Extraction terminologique : vers l'élaboration d'une méthode de corrélation

Terminology extraction : towards the elaboration of a correlation method

RAPPORT CRI : A-346
C. Chotteau
Centre de Recherche en Informatique,
Ecole Nationale Supérieure des Mines de Paris
35 rue Saint Honoré, 77 305 Fontainebleau, France
chotteau@cri.ensmp.fr

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. Après avoir effectué un tour d'horizon rapide sur les techniques existantes, cet article expose l'ébauche d'une méthode de corrélation reposant sur une extraction terminologique utilisant des méthodes statistiques de pondération. Nous comparons deux algorithmes d'extraction terminologique ainsi que notre algorithme expérimentale, T_{ifr}, avec les résultats d'une extraction manuelle. Cet article montre qu'il est possible de faire reposer une méthode de corrélation sur une extraction terminologique. Il montre également qu'une extraction opérée via T_{ifr} donne de meilleurs résultats que les méthodes concurrentes dans le cadre d'une extraction de syntagmes nominaux et le classement des termes, et que les résultats sont proches des autres méthodes pour la recherche de mots clefs.

Mots Clés

Corrélation de documents, extraction terminologique, méthode de pondération, $T_f.I_{df}$, calcul d'entropie.

Abstract

Among the large amounts of information retrieval techniques available on the Internet, the documents correlation technique appears like a complementary tool which allows net surfers to increase their knowledge about a document without asking any question. After glancing at existing techniques, this paper presents the outline of a new correlation method based on terminology extraction using statistical term weighting. We compared two different terminology extraction algorithms and our empirical approch, T_{ifr}, with the results of a manual extraction. This paper shows that terminology extraction can be used to built a correlation method. It also shows that our terminology extraction algorithm gives better results to detect noun phrases and rank terms, and it is as good as others to retrieve significant nouns.

Keywords

Documents correlation technique, terminology extraction, term weighting, $T_f.I_{df}$ weighting, entropy weighting.

1 Introduction

La recherche d'informations sur Internet débute le plus souvent par l'utilisation d'un moteur de recherche. Malheureusement dans certaines situations, cette démarche peut être pénalisante car elle implique de formuler correctement sa question alors qu'il existe d'autres méthodes d'accès à l'information telle que la corrélation entre documents. Cette méthode permet à un utilisateur possédant un article électronique d'obtenir des documents similaires dans lesquels on aborde les mêmes thèmes pour compléter ou approfondir ses connaissances du sujet.

Dans le domaine de la corrélation, il existe deux principales voies de recherche. La première est topologique : elle étudie la structure interne des documents électroniques en HTML, recherche et utilise les liens hypertextes existants. De manière schématique, ce procédé rapproche les documents grâce aux références et aux citations qui les lient les uns aux autres. Ainsi une méthode de corrélation fondée sur l'algorithme de co-citation [7] utilise les liens hypertextes pour associer des documents souvent co-cités. L'utilisation des liens hypertextes dans le cadre de la recherche d'informations [1, 11] se justifie par le fait qu'ils révèlent une volonté de l'auteur de mettre en relation plusieurs documents [13].

La deuxième voie de corrélation est terminologique : elle s'intéresse aux attributs du texte que sont les termes (désigne indifféremment des mots ou des syntagmes nominaux notés SNs). Cette forme de corrélation se rapproche des travaux de résumés automatiques par extraction de termes clefs, qui se reposent sur “ l’hypothèse qu’il existe, dans tout texte, des unités textuelles saillantes ” [8]. Une corrélation terminologique tente alors d’identifier les termes saillants dans les documents pour ensuite les rapprocher en fonction des similitudes trouvées dans le vocabulaire. L’idée est d’extraire une signature sémantique pour chaque texte et de la comparer aux autres.

Les voies topologiques posent plusieurs problèmes liés à l’utilisation des liens hypertextes. Tout d’abord se pose la question de leur choix car même s’ils résultent nécessairement d’une volonté de l’un des auteurs [3], cela n’assure pas qu’ils sont liés sémantiquement avec le document (exemple des liens publicitaires).

J. Savoy et J. Picard [18] soulignent un autre point négatif : Les liens hypertextes facilitent grandement la navigation et les recherches dans les documents, mais malheureusement ils omettent une grande partie de la somme d’informations disponible ce qui est pénalisant.

