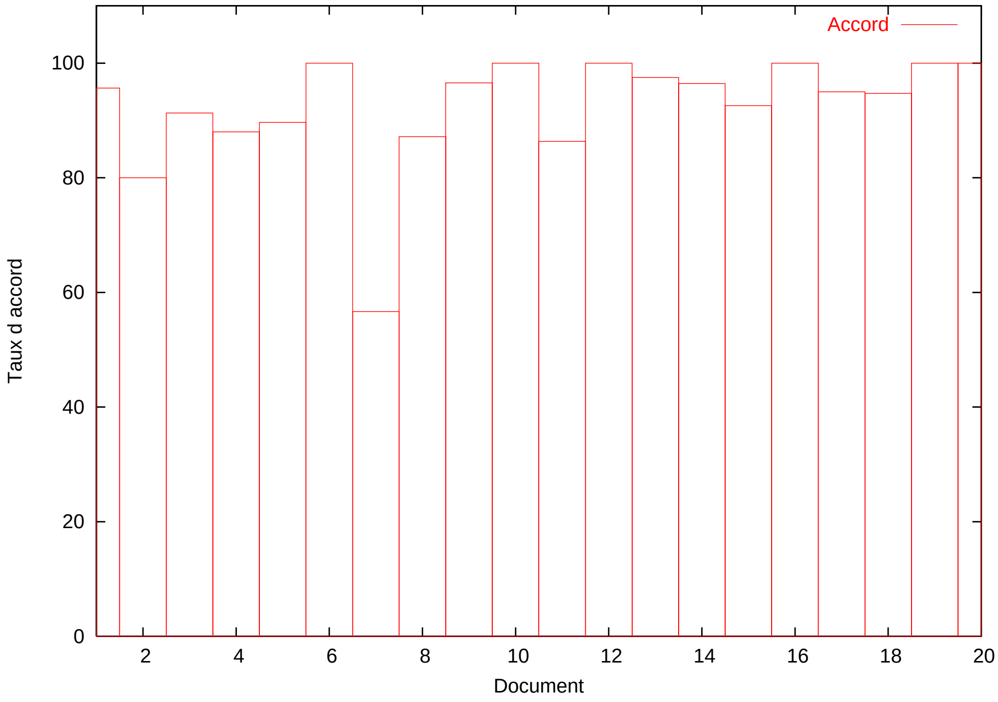


FIG. 8.2 – *Taux d'accord entre évaluateurs*

Pour mesurer un taux d'accord dans le cas présent, nous nous inspirons de la méthode de la section 5.4 et considérons que les évaluateurs s'accordent entre eux pour un document corrélé donné si :

[2] `http://evaluation.w3sites.net`

- tous les évaluateurs lui ont attribué le score " ?" ;
- tous les évaluateurs lui ont attribué le score " * ", " ** " ou " *** ".

Les résultats exprimés par la figure 8.2 nous révèlent un taux d'accord moyen de 91,2%. Ce taux souligne l'existence d'un réel consensus entre évaluateurs, permettant d'exploiter les résultats obtenus.

8.3 Comparaison des méthodes de corrélation

Notre évaluation terminée et la qualité des résultats ayant été validée au travers de la section précédente, nous nous attachons à définir le protocole de comparaison mis en place (section 8.3.1) et à étudier les résultats obtenus, présentés dans la section 8.3.2.

8.3.1 Protocole de comparaison

La validation et la notation manuelle des documents nous permettent d'obtenir deux types de renseignements sur les résultats corrélés obtenus grâce aux différentes méthodes de pondération expérimentées :

- leur précision ;
- leur qualité.

8.3.1.1 La précision des résultats

La précision des résultats pour un rang donné a été maintes fois utilisée (voir définition section 7.2.2) et nous permet dans le cas présent de quantifier le nombre de résultats correctement corrélés par les différentes méthodes étudiées. Dans le cas présent, la précision $P_f(n)$ est valable pour une méthode de corrélation f et pour un rang n donnés. Dans cette section 8.3.1, une méthode f est définie par un algorithme de pondération, un type de termes et une condition sur F (F_{max} ou F_{90}). Ainsi deux groupes de documents corrélés générés grâce à la pondération $T_f.I_{df}$ appliquée sur des mots mais dans les conditions F_{max} pour l'un et F_{90} pour l'autre auront été créés via deux méthodes de corrélation différentes.

D'une manière générale la précision des résultats $P_f(n)$ s'obtient grâce aux instructions suivantes :

1. Soit D un document de Δ et $C_D = [c_1, c_2, \ldots, c_r]$ la liste des documents à évaluer (voir section 8.2.1).
2. Soit $\chi_{D,f} = [\delta_1, \delta_2, \delta_3, \delta_4, \delta_5]$ les cinq résultats corrélés retrouvés par la méthode f, résultats classés par ordre de pertinence décroissante grâce à l'algorithme présenté

dans la section 7.3.1.

3. Soit $S_D = [c'_1, c'_2, \ldots, c'_{r'}]$ l'ensemble des documents considérés comme corrélés pour un évaluateur donné, c'est-à-dire les documents n'ayant pas reçu le score « ? ». On a $S_D \subset C_D$ et $r' \leq r$.

4. Soit E l'ensemble des évaluateurs et $e \in E$. On note $P_{f,e}(n)$ la précision partielle relative à e de f au rang n. $P_{f,e}(n)$ vaut 1 si $\delta_n \in S_D$ et 0 dans le cas contraire, avec $1 \leq n \leq 5$.

5. L'équation 8.1 définit alors la précision moyenne d'une méthode de corrélation f au rang n pour l'ensemble des documents de Δ et des évaluateurs de E.

$$P_f(n) = \frac{1}{\|\Delta\|.\|E\|} \sum_{d \in \Delta} \sum_{e \in E} P_{f,e}(n) \tag{8.1}$$

Les figures 8.3 et 8.4 décrivent la précision moyenne des résultats corrélés respectivement pour des signatures composées de mots et de SN dans le cas où $F = F_{max}$. Les figures C.1 et C.2 proposées en Annexe C décrivent la précision moyenne des résultats corrélés respectivement pour les signatures composées de mots et de SN dans le cas où $F = F_{90}$.

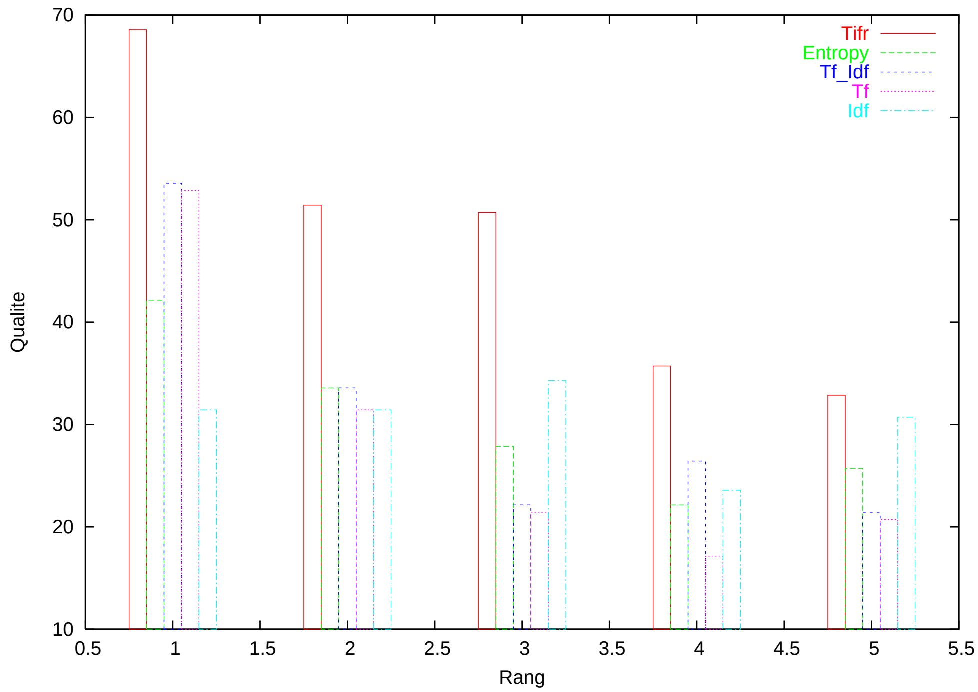


FIG. 8.3 – *Précision moyenne des résultats corrélés en fonction du rang de la réponse. Cas des mots. Cas de la performance optimale $F = F_{max}$ (cf analyse des résultats section 8.3.2).*

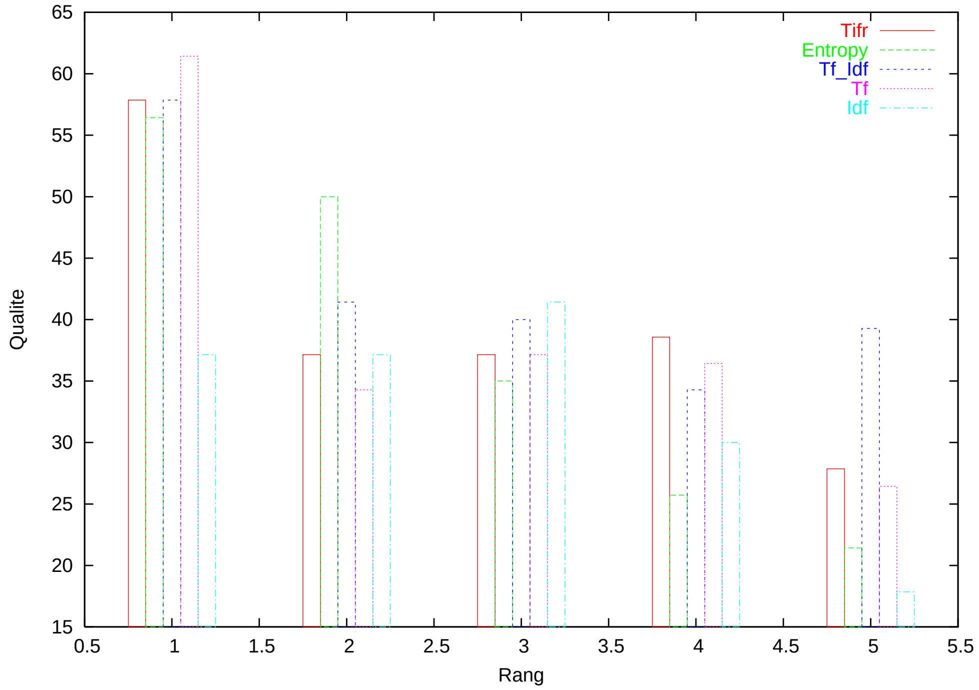


FIG. 8.4 – *Précision moyenne des résultats corrélés en fonction du rang de la réponse. Cas des SN. Cas de la performance optimale* $F = F_{max}$ *(cf analyse des résultats section 8.3.2).*

8.3.1.2 La qualité des résultats

Autre critère étudié, la qualité. Contrairement à la précision, le rappel ou la performance, la qualité n'est pas un critère standard de notre domaine d'étude. Il s'agit d'un critère défini spécifiquement pour nos travaux pour évaluer la pertinence sémantique des documents retrouvés. Cette qualité est la traduction des notations moyennes attribuées par les évaluateurs. Le tableau 8.3 illustre cette notion de qualité pour un document corrélé. On y trouve un document initial accompagné des cinq premiers documents corrélés classés par qualité décroissante. Pour évaluer la pertinence des méthodes de corrélation exposées nous comparons nos résultats expérimentaux avec les listes issues des évaluateurs comme celle présentée dans le tableau 8.3.

Comme pour la précision précédemment définie, cette qualité notée $Q_f(n)$ est relative à une méthode de corrélation f pour un rang n. Pour définir cette qualité, nous nous servons des notations des résultats corrélés assignés par chaque évaluateur. Pour un document $D \in \Delta$, chaque document corrélé de $S_D = [c'_1, c'_2, \ldots, c'_{r'}]$ reçoit 3 points à chaque fois qu'un évaluateur $e \in E$ l'a gratifié de “ *** ”, 2 points pour “ ** ” et 1 point pour “ * ”. On définit par W la fonction qui à un document c de S_D lui associe son score moyen noté $W(c)$. On admet que $W(c) = 0$ si c n'appartient pas à S_D.

À titre d'exemple, un document corrélé c ayant été validé par quatre évaluateurs lui

Document initial : Arrêté YYY relatif aux conditions techniques d'exploitation d'hélicoptères par une entreprise de transport aérien public (OPS 3).

1. Instruction YYY relative aux conditions techniques d'exploitation d'hélicoptères par une entreprise de transport aérien public (OPS 3), **Qualité : 3**.
2. Arrêté YYY fixant le programme et le régime des examens pour l'obtention du brevet et de la licence de pilote de ligne hélicoptère..., **Qualité : 2,4**.
3. Arrêté YYY modifiant divers textes réglementaires relatifs aux brevets, licences et qualifications des navigants de l'aéronautique civile, **Qualité : 0,9**.
4. Arrêté YYY relatif au régime de l'examen d'aptitude à la langue anglaise pour les navigants de l'aéronautique civile candidats à la qualification de vol aux instruments, **Qualité : 0,8**.
5. Décret YYY portant réglementation des services de la circulation aérienne militaire, **Qualité : 0,4**.

TAB. 8.3 – *Exemple de documents corrélés accompagnés des notations moyennes attribuées par les évaluateurs*

ayant successivement attribué les notes " *** ", " ** ", " * " et " ? " aura au final un score moyen de : $W(c) = \frac{3+2+1}{4} = 1,5$. Les scores présentés dans le tableau 8.3 illustrent ce calcul.

Afin de poursuivre, nous rappelons que $\chi_{D,f} = [\delta_1, \delta_2, \delta_3, \delta_4, \delta_5]$ représente les cinq résultats corrélés retrouvés par la méthode f, résultats classés par ordre de pertinence décroissante grâce à l'algorithme présenté dans le section 7.3.1.

Nous calculons la qualité partielle notée $Q_{f,e}(n)$ pour un évaluateur $e \in E$, relative à f et fonction du rang n : $Q_{f,e}(n) = W(\delta_n)$, avec $1 \leq n \leq 5$. L'équation 8.2 définit la qualité moyenne d'une méthode de corrélation f au rang n pour l'ensemble des documents de Δ et des évaluateurs de E.

$$Q_f(n) = \frac{1}{\|\Delta\|.\|E\|} \sum_{d \in \Delta} \sum_{e \in E} Q_{f,e}(n) = \frac{1}{\|\Delta\|.\|E\|} \sum_{d \in \Delta} \sum_{e \in E} W(\delta_n) \tag{8.2}$$

Les résultats sont alors publiés au travers des figures 8.5 et 8.6 qui représentent la qualité moyenne des documents respectivement pour des signatures composées de mots et de SN avec $F = F_{max}$. Les figures C.3 et C.4 reportées en Annexe C sont leurs pendants mais pour $F = F_{90}$.

Sur ces figures, on peut également distinguer une sixième courbe que nous qualifions de courbe de référence. Elle représente la qualité moyenne obtenue par les cinq premiers résultats corrélés désignés par les évaluateurs ; il s'agit donc de la courbe W. À ce titre elle permet de relativiser les scores des différentes méthodes de pondérations avec le score maximal qu'il est probablement envisageable d'atteindre.

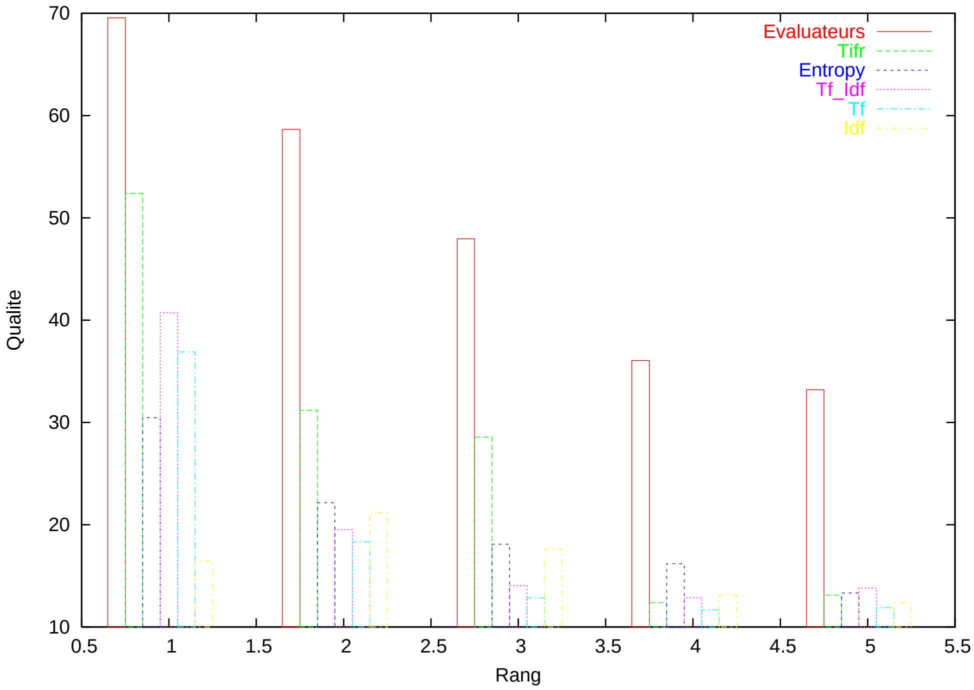


FIG. 8.5 – *Qualité moyenne des résultats corrélés en fonction du rang de la réponse. Cas des mots. Cas de la performance optimale* $F = F_{max}$ *(cf analyse des résultats section 8.3.2).*

8.3.2 Résultats

L'étude de la figure 8.3 nous montre qu'à l'image des résultats obtenus dans le chapitre 7, il existe une différence significative entre les résultats de T_{ifr} et les pondérations classiques. On peut ainsi constater des écarts de la précision moyenne allant jusqu'à 15% entre les méthodes utilisant des signatures lexicales composées de mots. La figure 8.4 nous apprend que cette différence n'est valable que pour les mots puisque ces écarts s'estompent avec des signatures composées de SN. On constate également que, mis à part pour l'indice T_{ifr}, la précision moyenne des résultats est meilleure lors de l'utilisation de SN.

La comparaison des figures 8.3 et 8.4 avec les figures 7.3 et 7.4 indique une perte de précision moyenne de l'ordre de 25% entre ces deux expérimentations en défaveur de notre étude qualitative. Cette comparaison nous montre également que les écarts constatés entre les différentes méthodes de pondération, qui pouvaient aller jusqu'à 40% (figures 7.3 et 7.4) lors de la précédente étude quantitative menée au chapitre 7, se sont considérablement restreints, en particulier lors de l'emploi de SN.