Le dernier problème rencontré avec des approches de corrélations topologiques est qu’elles ont une vision macroscopique du Web : elles permettent de relier des sites entre eux mais ne savent pas le faire pour les pages qui les composent. La majeure partie des informations intéressantes étant contenue dans les pages du site, il faut descendre à un niveau inférieur et corréler des pages entre elles.

Pour ces raisons, nous nous dirigeons vers une voie terminologique qui nous offre plus de souplesse et nous permet de mieux appréhender l’aspect sémantique du texte qui réside dans la création d’une signature.

L’approche que nous envisageons se scinde en trois phases : une phase d’extraction terminologique pour extraire tous les termes du document que l’on nommera par la suite “ candidats termes ”; une phrase de sélection pour ne garder que les termes les plus pertinents que l’on appellera “ termes valides ”; la dernière étape permet de retrouver les documents du corpus ayant une signature similaire.

L’objet de cet article est d’aborder la problématique liée aux deux premières phases. Ainsi dans une première partie, nous allons traiter du problème de l’extraction des “ candidats termes ” et de la recherche des “ termes valides ” par l’utilisation de méthodes de pondérations classiques. Dans une deuxième partie, nous montrerons les insuffisances de ces méthodes classiques qui nous poussent à élaborer un indice de pondération expérimental. La troisième partie détaille une expérience visant à confronter les méthodes classiques et notre indice expérimental avec une pondération manuelle. Enfin, la dernière partie permet de comparer les résultats obtenus avant de pouvoir conclure.

2 Sélection des termes valides

Retrouver les “ termes valides ” passe par une étape préliminaire de sélection des “ candidats termes ”. Pour cette étape d’extraction, nous avons utilisé l’API d’extraction terminologique Genet [1]. Cet outil de traitement automatique du langage qui dérive dans sa conception de l’extracteur terminologique Sylex [5] a été retenu pour sa grande souplesse d’utilisation. Son paramétrage nous a permis, notamment, de filtrer et éliminer tous les “ candidats termes ” considérés comme trop fréquents donc porteurs de bruit pour la corrélation, ceux présents dans plus de 10% du corpus.

2.1 Méthodes de pondérations

La sélection des termes saillants parmi un ensemble de “ candidats termes ” est un problème souvent évoqué dans la littérature. Pour réaliser cette opération deux voies sont couramment suivies : la première est linguistique, elle tente de faire émerger les “ termes valides ” en utilisant des filtres successifs, le plus souvent morpho-syntaxiques, pour statuer sur l’importance qu’occupe un terme au sein d’une partie du corpus. La deuxième approche est statistique et ne prend en compte que des considérations fréquentielles pour choisir les “ termes valides ”.

La voie que nous allons suivre est statistique, car elle permet de conserver une approche générale ce qui évite de devenir dépendant du corpus étudié. Les approches statistiques proposent de multiples techniques de pondérations pour quantifier la valeur informationnelle apportée par un terme. On peut évoquer des pondérations tel que le concept de mutuelle information [4], le coefficient Φ^2 [12] ou encore le coefficient Loglike [9], des indices étudiés par B. Daille [6] pour extraire des SNs pertinents dans des documents. Malgré les bons résultats obtenus par certains indices, nous ne pouvons pas utiliser des techniques similaires car elles ne s’appliquent qu’à la pondération des SNs. Or nous souhaitons pondérer de la même manière mots et SNs pour ainsi pouvoir comparer les résultats obtenus.

L’extraction des “ termes valides ” a également été abordée par P. Lafon [15] qui propose d’appliquer une distribution hypergéométrique à la répartition des termes dans le corpus. La méthode permet de retrouver les spécificités du vocabulaire propre à chaque partie du corpus et donc d’en déduire les termes saillants. Toutefois cette approche impose de nombreuses restrictions [14]. Ainsi, il faut, à titre d’exemple, étudier un corpus découpé en parties de tailles sensiblement égales ou encore comparer des termes de même nature (noms, verbes, pronoms,...). Toutes ces restrictions nous empêche d’utiliser cette approche car ces conditions initiales ne peuvent être respectées de par la nature même de notre corpus d’étude.

Pour retrouver les “ termes valides ” parmi les “ candidats termes ”, nous allons donc utiliser des méthodes classiques de pondérations qui satisfont aux caractéristiques de notre corpus. Notre démarche est d’utiliser ces méthodes statistiques, exposées ci-après, pour extraire les “ termes valides ” et comparer leurs résultats avec une méthode expé-

1. Développée par la société Ogmios/Systal.

rimentale de pondération.

2.2 Indice $T_f.I_{df}$

Dans l'optique de l'élaboration d'une méthode de corrélation, on peut qualifier de termes pertinents des " candidats termes " qui sont peu présents dans le corpus d'étude noté Γ mais qui sont occurants dans les documents où ils apparaissent. Un moyen simple de retrouver ce type de terme est d'utiliser une combinaison de deux critères de pondération que sont T_f [16] et I_{df} [19]. L'indice résultant noté $T_f.I_{df}$ donne un moyen simple de mettre en évidence des termes fréquents mais peu présents dans Γ [20] :

$$T_f.I_{df}(D_i,\mathcal{T}_j) = \frac{N(D_i,\mathcal{T}_j)}{Max_{d\epsilon\Gamma}(N(d,\mathcal{T}_j))}.\frac{\|\Gamma\|}{n_\Gamma(\mathcal{T}_j)} \quad (1)$$

avec $N(D_i,\mathcal{T}_j)$ le nombre d'occurrences du terme $\mathcal{T}_j$ dans le document D_i et $n_\Gamma(\mathcal{T}_j)$ le nombre de documents distincts de Γ contenant $\mathcal{T}_j$

2.3 Entropie d'un terme

Pour mesurer l'importance d'un terme dans un texte, on peut aussi regarder la quantité d'informations 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[2]. 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 (2); elle est aussi appelée ratio d'indication de bruit par Salton [17], qui l'utilise pour sélectionner des " termes valides " dans un document :

$$\mathcal{H}_\Gamma(\mathcal{T}_j) = -\sum_{\substack{d\epsilon\Gamma \\ N(d,\mathcal{T}_j)>0}} P(d,\mathcal{T}_j).\log_2(P(d,\mathcal{T}_j)) \quad (2)$$

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

Si maintenant on remplace la probabilité $P(D_i,\mathcal{T}_j)$ par la fréquence d'apparition de $\mathcal{T}_j$ dans Γ, on obtient l'équation (3).

$$\mathcal{H}_\Gamma(\mathcal{T}_j) = -\sum_{\substack{d\epsilon\Gamma \\ N(D_i,\mathcal{T}_j)>0}} N(D_i,\mathcal{T}_j).\log_2(\frac{\|\Gamma\|}{df(D_i)}) \quad (3)$$

Cette deuxième forme du ratio d'indication est plus souvent utilisée, car elle permet de prendre en compte les statistiques sur Γ et $\mathcal{T}_j$. Elle était, notamment, employée par C. Fluhr [10] lors de la mise au point du système de recherche documentaire SPIRIT. Il s'agit de la deuxième méthode de pondération que nous allons expérimenter.

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 : Elles ne tiennent pas assez compte de la répartition des termes dans le corpus.

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. Lafon [15]. 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 occurant dans le document analysé alors que T_{ir} met en avant des termes discriminants en étudiant leur répartition dans le corpus.

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 :

$$iR(\mathcal{T}_j) = \sum_{\substack{d\epsilon\Gamma \\ N(d,\mathcal{T}_j)>0}} \frac{N(\mathcal{T}_j)}{N(d,\mathcal{T}_j)} \quad (4)$$

avec $N(\mathcal{T}_j)$ le nombre d'occurrences de $\mathcal{T}_j$ dans Γ

D'après l'équation (4), 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, prenons en exemple un terme quelconque apparaissant 30 fois dans Γ. Ce terme peut être réparti de multiples façons : 30 fois dans le même document, 1 fois dans 30 documents et bien entendu il reste toutes les étapes intermédiaires.

Pour trouver toutes les valeurs iR, il faut envisager toutes les répartitions possibles dans Γ, ce qui 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 qui nous permet de trouver et de dénombrer toutes les décompositions possibles. Dans l'exemple pris, il y a ainsi 5.604 décompositions différentes du nombre 30 en somme d'entiers positifs non nuls.