Les enseignement tirés de l'étude des figures 8.5 et 8.6 sont sensiblement les mêmes

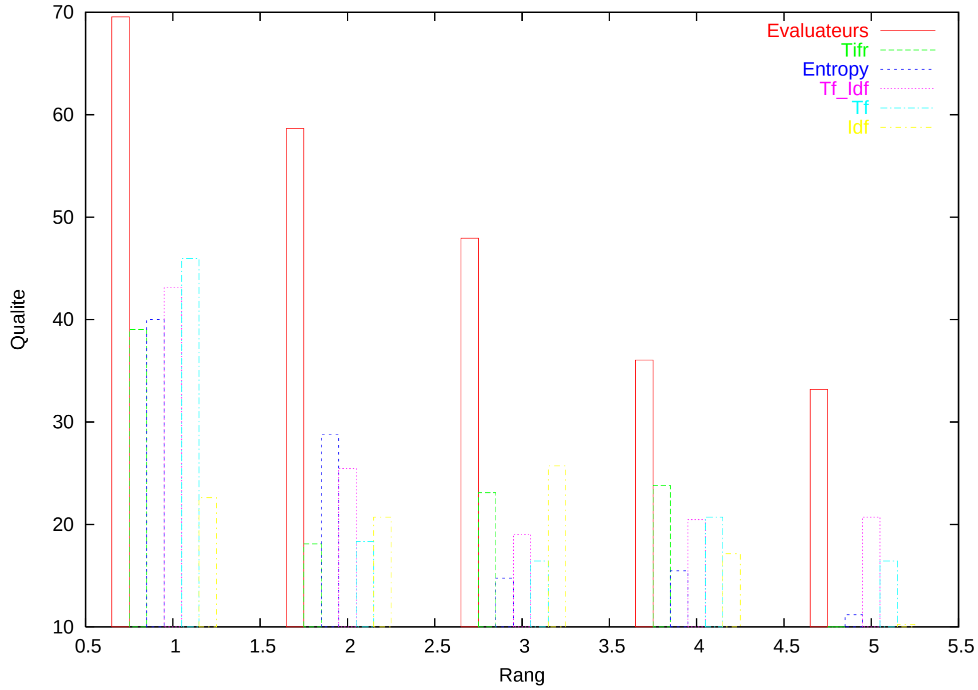


FIG. 8.6 – *Qualité moyenne des résultats corrélés en fonction du rang de la réponse. Cas des SN. Cas de la performance optimale* $F = F_{max}$ *(cf analyse des résultats section 8.3.2).*

que précédemment. Alors que pour le cas des mots, l'indice T_{ifr} arrive à se démarquer légèrement, l'étude de la figure 8.6 nous montre qu'il y a un resserrement entre les méthodes et que globalement, dans le cas des SN, il est difficile de les différencier.

L'étude de la courbe traduisant le choix des évaluateurs représente la valeur maximale qu'il est possible d'obtenir. Elle nous montre que, contrairement à toute attente, les deux premiers rangs ont des valeurs relativement faibles de l'ordre de 60% à 70%. La signification est que les évaluateurs ont noté globalement peu de documents “ *** ”, une qualité de l'ordre de 60% correspondant à des textes notés “ ** ”.

8.4 Enseignements

Cette expérimentation a permis de mettre en évidence plusieurs enseignements concernant la qualité des résultats corrélés extraits automatiquement. En premier lieu, le taux d'accord élevé entre les évaluateurs (figure 8.2) permet de penser qu'il existe un réel consensus sur le choix des résultats corrélés. Ce consensus montre également que les résultats corrélés n'ont que très rarement été qualifiés de *très pertinents* (“ *** ”). Cette constatation nous enseigne que, quelle que soit la méthode de corrélation employée, nos résultats restent éloignés des attentes des évaluateurs et donc de ceux des utilisateurs finaux puisqu'au vu des figures 8.5 et 8.6 nous retrouvons principalement des résultats étiquetés “ * ”.

De la même manière, la précision des résultats, même si elle reste convenable, reste en retrait de ce qui a été constaté dans le chapitre 7. L'effet conjugué d'une qualité réduite (telle qu'elle a été définie section 8.3.1) ainsi qu'une précision en retrait nous laisse entendre qu'une telle méthode de corrélation n'a de sens que si elle propose à l'utilisateur un panel réduit de résultats corrélés. Ce chapitre nous prouve, si besoin en était, que qualité et précision des résultats décroissent trop rapidement pour pouvoir espérer les retrouver en grand nombre. De ce point de vue, nous arrivons aux même conclusions que Park & al [PPGK02] et Phelps & Wilensky [PW00] qui utilisent les signatures lexicales pour obtenir un nombre très restreint de documents corrélés, de l'ordre d'une dizaine au maximum.

En terme de comparaison des pondérations en présence, cette expérience nous apprend deux choses : en premier lieu, que conformément aux chapitres antérieurs, l'indice T_{ifr}, fait mieux que les méthodes classiques de pondération (cas des mots) ou au moins aussi bien (cas des SN). Ensuite, cette expérience aura révélé l'importance de l'utilisation des SN dans la création des signatures lexicales. En effet, qu'il s'agisse des figures 8.4 ou 8.6, les résultats obtenus sont globalement meilleurs que lorsque des mots sont employés.

Ce constat, qui ne contredit pas les conclusions apportées au chapitre 7, montre que l'utilisation de SN en petit nombre (voir tableau 7.5) est suffisant pour améliorer la qualité et la pertinence des résultats corrélés. Les SN ne doivent donc pas être écartés de la composition des signatures lexicales, bien au contraire.

La dernière remarque que nous formulerons concerne l'étude des figures de l'Annexe C. Nous constatons que, quelque soit le critère choisi (précision ou qualité) ou le type de terme utilisé (mot ou SN), les résultats obtenus dans le cas ou $F = F_{90}$ sont toujours dégradés par rapport à ceux obtenus pour $F = F_{max}$. Cette dégradation peut être qualifiée de relativement lente et montre que la méthode de corrélation employée est relativement stable même si F est éloigné de sa valeur optimale (F_{max}). Néanmoins, les performances du système étant inférieures lorsque nous ne sommes pas dans les conditions optimales, le choix d'une méthode de pondération dédiée à la corrélation doit être guidé par plusieurs critères dont celui de pouvoir trouver une performance optimale F_{max}.

Quatrième partie

Synthèse

Chapitre 9

Description d'une méthode de corrélation entre documents

Dans ce chapitre, nous synthétisons les précédentes informations recueillies avec comme objectifs : de décrire une méthode de corrélation adaptée à notre corpus de référence (section 9.1), d'utiliser nos connaissances pour réaliser un outil de corrélation (section 9.2) et de discuter de la méthode employée, section 9.3 , en illustrant nos propos d'exemples de documents corrélés.

9.1 Bilan sur la corrélation

Cette section revoit les conclusions apportées par les chapitres 5, 7 et 8 dans le but de désigner, section 9.1.1, non pas la meilleure voie de corrélation dans l'absolu, mais le meilleur compromis dans la recherche de résultats corrélés. Une fois cette méthode de corrélation choisie, la section 9.1.2 rappelle ses limites et énonce les options retenues pour réaliser un outil de corrélation.

9.1.1 Entre performance et qualité des résultats

L'analyse des chapitres 5, 7 et 8 montre qu'il existe trois principaux aspects ou critères pour juger de la pertinence d'une méthode de corrélation :

- un aspect sémantique (chapitre 5) ;
- un aspect quantitatif (chapitre 7) ;
- un critère qualitatif (chapitre 8).

L'aspect sémantique a été introduit dans le but de comparer la proximité des signatures lexicales produites automatiquement et celles réalisées manuellement par des évaluateurs. Cette comparaison montre que l'extraction automatique de termes se rapproche d'une extraction manuelle et donc d'une extraction sémantique.

L'aspect quantitatif a été étudié en mettant en concurrence les cinq méthodes décrites afin de trouver celles qui maximisent le rapport précision/rappel, obtiennent les meilleures performances et retrouvent rapidement des résultats précis. Cet aspect abordé au chapitre 7 est en fait la synthèse des informations étudiées sections 7.2 et 7.3 : la performance et la précision moyenne des résultats corrélés.

À l'opposé l'aspect qualitatif étudié dans le précédent chapitre met en avant les méthodes qui retrouvent les résultats corrélés les plus proches des attentes des utilisateurs. De la même manière que pour l'aspect quantitatif, celui-ci est la combinaison de deux facteurs étudiés section 8.3.1.1 et 8.3.1.2 qui évoquent : la précision et la qualité moyenne des résultats corrélés triés par pertinence décroissante.

La méthode que nous souhaitons utiliser pour réaliser un outil de corrélation doit être un compromis entre ces trois différents critères. Afin de la dégager, nous avons résumé dans les tableaux 9.1 et 9.2 la pertinence des pondérations étudiées par rapport aux trois critères énoncés. Les aspects quantitatif et qualitatif ont tous deux été subdivisés en deux sous-aspects précédemment énoncés. Ces tableaux représentent le comportement des méthodes de pondérations pour des signatures lexicales composées respectivement pour des mots et des SN. Les notations qui y sont portées symbolisent le classement obtenu par les pondérations pour l'aspect ou le sous-aspect étudié. Ces chiffres vont de 1 à 5, 1 étant attribué à la meilleure pondération et 5 à la moins bonne. Des classements *ex aequo* apparaissent dès que les différences entre les critères étudiés sont trop minimes à nos yeux pour être sanctionnés par des valeurs différentes. Le tableau 9.2 donne des exemples de classements *ex aequo*, notamment pour l'aspect qualitatif.

critères	sémantique	quantitatif		qualitatif		rang moyen
		performance	précision	précision	qualité	
T_{ifr}	1	1	1	1	1	1
$T_f.Idf$	1	3	2	2	2	2
entropie	1	3	2	2	2	2
T_f	4	1	5	2	2	2.8
Idf	4	3	2	2	2	2.6

TAB. 9.1 – *Synthèse de la pertinence des méthodes de pondération employées avec des signatures lexicales composées de mots*

critères	sémantique	quantitatif		qualitatif		rang moyen
		performance	précision	précision	qualité	
T_{ifr}	1	1	1	1	1	1
$T_f.Idf$	2	3	2	1	1	1.8
entropie	2	3	3	1	1	2
T_f	2	1	3	1	1	1.6
Idf	5	3	3	1	1	2.6

TAB. 9.2 – *Synthèse de la pertinence des méthodes de pondération employées avec des signatures lexicales composées de SN*

Les enseignements que l'on peut tirer de ces tableaux sont multiples. On peut tout d'abord remarquer que le rang moyen obtenu par les différentes pondérations est à peu près constant, que l'on prenne en considération de signatures composées de mots ou des SN, à l'exception de T_f.

Ensuite on peut voir apparaître deux regroupements de pondérations : I_{df} et T_f d'une part et T_{ifr}, $Tf.I_{df}$ et l'entropie d'autre part. Les secondes étant plus évoluées que les premières, elles obtiennent naturellement un meilleur classement au final (le cas particulier de T_f pour les SN a été traité au chapitre 7).

L'enseignement qui se dégage de cette synthèse est que parmi les pondérations énoncées, T_{ifr} permet de retrouver des documents corrélés dans les meilleures circonstances quel que soit le critère étudié (sémantique, quantitatif ou qualitatif). Cette pondération expérimentale obtient, à l'issue de ce comparatif, le meilleur classement ; cela ne traduit pourtant pas une écrasante supériorité mais plutôt une constance face aux résultats obtenus. L'utilisation de T_{ifr} pour extraire des SN en est la preuve : elle ne fait pas mieux mais elle reste dans la moyenne des méthodes utilisées.

En conclusion, notre choix pour élaborer une méthode de corrélation entre documents fondée sur la création de signatures lexicales est d'utiliser T_{ifr} pour extraire indifféremment mots et syntagmes nominaux.

9.1.2 Description de la méthode de corrélation employée

Le choix d'un indice de pondération étant fixé, il reste encore à préciser les limites d'application d'une telle méthode et énoncer les éventuelles options prises.

La principale limite de cette méthode concerne le nombre de résultats corrélés que nous souhaitons présenter aux utilisateurs. Le précédent chapitre, en accord avec les résultats de Park & al [PPGK02] et Phelps & Wilensky [PW00], montre que l'on ne peut pas obtenir une grande quantité de résultats corrélés sans qu'une baisse rapide de leur qualité ne soit constatée. Le chapitre précédent permet d'étudier les résultats corrélés jusqu'au cinquième rang et met en avant une dégradation de la qualité non négligeable. Pour limiter cette dégradation mais également permettre à l'utilisateur d'avoir un panel de documents suffisant, nous adoptons un seuil empirique qui fixe à dix le nombre de résultats corrélés recherchés, les dix les plus pertinents.

Jusqu'à présent nous avons toujours poursuivi nos expérimentations sur des signatures de type homogène pour mettre en avant les qualités et les faiblesses de l'emploi des deux types en présence (mot ou SN). La conclusion de cette étude est que tous les deux ont leurs avantages, les mots améliorent l'aspect quantitatif de la recherche de résultats corrélés alors que les SN font progresser l'aspect qualitatif. L'utilisation de signatures lexicales mixtes pour la recherche de résultats corrélés ne peut donc que bénéficier de cette complémentarité. Pour réaliser notre méthode de corrélation, nous utiliserons des signatures mixtes.

L'utilisation de signatures mixtes soulève une dernière question : le choix des facteurs F_{sem} et F_{lon} vitaux à la recherche de résultats corrélés. Pour choisir la valeur de F_{lon}, deux cas de figure se présentent à nous : Nous pouvons tout d'abord considérer que nous extrayons séparemment les F_{lon} premiers mots une fois pondérés et les F'_{lon} premiers SN pour constituer nos signatures (voir tableaux 7.2 et 7.3 pour les valeurs de F_{lon} et F'_{lon}). Nous pouvons également extraire les $F_{lon} + F'_{lon}$ premiers termes quel que soit leur type.

Nous préférons extraire les termes séparément en utilisant la première méthode car nous n'avons pas étudié préalablement la valeur moyenne prise par T_{ifr} pour les mots et les SN. Cette pondération, comme toutes les autres, risque d'avantager un type de terme par rapport à un autre, ce qui serait au détriment de la pertinence des signatures lexicales.

Pour le choix de F_{sem} l'homogénéité ou la mixité des signatures influence directement la valeur de cet indice, valeur qui initialement était de 3 indifféremment pour les mots ou les SN. L'examen des tableaux 7.2 et 7.3 nous montre que les valeurs de F_{sem} restent faibles quelle que soit la pondération utilisée et les valeurs de F_{lon} considérées. En toute logique, la nouvelle valeur de F_{sem} à prendre en considération ne doit pas être très éloignée de la précédente. Comme nous avons presque doublé la taille des signatures lexicales, nous préconisons d'utiliser $F_{sem} = 5$ comme nouvelle valeur seuil.

Toute modification de cette valeur peut paraître hasardeuse, néanmoins les chapitres 7 et 8 montrent que la pondération retenue est relativement stable aux variations de (F_{lon}, F_{sem}) par rapport à la performance optimale F_{max}. On peut donc raisonnablement penser que cette nouvelle valeur de F_{sem} n'affectera, dans le pire des cas, que peu les résultats corrélés. D'autre part, une valeur de F_{sem} trop élevée dégrade rapidement le taux de rappel des résultats mais n'affecte que modérément leur précision. La limitation du nombre de résultats corrélés à dix fait donc passer la dégradation du taux de rappel au second plan.

9.2 Un outil de corrélation entre documents

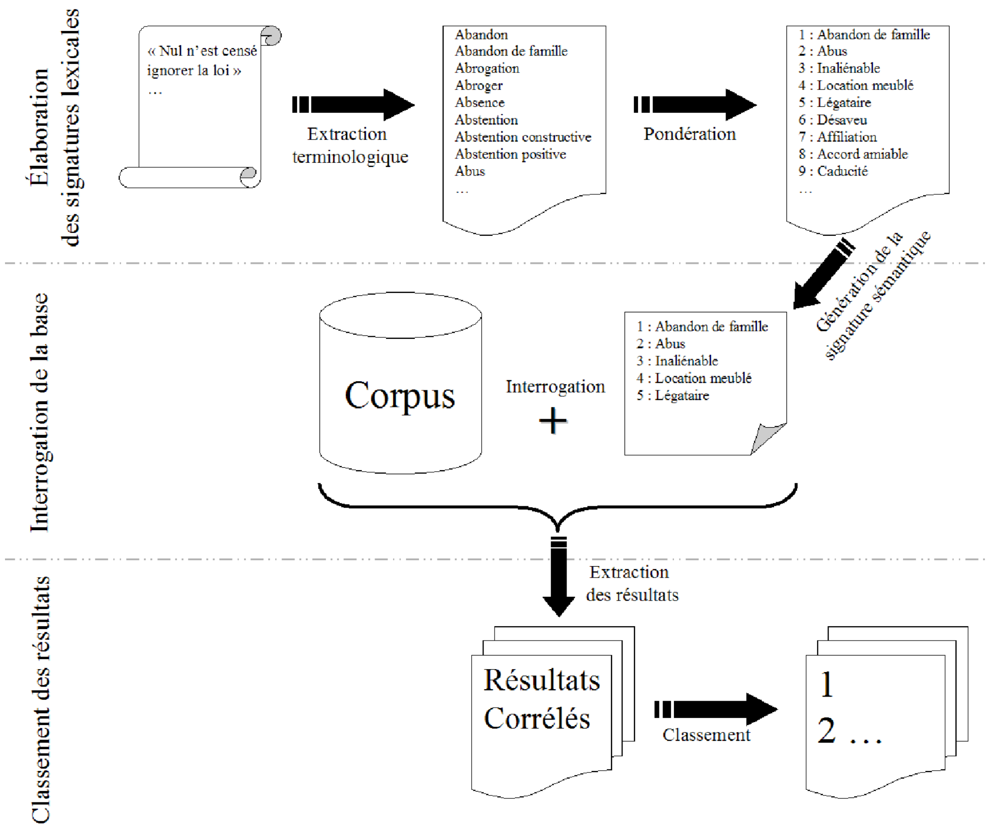


FIG. 9.1 – *Les étapes de la corrélation*

Cette section, contrairement à la précédente et aux chapitres antérieurs, montre la corrélation sous un angle nouveau, moins théorique et plus applicatif. La motivation de cette partie est de décrire d'un point de vue technique notre méthode de corrélation et de l'illustrer au travers de nos réalisations. La finalité de notre méthode de corrélation est de rechercher des résultats corrélés sur notre corpus de référence mais également dans l'ensemble des documents détenus par un portail juridique[1] regroupant des documents tels que les JO, les codes de la république française ou encore des textes législatifs européens.