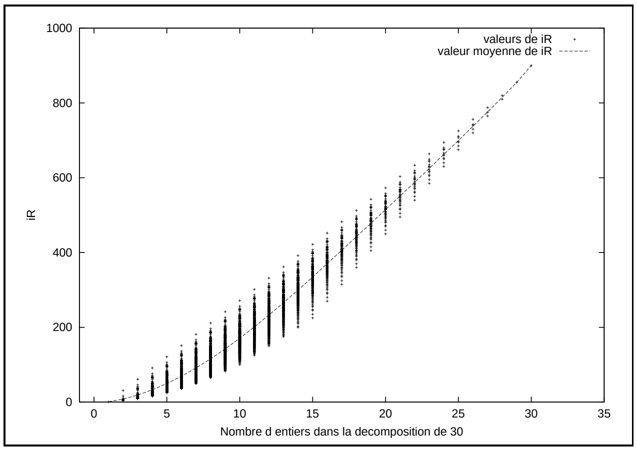

FIG. 1 – *Comportement de iR.*

La figure 1 illustre toutes les valeurs prises par iR dans le cas d'un terme présent 30 fois, valeurs que l'on retrouve en abscisse alors qu'en ordonnée apparaît le nombre d'entiers strictement positifs participant à la décomposition, c'est à dire le nombre de documents dans lequel est réparti le terme.

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 à 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) en l'encadrant par ses bornes atteintes dans le cas ou $\mathcal{T}_j$ est un hapax et ou il est présent dans tous les documents.

$$1 \leq iR(\mathcal{T}_j) \leq N(\mathcal{T}_j)^2 \tag{5}$$

La deuxième étape consiste à normaliser la formule (5) pour obtenir l'inéquation (6) :

$$\frac{1}{N(\mathcal{T}_j)^2} \leq \frac{iR(\mathcal{T}_j)}{N(\mathcal{T}_j)^2} \leq 1 \tag{6}$$

On retranche ensuite à l'inéquation (6) sa borne inférieure pour obtenir la formule (7). Cela permet de dilater les écarts de poids obtenus par les termes et de faciliter la comparaison entre des termes provenant de documents différents.

$$0 \leq \frac{iR(\mathcal{T}_j)}{N(\mathcal{T}_j)^2} - \frac{1}{N(\mathcal{T}_j)^2} \leq 1 - \frac{1}{N(\mathcal{T}_j)^2} \tag{7}$$

Dernière étape, obtenir un indice qui mesure le désordre amené par un terme. On prend alors l'opposé de l'équation (7) 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 (8) que nous allons utiliser pour rechercher des termes porteurs d'informations et donc à priori discriminants :

$$T_{ir}(\mathcal{T}_j) = 1 - (\frac{iR(\mathcal{T}_j)}{N(\mathcal{T}_j)^2} - \frac{1}{N(\mathcal{T}_j)^2}) \tag{8}$$

3.2 Lissage par T_{if}

Pour faire le parallèle avec le concept énoncé par Luhn [16] 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 occurant dans un document. Nous avons alors retenu l'équation (9) pour T_{if}.

$$T_{if}(D_i,\mathcal{T}_j) = 1 + (\frac{N(D_i,\mathcal{T}_j)}{N(\mathcal{T}_j)}) \tag{9}$$

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.

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 deux documents, or en terme de corrélation il n'apporte pas d'information supplémentaire, au contraire. 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.

2. Terme uniformément réparti dans le corpus
3. Hapax

C'est le rôle jouer par le Dirac introduit dans l'équation (10) qui représente la formule de T_{ifr} valable pour un terme $\mathcal{T}_j$ dans un document D_i.

$$T_{ifr}(D_i,\mathcal{T}_j) = (1-\delta_1(n_\Gamma(\mathcal{T}_j)))(1+(\frac{N(D_i,\mathcal{T}_j)}{N(\mathcal{T}_j)}))(1-(\frac{iR(\mathcal{T}_j)}{N(\mathcal{T}_j)^2}-\frac{1}{N(\mathcal{T}_j)^2})) \tag{10}$$

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

La figure 2 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 lequel 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 dans la figure 1, car à un point particulier de cette figure correspond plusieurs valeurs possibles de T_{if} et donc plusieurs valeurs de T_{ifr}.