Cette base documentaire est, de par sa constitution, en perpétuelle expansion grâce notamment à la parution quotidienne de nouveaux JO. La méthode de corrélation doit conformément aux attentes formulées au chapitre 4, être interactive et permettre de prendre en compte, par exemple, les nouveaux JO dès leur parution.

Pour comprendre et localiser la difficulté technique, prenons la figure 9.1 qui illustre la recherche de résultats corrélés. Celle-ci se scinde en trois étapes :

[1] `http://www.admi.net`

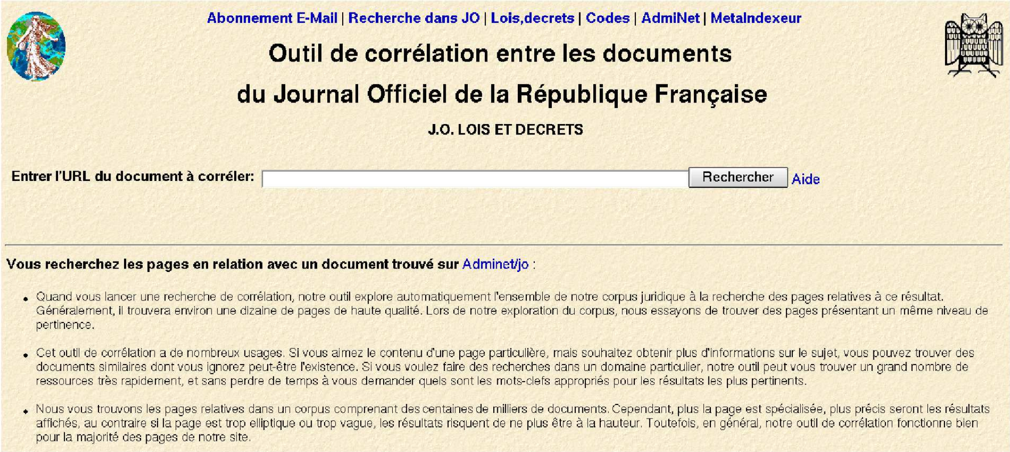


FIG. 9.2 – *Recherche de résultats corrélés sur les JO*

- l'élaboration des signatures lexicales ;
- l'interrogation de la base documentaire ;
- le classement des résultats corrélés.

Parmi ces étapes, les deux premières sont directement influencées par la modification de la base documentaire existante. Son interrogation a été confiée au moteur de recherche Pertimm [Per] qui prend en compte les modifications apparues sur notre corpus (en terme d'ajouts ou de suppressions de documents).

La première étape qui concerne l'élaboration des signatures lexicales est délicate, à cause notamment de la pondération. En effet, une fois l'extraction terminologique effectuée et les mots ou SN inutiles soustraits, nous nous retrouvons avec une liste de termes que nous devons pondérer. Cette liste est plus ou moins longue et à cause des hapax, dus principalement aux noms propres présents dans les documents et aux SN, sa taille est le plus souvent proportionnelle à la longueur du document considéré.

D'autre part, pour pondérer ces termes, nous devons posséder leurs statistiques (nombre d'occurrence, nombre de documents,...), statistiques qui évoluent au fil du temps car soumises aux variations du corpus ; nous devons donc les déterminer dynamiquement. Pour les obtenir, nous nous référons aux indications fournies par Pertimm [Per]. Malgré sa rapidité, nous devons en moyenne attendre 300 ms par terme pour les obtenir. Ce délai s'explique par le fait que l'obtention de ces statistiques nous oblige à détourner l'outil de sa fonction première et altère ses performances.

La conséquence directe est que pour des documents de taille raisonnable (plus de 100 termes extraits), on peut attendre plus de 30 secondes avant d'obtenir les informations souhaitées. À ce stade nous n'avons même pas encore abordé la recherche de résultats corrélés que nous avons déjà perdu notre internaute depuis longtemps.

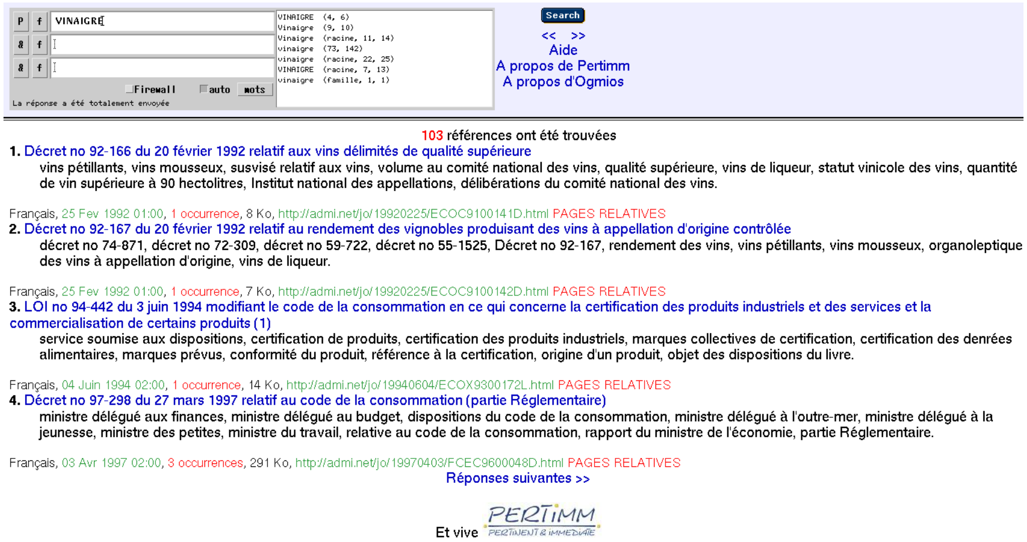


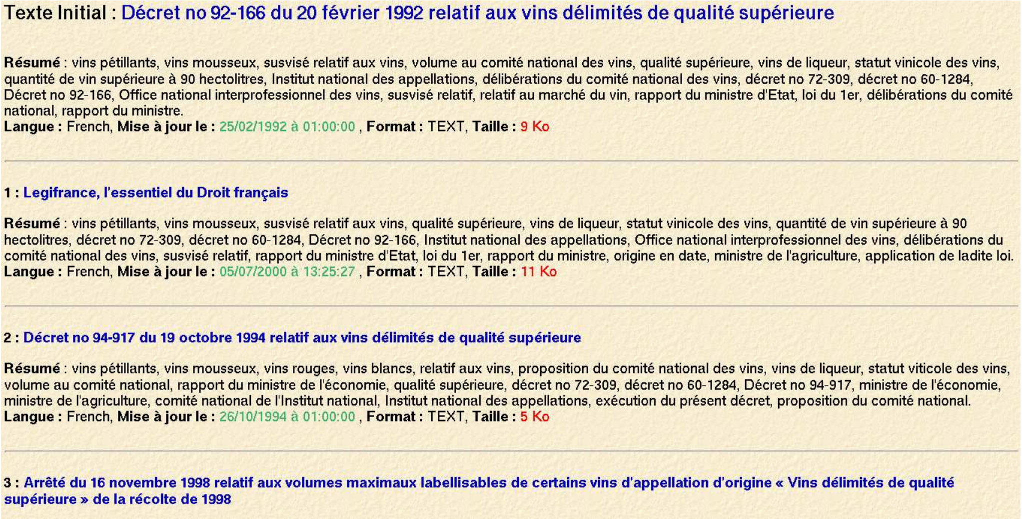
FIG. 9.3 – *Intégration de la recherche de résultats corrélés à l'interface du moteur de recherche Pertimm [Per]*

Pour accélérer notre procédé, nous sommes obligés de construire préalablement les signatures lexicales afin de rechercher directement les documents corrélés. Néanmoins nos signatures ne sont pas définitives, nous leur octroyons une durée de vie d'un mois avant de les régénérer (dans la pratique, une partie des signatures est régénérée chaque jour). Sur une période d'un mois, 2.000 nouveaux documents auront, en moyenne, été publiés, un chiffre à comparer aux 210.000 documents actuellement présents sur notre portail. Certes les signatures peuvent changer en un mois mais nous postulons que l'approximation n'est pas très éloignée de la réalité. Cette façon de procéder ne nous prive pas d'obtenir des articles parus le jour même, puisque la recherche et l'extraction des documents corrélés reste une étape dynamique.

Concrètement et techniquement, notre méthode de corrélation se scinde en deux étapes : une étape de pré-traitement et une étape interactive.

Durant l'étape de pré-traitement, nous générons une partie des signatures lexicales valables un mois durant. Les termes, mots et SN, sont extraits grâce à une version révisée de Sylex nommée Genet incorporée au moteur de recherche Pertimm [Per]. La pondération des termes se fait grâce à l'indice T_{ifr} et aux statistiques extraites de Pertimm [Per]. Des signatures mixtes sont ainsi constituées à partir des 20 premiers mots et 10 premiers SN trouvés conformément aux valeurs reportées dans les tableaux 7.2 et 7.3.

L'étape interactive consiste à interroger le moteur de recherche Pertimm [Per] en formulant, à partir de la signature lexicale, une requête conceptuelle avec $F_{sem} = 5$. Simultanément les documents corrélés sont triés grâce à l'algorithme décrit section 7.3.1. Les 10



Texte Initial : Décret no 92-166 du 20 février 1992 relatif aux vins délimités de qualité supérieure

Résumé : vins pétillants, vins mousseux, susvisé relatif aux vins, volume au comité national des vins, qualité supérieure, vins de liqueur, statut vinicole des vins, quantité de vin supérieure à 90 hectolitres, Institut national des appellations, délibérations du comité national des vins, décret no 72-309, décret no 60-1284, Décret no 92-166, Office national interprofessionnel des vins, susvisé relatif, relatif au marché du vin, rapport du ministre d'Etat, loi du 1er, délibérations du comité national, rapport du ministre.
Langue : French, **Mise à jour le :** 25/02/1992 à 01:00:00 , **Format :** TEXT, **Taille :** 9 Ko

1 : Legifrance, l'essentiel du Droit français

Résumé : vins pétillants, vins mousseux, susvisé relatif aux vins, qualité supérieure, vins de liqueur, statut vinicole des vins, quantité de vin supérieure à 90 hectolitres, décret no 72-309, décret no 60-1284, Décret no 92-166, Institut national des appellations, Office national interprofessionnel des vins, délibérations du comité national des vins, susvisé relatif, rapport du ministre d'Etat, loi du 1er, rapport du ministre, origine en date, ministre de l'agriculture, application de ladite loi.
Langue : French, **Mise à jour le :** 05/07/2000 à 13:25:27 , **Format :** TEXT, **Taille :** 11 Ko

2 : Décret no 94-917 du 19 octobre 1994 relatif aux vins délimités de qualité supérieure

Résumé : vins pétillants, vins mousseux, vins rouges, vins blancs, relatif aux vins, proposition du comité national des vins, vins de liqueur, statut viticole des vins, volume au comité national, rapport du ministre de l'économie, qualité supérieure, décret no 72-309, décret no 60-1284, Décret no 94-917, ministre de l'économie, ministre de l'agriculture, comité national de l'Institut national, Institut national des appellations, exécution du présent décret, proposition du comité national.
Langue : French, **Mise à jour le :** 26/10/1994 à 01:00:00 , **Format :** TEXT, **Taille :** 5 Ko

3 : Arrêté du 16 novembre 1998 relatif aux volumes maximaux labellisables de certains vins d'appellation d'origine « Vins délimités de qualité supérieure » de la récolte de 1998

FIG. 9.4 – *Présentation de quelques documents corrélés*

premiers résultats sont finalement présentés à l'utilisateur.

Des exemples d'intégration de cette méthode de corrélation au sein de notre portail juridique sont donnés par les figures 9.2 et 9.3. La figure 9.2 représente un formulaire d'interrogation qui, à partir d'une URL donnée, renvoie une série de documents corrélés (valable pour les JO uniquement). La figure 9.3 illustre l'intégration de la corrélation au cœur même du moteur de recherche Pertimm [Per]. La figure 9.4 donne un exemple de présentation des documents corrélés suite à une recherche.

9.3 Discussion autour d'un panel de document corrélés

La description de notre approche de la corrélation ne serait pas complète sans des exemples de résultats corrélés. Cette section présente un florilège de résultats choisis pour mettre en avant quelques caractéristiques propres à la méthode utilisée. Cette partie illustre, à partir de cas concrets, quelques avantages et inconvénients propres à l'utilisation de la méthode de corrélation.

Chaque exemple, détaillé en annexe D, est composé de deux parties : Un document initial servant de base à la corrélation ainsi qu'un tableau regroupant les documents corrélés et la signature lexicale à l'origine de ces résultats.

9.3.1 Les termes du domaine juridique

Au sein de notre corpus composé de 210.000 documents juridiques, on peut dégager différentes thématiques. À titre d'exemple, voici quelques thématiques du tableau 7.1 : les télécommunications, le vin, les nominations. Les 210.000 documents, bien qu'étant spécifiques au domaine juridique appartiennent également à d'autres thématiques. La méthode de corrélation exposée permet justement de regrouper les documents en fonction de leurs spécificités non juridiques.

Prenons l'exemple des documents illustrés par les figures D.3 et D.5. Ces documents diffèrent tout d'abord par la taille mais également par les thèmes qui les composent. Alors que le document D.5 fait référence à la *délégation de signature*, le document D.3 traite de thématiques propres au domaine des *télécommunications*. L'étude de leurs signatures lexicales respectives montre que les termes qui les constituent sont principalement des termes usuels du langage sans connotation juridique particulière. Conformément aux dires de la section 4.3.2, l'utilisation de termes ordinaires n'est pas un handicap en terme de corrélation. En pratique on s'aperçoit que ceux-ci sont même majoritaires. Ainsi dans les exemples précédents des termes comme *délégation de signature* ou *autorité de régulation des télécommunications* sont absents des signatures alors qu'il s'agit de termes du domaine pertinents.

Malgré l'absence de tels termes, les documents corrélés respectifs, présentés dans les tableaux D.3 et D.5, montrent que la thématique de chacun est respectée. Des documents traitant de *télécommunications* sont associés au document D.3 et des documents traitant de *délégation de signature* sont associés au document D.5. Cela pourrait laisser croire que ces termes du domaine ne sont pas nécessaires à la recherche des résultats corrélés.

Néanmoins leur quasi absence peut être parfois gênante. Considérons le document D.1 associé aux résultats du tableau D.1. Cet exemple nous présente un document traitant de *nominations au comité régional vins et eaux-de-vie* et des *appelations d'origine*. Là encore, peu de termes du domaine sont présents dans la signature à l'exception de *relatif à l'appellation d'origine* ou de *code général des impôts*. La conséquence est qu'aucun des documents corrélés ne reprend la thématique exacte du document initial, les *nominations dans le domaine viticole*. Malgré tout, les résultats corrélés restent proches des thématiques *viticole* ou des *appellations d'origine*.

Dans un cadre plus général, la méthode de corrélation décrite retrouve globalement des résultats corrélés de même nature. Néanmoins ceux-ci peuvent éventuellement manquer de précision sémantique car l'utilisation des termes du domaine juridique reste limitée.

9.3.2 Taille des signatures et taille de documents

Dans le chapitre 5, nous avions montré que la taille de la signature lexicale pouvait être indépendante de la taille du document ; mais qu'en est-il réellement pour les documents corrélés. Pour illustrer cette question nous étudions les documents D.1 et D.2 qui traitent tous deux de problématiques *viticoles* en plus de leur composante juridique.

Les résultats corrélés, présentés respectivement dans les tableaux D.1 et D.2, nous montrent que malgré des différences évidentes de taille (quelques lignes contre deux pages), les documents corrélés n'en souffrent pas. Les résultats restent dans les deux cas proches de la thématique initiale sans qu'il soit constaté de dérive grave. Le postulat, démontré au chapitre 5, qui affirmait que la taille du document n'influence pas la signature, se vérifie sur notre exemple.

Dans le cadre général, nous avons pu vérifier que, sauf cas extrêmes (documents constitués de plusieurs dizaines de pages ou d'une unique phrase), la taille du document n'influence que très peu la qualité des résultats corrélés.

9.3.3 L'aspect linguistique

Le chapitre 4 consacré en partie au positionnement et à la définition de notre méthode de corrélation nous a laissé entrevoir de nombreuses possibilités quant à la recherche de résultats corrélés. Parmi ces possibilités, nous avions initialement rejeté toute utilisation de quelque méthode linguistique que ce soit pour élaborer nos signatures lexicales. Malgré le constat de résultats corrélés probants permettant à l'internaute d'avoir davantage d'informations sur un document initial, le tout statistique atteint quand même ses limites.

L'exemple le plus flagrant est celui des résultats présentés dans le tableau D.4 associé au document illustré par la figure D.4. Ce document relativement long en nombre de termes mais restreint par la quantité d'information possède une signature lexicale peu convaincante car composée en grande partie de prénoms.

Les documents corrélés retrouvés sont tous dans la même thématique, *promotion et nomination*, et ne sont pas très éloignés de celle du document initial. Néanmoins ils n'ont pas été retrouvés grâce aux termes porteurs de sens qui constituent la signature lexicale : *chasse, faune, sauvage, mandat, écologie, titulaires, faune sauvage, conseil national, membres titulaires, mandat des membres, membres du conseil national, mandats des membres, conseil national de la chasse, durée du mandat.* Les documents corrélés retrouvés l'ont été uniquement sur la base des prénoms qui les composent : *henri, frédéric, raymond, yves, gilbert, bruno, guy, vincent, paul, mathieu.*

Nous pouvons appréhender ce résultat de deux façons diamétralement opposées : la première est de constater que dans notre corpus seuls les documents traitant de *nomination et promotion* sont largement constitués de prénoms et que de ce point de vue, la participation de ce type de termes dans la composition de la signature a été utile. D'un autre côté on peut considérer qu'un prénom seul est une notion sans intérêt d'un point de vue sémantique et qu'à ce titre nous n'aurions pas dû le prendre en compte. Dans ce deuxième cas de figure, qui reflète notre pensée, il faut pouvoir faire appel à un traitement supplémentaire pour éliminer certaines catégories de termes telles que les prénoms, les verbes, etc. Une solution envisageable est de faire appel à des méthodes linguistiques, à l'utilisation de taxonomie, de thésaurus ou d'ontologie pour ne garder dans nos signatures que des termes sémantiquement pertinents.

Le point sur lequel nous souhaitions insister est que notre approche intégralement statistique donne certes de bons résultats, mais en étudiant de plus près les signatures lexicales générées, on peut s'apercevoir, comme dans l'exemple précédent que nous arrivons parfois au bon résultat mais pour de mauvaises raisons. La coopération avec des procédés linguistiques permettrait, peut être, d'améliorer les résultats mais également la génération des signatures.

Chapitre 10

Conclusions et perspectives

Nous avons présenté au cours de cette thèse les étapes qui ont prévalu à l'élaboration et à la réalisation d'une méthode de corrélation entre documents dédiée à la recherche d'information sur un corpus juridique afin de rendre cette information plus accessible vis-à-vis des internautes.

Notre approche a été de considérer une voie de corrélation linguistique qui, au travers de l'élaboration de signatures lexicales, nous permet de rechercher des résultats corrélés. Au cours de cette thèse, les différents degrés de liberté énoncés au chapitre 4 ont été progressivement levés le plus souvent grâce à l'expérimentation et à la mise en concurrence des différentes approches possibles, par exemple, pour le choix de la pondération ou la définition d'une unité de recherche.

Le domaine de recherche de la corrélation est une voie peu souvent explorée contrairement à d'autres thématiques telles que la catégorisation ou la classification de documents. Cette thèse permet d'ouvrir de nouvelles perspectives en terme de corrélation linguistique et pose les bases d'une méthodologie de corrélation entre documents. Elle s'achève sur la réalisation d'une méthode de corrélation et son intégration au sein d'outils de recherche d'information spécifiques à notre portail juridique. Néanmoins nous préférons concevoir notre contribution comme un travail préliminaire, comme une méthodologie qui doit encore évoluer et être remaniée afin d'être améliorée. Le but qui nous a animé pour réaliser cette méthode de corrélation a été de définir avec le plus de soin possible toutes les étapes intermédiaires nécessaires et de dégager les problématiques liées. À chaque étape, nous avons justifié nos choix sur un axe théorique et/ou expérimental qui nous a permis en fonction des éléments présents de dessiner au fur et à mesure les contours de notre méthode. Cette démarche ne doit pas être considérée comme une fin en soi, il s'agit avant tout d'un travail qui apporte une nouvelle dimension aux recherches visant à l'élaboration de méthodes de corrélation linguistique entre documents mais elle comporte également des limites qui ne doivent pas être négligées.

10.1 Contributions

Nous avons étudié plusieurs voies de corrélation et mis en concurrence cinq indices de pondération différents qui nous ont permis d'obtenir les résultats suivants :

10.1.1 Une corrélation sémantique

Une première contribution concerne l'aspect sémantique de notre méthode de corrélation. Parmi les objectifs initiaux figure l'aspect sémantique de la méthode de corrélation. Nous avons constaté que cet aspect sémantique se manifeste aux travers de deux caractéristiques : la génération des signatures lexicales et la définition d'une unité de recherche.

Ces deux aspects sont traités dans les chapitres 5 et 6 qui montrent :

- d'une part, qu'une signature lexicale générée automatiquement est relativement proche d'une extraction terminologique sémantique (cas d'une extraction manuelle, voir chapitre 5) ;
- d'autre part, la définition du paragraphe comme unité de recherche rend le nombre de résultats corrélés intuitivement acceptables et s'accorde avec la conception commune qu'un paragraphe traduit une idée particulière de l'auteur et forme ainsi une unité sémantique.

Grâce à ces deux résultats, la méthode de corrélation ainsi réalisée peut être qualifiée de sémantique malgré l'absence totale de décision d'expert ou d'intervention humaine comme lors de la réalisation d'ontologies, par exemple.

10.1.2 Élaboration d'un nouvel indice de pondération

Le chapitre 5 a défini et expérimenté un nouvel indice de pondération nommé T_{ifr}. Sa comparaison avec des pondérations classiques montre qu'il est mieux adapté à la recherche de termes consensuels (mots ou SN) et améliore la précision des signatures lexicales générées. Ce nouvel indice dont la conception avait pour principal objectif d'être dédié à la recherche de termes discriminants pour la corrélation, peut néanmoins être expérimenté dans d'autres travaux utilisant des méthodes statistiques de pondération. Notre contribution inclut l'élaboration de cet indice de pondération statistique dont les champs d'application restent encore à définir.

10.1.3 Des signatures lexicales dédiées à la corrélation

Le chapitre 7 présente une méthodologie destinée à optimiser la création de signatures lexicales et soulève plusieurs problématiques non encore abordées, à notre connaissance, par les travaux de recherche contemporains. Plusieurs axes pour améliorer la pertinence des signatures lexicales sont abordés. Nous avons notamment travaillé à la recherche des types de termes les mieux adaptés à la fabrication des signatures, à leur longueur optimale (F_{lon}) ou encore au type d'utilisation au niveau des requêtes.

Toutes ces problématiques ont été abordées dans un cas particulier, celui de l'utilisation d'un corpus juridique en langue française. Leur solution montre qu'il est possible d'optimiser les signatures pour un corpus défini pour les rendre plus efficaces lors de la recherche de résultats corrélés.

10.1.4 Description d'une méthodologie dédiée à la corrélation

L'objectif principal de ce mémoire est de dégager une méthode de corrélation linguistique sur un corpus quelconque et pas seulement sur des textes législatifs français. En effet, nous avons défini plusieurs étapes clefs que nous avons appelées degrés de liberté. La détermination de ces degrés de liberté dépend du corpus étudié. Néanmoins, les choix d'utiliser, comme termes, des mots et des SN, de prendre comme unité de recherche le paragraphe, d'utiliser les signatures comme des requêtes conceptuelles, peuvent être transposés aisément à de nombreux corpus spécialisés. Ces corpus spécialisés sont ceux pour lesquels une thématique technique est prédominante (le corpus des JO est un corpus spécialisé dans le domaine jurique).

Notre méthode est générique et permet d'analyser un corpus spécialisé quelconque et d'y retrouver des documents corrélés. Le principal paramètre de la méthode est la pondération à utiliser, T_{ifr} dans notre cas, mais celle-ci peut être différente pour un autre corpus. Cette pondération peut être déterminée par l'expérimentation, le critère de sélection étant de maximiser F. La détermination de ce critère donne également les valeurs de F_{lon} et F_{sem} à utiliser.

10.1.5 Implémentation et réalisation

Notre dernière contribution, abordée au travers du chapitre 9, concerne la réalisation technique de la corrélation sur le portail adminet[1]. La méthode a été également intégrée à des outils existants de recherche d'information comme par exemple au moteur de recherche Pertimm.

[1] `http://www.admi.net/jo`

10.2 Perspectives

Notre approche de la corrélation a permis de soulever de nombreuses problématiques dont certaines ont été résolues comme en témoigne la section précédente mais certaines améliorations doivent encore être expérimentées.

10.2.1 Vers une meilleure intégration

La première application pratique consiste à achever l'implémentation et la mise à disposition de notre méthode de corrélation sur l'ensemble des documents contenus sur notre portail juridique (pour l'instant le déploiement est limité aux JO).

Une amélioration consisterait à automatiser la détermination de certains degrés de liberté comme F_{lon} et F_{sem} qui pour l'instant nécessite une intervention humaine. L'objectif serait de fournir notre méthode de corrélation sous la forme d'une application, où les degrés de liberté seraient paramétrables ou éventuellement déterminés automatiquement.

10.2.2 Des signatures plus compactes

Il est impératif de déterminer une performance optimale F. Il est nécessaire de pouvoir juger de la validité d'une pondération et, d'autre part, de trouver les meilleures valeurs pour F_{lon} et F_{sem}. Nous souhaitons en effet retrouver le plus grand nombre de documents corrélés tout en gardant une pertinence élevée.

Le chapitre 8 montre que, malgré nos efforts, la qualité des résultats décroît rapidement. De ce fait, nous préconisons, au chapitre 9, de ne pas rechercher plus de 10 résultats corrélés sous peine d'une dégradation trop prononcée de leur pertinence.

La recherche d'une condition optimale $F = F_{max}$ a montré son intérêt pour l'obtention de résultats corrélés, mais peut être améliorée. En partant du principe que la recherche d'un grand nombre de résultats corrélés pertinents reste une question difficile, on peut imaginer de dégrader les valeurs des indices (avec une taille de signature F_{lon} plus petite qu'actuellement et un nombre de termes communs F_{sem} plus grand) pour ne retrouver à chaque fois qu'un nombre restreint de résultats mais d'une qualité plus grande (voir la définition de la qualité au chapitre 8).

10.2.3 Une mixité mal évaluée

Parmi les choix du chapitre 9, on peut éventuellement critiquer l'utilisation de signatures mixtes. De telles signatures combinent à l'évidence les avantages des deux types de termes en présence : mot et SN. Néanmoins, en toute rigueur, nous aurions dû expérimenter les effets de leur mixité sur la recherche d'une performance optimale F et le choix des valeurs de F_{lon} et F_{sem}.

Cette expérimentation n'a pas été envisagée faute de temps. Pour déterminer la valeur optimale de F, nous avions été obligés de faire varier simultanément les valeurs de F_{lon} et F_{sem} (chapitre 7). Le choix de signatures mixtes multiplie par deux le nombre de valeurs à faire varier simultanément. La recherche de l'optimalité de F dans le cadre de signatures mixtes est un travail exigeant qui peut à terme améliorer la performance des signatures générées et la qualité des résultats corrélés.

10.2.4 Utiliser les caractéristiques du corpus

Comme évoqué au chapitre 9, notre approche de la corrélation est purement statistique. Le principal avantage de cette approche est d'être robuste, c'est à dire indépendante du corpus d'étude à condition toutefois que celui-ci soit spécialisé. Cette robustesse devrait permettre d'expérimenter notre corrélation sur d'autres corpus spécialisés.

Afin d'améliorer la qualité des résultats corrélés, une exploitation plus fine des caractéristiques du corpus semble nécessaire. À ce titre, le filtrage et l'utilisation de termes spécifiques au domaine juridique permettraient d'améliorer les recherches de résultats corrélés [Lam02]. Plus généralement, l'utilisation de méthodes et de filtrages linguistiques complémentaires lors de la création des signatures lexicales serait une aide non négligeable pour notre recherche de résultats corrélés.

Table des figures

2.1 *La croissance du Web entre 1999 et 2000 [Cyv00]* 19

2.2 *Le Web invisible par type de document [Pla00]* 20

2.3 *Fonctionnement interne d'un moteur de recherche* 22

2.4 *Typologie des requêtes des internautes [Ad']* 25

3.1 *Recherches sur* `www.voila.fr` 36

3.2 *Exemple de calcul des H_i et A_i* 41

3.3 *Construction de S pour l'algorithme Companion* 43

3.4 *Construction de S pour l'algorithme Co-citation* 45

4.1 *Illustration des travaux de Lame : typologie de termes juridiques [Lam02].* 78

4.2 *Les différentes étapes de la corrélation* 79

4.3 *Les différents degrés de liberté* 80

5.1 *Valeur moyenne de T_f pour un terme présent 30 fois dans le corpus (moyenne sur 5604 décompositions possibles)* 90

5.2 *Valeur moyenne de I_{df} pour un terme présent 30 fois dans le corpus (moyenne sur 5604 décompositions possibles)* 91

5.3 *Valeur moyenne de $T_f.I_{df}$ pour un terme présent 30 fois dans le corpus (moyenne sur 5604 décompositions possibles)* 92

5.4 *Valeur moyenne de l'entropie d'un terme présent 30 fois dans le corpus (moyenne sur 5604 décompositions possibles)* 94

5.5 Valeur moyenne de iR pour un terme présent 30 fois dans le corpus (moyenne sur 5604 décompositions possibles) . . . 95

5.6 Valeur moyenne T_{ifr} pour un terme présent 30 fois dans le corpus (moyenne sur 5604 décompositions possibles) . . . 98

5.7 Illustration d'un questionnaire d'évaluation . . . 100

5.8 Renseignements demandés aux évaluateurs . . . 100

5.9 Caractéristiques du corpus expérimental au regard de la taille des documents ($T1$), le nombre de termes validés ($T2$) et consensuels ($T3$) . . . 102

5.10 Corrélation entre taille des documents et nombre de termes validés. En abscisse, les documents sont triés par taille croissante. . . . 103

5.11 Taux d'accord entre évaluateurs . . . 104

5.12 Comparaison des méthodes de pondération pour les mots. En abscisse, le nombre de termes comparés. . . . 108

5.13 Comparaison des méthodes de pondération pour les SN. En abscisse, le nombre de termes comparés. . . . 109

6.1 Parcours de l'ensemble des questions possibles grâce à un arbre binaire . . . 117

6.2 Parcours de l'arbre des questions après optimisation . . . 118

6.3 Nombre de requêtes valides contenant le mot de rang N avec le document comme unité de recherche. Les mots sont ordonnés selon la pondération décroissante de l'un des 3 indices. . . . 124

6.4 Nombre de requêtes valides contenant le mot de rang N avec le paragraphe comme unité de recherche. Les mots sont ordonnés selon la pondération décroissante de l'un des 3 indices. . . . 125

6.5 Taux d'utilisation des mots dans les questions en prenant le document comme unité de recherche. Les mots sont ordonnés selon la pondération décroissante de l'un des 3 indices. . . . 126

6.6 Taux d'utilisation des mots dans les questions en prenant le paragraphe comme unité de recherche. Les mots sont ordonnés selon la pondération décroissante de l'un des 3 indices. . . . 127

6.7 *Rang moyen du dernier mot constituant une question valide en prenant le document comme unité de recherche. Les mots sont ordonnés selon la pondération décroissante de l'un des 3 indices.* . . . 128

6.8 *Rang moyen du dernier mot constituant une question valide en prenant le paragraphe comme unité de recherche. Les mots sont ordonnés selon la pondération décroissante de l'un des 3 indices.* . . . 129

7.1 *Performance moyenne en utilisant des mots, en fonction du nombre de termes dans la requête* . . . 139

7.2 *Performance moyenne en utilisant des SN, en fonction du nombre de termes dans la requête* . . . 140

7.3 *Précision moyenne des résultats de la corrélation en utilisant des mots, en fonction du rang du document dans la liste des réponses* . . . 143

7.4 *Précision moyenne des résultats de la corrélation en utilisant des SN, en fonction du rang du document dans la liste des réponses* . . . 144

8.1 *Illustration d'un formulaire d'interrogation* . . . 154

8.2 *Taux d'accord entre évaluateurs* . . . 155

8.3 *Précision moyenne des résultats corrélés en fonction du rang de la réponse. Cas des mots. Cas de la performance optimale $F = F_{max}$ (cf analyse des résultats section 8.3.2).* . . . 157

8.4 *Précision moyenne des résultats corrélés en fonction du rang de la réponse. Cas des SN. Cas de la performance optimale $F = F_{max}$ (cf analyse des résultats section 8.3.2).* . . . 158

8.5 *Qualité moyenne des résultats corrélés en fonction du rang de la réponse. Cas des mots. Cas de la performance optimale $F = F_{max}$ (cf analyse des résultats section 8.3.2).* . . . 160

8.6 *Qualité moyenne des résultats corrélés en fonction du rang de la réponse. Cas des SN. Cas de la performance optimale $F = F_{max}$ (cf analyse des résultats section 8.3.2).* . . . 161

9.1 *Les étapes de la corrélation* . . . 169

9.2 *Recherche de résultats corrélés sur les JO* . . . 170

9.3 *Intégration de la recherche de résultats corrélés à l'interface du moteur de recherche Pertimm [Per]* . . . 171

9.4 *Présentation de quelques documents corrélés* . . . 172

C.1 *Précision moyenne des résultats corrélés en fonction du rang de la réponse. Cas des mots. Cas d'une performance dégradée* $F = F_{90}$. . . . 212

C.2 *Précision moyenne des résultats corrélés en fonction du rang de la réponse. Cas des SN. Cas d'une performance dégradée* $F = F_{90}$. . . . 212

C.3 *Qualité moyenne des résultats corrélés en fonction du rang de la réponse. Cas des mots. Cas d'une performance dégradée* $F = F_{90}$. . . . 213

C.4 *Qualité moyenne des résultats corrélés en fonction du rang de la réponse. Cas des SN. Cas d'une performance dégradée* $F = F_{90}$. . . . 213

D.1 *Document N°AGRP0200289A* . . . 216

D.2 *Document N°AGRP0200303A* . . . 218

D.3 *Document N°ARTL0200455S* . . . 220

D.4 *Document N°DEVN0210260A* . . . 222

D.5 *Document N°DEFD0202303A* . . . 224

Liste des tableaux

2.1 *Taille des principaux annuaires mondiaux [Bak00]* 21

2.2 *Principaux moteurs de recherche mondiaux. Taille exprimée en nombre de pages indexées et fréquentation représentant le pourcentage de requêtes traitées par chaque moteur par rapport au nombre total de requêtes posées sur le Web. Enquête de Septembre 2002 [dN].* 25

4.1 *Comparaisons entre les différentes voies de corrélation* 76

5.1 *Différentes catégories de syntagmes* 86

5.2 *Les différents catégories d'unitermes* 86

5.3 *Exemple de valeurs de* T_{ifr} 98

5.4 *Taille moyenne des documents* 102

5.5 *Accord moyen entre les évaluateurs.* 105

6.1 *Première version de notre algorithme de parcours d'arbre* 120

6.2 *Version améliorée de notre algorithme de parcours d'arbre* 121

6.3 *Exemple de requêtes sans terme pivot commun issue d'une signature* S *composée de cinq termes* 123

6.4 *Exemple de requêtes avec un terme pivot commun issue d'une signature* S *composée de cinq termes* 123

7.1 *Liste des thèmes et nombre de documents* 135

7.2 *Valeurs des paramètre en utilisant des mots* 141

7.3 *Valeurs des paramètre en utilisant des SN* 141

7.4 *Précision moyenne (%) en utilisant des mots* 145

7.5 *Précision moyenne (%) en utilisant des SN* 145

8.1 *Exemple de document extrait de la campagne TREC-6* 149

8.2 *Thématiques du corpus TREC 2001* 152

8.3 *Exemple de documents corrélés accompagnés des notations moyennes attribuées par les évaluateurs* 159

9.1 *Synthèse de la pertinence des méthodes de pondération employées avec des signatures lexicales composées de mots* 166

9.2 *Synthèse de la pertinence des méthodes de pondération employées avec des signatures lexicales composées de SN* 167

D.1 *Documents corrélés pour document N°AGRP0200289A* 217

D.2 *Documents corrélés pour document N°AGRP0200303A* 219

D.3 *Documents corrélés pour document N°ARTL0200455S* 221

D.4 *Documents corrélés pour document N°DEVN0210260A* 223

D.5 *Documents corrélés pour document N°DEFD0202303A* 225

Bibliographie

[Abo] Réseau Abondance. Informations sur internet. *http ://www.abondance.com.*

[ACGM+01] A. Arasu, J. Cho, H. Garcia-Molina, A. Paepcke, and S. Raghavan. Searching the web. *ACM Transactions on Internet Technologies*, 2001.