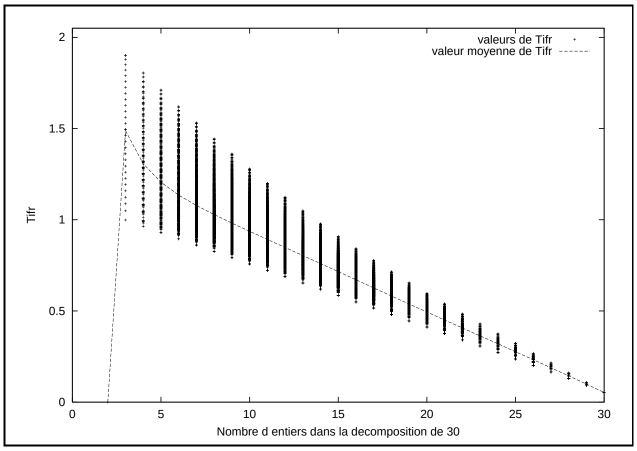

FIG. 2 – *Comportement de T_{ifr}.*

Sur la figure 2, 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 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 1 nous donne les valeurs de T_{ifr} pour les deux termes dans chaque document.

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.

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. 1 – *Exemple de valeurs de T_{ifr}*

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.

4 Expérimentation

Cette partie s'intéresse à la mise en place d'une expérimentation permettant de comparer les trois méthodes statistiques de pondérations décrites précédemment, à savoir l'indice $T_f.I_{df}$, le calcul d'entropie donné par l'équation (3) 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.

4.1 Corpus juridique expérimental

Notre méthode a pour but d'améliorer la navigation dans un site Internet, représentant l'un des portails juridiques du droit Français [4]. Cette base documentaire regroupe de multiples documents juridiques provenant des publications du Journal Officiel de la République française (notés J.O.) ainsi que de différents codes juridiques. Notre corpus d'expérimentation puise dans cette base; il s'agit des J.O. de l'année 2000 qui représente un corpus figé dans le temps. Cette année de parution représente un total de 304 journaux découpés en 24.178 documents (lois, décrets, articles, ...). La taille moyenne des documents est de 5 Ko, ce qui

4. http://www.droit.org

cache une très grande hétérogénéité, les plus petits documents étant de l'ordre de 1 Ko et les plus grands dépassant les 200 Ko.

4.2 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 $[D_{exp}(1) \ldots D_{exp}(50)]$. Les documents ont été sélectionnés aléatoirement dans notre corpus de référence. A partir de ces textes sont créés des formulaires d'évaluation qui se découpent en deux parties. Une partie qui permet de lire le document d'expérimentation provenant de la liste $[D_{exp}(1) \ldots D_{exp}(50)]$ et la seconde qui permet de valider la liste des " candidats termes " extraits de ce document grâce à notre extracteur terminologique et ordonnés 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.

Chaque évaluateur ayant été pré-sélectionné et souhaitant effectuer la validation doit fournir quelques informations nous permettant de le catégoriser comme candidat candide ou expert dans le domaine juridique. 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.

4.3 Résultats

La mise en place de l'évaluation à débuté en février 2002 par l'ouverture d'une enquête menée sur le Web [5]. Dans ce cadre, 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. Jusqu'à présent 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 bien 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.

Nombre de termes validés :
Nous commençons notre analyse des résultats par l'étude de la figure 3 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.

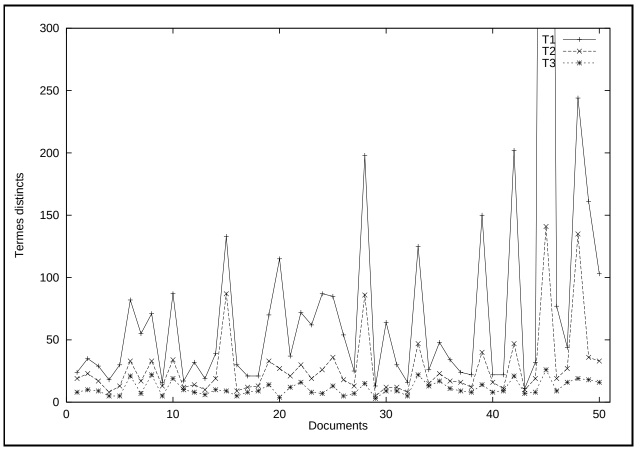

FIG. 3 – *Le corpus expérimental.*

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 étant 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 [6]. 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 effectue un test de χ^2 entre ces courbes. 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, largement supérieure à la valeur 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.

5. http://evaluation.w3sites.net

6. Termes validés par au moins deux évaluateurs

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

TAB. 2 – *Taille moyenne des documents*

Le tableau 2 est une synthèse des résultats de la courbe 3, 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 SNs.

Le but est de vérifier qu'il n'y a pas de différences fondamentales entre la taille des textes pour les mots et les SNs, colonne $T1$. On constate que les deux catégories ont à peu près 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 SNs. La seule différence visible, colonne $T3$, est constatée entre le nombre de mots et de SNs 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 plus souvent à l'établissement d'un consensus.

Accord entre évaluateurs :
La figure 4 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 alors à vérifier qu'il existe une concordance suffisante entre les réponses.

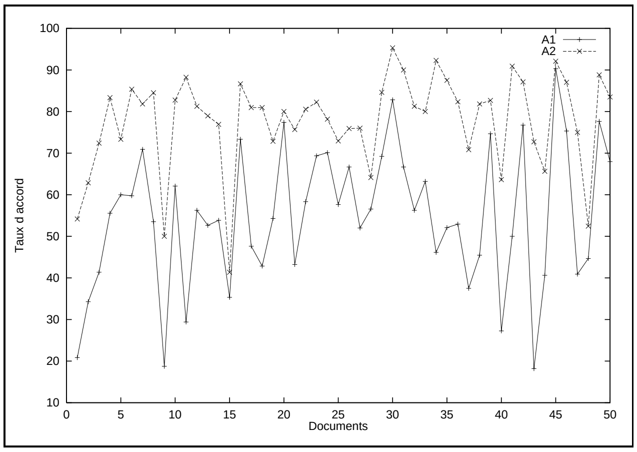

FIG. 4 – *Taux d'accords entre évaluateurs*

La courbe $A1$ que l'on peut tracer en utilisant la formule (11) 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.

La courbe $A2$ qui peut se tracer grâce à la formule (12), 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 unique évaluateur et il y a accord dans tous les autres cas de figure.

$$f_1(d) = \frac{\mathcal{N}_{tnc}(d) + \mathcal{N}_{tvt}(d)}{\mathcal{N}_t(d)} \tag{11}$$

$$f_2(d) = \frac{\mathcal{N}_t(d) - \mathcal{N}_{tu}(d)}{\mathcal{N}_t(d)} \tag{12}$$

avec $\mathcal{N}_{tnc}(d)$: ||termes non cochés par les évaluateurs||
$\mathcal{N}_{tvt}(d)$: ||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 3 n'est pas significatif, car il représente un taux d'accord moyen de 54% qui masque des irrégularités profondes sur la figure 4.

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'à 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%
SNs	51.51%	82.20%
Termes	54.61%	77.83%

TAB. 3 – *Accord moyen entre les évaluateurs.*

Le reste des valeurs du tableau 3 est là pour témoigner des différences minimes qui existent pour les évaluateurs entre les taux d'accords concernant les mots et les SNs; la colonne $A1$ représente les taux d'accords moyens trouvés à l'aide de l'équation (11). La colonne $A2$ représente ceux trouvés à partir de la formule (12).

Cette partie de vérification nous permet de confirmer notre hypothèse, à savoir que seuls les termes consensuels sont à prendre en compte. Une hypothèse qui se justifie en étudiant les courbes 3 et 4. 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 : un 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 Comparaison des pondérations

La finalité de cette dernière partie 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.

On peut toutefois s'interroger préalablement sur la justesse de cette comparaison. En effet, des indices tels que $T_f.I_{df}$ ou des mesures entropiques sont utilisés, entre autre, pour extraire des termes techniques spécifiques au domaine d'application du corpus, de ce fait il est tout à fait envisageable de les comparer entre eux ainsi qu'avec une extraction manuelle. A l'opposé, l'indice T_{ifr} a été, à priori, élaboré pour retrouver des termes intéressants pour la corrélation. Les termes doivent avoir des répartitions particulières, ce qui n'assure pas qu'ils sont spécifiques au domaine. Pourtant par construction, ces termes doivent être porteurs d'informations, l'hypothèse que nous formulons et que nous vérifierons, est qu'il existe une part non négligeable de termes retournés par T_{ifr} qui sont en relation directe avec la thématique du document considéré.