[Ad'] Ad'oc. Le baromètre ad'oc : Référencement et traffic. *http ://www.barometre.adoc.fr.*

[AMM97] G. O. Arocena, A. O. Mendelzon, and G. A. Mihaila. Applications of a web query language. In *Proceedings of the 6th WWW Conference*, pages 587–595, 1997.

[Bak00] J. Baker. Selected subject directories, 2000. *http ://www.lib.berkeley.edu/TeachingLib/Guides/Internet/SubjDirectories.html.*

[Bax58] P. B. Baxendale. Machine-made index for technical literature – an experiment. *IBM Journal of Research and Development*, 2 :354–361, 1958.

[BB99] K. Bharat and A. Broder. Mirror, mirror on the web : a study of host pairs with replicated content. In *Proceedings of the Eighth International World Wide Web Conference*, 1999.

[BGMZ97] A. Z. Broder, S. C. Glassman, M. S. Manasse, and G. Zweig. Syntactic clustering of the web. In *Proceedings of the 6 th International WWW Conference*, pages 391–404, 1997.

[BH80] C. Busha and S. Harter. *Research methods in librarianship. Techniques and interpretation.* Academic Press, 1980.

[BH98] K. Bharat and M. R. Henzinger. Improved algorithms for topic distillation in hyperlinked environments. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 104–111, 1998.

[BHB01] J. Budzik, K. Hammond, and L. Birnbaum. Information access in context. Knowledge based systems. *Elsevier Science*, 14(1) :37–53, 2001.

[BKR98] A. Bookstein, S. T. Klein, and T. Raita. Clumping properties of content-bearing words. *Journal of the American Society for Information Science*, 1998.

[Bot93] R. A. Botafogo. Cluster analysis for hypertext systems. In *ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 116–125, 1993.

[Bou99] F. Bourdoncle. Panorama et perspectives des outils de recherche d'information textuelle sur internet. In *Actes du colloque IDT'99*, 1999.

[BP98] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. In *Proceedings of the 7 th International World Wide Web Conference*, pages 107–117, 1998.

[BRS92] R. A. Botafogo, E. Rivlin, and B. Shneiderman. Structural analysis of hypertexts : Identifying hierarchies and useful metrics. *Information Systems*, 10(2) :142–180, 1992.

[BS74] A. Bookstein and D. Swanson. Probabilistic models for automatic indexing. *Journal of the American Society for Information Science*, 25 :312–318, 1974.

[BS91] R. A. Botafogo and B. Shneiderman. Identifying aggregates in hypertext structures. In *ACM Conference on Hypertext*, volume 3, pages 63–74, 1991.

[Buc93] C. Buckley. The importance of proper weighting methods. In *Proceedings of a workshop held at Plainsboro, New Jersey*, pages 349–352, 1993.

[BYBC+00] Z. Bar-Yossef, A. Berg, S. Chien, J. Fakcharoenphol, and Dror. Weitz. Approximating aggregate queries about web pages via random walks. In *VLDB 2000, Proceedings of 26th International Conference on Very Large Data Bases*, pages 535–544, 2000.

[CCA89] D. B. Crouch, C. J. Crouch, and G. Andreas. The use of cluster hierarchies in hypertext information retrieval. In *Proceedings of the 2 nd ACM Conference on Hypertext*, pages 225–237, 1989.

[CCC92] F. Ciravegna, P. Campia, and A. Colognese. Knowledge extraction from texts by SINTESI. In *Proceedings of the 14th International Conference on Computational Linguistics*, pages 1244–1248, 1992.

[CDR+98] S. Chakrabarti, B. Dom, P. Raghavan, S. Rajagopalan, D. Gibson, and J. Kleinberg. Automatic resource compilation by analyzing hyperlink structure and associated text. *Computer Networks and ISDN Systems*, 30(1-7) :65–74, 1998.

[CGM00] J. Cho and H. Garcia-Molina. The evolution of the web and implications for an incremental crawler. In *Proceedings of the Twenty-sixth International Conference on Very Large Databases*, 2000.

[CH90] K. W. Church and P. Hanks. Word association norms, mutual information, and lexicography. In *Computational Linguistics*, volume 16, pages 22–29, 1990.

[Cha01] G. Chalendar. *Svetlan' : un système de structuration du lexique guidé par la détermination automatique du contexte thématique.* PhD thesis, Université Paris XI Orsay, 2001.

[Cho02] C. Chotteau. Extraction terminologique : Vers l'élaboration d'une méthode de pondération. In *Proceedings of the CIFT'02 Colloque International sur la Fouile de Texte*, pages 159–178, 2002.

[CK97] J. Carriere and R. Kazman. Webquery : Searching and visualizing the web through connectivity. In *Proceedings of the 7th WWW Conference*, pages 701–711, 1997.

[CKPT92] D. R. Cutting, D. R. Karger, J. O. Pedersen, and J. W. Tukey. Scatter/gather : A cluster-based approach to browsing large document collections. In *ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 318–329, 1992.

[Con91] P. Constant. *Analyse syntaxique par couche.* PhD thesis, École Nationale Supérieure des Télécommunications, Paris, 1991.

[Con95] P. Constant. L'analyseur linguistique sylex. *École d'été du CNET*, 1995.

[Cora] Alexa Corporation. The web information company. *http ://www.alexa.com.*

[Corb] Google Corporation. Moteur de recherche : Google. *http ://www.google.com.*

[Corc] Netscape Corporation. Netscape. *http ://www.netscape.com.*

[Cord] Yahoo Corporation. Annuaire : Yahoo. *http ://www.yahoo.com.*

[Cor00] G. Cornu. *Linguistique juridique.* Dunod, Paris, 2000.

[CTL91] W. B. Croft, H. R. Turtle, and D. D. Lewis. The use of phrases and structured queries in information retrieval. In *ACM SIGIR*, pages 32–45, 1991.

[Cyv00] Cyveillance. Sizing the internet. Technical report, Cyveillance Inc, July 2000. *http ://www.cyveillance.com.*

[Dai96] B. Daille. *Study and Implementation of Combined Techniques for Automatic Extraction of Terminology* dans *The Balancing Act : Combining Symbolic and Statistical Approaches to Language*, pages 46–66. MIT Press, 1996.

[Den67] S. F. Dennis. *The design and testing of a fully automatic indexing-searching system for documents consisting of expository text.* Thompson Book Co., Washington, D.C., 1967. Information retrieval : A critical review.

[DFKM97] F. Douglis, A. Feldmann, B. Krishnamurthy, and J. C. Mogul. Rate of change and other metrics : a live study of the world wide web. In *USENIX Symposium on Internet Technologies and Systems*, 1997.

[DH99] J. Dean and M. R. Henzinger. Finding related pages in the world wide web. *WWW8 / Computer Networks*, 31(11-16) :1467–1479, 1999.

[DM00] J-P. Desclés and J-L. Minel. *Résumé Automatique et Filtrage des textes* dans *Ingénierie des langues*, pages 253–270. Editions Hermès, Paris, 2000.

[dN] Le Journal du Net. Journal d'information sur internet. *http ://www.journaldunet.com.*

[Dun93] T. Dunning. Accurate methods for the statistics of surprise and coincidence. In *Computational Linguistics*, volume 19, 1993.

[ea55] Berry et al. Operational criteria for designing information retrieval systems. *American Documentation*, 6 :93–101, 1955.

[Edm69] H. P. Edmundson. New methods in automatic abstracting. *Journal of the ACM*, 16(2) :264–285, 1969.

[EHW89] A. El-Hamdouchi and P. Willett. Comparison of hierarchical agglomerative clustering methods for document retrieval. *The Computer Journal*, 32(3), 1989.

[Exa] Exalead. Moteur de recherche. *http ://www.exalead.com.*

[Fag89] J. L. Fagan. The effectiveness of a nonsyntactic approach to automatic phrase indexing for document retrieval. *Journal of the American Society for Information Science*, 40(2) :115–139, 1989.

[FC85] C. Faloutsos and S. Christodoulakis. Design of a signature method that accounts for non-uniform occurrence and query frequencies. In *Proceedings of the 11th Int'l Conference on Very Large Data Bases*, 1985.

[Fer98] O. Ferret. *ANTHAPSI : un système d'analyse thématique et d'apprentissage de connaissances progmatiques fondé sur l'amorçage.* PhD thesis, Université Paris Sud, U.F.R. Scientifique d'Orsay, 1998.

[Flu77] C. Fluhr. *Algorithmes à apprentissages et traitement automatique des langues.* PhD thesis, Université Paris Sud, Centre d'Orsay, 1977.

[Fri88] M. E. Frisse. Searching for information in a hypertext medical handbook. *Communications of the ACM*, 31(7) :880–886, 1988.

[Fuh92] N. Fuhr. Probabilistic models in information retrieval. *The Computer Journal*, 35(3) :243–255, 1992.

[GB01] N. Grabar and S. Berland. Construire un corpus web pour l'acquisition terminologique. In Unité de recherche et innovation INIST-CNRS, editor, *Actes de la conférence TIA'2001, Terminologie et intelligence artificielle*, pages 44–54, Nancy, France, 2001.

[GC91] W. Gale and K. Church. Concordances for parallel texts. In *Conference of the UW Centre for the New OED and Text Research, Using Corpora*, pages 40–62, 1991.

[GLW86] A. Griffiths, H. C. Lackhurst, and P. Willett. Using inter-document similarity information in document retrieval systems. *Journal of the American Society for Information Science*, 37 :3–11, 1986.

[Gre86] A. Greimas. *Sémantique structurale.* PUF, 1986.

[GRW84] A. Griffiths, L. A. Robinson, and P. Willett. Hierarchic agglomerative clustering methods for automtic document classification. *Journal of Documentation*, 40 :175–205, 1984.

[Har68] Z. Harris. *Mathematical Structures of Language.* Wiley, 1968.

[Har86] D. Harman. An experimental study of factors important in document ranking. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 186–193, 1986.

[Her02] N. Hernandez. Analyse thématique du discours : segmentation, structuration, description et représentation. In *Proceedings of the CIFT'02 Colloque International sur la Fouile de Texte*, 2002.

[HKP95] M. Hearst, D. R. Karger, and J. Pederson. Scatter/gather as a tool for the navigation of retrieval results. In *Proceedings of AAAI Fall Symposium on Knowledge Navigation*, 1995.

[HP96] M. Hearst and J. O. Pedersen. Reexamining the cluster hypothesis : Scatter/gather on retrieval results. In *Proceedings of the ACM SIGIR 96 International Conference on Research and Development in Information Retrieval*, 1996.

[HR91] D. Hindle and M. Rooth. Structural ambiguity ans lexical relations. In *Association for Computational Linguistics*, 1991.

[Jin00] H. Jing. Sentence reduction for automatic text summarization. In *Proceedings of the 6th Applied Natural Language Processing Conference ANLP'00*, 2000.

[JT99] H. Jing and E. Tzoukermann. Information retrieval based on context distance and morphology. In *SIGIR'99*, volume 22, 1999.

[JvR71] N. Jardin and C. J. van Rijsbergen. The use of hierarchic clustering in information retrieval. *Information Storage and Retrieval*, 7 :217–240, 1971.

[Kah91] B. Kahle. Wide area information server concepts. Technical Report TR-202, 1991.

[Kar] Kartoo. Moteur de recherche. *http ://www.kartoo.com.*

[Kle99] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5) :604–632, 1999.

[KRRT99a] S. R. Kumar, P. Ragavan, S. Rajagopalan, and A. Tomkins. Extracting large-scale knowledge bases from the web. *The VLDB Journal*, pages 639–650, 1999.

[KRRT99b] S. R. Kumar, P. Ragavan, S. Rajagopalan, and A. Tomkins. Trawling the web for emerging cyber-communities. *WWW8 / Computer Networks*, 31(11-16) :1481–1493, 1999.

[Laf80] P. Lafon. Sur la variabilité de la fréquence des formes dans un corpus. *Mots*, 1 :128–165, 1980.

[Lam00] G. Lame. Acquisition de connaissances à partir de textes, vers l'élaboration d'une ontologie du droit. In M. Ayel and J.-M. Fouet, editors, *Actes des RJCIA'2000. Cinquièmes rencontres nationales des jeunes chercheurs en intelligence artificielle*, pages 211–221, Lyon, 2000.

[Lam01a] G. Lame. A categorization method for french legal documents on the web. In H. Prakken, editor, *Proceedings of the 8th International Conference on Artificial Intelligence and Law*, pages 219–220, Saint-Louis MO USA, 2001. ACM, ACM Press.

[Lam01b] G. Lame. Classement automatique de documents et analyse terminologique de corpus. In Unité de recherche et innovation INIST CNRS, editor, *Actes de la conférence TIA'2001, quatrièmes rencontres Terminologie et Intelligence Artificielle*, pages 149–158, Nancy, France, 2001.

[Lam02] G. Lame. *Construction d'ontologies à partir de textes : Une ontologie du droit dédiée à la recherche d'informations sur le Web.* PhD thesis, École des Mines de Paris, 2002.

[Lar92] R. Larson. Experiment in automatic library of congress classification. *Journal of the American Society for Information Science*, 43(2) :130–148, 1992.

[Lee95] J. H. Lee. Combining multiple evidence from different properties of weighting schemes. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 180–188, 1995.

[Lef00] P. Lefevre. *La recherche d'information, du texte intégral au thesaurus.* Hermes science, Paris, 2000.

[LG98a] S. Lawrence and L. Giles. Context and page analysis for improved web search. *Journal of the IEEE*, 2(4) :38–46, 1998.

[LG98b] S. Lawrence and L. Giles. Searching the world wide web. *Science*, 280(5360), 1998.

[LG99] S. Lawrence and L. Giles. Accessibility of information on the web. *Nature*, 400(6740) :107–109, 1999.

[LH97] C-Y. Lin and E. Hovy. Identify topics by position. In *Proceedings of the 5th Conference on Applied Natural Language Processing*, 1997.

[LKSS99] K. Lespinasse, P. Kremer, D. Schibler, and L. Schmitt. Évaluation des outils d'accès l'information textuelle, les expériences américaine (trec) et française (amaryllis). *Langues*, 2(2) :100–109, 1999.

[LL00] C. Labbé and D. Labbé. Que mesure la spécificité du vocabulaire ? *Lexicometrica*, 3, 2000.

[LSJ96] D. D. Lewis and K. Sparck Jones. Natural language processing for information retrieval. *Communications of the ACM*, 39(1) :92–101, 1996.

[Luh57] H. P. Luhn. A statistical approach to mechanized encoding and searching of literary information. *IBM Journal of Research and Development*, 1 :309–317, 1957.

[Mar92] E. L. Margulis. N-Poisson document modelling. In *the 15th International ACM SIGIR Conference on R&D in Information Retrieval*, pages 177–189, 1992.

[MC91] G. A. Miller and W. G. Charles. Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6 :1–28, 1991.

[Mel95] C. S. et al. Mellish. The TIC message analyser. *Computational Linguistics*, 1995.

[Mie89] B. Miege. *La société conquise par la communication*. Presses Universitaires de Grenoble, 1989.

[MKS94] K. McKeown, K. Kukich, and J. Shaw. Practical issue in automatic documentation generation. In *Proceedings of the 4th Conference on Applied Natural Language Processing*, pages 7–14, 1994.

[MML+93] I. Mani, T. R. MacMillan, S. Luperfoy, E. P. Lusher, and S. J. Laskowski. Identifying unknown proper names in newswire text. In *Proceedings of the Workshop on Acquisition of Lexical Knowledge from Text*, pages 44–54, 1993.

[MP01] A. Moreno and C. Pérez. From text to ontology : Extraction and representation of conceptual information. In CNRS Unité de Recherche et Innovation, INIST, editor, *Actes de la conférence TIA'2001, Terminologie et Intelligence Artificielle*, pages 233–242, Nancy, France, 2001.

[MRY73] B. Mathis, J. Rush, and C. Young. Improvement of automatic abstracts by the use of structural analysis. *JASIS*, 24(2) :101–109, 1973.

[MS00] D. Modha and W. S. Spangler. Clustering hypertext with applications to web searching. In *Proceedings of the 11 th ACM Hypertext Conference*, pages 143–152, 2000.

[Muk00a] S. Mukherjea. Organizing topic-specific web information. In *Proceedings of the 11 th ACM Conference on Hypertext*, pages 133–141, 2000.

[Muk00b] S. Mukherjea. Wtms : a system for collecting and analyzing topic-specific web information. In *Proceedings of 9 th International World Wide Web Conference*, pages 457–471, 2000.

[MWZ72] J. Minker, G. Wilson, and B. H. Zimmerman. An evaluation of query expansion by the addition of clustered terms for a document retrieval system. *Information Storage and Retrieval*, 8 :329–348, 1972.

[Ouf01] R. Ouf. Le dynamisme du world wide web : taille, croissance, visibilité, distribution et accessibilité de l'information. Master's thesis, ENSSIB, École des Sciences de l'Information et des Bibliothèques, 2001. *http ://www.enssib.fr/bibliotheque/documents/dpssib/rrbouf.pdf*.

[PBMW98] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking : Bringing order to the web. Technical report, Stanford Digital Library Technologies Project, 1998. *citeseer.nj.nec.com/article/page98pagerank.html*.

[Pea98] J. Pearson. Terms in context. *Studies in Corpus Linguistics*, 1, 1998.

[Per] Pertimm. Solutions de recherches PERTinentes et IMMédiates. *http ://www.pertimm.com*.

[PJ93] C. D. Paice and P. A. Jones. The identification of important concepts in highly structured technical papers. In *Proceedings of the Sixteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 69–78, 1993.

[Pla00] Bright Planet. The deep web : Surfacing hidden value. Technical report, Bright Planet Corp., the Internet content compagny, July 2000. *http ://www.brightplanet.com*.

[PP97] J. Pitkow and P. Pirolli. Life, death and lawfulness on the electronic frontier. In *Proceedings of Human Factors in Computing Systems (CHI 97)*, pages 383–390, 1997.

[PPGK02] S-T. Park, D. M. Pennock, C. L. Giles, and R. Krovets. Analysis of lexical signatures for finding lost or related. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 11–18, 2002.

[PPR96] P. Pirolli, J. Pitkow, and R. Rao. Silk from a sow's ear : Extracting useable structures from the web. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI 96)*, pages 13–18, 1996.

[PSHD96] P. Pirolli, P. Schank, M. Hearst, and C. Diehl. Scatter/ gather browsing communicates the topic structure of a very large text collection. In *Proceedings of the Conference on Human Factors in Computing Systems*, pages 213–220, 1996.

[PW91] H. J. Peat and P. Willett. The limitations of term co-occurence data for query expansion in document retrieval systems. *Journal of the American Society for Information Science*, 42(5) :378–383, 1991.

[PW00] T. A. Phelps and R. Wilensky. Robust hyperlinks : Cheap, everywhere, now. In *Proceedings of Digital Documents and Electronic Publishing*, 2000.

[PY94] L. W. Paik and E. Yu. Text categorisation for multiple users based on semantic features from a machine-readable dictionary. *ACM Transactions on Information Systems*, 12(3) :278–295, 1994.

[Ras89] F. Rastier. *Sémantique et recherche cognitive. Formes Sémiotiques.* PUF, 1989.

[Ras92] E. Rasmussen. Information retrieval : Data structures and algorithms, 1992.

[RAS94] E. Rivlin, Botafogo R. A., and B. Shneiderman. Navigating in hyperspace : designing a structure-based toolbox. *Communications of the ACM*, 37(2) :87–96, 1994.

[RSJ76] S. E. Robertson and K. Sparck Jones. Relevance weighting of search terms. *Journal of the American Society for Information Science*, 27 :129–146, 1976.

[S.A] Samaris S.A. Annuaire dynamique. *http ://www.netboussole.com.*

[Sab00] G. Sabah. *Ingénierie des Langues, chapitre Sens et traitements automatiques des langues.* HERMES Science, 2000.

[Sal71] G. Salton. *The SMART Retrieval System - Experiments in Automatic Document Processing.* Prentice-Hall, 1971.

[Sal72] G. Salton. Experiments in automatic thesaurus construction for information retrieval. *Information Processing*, 71 :115–123, 1972.

[Sal86] G. Salton. On the use of term associations in automatic information retrieval. In *Proceedings of the 11 th International Conference on Computational Linguistics*, pages 380–386, 1986.

[Sal89] G. Salton. *Automatic text processing : the transformation, analysis, and retrieval of information by computer.* Addison-Wesley, 1989.

[SB88] G. Salton and C. Buckley. Term weighting approaches in automatic text retrieval. *Information Processing and Management*, 24 :513–523, 1988.

[SBY83] G. Salton, C. Buckley, and C. T. Yu. *An Evaluation of Term Dependence Models in Information Retrieval.* Springer-Verlag, 1983.

[SGPC$^+$97] T. Strzalkowski, L. Guthrie, J. Perez-Carballo, F. Lin, J. Wang, J. Leistensnider, J. Kalgren, J. Wilding, and T. Straszheim. Natural language information retrieval : Trec-5 report, 1997.

[SHMM98] C. Silverstein, M. Henzinger, H. Marais, and M. Moricz. Analysis of a very large altavista query log. SRC Technical note 1998-14, 1998.

[Sil93] M. Silberztein. Les groupes nominaux productifs et les noms composés lexicalisés. *Linguisticae Investigationes*, 17(2), 1993.

[SJ71] K. Sparck Jones. *Automatic Keyword Classification for Information Retrieval.* Butterworth, 1971.

[SJ72] K. Sparck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28 :11–21, 1972.

[SJ73] K. Sparck Jones. Index term weighting. *Information Storage and Retrieval*, 9 :619–633, 1973.

[SM83] G. Salton and J. M. McGill. *Introduction to modern information retrieval.* McGraw-Hill, 1983.

[Sma73] H. Small. Co-citation in the scientific literature : a new measure of the relationship between scientific documents. *Journal of the American Society for Information Science*, 24 :265–269, 1973.

[SOK94] A. F. Smeaton, R. O'Donnell, and F. Kelledy. Indexing structures derived from syntax in TREC-3 : System description. In *Text REtrieval Conference*, volume 3, pages 55–63, 1994.

[SP01] J. Savoy and J. Picard. Retrieval effectiveness on the Web. *Information Processing and Management*, 37(4) :543–569, 2001.

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

[SS77] J. M. Smith and D. C. P. Smith. Database abstractions : Aggregation and generalization. *TODS*, 2(2) :105–133, 1977.

[Str94] T. Strzalkowski. Document indexing and retrieval using natural language processing. In *Proceedings of RIAO'94, New York*, 1994.

[SY73] G. Salton and C. Yang. A theory of term importance in automatic text analysis. *Journal of the ASIS*, 26(1) :33–34, 1973.

[SYY75] G. Salton, C. S. Yang, and C. T. Yu. A theory of term importance in automatic text analysis. *Journal of the American Society for Information Science*, 26(1) :33–44, 1975.

[TH98a] L. Terveen and W. Hill. Finding and visualizing inter-site clan graphs. In *Proceedings of CHI'98*, pages 448–455, 1998.

[TH98b] L. G. Terveen and W. C. Hill. Evaluating emergent collaboration on the web. In *Computer Supported Cooperative Work*, pages 355–362, 1998.

[TK01] M. Toyoda and M. Kitsuregawa. Creating a web community chart for navigating related communities. In *Hypertext 2001.*, 2001.

[Voo86] E. M. Voorhees. *The effectiveness and efficiency of agglomerative hierarchical clustering in document retrieval.* PhD thesis, Cornell University, 1986.

[Voo93] E. M. Voorhees. Using WordNet to disambiguate word senses for text retrieval. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 171–180, 1993.

[vR79] C. J. van Rijsbergen. *Information retrieval.* Butterworths, 2 edition, 1979.

[vRC75] C. J. van Rijsbergen and W. B. Croft. Document clustering : An evaluation of some experiments with the cranfield 1400 collection. *Information Processing and Management*, 11 :171–182, 1975.

[Wil88] P. Willett. Recent trends in hierarchic document clustering : A critical review. *Information Processing and Management*, 24 :577–597, 1988.

[WM89] H. D. White and K. W. McCain. Bibliometrics. *Annual Review of Information Science and Technology*, pages 119–186, 1989.

[WM99] C. E. Wills and M. Mikhailov. Towards a better understanding of Web resources and server responses for improved caching. *Computer Networks (Amsterdam, Netherlands : 1999)*, 31(11-16) :1231–1243, 1999.

[WVS+96] R. Weiss, B. Velez, M. A. Sheldon, C. Nemprempre, P. Szilagyi, A. Duda, and D. K. Gifford. Hypursuit : A hierarchical network search engine that exploits content-link hypertext clustering. In *Proceedings of the 7 th ACM Conference on Hypertext*, pages 180–193, 1996.

[WW01] K. Wegrzyn-Wolska. *Étude et réalisation d'un méta-indexeur pour la recherche sur le Web de documents produits par l'administration Française.* PhD thesis, École des Mines de Paris, 2001.

[Zip49] G. G. Zipf. *Human Behavior and the Principle of Least Effort.* Addison-Wesley, 1949.

Annexe A

Panel de documents juriques

Les cinquante documents listés ci-dessous ont été extraits aléatoirement des JO de l'année 2000 et servent de base à nos expérimentations. Pour pouvoir les consulter ou les rechercher sur une base documentaire juridique, nous founissons le titre du document, son numéro NOR (identifiant unique d'archivage des documents juriques français) ainsi qu'une URL pour la consultation en ligne.

1. *Décrets du 28 décembre 1999 portant maintien en détachement de magistrats*
   NOR : JUSB9910437D,
   URL : `http://admi.net/jo/20000101/JUSB9910437D.html`
2. *Liste des équipements terminaux de télécommunications ayant fait l'objet de prorogations d'attestations de conformité au cours du mois de novembre 1999*
   NOR : ARTT9900473S,
   URL : `http://admi.net/jo/20000105/ARTT9900473S.html`
3. *Décret du 10 janvier 2000 portant nomination et affectation (chambres régionales des comptes)*
   NOR : CPTE9900041D,
   URL : `http://admi.net/jo/20000111/CPTE9900041D.html`
4. *Arrêtés portant admission à la retraite (services déconcentrés)*
   NOR : EQUP9901934A,
   URL : `http://admi.net/jo/20000114/EQUP9901934A.html`
5. *Arrêté du 6 janvier 2000 portant nomination (inspection des installations nucléaires de base)*
   NOR : ECOI0000012A,
   URL : `http://admi.net/jo/20000120/ECOI0000012A.html`
6. *Arrêtés du 17 janvier 2000 portant homologationd'avenants aux cahiers des charges de labels agricoles*

NOR : AGRP0000126A,

URL : `http://admi.net/jo/20000127/AGRP0000126A.html`

7. *LOI no 2000-65 du 27 janvier 2000 autorisant l'adhésion de la République française à la convention sur les privilèges et immunités des institutions spécialisées approuvée par l'assemblée générale des Nations unies le 21 novembre 1947 (ensemble dix-sept annexes approuvées par les institutions spécialisées) (1)*
   NOR : MAEX9800082L,
   URL : `http://admi.net/jo/20000128/MAEX9800082L.html`
8. *Décret no 2000-107 du 8 février 2000 relatif à la charge maximale des groupes de deux essieux des véhicules à moteur et modifiant l'article R. 58 du code de la route*
   NOR : EQUS0000222D,
   URL : `http://admi.net/jo/20000210/EQUS0000222D.html`
9. *Arrêtés du 15 février 2000 relatifs à des régies d'avances*
   NOR : JUSB0010084A,
   URL : `http://admi.net/jo/20000224/JUSB0010084A.html`
10. *Arrêté du 23 février 2000 portant désignation des assesseurs des tribunaux pour enfants (2e liste)*
    NOR : JUSF0050009A,
    URL : `http://admi.net/jo/20000229/JUSF0050009A.html`
11. *Arrêté du 16 février 2000 relatif au budget de l'agence régionale de l'hospitalisation de Guyane pour l'exercice 2000*
    NOR : MESG0020629A,
    URL : `http://admi.net/jo/20000229/MESG0020629A.html`
12. *Arrêtés du 24 février 2000 conférant le titre de docteur honoris causa*
    NOR : MENR0000451A,
    URL : `http://admi.net/jo/20000303/MENR0000451A.html`
13. *Arrêtés du 14 mars 2000 portant nomination de notaires (officiers publics ou ministériels)*
    NOR : JUSC0020165A,
    URL : `http://admi.net/jo/20000322/JUSC0020165A.html`
14. *Arrêté du 17 mars 2000 portant ouverture de l'examen de sélection professionnelle en vue de l'établissement du tableau d'avancement au titre de l'année 2001 pour l'accès au deuxième grade du corps des greffiers en chef des services judiciaires*
    NOR : JUSB0010144A,
    URL : `http://admi.net/jo/20000323/JUSB0010144A.html`
15. *Arrêté du 14 avril 2000 relatif à l'agrément des médecins pour la mise en oeuvre des contrôles prévus par la loi no 99-223 du 23 mars 1999 relative à la protection de la santé des sportifs et à la lutte contre le dopage*

NOR : MJSK0070040A,
URL : `http://admi.net/jo/20000421/MJSK0070040A.html`

16. *Arrêté du 18 avril 2000 admettant des magistrats à faire valoir leurs droits à la retraite*
    NOR : JUSB0010167A,
    URL : `http://admi.net/jo/20000428/JUSB0010167A.html`
17. *Arrêté du 22 mars 2000 portant nomination (aviation civile)*
    NOR : EQUA0000754A,
    URL : `http://admi.net/jo/20000520/EQUA0000754A.html`
18. *Arrêté du 10 mai 2000 relatif à une régie d'avances*
    NOR : DEFF0001567A,
    URL : `http://admi.net/jo/20000525/DEFF0001567A.html`
19. *Arrêté du 18 mai 2000 portant nomination au conseil d'école de l'Ecole nationale supérieure des télécommunications de Bretagne*
    NOR : ECOI0020144A,
    URL : `http://admi.net/jo/20000527/ECOI0020144A.html`
20. *Décrets du 29 mai 2000 portant délégation de signature*
    NOR : MESG0021417D,
    URL : `http://admi.net/jo/20000531/MESG0021417D.html`
21. *Textes généraux - 3 Juin 2000*
    NOR : CREX0004131S,
    URL : `http://admi.net/jo/20000603/CREX0004131S.html`
22.
    NOR : ECOP0000371A,
    URL : `http://admi.net/jo/20000606/ECOP0000371A.html`
23. *Arrêté du 5 mai 2000 portant inscription à un tableau d'avancement (corps des ingénieurs de l'industrie et des mines)*
    NOR : MESS0021746A,
    URL : `http://admi.net/jo/20000614/MESS0021746A.html`
24. *Arrêté du 25 mai 2000 modifiant le titre Ier du tarif interministériel des prestations sanitaires et relatif aux nutriments pour supplémentation*
    NOR : MESH0021651A,
    URL : `http://admi.net/jo/20000615/MESH0021651A.html`
25. *Avis relatif à l'extension d'un avenant à la convention collective nationale de travail concernant les coopératives agricoles, unions de coopératives agricoles et sociétés d'intérêt collectif agricole de fleurs, de fruits et légumes et de pommes de terre*
    NOR : AGRS0001167V,
    URL : `http://admi.net/jo/20000621/AGRS0001167V.html`

26. *Décret du 23 juin 2000 portant délégation de signature*
    NOR : MCCB0000435D,
    URL : `http://admi.net/jo/20000625/MCCB0000435D.html`
27. *Décret du 23 juin 2000 portant admission à la retraite et maintien en activité de magistrats*
    NOR : JUSB0010215D,
    URL : `http://admi.net/jo/20000630/JUSB0010215D.html`
28. *Liste d'admissibilité aux concours d'admission à l'Ecole spéciale militaire de Saint-Cyr en 2000 (concours ouverts aux diplômés de l'enseignement supérieur)*
    NOR : DEFT0001732K,
    URL : `http://admi.net/jo/20000701/DEFT0001732K.html`
29. *Arrêté du 2 juin 2000 portant admission à la retraite (administration préfectorale)*
    NOR : INTA0000347A,
    URL : `http://admi.net/jo/20000701/INTA0000347A.html`
30. *Avis relatifs à l'extension d'avenants à la convention collective nationale des maisons d'étudiants*
    NOR : MEST0010911V,
    URL : `http://admi.net/jo/20000726/MEST0010911V.html`
31. *Arrêté du 4 mai 2000 relatif à la reconnaissance d'organisations de producteurs dans le secteur des fruits et légumes*
    NOR : AGRP0001129A,
    URL : `http://admi.net/jo/20000809/AGRP0001129A.html`
32. *Arrêté du 27 juillet 2000 portant nomination (régisseurs de recettes)*
    NOR : EQUG0001262A,
    URL : `http://admi.net/jo/20000817/EQUG0001262A.html`
33. *Modification du règlement du jeu de loterie instantanée de La Française des jeux dénommé Millionnaire*
    NOR : ECOZ0099151X,
    URL : `http://admi.net/jo/20000903/ECOZ0099151X.html`
34. *Arrêté du 28 août 2000 relatif à une régie d'avances*
    NOR : DEFF0002047A,
    URL : `http://admi.net/jo/20000907/DEFF0002047A.html`
35. *Décisions relatives à des demandes de création, d'extension d'établissements sanitaires et d'installation d'équipements matériels lourds*
    NOR : MESH0022507S,
    URL : `http://admi.net/jo/20000908/MESH0022507S.html`
36. *Arrêté du 1er septembre 2000 portant promotion (ponts et chaussées)*
    NOR : EQUP0001240A,
    URL : `http://admi.net/jo/20000912/EQUP0001240A.html`

37. *Arrêté du 4 septembre 2000 portant admission à la retraite (services vétérinaires)*
NOR : AGRA0001816A,
URL : `http://admi.net/jo/20000916/AGRA0001816A.html`
38. *Arrêté du 8 septembre 2000 modifiant l'arrêté du 16 mars 2000 portant ouverture en 2000 de concours pour le recrutement d'ingénieurs subdivisionnaires territoriaux*
NOR : FPPT0000119A,
URL : `http://admi.net/jo/20001005/FPPT0000119A.html`
39. *Arrêté du 8 septembre 2000 modifiant l'arrêté du 16 mars 2000 portant ouverture en 2000 de concours pour le recrutement d'ingénieurs subdivisionnaires territoriaux*
NOR : JUSA0000116D,
URL : `http://admi.net/jo/20001008/JUSA0000116D.html`
40. *Décrets du 18 octobre 2000 portant nomination et titularisation (corps de contrôle des assurances)*
NOR : ECOP0000694D,
URL : `http://admi.net/jo/20001022/ECOP0000694D.html`
41. *Arrêté du 16 octobre 2000 modifiant l'arrêté du 11 septembre 2000 portant nomination au Conseil supérieur de l'éducation*
NOR : MENG0002727A,
URL : `http://admi.net/jo/20001027/MENG0002727A.html`
42. *Arrêté du 6 novembre 2000 relatif à la création d'un site sur internet intitulé service-public.fr*
NOR : PRMX0004473A,
URL : `http://admi.net/jo/20001108/PRMX0004473A.html`
43. *Décret du 9 novembre 2000 portant nomination (tribunaux administratifs et cours administratives d'appel)*
NOR : JUSA0000339D,
URL : `http://admi.net/jo/20001111/JUSA0000339D.html`
44. *Arrêtés du 6 novembre 2000 relatifs à des sociétés civiles professionnelles (officiers publics ou ministériels)*
NOR : JUSC0020691A,
URL : `http://admi.net/jo/20001114/JUSC0020691A.html`
45. *Arrêté du 24 octobre 2000 autorisant la société riodata n.v. à établir et exploiter un réseau expérimental de télécommunications ouvert au public*
NOR : ECOI0020352A,
URL : `http://admi.net/jo/20001124/ECOI0020352A.html`
46. *Arrêté du 10 novembre 2000 portant extension d'un avenant à la convention collective nationale du personnel au sol des entreprises de transport aérien*
NOR : MEST0011543A,
URL : `http://admi.net/jo/20001125/MEST0011543A.html`

47. *Décret du 5 décembre 2000 portant désignation et cessation de fonctions de commissaires du Gouvernement (tribunaux administratifs et cours administratives d'appel)*
NOR : JUSA0000371D,
URL : `http://admi.net/jo/20001208/JUSA0000371D.html`
48. *Arrêté du 21 décembre 2000 portant constatation de l'état de catastrophe naturelle*
NOR : INTE0000791A,
URL : `http://admi.net/jo/20001222/INTE0000791A.html`
49. *Décret no 2000-1321 du 26 décembre 2000 fixant des modalités exceptionnelles de recrutement dans les corps d'adjoints administratifs d'administration centrale et d'adjoints administratifs des services déconcentrés du ministère de l'emploi et de la solidarité*
NOR : MESG0023085D,
URL : `http://admi.net/jo/20001230/MESG0023085D.html`
50. *Arrêté du 28 décembre 2000 autorisant au titre de l'année 2000 l'ouverture de concours pour le recrutement de techniciens des services déconcentrés de l'administration pénitentiaire (femmes et hommes)*
NOR : JUSE0040100A,
URL : `http://admi.net/jo/20001231/JUSE0040100A.html`

Annexe B

Panel réduit de documents juridiques

Parmi les documents listés en Annexe A, nous avons gardé dans ce panel ceux composés d'au moins deux paragraphes. Afin de pouvoir les consulter ou les rechercher sur une base documentaire juridique, nous founissons le titre du document, son numéro NOR (identifiant unique d'archivage des documents juriques français), le nombre de paragraphes qui le composent d'après la définition donnée chapitre 6 ainsi qu'une URL pour la consultation en ligne.