Pour effectuer cette comparaison, nous allons étudier et relever les similitudes existantes entre les listes des évaluateurs et nos listes expérimentales.

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 SNs. Ces deux listes sont ordonnées en fonction des choix de termes faits par les évaluateurs. Les termes placés au sommet des listes étant 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 partie précédente, 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 SNs. 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.

5.2 Protocole de comparaison

La comparaison des résultats entre les listes de termes de références 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 procéder à ce type de comparaison et confronter deux listes de termes, nous calculons leur précision (formule (13)) à chaque rang, en allant jusqu'à épuisement des termes de la plus grande des deux listes.

$$P(\mathcal{L}_1,\mathcal{L}_2,n) = \frac{\mathcal{N}_{tc}(\mathcal{L}_1,\mathcal{L}_2,n)}{n} \qquad (13)$$

avec $\mathcal{N}_{tc}(\mathcal{L}_1,\mathcal{L}_2,n) = \|$termes communs entre $\mathcal{L}_1$ et $\mathcal{L}_2$ 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 et 6 qui expriment la précision moyenne obtenue par les méthodes de pondérations. La figure 5 exprime les résultats pour les mots et la figure 6 pour les SNs. Pour les deux figures, la courbe $M1$ désigne la comparaison entre la liste idéale et la liste pondérée via T_{ifr}, la courbe $M2$ en comparant avec $T_f.I_{df}$ et $M3$ en comparant avec le calcul d'entropie vu précédemment.

En complément du taux de précision, le taux de rappel à été étudié; malheureusement il tend à être optimal 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 quel que soit la méthode de pondération.

5.3 Résultats

Les résultats proposés par la courbe 5 montrent que les trois méthodes obtiennent des taux similaires 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 $M1$ croît légèrement plus rapidement que ses concurrentes pour atteindre, un rang plus tôt, sa précision maximale.

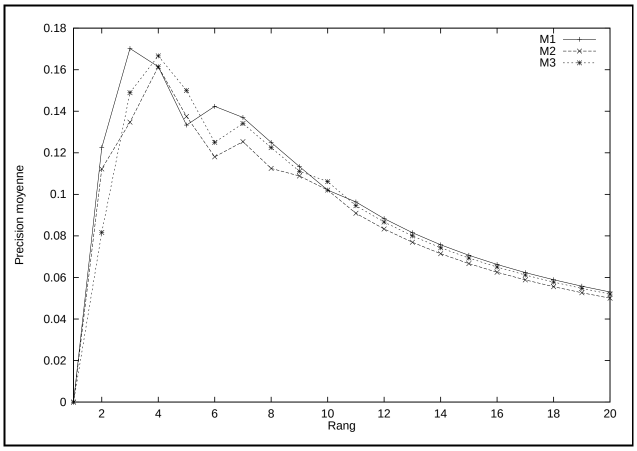

FIG. 5 – *Comparaison des méthodes de pondération pour les mots.*

Les résultats des SNs illustrés par la courbe 6 indiquent que la précision maximale est obtenue pour un rang supérieur valant 5 avec une précision de 30%, ceci pour la courbe $M1$. 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 ces deux concurrentes en atteignant, comme pour le cas des mots, plus rapidement sa précision maximale et en la maintenant quelque temps.

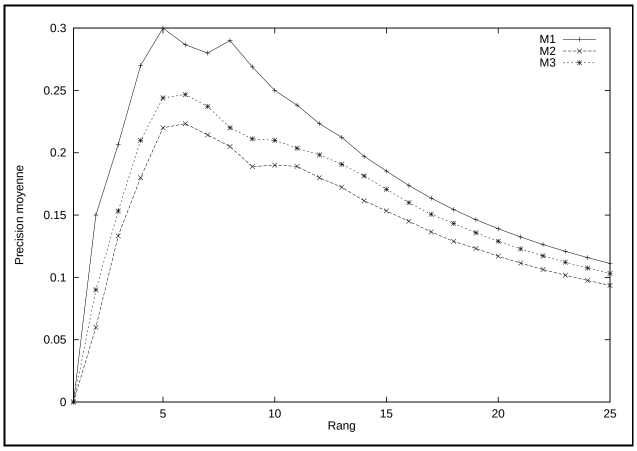

FIG. 6 – *Comparaison des méthodes de pondération pour les SNs.*

5.4 Analyse

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

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 2. 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 jugés comme pertinents par les évaluateurs. 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 SNs 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 SNs d'après le tableau 2 pour une longueur moyenne des textes de 41,46 SNs, ce qui nous permet statistiquement d'espérer de meilleurs résultats.