1. *Liste des équipements terminaux de télécommunications ayant fait l'objet de prorogations d'attestations de conformité au cours du mois de novembre 1999*
   NOR : ARTT9900473S,
   URL : `http://admi.net/jo/20000105/ARTT9900473S.html`
2. *LOI no 2000-65 du 27 janvier 2000 autorisant l'adhésion de la République française à la convention sur les privilèges et immunités des institutions spécialisées approuvée par l'assemblée générale des Nations unies le 21 novembre 1947 (ensemble dix-sept annexes approuvées par les institutions spécialisées) (1)*
   NOR : MAEX9800082L,
   URL : `http://admi.net/jo/20000128/MAEX9800082L.html`
3. *Décret no 2000-107 du 8 février 2000 relatif à la charge maximale des groupes de deux essieux des véhicules à moteur et modifiant l'article R. 58 du code de la route*
   NOR : EQUS0000222D,
   URL : `http://admi.net/jo/20000210/EQUS0000222D.html`
4. *Arrêté du 23 février 2000 portant désignation des assesseurs des tribunaux pour enfants (2e liste)*
   NOR : JUSF0050009A,
   URL : `http://admi.net/jo/20000229/JUSF0050009A.html`
5. *Arrêté du 14 avril 2000 relatif à l'agrément des médecins pour la mise en oeuvre des*

contrôles prévus par la loi no 99-223 du 23 mars 1999 relative à la protection de la santé des sportifs et à la lutte contre le dopage
NOR : MJSK0070040A,
URL : `http://admi.net/jo/20000421/MJSK0070040A.html`

6. *Arrêté du 22 mars 2000 portant nomination (aviation civile)*
NOR : EQUA0000754A,
URL : `http://admi.net/jo/20000520/EQUA0000754A.html`
7. *Arrêté du 18 mai 2000 portant nomination au conseil d'école de l'Ecole nationale supérieure des télécommunications de Bretagne*
NOR : ECOI0020144A,
URL : `http://admi.net/jo/20000527/ECOI0020144A.html`
8. *Décrets du 29 mai 2000 portant délégation de signature*
NOR : MESG0021417D,
URL : `http://admi.net/jo/20000531/MESG0021417D.html`
9. *Arrêté du 5 mai 2000 portant inscription à un tableau d'avancement (corps des ingénieurs de l'industrie et des mines)*
NOR : MESS0021746A,
URL : `http://admi.net/jo/20000614/MESS0021746A.html`
10. *Arrêté du 25 mai 2000 modifiant le titre Ier du tarif interministériel des prestations sanitaires et relatif aux nutriments pour supplémentation*
NOR : MESH0021651A,
URL : `http://admi.net/jo/20000615/MESH0021651A.html`
11. *Décret du 23 juin 2000 portant délégation de signature*
NOR : MCCB0000435D,
URL : `http://admi.net/jo/20000625/MCCB0000435D.html`
12. *Liste d'admissibilité aux concours d'admission à l'Ecole spéciale militaire de Saint-Cyr en 2000 (concours ouverts aux diplômés de l'enseignement supérieur)*
NOR : DEFT0001732K,
URL : `http://admi.net/jo/20000701/DEFT0001732K.html`
13. *Modification du règlement du jeu de loterie instantanée de La Française des jeux dénommé Millionnaire*
NOR : ECOZ0099151X,
URL : `http://admi.net/jo/20000903/ECOZ0099151X.html`
14. *Arrêté du 8 septembre 2000 modifiant l'arrêté du 16 mars 2000 portant ouverture en 2000 de concours pour le recrutement d'ingénieurs subdivisionnaires territoriaux*
NOR : JUSA0000116D,
URL : `http://admi.net/jo/20001008/JUSA0000116D.html`
15. *Arrêté du 6 novembre 2000 relatif à la création d'un site sur internet intitulé service-public.fr*

NOR : PRMX0004473A,

URL : http://admi.net/jo/20001108/PRMX0004473A.html

16. *Arrêté du 24 octobre 2000 autorisant la société riodata n.v. à établir et exploiter un réseau expérimental de télécommunications ouvert au public*
NOR : ECOI0020352A,
URL : http://admi.net/jo/20001124/ECOI0020352A.html

17. *Arrêté du 10 novembre 2000 portant extension d'un avenant à la convention collective nationale du personnel au sol des entreprises de transport aérien*
NOR : MEST0011543A,
URL : http://admi.net/jo/20001125/MEST0011543A.html

18. *Décret du 5 décembre 2000 portant désignation et cessation de fonctions de commissaires du Gouvernement (tribunaux administratifs et cours administratives d'appel)*
NOR : JUSA0000371D,
URL : http://admi.net/jo/20001208/JUSA0000371D.html

19. *Arrêté du 21 décembre 2000 portant constatation de l'état de catastrophe naturelle*
NOR : INTE0000791A,
URL : http://admi.net/jo/20001222/INTE0000791A.html

20. *Décret no 2000-1321 du 26 décembre 2000 fixant des modalités exceptionnelles de recrutement dans les corps d'adjoints administratifs d'administration centrale et d'adjoints administratifs des services déconcentrés du ministère de l'emploi et de la solidarité*
NOR : MESG0023085D,
URL : http://admi.net/jo/20001230/MESG0023085D.html

Annexe C

Évaluation des résultats corrélés

Les résultats présentés ci-après complètent les informations du chapitre 8 et notamment la section 8.3.

Les figures C.1 et C.2 représentent la précision moyenne des résultats corrélés respectivement pour des signatures composées de mots et de SN dans le cas où $F = F_{90}$.

Les figures C.3 et C.4 représentent la qualité moyenne des documents respectivement pour des signatures composées de mots et de SN avec $F = F_{90}$.

Ces figures montrent que, quel que soit le critère choisi (précision ou qualité) ou le type de terme utilisé (mot ou SN), les résultats obtenus dans le cas où $F = F_{90}$ sont toujours dégradés par rapport à ceux obtenus pour $F = F_{max}$ (section 8.3).

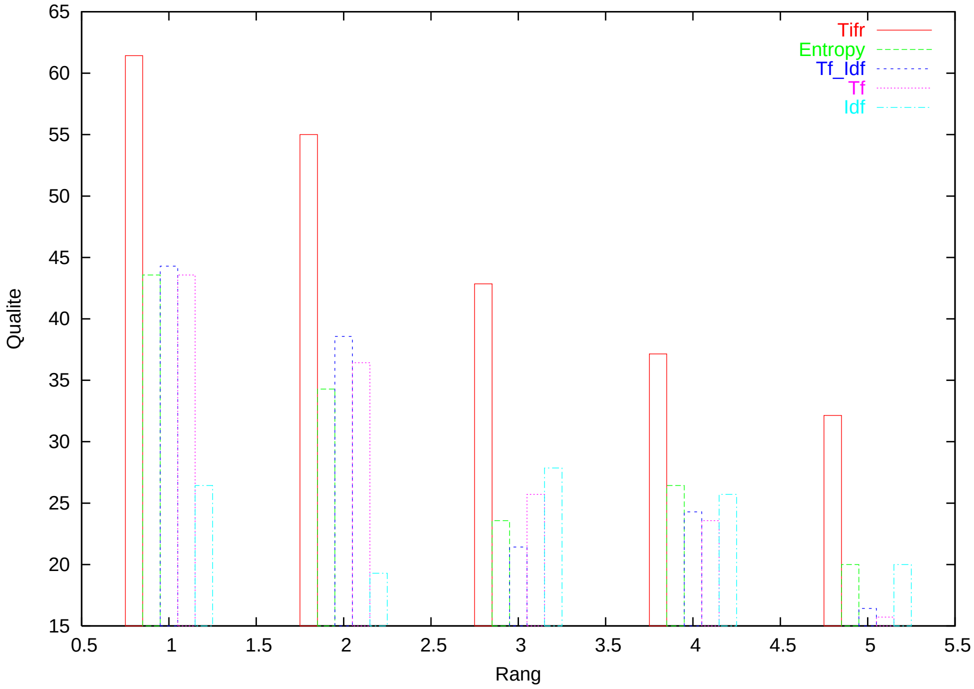


Fig. C.1 – *Précision moyenne des résultats corrélés en fonction du rang de la réponse. Cas des mots. Cas d'une performance dégradée* $F = F_{90}$.

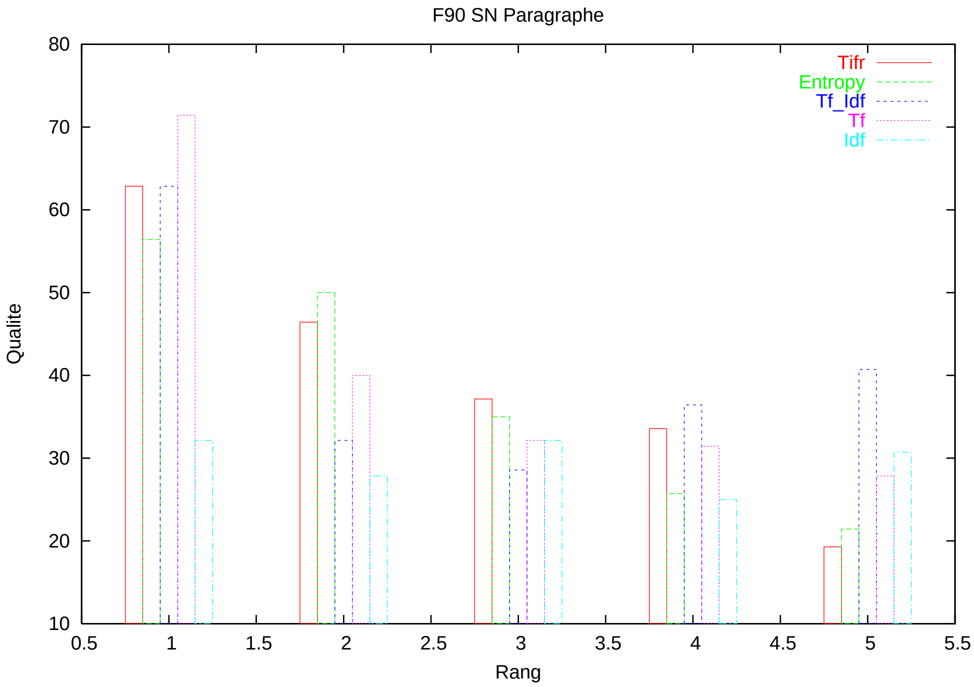


Fig. C.2 – *Précision moyenne des résultats corrélés en fonction du rang de la réponse. Cas des SN. Cas d'une performance dégradée* $F = F_{90}$.

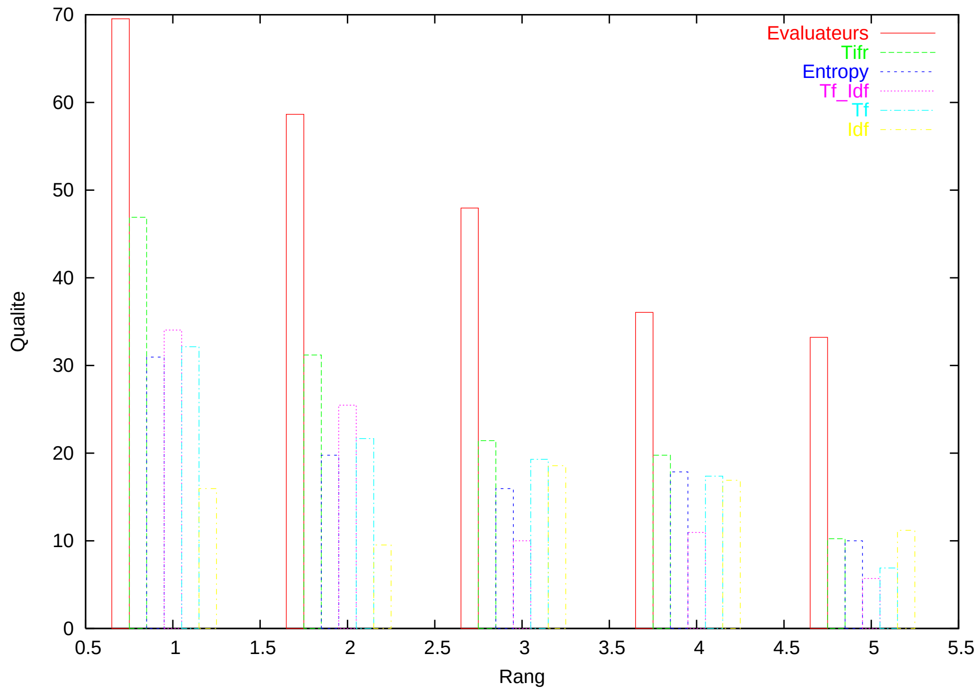


FIG. C.3 – *Qualité moyenne des résultats corrélés en fonction du rang de la réponse. Cas des mots. Cas d'une performance dégradée* $F = F_{90}$.

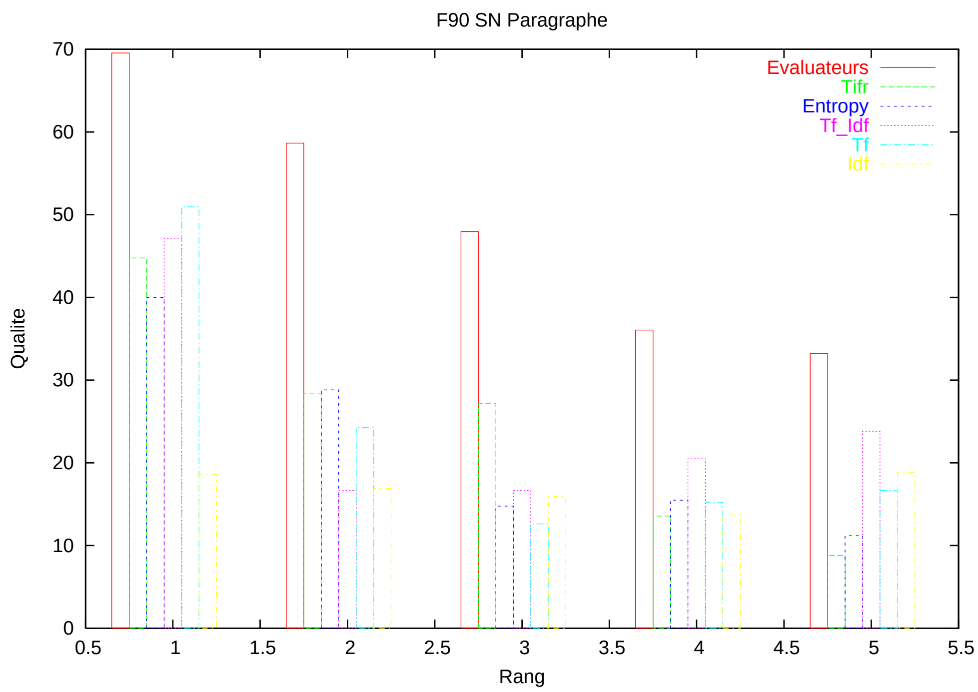


FIG. C.4 – *Qualité moyenne des résultats corrélés en fonction du rang de la réponse. Cas des SN. Cas d'une performance dégradée* $F = F_{90}$.

Annexe D

Exemple de documents corrélés issus du JO

Les documents présentés ci-après sont de deux types : des documents intiaux servant de base à la corrélation et des tableaux regroupant les signatures lexicales extraites ainsi que les documents corrélés retrouvés.

Les documents initiaux sont extraits du JO ; pour des aspects pratiques ceux-ci ont été tronqués pour n'obtenir au final qu'une unique page format A4. Il se peut donc que pour certains textes la fin soit manquante, la finalité recherchée étant d'appréhender la sémantique du document pour juger de la pertinence des résultats corrélés présentés, cette troncature n'apparait pas comme génante.

Dans les tableaux (voir par exemple le tableau D.1), on présente en premier lieu le titre du document initial, puis la signature lexicale qui en a été extraite, avec les mots et les SN qui sont classés séparemment par pondération décroissante, et enfin les résultats corrélés retrouvés, également classés par pertinence décroissante.

J.O. Numéro 43 du 20 Février 2002 J.O. disponibles Alerte par mail Lois,décrets codes AdmiNet

Ce document peut également être consulté sur le site officiel Legifrance

Arrêté du 8 février 2002 portant nomination au comité régional vins et eaux-de-vie pour la région Cognac de l'Institut national des appellations d'origine

---

NOR : AGRP0200289A

Par arrêté du ministre de l'agriculture et de la pêche en date du 8 février 2002, M. Jean-Marie Beulque-Schaub, à Cognac (Charente), est nommé en qualité de représentant du secteur négoce au comité régional vins et eaux-de-vie de l'Institut national des appellations d'origine pour la région Cognac jusqu'au 20 juillet 2004, en remplacement de M. Alain Braastad, à Jarnac (Charente).

FIG. D.1 – *Document N°AGRP0200289A*

Document initial : Arrêté du 8 février 2002 portant nomination au comité régional vins et eaux-de-vie pour la région Cognac de l'Institut national des appellations d'origine (voir la source figure D.1)
Signature lexicale : braastad, appellations, négoce, cognac, jarnac, charente, schaub, beulque, région cognac de l'institut national, comité régional vins, nomination au comité régional vins, région cognac, appellations d'origine, institut national des appellations, représentant du secteur négoce, secteur négoce, secteur négoce au comité régional vins, qualité de représentant

1. Décret n°2002-1325 du 5 novembre 2002 relatif aux conditions de production et au rendement des vignobles produisant des vins à appellation d'origine contrôlée
2. Arrêté du 20 décembre 2002 relatif à l'agrément de l'appellation d'origine contrôlée «Domfront»
3. Décret du 20 décembre 2002 portant reconnaissance de l'appellation d'origine contrôlée «Domfront»
4. Arrêté du 20 mars 2002 relatif aux conditions d'attribution de l'aide à la restructuration et à la reconversion du vignoble pour la campagne 2001-2002
5. Décret du 22 février 2002 relatif à l'appellation d'origine contrôlée «Noix de Grenoble»
6. Arrêté du 30 août 2002 modifiant l'arrêté du 21 octobre 1993 modifié agréant les agents de l'Institut national des appellations d'origine à rechercher et à constater les infractions prévues aux articles L.115-16 et L.213-1 et suivants du code de la consommation
7. Arrêté du 18 décembre 2001 relatif aux volumes substituables individuels pour certaines appellations d'origine contrôlées de la récolte 2001
8. Décret n°2002-1486 du 20 décembre 2002 relatif à la gestion du potentiel de production viticole
9. Décret du 12 novembre 2002 relatif à l'appellation d'origine contrôlée «Tome des Bauges»
10. Décret du 22 février 2002 relatif à l'appellation d'origine contrôlée «Fourme d'Ambert»

TAB. D.1 – *Documents corrélés pour document N°AGRP0200289A*

J.O. Numéro 39 du 15 Février 2002 J.O. disponibles Alerte par mail Lois,décrets codes AdmiNet

Texte paru au JORF/LD *page 02992*

Ce document peut également être consulté sur le site officiel Legifrance

Arrêté du 8 février 2002 relatif à l'appellation d'origine contrôlée « Armagnac »

NOR : AGRP0200303A

Le ministre de l'agriculture et de la pêche, la secrétaire d'Etat au budget et le secrétaire d'Etat aux petites et moyennes entreprises, au commerce, à l'artisanat et à la consommation,
Vu le code général des impôts ;
Vu le code des douanes ;
Vu le code rural ;
Vu le code de la consommation ;
Vu le décret du 6 août 1936 modifié relatif à l'appellation d'origine contrôlée « Armagnac » ;
Vu le décret no 2001-510 du 12 juin 2001 portant application du code de la consommation en ce qui concerne les vins, vins mousseux, vins pétillants et vins de liqueur ;
Vu la proposition du comité national des vins et eaux-de-vie de l'Institut national des appellations d'origine des 11 et 12 décembre 2001,
Arrêtent :

Art. 1er. - La distillation des vins de la récolte 2001 destinés à la production des appellations d'origine contrôlées « Armagnac », « Bas-Armagnac », « Armagnac-Ténarèze » et « Haut-Armagnac » doit être effectuée au plus tard le 15 février 2002.

Art. 2. - Le directeur général de la concurrence, de la consommation et de la répression des fraudes, le directeur général des douanes et droits indirects au ministère de l'économie, des finances et de l'industrie et le directeur des politiques économique et internationale au ministère de l'agriculture et de la pêche sont chargés, chacun en ce qui le concerne, de l'exécution du présent arrêté, qui sera publié au Journal officiel de la République française.

FIG. D.2 – *Document N°AGRP0200303A*

Document initial : Arrêté du 8 février 2002 relatif à l'appellation d'origine contrôlée «Armagnac» (voir la source figure D.2)
Signature lexicale : vins, concurrence, armagnac, appellation, contrôlée, appellations, douanes, distillation, production, récolte, répression, agronomie, fraudes, eaux, haut, indirects, internationale, proposition, politiques, destinés, appellation d'origine, relatif à l'appellation, appellations d'origine, relatif à l'appellation d'origine, comité national, institut national des appellations, distillation des vins, droits indirects, code général des impôts, code de la consommation

1. Décret n°2002-1325 du 5 novembre 2002 relatif aux conditions de production et au rendement des vignobles produisant des vins à appellation d'origine contrôlée
2. Arrêté du 22 mars 2002 portant habilitation de fonctionnaires de catégorie A du ministère chargé de l'économie à recevoir, des juges d'instruction, des commissions rogatoires
3. Décret du 20 décembre 2002 portant reconnaissance de l'appellation d'origine contrôlée «Domfront»
4. Arrêté du 20 mars 2002 relatif aux conditions d'attribution de l'aide à la restructuration et à la reconversion du vignoble pour la campagne 2001-2002
5. Décret du 22 février 2002 relatif à l'appellation d'origine contrôlée «Noix de Grenoble»
6. Arrêté du 20 décembre 2002 relatif à l'agrément de l'appellation d'origine contrôlée «Domfront»
7. Arrêté du 18 décembre 2001 relatif aux volumes substituables individuels pour certaines appellations d'origine contrôlées de la récolte 2001
8. Décret du 29 août 2002 relatif à l'appellation d'origine contrôlée «Coteaux de Saumur»
9. Arrêté du 18 décembre 2001 relatif aux appellations d'origine contrôlées «Champagne», «Coteaux champenois» et «Rosé des Riceys» de la récolte 2001

TAB. D.2 – *Documents corrélés pour document N°AGRP0200303A*

J.O. 257 du 3 novembre 2002 J.O. disponibles Alerte par mail Lois,décrets codes AdmiNet

Texte paru au JORF/LD *page 18213*

Ce document peut également être consulté sur le site officiel Legifrance

Décision n° 2002-731 de l'Autorité de régulation des télécommunications en date du 5 septembre 2002 modifiant la décision n° 2000-822 en date du 28 juillet 2000 portant attribution de fréquences dans les bandes 3,5 GHz et 26 GHz à la société FirstMark Communications France

---

NOR : ARTL0200455S

L'Autorité de régulation des télécommunications,

Vu le code des postes et télécommunications, et en particulier l'article L. 36-7 (6°) ;

Vu le décret du 3 février 1993 modifié relatif aux redevances de mise à disposition de fréquences radioélectriques et de gestion dues par les titulaires des autorisations délivrées en application des articles L. 33-1 et L. 33-2 du code des postes et télécommunications ;

Vu l'arrêté du 6 mars 2001 relatif au tableau national de répartition des bandes de fréquences ;

Vu la décision n° 99-830 de l'Autorité de régulation des télécommunications en date du 6 octobre 1999 fixant les conditions techniques et d'exploitation générales de la bande de fréquences 3,4-3,6 GHz pour les liaisons de transmission point à multipoint du service fixe, homologuée par l'arrêté du 26 novembre 1999 ;

Vu la décision n° 99-831 de l'Autorité de régulation des télécommunications en date du 6 octobre 1999 fixant les conditions techniques et d'exploitation générales de la bande de fréquences 24,5-26,5 GHz pour les liaisons de transmission du service fixe et abrogeant la décision n° 98-283 en date du 30 avril 1998, homologuée par l'arrêté du 26 novembre 1999 ;

Vu la décision n° 2000-822 de l'Autorité de régulation des télécommunications en date du 28 juillet 2000 portant attribution de fréquences dans les bandes 3,5 GHz et 26 GHz à la société FirstMark Communications France ;

Vu l'arrêté du 4 août 2000 autorisant la société FirstMark Communications France à établir et à exploiter un réseau ouvert au public et à fournir le service téléphonique au

FIG. D.3 – *Document N°ARTL0200455S*

Document initial : Décision n°2002-731 de l'Autorité de régulation des télécommunications ... portant attribution de fréquences dans les bandes 3,5 GHz et 26 GHz à la société FirstMark Communications France (voir la source figure D.3)
Signature lexicale : ldcom, boucle, abonnés, clients, opérateurs, appels, corse, fournir, cahier, raccordement, multipoint, transmission, redevances, homologuée, liaisons, point, déploiement, basse-normandie, opération, pays, boucle locale radio, obligations de déploiement, firstmark communications france, service téléphonique au public, régions bourgogne, ghz à la société firstmark communications france, code des postes, cahier des charges, réseau ouvert au public, réseau de télécommunications

1. Décision no 2002-147 du 12 février 2002 se prononçant sur le différend opposant MFS Communications à France Télécom relatif à la fourniture par France Télécom de liaisons louées aux opérateurs tiers
2. Décision no 2002-508 de l'Autorité de régulation des télécommunications en date du 27 juin 2002 prise au terme de la procédure engagée à l'encontre de la société Landtel France SAS en application de l'article L. 36-11 du code des postes et télécommunications
3. Décision no 2002-507 de l'Autorité de régulation des télécommunications en date du 27 juin 2002 prise au terme de la procédure engagée à l'encontre de la société Broadnet France SAS en application de l'article L. 36-11 du code des postes et télécommunications
4. Décision no 2001-1235 du 21 décembre 2001 se prononçant sur un différend entre les sociétés UPC France et France Télécom
5. Arrêté du 29 août 2002 modifiant l'arrêté du 17 novembre 1998 modifié autorisant la société Completel SAS à établir et exploiter un réseau de télécommunications ouvert au public et à fournir le service téléphonique au public
6. Arrêté du 31 juillet 2002 autorisant la société Dauphin Télécom à établir et exploiter un réseau de télécommunications ouvert au public et à fournir le service téléphonique au public
7. Arrêté du 30 août 2002 autorisant la société Outremer Télécom à établir et exploiter un réseau de télécommunications ouvert au public par satellite et à fournir le service téléphonique au public
8. Décision no 2002-278 du 28 mars 2002 se prononçant sur le différend entre les sociétés LDCOM et France Télécom relatif à certaines conditions techniques et tarifaires de la convention d'accès à la boucle locale
9. Arrêté du 14 novembre 2002 autorisant la société Globalstar Europe à établir et exploiter un réseau de télécommunications ouvert au public et à fournir le service téléphonique au public
10. Avis n° 2002-291 en date du 4 avril 2002 définissant le contenu de l'offre de services avancés de téléphonie vocale et les indicateurs de qualité du service téléphonique au public

TAB. D.3 – *Documents corrélés pour document N° ARTL0200455S*

J.O. Numéro 165 du 17 Juillet 2002 J.O. disponibles Alerte par mail Lois,décrets codes AdmiNet

Texte paru au JORF/LD *page 12203*

Ce document peut également être consulté sur le site officiel Legifrance

Arrêté du 25 juin 2002 relatif à la durée du mandat des membres du Conseil national de la chasse et de la faune sauvage

NOR : DEVN0210260A

Par arrêté de la ministre de l'écologie et du développement durable en date du 25 juin 2002, la durée des mandats des membres du Conseil national de la chasse et de la faune sauvage est la suivante :
Le mandat des membres titulaires et suppléants désignés ci-dessous prendra fin le 11 mars 2005 :
MM. Mathieu (Bernard), titulaire, et Esclope (Alain), suppléant ;
MM. Dobremez (Jean-François), titulaire, et Jean (Alain), suppléant ;
MM. Dulac (Philippe), titulaire, et Molina (Patrick), suppléant ;
MM. Athanaze (Pierre), titulaire, et Terrasse (Jean-François), suppléant ;
MM. Baudin (Bernard), titulaire, et Lavallart (Philippe), suppléant ;
M. Moutou (François), titulaire, et Mme Terrier (Marie-Eve), suppléante ;
MM. Dufranc (Michel), titulaire, et Deluga (François), suppléant ;
MM. De Turkheim (Gilbert), titulaire, et Marcotte (Michel), suppléant ;
MM. Giraud (Yves), titulaire, et Peyrin (Christian), suppléant ;
MM. Kittler (Daniel), titulaire, et Aurange (Jacques), suppléant ;
MM. Fuzies (Pierre), titulaire, et Lecha (Jean-Raymond), suppléant.
Le mandat des membres titulaires et suppléants désignés ci-dessous prendra fin le 11 mars 2008 :
MM. Bidault (Edouard-Alain), titulaire, et Bonnefous (Guy), suppléant ;
MM. Pouget (Raymond), titulaire, et Olivier (Guy-Noël), suppléant ;
MM. Barbedienne (Philippe), titulaire, et Gratadour (Vincent), suppléant ;
MM. Boidot (Jean-Paul), titulaire, et Justeau (Philippe), suppléant ;
MM. Brayer (Philippe), titulaire, et Delavenne (Bruno), suppléant ;
M. Chazalet (Jacques), titulaire, et Mme Dubanchet (Sandrine), suppléante ;
MM. Petitpas (Henry), titulaire, et Poulain (Jean-Luc), suppléant ;
MM. Plauche-Gillon (Henry), titulaire, et de Montgascon (Henri), suppléant ;

FIG. D.4 – *Document N°DEVN0210260A*

Document initial : Arrêté du 25 juin 2002 relatif à la durée du mandat des membres du Conseil national de la chasse et de la faune sauvage (voir la source figure D.4)
Signature lexicale : chasse, faune, henri, suppléant, sauvage, frédéric, mandat, raymond, yves, durable, suppléants, écologie, titulaires, gilbert, bruno, guy, vincent, paul, mathieu, mandats, faune sauvage, conseil national, relatif à la durée, ministre de l'écologie, membres titulaires, mandat des membres, membres du conseil national, mandats des membres, conseil national de la chasse, durée du mandat

1. Décision du 20 décembre 2001 portant inscription au tableau d'avancement pour l'année 2002 (armée active)
2. Décret du 9 juillet 2002 portant nomination et promotion dans l'armée active
3. Décret du 3 avril 2002 portant concession de la médaille militaire
4. Décret du 22 mai 2002 portant délégation de signature
5. Arrêté du 10 juillet 2002 portant attribution de la qualité d'officier de police judiciaire à des militaires de la gendarmerie
6. Décret du 23 octobre 2002 portant nomination et promotion dans l'armée active
7. Décret du 19 mars 2002 portant nomination et promotion dans l'armée active
8. Décret du 12 juin 2002 portant promotion et nomination
9. Décret du 6 novembre 2002 portant promotion et nomination
10. Décret du 5 novembre 2002 portant nomination et promotion dans l'armée active
11. Décret du 18 janvier 2002 portant promotion et nomination dans les cadres des officiers de réserve

TAB. D.4 – *Documents corrélés pour document N°DEVN0210260A*

J.O. 246 du 20 octobre 2002 J.O. disponibles Alerte par mail Lois,décrets codes AdmiNet

Texte paru au JORF/LD *page 17429*

Ce document peut également être consulté sur le site officiel Legifrance

Arrêté du 17 octobre 2002 portant délégation de signature

NOR : DEFD0202303A

La ministre de la défense,
Vu le décret n° 88-91 du 27 janvier 1988 modifié autorisant le ministre de la défense à déléguer, par arrêté, sa signature ;
Vu le décret n° 2000-1178 du 4 décembre 2000 portant organisation de l'administration centrale du ministère de la défense, modifié par les décrets n° 2001-1125 du 29 novembre 2001 et n° 2002-503 du 10 avril 2002 ;
Vu le décret du 17 juin 2002 portant nomination du Premier ministre ;
Vu le décret du 17 juin 2002 relatif à la composition du Gouvernement ;
Vu l'arrêté du 16 mai 2002 modifié portant délégation de signature,
Arrête :

Article 1

Le I de l'article 3 (Directions relevant du chef d'état-major des armées) du titre Ier (Etat-major des armées) de l'arrêté du 16 mai 2002 susvisé est modifié comme suit :
Au A (Direction du renseignement militaire), tableau, à la colonne Suppléants, l'alinéa : « Mme Annie Grange, attachée de service administratif. » est abrogé.

Article 2

Le présent arrêté sera publié au Journal officiel de la République française.

Fait à Paris, le 17 octobre 2002.

Michèle Alliot-Marie

FIG. D.5 – *Document N°DEFD0202303A*

Document initial : Arrêté du 17 octobre 2002 portant délégation de signature (voir la source figure D.5)
Signature lexicale : major, armées, suppléants, colonne, attachée, relevant, directions, annie, grange, renseignement, michèle, abrogé, déléguer, alliot, colonne suppléants, renseignement militaire, direction du renseignement militaire, service administratif, mme annie grange, organisation de l'administration centrale

1. Arrêté du 25 juillet 2002 portant organisation de la direction du renseignement militaire
2. Arrêté du 19 juillet 2002 portant délégation de signature
3. Arrêté du 27 août 2002 portant délégation de signature
4. Arrêté du 8 juillet 2002 portant délégation de signature
5. LOI no 2002-73 du 17 janvier 2002 de modernisation sociale (1)
6. Arrêté du 13 août 2002 portant délégation de signature
7. Arrêté du 12 février 2002 portant délégation de signature
8. LOI no 2002-2 du 2 janvier 2002 rénovant l'action sociale et médico-sociale (1)
9. Décret no 2002-105 du 25 janvier 2002 portant actualisation et adaptation du droit électoral applicable outre-mer
10. Rapport au Président de la République relatif à l'ordonnance n° 2002-1476 du 19 décembre 2002 portant extension et adaptation de dispositions de droit civil à Mayotte et modifiant son organisation judiciaire

TAB. D.5 – *Documents corrélés pour document N°DEFD0202303A*

Corrélation sémantique entre documents.

Résumé

Parmi les nombreuses méthodes d'accès à l'information présentes sur Internet, la corrélation de documents apparaît comme un outil complémentaire permettant aux internautes d'enrichir leurs connaissances sur un document sans avoir à formuler de question. L'objectif de nos travaux est de réaliser une méthode de corrélation sémantique dédiée à la recherche d'information juridique.

La méthode que nous dégageons vise à appliquer des outils et techniques d'ingénierie linguistique sur des textes préalablement choisis. Les unités textuelles saillantes les constituant sont alors dégagées, définissant pour chaque document analysé ce que nous appelons une signature lexicale. Ces signatures lexicales servent ensuite d'éléments clefs pour interroger un moteur de recherche dont les résultats représentent l'ensemble des documents corrélés. Cette méthode de corrélation est utilisée et évaluée dans un contexte de recherche d'information sur Internet et plus spécifiquement est intégrée aux développements d'un moteur de recherche.

Les principaux apports de nos travaux sont (1) un renouvellement des méthodes de recherche de documents corrélés par l'optimisation de signatures lexicales dédiées, (2) l'élaboration et l'évaluation d'un nouvel indice de pondération statistique noté T_{ifr}, (3) une réflexion sur l'aspect sémantique de la méthode de corrélation exposée, et enfin (4) une proposition concrète de réponse à la problématique de l'accès à l'information dans un contexte juridique.

Mots clefs : ingénierie des connaissances, recherche d'information, corrélation de documents, signature lexicale, pondération statistique, indice T_{ifr}

A new semantic relative pages algorithm.

Abstract

There are many ways to find information on the Web and search engines are the most frequently used tools. In this context, relative pages algorithms are complementary techniques providing more information about one specific document without asking any question. The goal of our work is to define a new semantic relative pages algorithm to perform search on a law oriented corpus.

To reach that goal, we defined a method that applies linguistic tools and techniques on previously selected documents. Relevant text units are extracted from our documents' corpus and are called lexical signatures. We use those lexical signatures as requests to search engine ; the results correspond to the pool of relative pages. Our relative pages algorithm is used and evaluated in an information retrieval context, being included in the development of a search engine.

The main contributions of our work are (1) a new perspective for building lexical signatures to perform relative pages searches, (2) the definition and the evaluation of a new relative pages algorithm called T_{ifr}, (3) a discussion on the semantic aspect of our method and finally, (4) a practical answer to the challenge of information retrieval in a law oriented context.

Keywords : information retrieval, relative pages algorithms, lexical signatures, T_{ifr} weight.