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

6 Conclusion

Dans cet article, 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 alors peu probable de pouvoir conclure correctement.

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 à priori ceux qui sémantiquement sont les plus proches des thématiques des documents étudiés; ils peuvent alors, être considérés comme de bons indicateurs pour la recherche de documents corrélés. Cette remarque aurait été tout autre si les évaluateurs n'avaient pas su s'accorder pour définir des termes consensuels.

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 vouloir 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 (13).

La dernière partie de l'expérimentation nous montre qu'il existe une différence significative entre les comparaisons de listes pondérées de mots et de SNs. 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 plus judicieux de concentrer ses efforts sur l'extraction et la pondération des SNs pour qu'ils acquièrent un rôle prépondérant dans la suite des travaux.

La comparaison des deux méthodes statistiques de pondération avec notre méthode empirique T_{ifr} montre que l'on obtient des résultats comparables avec des méthodes classiques (figures 5 et 6) 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. Les résultats qui en découlent ont donc plus de chance d'être pertinents. Malgré tout, pour vérifier la véracité des hypothèses engagées il est nécessaire de pouvoir tout d'abord comparer avec une extraction manuelle faite par des experts et d'attendre une évaluation des résultats corrélés.

Références

[1] R. A. Botafogo, E. Rivlin, and B. Shneiderman. Structural analysis of hypertexts: Identifying hierarchies and useful metrics. *Information Systems*, 10(2):142–180, 1992.

[2] C. Busha and S. Harter. *Research methods in librarianship. Techniques and interpretation*. Academic Press, 1980.

[3] S. Chakrabarti, B. Dom, S. Ravi Kumar, P. Raghavan, S. Rajagopalan, A. Tomkins, D. Gibson, and J. Kleinberg. Mining the Web's link structure. *Computer*, 32(8):60–67, 1999.

[4] K. W. Church and P. Hanks. Word association norms, mutual information, and lexicography. In *Computational Linguistics*, volume 16, pages 22–29, 1990.

[5] Patrick Constant. L'analyseur linguistique sylex. *Ecole d'été du CNET*, 1995.

[6] 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.

[7] J. Dean and M. R. Henzinger. Finding related pages in the world wide web. *WWW8 / Computer Networks*, 31(11-16):1467–1479, 1999.

[8] J-P. Desclés and J-L. Minel. *Résumé Automatique et Filtrage des textes* dans *Filtrage des textes dans Ingénierie des langues*, pages 253–270. Editions Hermès, Paris, 2000.

[9] T. Dunning. Accurate methods for the statistics of surprise and coincidence. In *Computational Linguistics*, volume 19, 1993.

[10] C. Fluhr. *Algorithmes à apprentissages et traitement automatique des langues*. PhD thesis, Université Paris Sud, Centre d'Orsay, 1977.

[11] M. E. Frisse. Searching for information in a hypertext medical handbook. *Communications of the ACM*, 31(7):880–886, 1988.

[12] 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.

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

[14] C. Labbe and D. Labbe. Que mesure la spécificité du vocabulaire ? *Lexicometrica*, 3, 2000.

[15] P. Lafon. Sur la variabilité de la fréquence des formes dans un corpus. *Mots*, 1:128–165, 1980.

[16] H. P. Luhn. A statistical approach to mechanized encoding and searching of literary information. *IBM Journal of Research and Development*, 1:309–317, 1957.

[17] G. Salton and C. Yang. A theory of term importance in automatic text analysis. *Journal of the ASIS*, 26(1):33–34, 1973.

[18] J. Savoy and J. Picard. Retrieval effectiveness on the Web. *Information Processing and Management*, 37(4):543–569, 2001.

[19] K. Sparck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28:11–21, 1972.

[20] K. Sparck Jones. Index term weighting. *Information Storage and Retrieval*, 9:619–633, 1973.