Assistance automatique pour l'homogénéisation d'un corpus Web de spécialité

Sophie Berland[1], Natalia Grabar[1,2]

[1] CRIM/INALCO – 1, rue de Lille – 75005 Paris – France – sophie.berland1@libertysurf.fr
[2] SIM/DIAM, DSI AP-HP – Université Paris 6 – 91, Bd de l'Hôpital – 75634 Paris cedex 13 – France – ngr@biomath.jussieu.fr

Abstract

Textual terminology aims at building terminological resources (glossaries, terminologies, terminological knowledge bases, etc) and is based on the analysis of textual data. The choice of relevant textual data therefore plays a crucial role in the task of acquiring terminological data. We claim that the more the data on which we work is homogeneous, the more the result will be reliable and representative of the domain being studied. In this paper, we present automatic processing means to help the homogenization of a corpus of web documents.

Résumé

La terminologie textuelle est une approche de construction de produits terminologiques (glossaires, terminologies, bases de connaissances terminologiques, etc.) qui est basée sur l'analyse de corpus textuels. Les données textuelles occupent donc une place cruciale pour la tâche d'acquisition terminologique. Nous faisons l'hypothèse que plus les données sur lesquelles on travaille sont homogènes plus le résultat de l'acquisition terminologique est fiable et représentatif du domaine traité. Dans cet article nous présentons les traitements automatiques mis en oeuvre pour aider à l'homogénéisation d'un corpus de documents Web.

Mots-clés : Corpus de spécialité, acquisition terminologique, homogénéité des données, document Web, filtrage, pertinence des documents.

1. Introduction

La terminologie textuelle (Bourigault and Slodzian, 1999) est une approche de construction de produits terminologiques (glossaires, terminologies, bases de connaissances terminologiques, etc.) qui repose sur l'analyse de corpus textuels. Les données textuelles occupent donc une place cruciale dans les travaux d'acquisition terminologique à partir de corpus, car ce sont ces données qui définissent (1) le type et la nature des traitements que l'on pourra effectuer et (2) la nature des données terminologiques que l'on pourra y trouver (Bachimont, 2000, p. 308-309). Le choix des textes à inclure dans un corpus pour l'acquisition terminologique doit ainsi répondre à la pertinence par rapport au domaine (textes représentatifs de ceux produits dans le domaine) et par rapport à l'application (textes représentatifs de ceux manipulés par l'application finale). Ces textes doivent en outre être suffisamment homogènes pour que les traitements automatiques appliqués fournissent un résultat fiable (Habert et al., 2000). Afin de disposer de données homogènes il peut être intéressant de mettre en oeuvre des filtrages automatiques qui permettraient de préparer et de faciliter la sélection manuelle de textes.

La réflexion que nous proposons dans cet article s'appuie sur le travail effectué dans le cadre

du projet SAFIR[1] sur la recherche d'information multilingue (français, anglais, allemand). La tâche que nous assumons dans ce projet est double : (1) proposer une méthode de construction de terminologie et (2) élaborer une terminologie de la cogénération[2], domaine choisi par EDF, un des partenaires du projet. La terminologie construite sera utilisée dans l'outil de recherche d'information multilingue SAFIR au lancement de la requête (aide à la formulation, expansion de la requête) et à la réception des réponses (filtrage et établissement de la pertinence des documents).

Afin de préparer le matériel nécessaire à la construction de la terminologie, nous avons été confrontés à la problématique de constitution d'un corpus. Nous avons décidé de chercher les documents sur la cogénération sur le Web. Ceci pour les raisons suivantes : (1) le domaine de la cogénération est relativement nouveau en France ; (2) souvent les documents sur cette technique sont destinés à l'usage privé des entreprises; l'accès à ces documents est donc restreint. (3) L'application finale est destinée à la recherche d'information sur le Web. Le Web se présente donc comme le contexte de l'application. (4) Le Web constitue une énorme base de documents provenant de sources multiples. Nous faisons l'hypothèse que des documents sur la cogénération s'y trouvent également. (5) Le Web fournit des documents sous format électronique, et donc assez facilement exploitables par des outils de traitement automatique des langues.

A la différence d'une base de données, créée souvent à des fins bien spécifiques et comportant donc des données structurées et propres, le Web permet d'accéder à toute sortes d'informations dont la quantité est difficilement imaginable et dont la qualité et la fiabilité varient beaucoup. Il en découle une autre différence souvent mentionnée : en interrogeant une base de données on trouve la réponse à la requête, tandis qu'en interrogeant le Web on trouve l'information qui porte sur le sujet de la requête.

Etant bien conscients de cette particularité des données du Web (profusion et hétérogénéité des données), nous avons proposé une méthode incrémentale à restriction progressive pour la constitution d'un corpus de spécialité à partir des données du Web (Grabar and Berland, 2001). C'est une méthode semi-automatique. La méthode est incrémentale dans le sens où le résultat final peut être obtenu en plusieurs étapes : les données dont on dispose déjà peuvent être réutilisées pour en fournir d'autres plus complètes et pertinentes (une première liste de mots-clés permet de trouver des documents. A partir de ces documents une liste plus complète et pertinente de mots clés est constituée. La nouvelle liste de mots-clés permet de compléter l'ensemble des documents déjà collectés, etc.). La méthode permet d'appliquer une restriction progressive à la masse de documents collectée sur le Web avec des moteurs de recherche généraux ou spécialisés et ensuite de passer la totalité de ces documents par des filtres successifs pour affiner et organiser les résultats. Cette méthode de constitution de corpus se décompose en quatre étapes : (1) recherche sur le Web de documents abordant le domaine, (2) sélection parmi ceux-ci des plus représentatifs du domaine, (3) regroupement des documents en sous-corpus homogènes, (4) formatage, normalisation et documentation du corpus. Dans (Grabar and Berland, 2001) nous avons présenté la première étape qui vise à collecter des documents concernant un domaine donné sur le Web. Maintenant nous présentons des filtrages automatiques mis en oeuvre pour préparer et aider à la tâche de sélection des documents représentatifs pour la constitution

1. Système d'Agents pour le Filtrage d'Information sur les Réseaux. La description du projet se trouve à l'adresse : `http://www-poleia.lip6.fr/~slodzian/safir.html`.

2. La cogénération est une technique de production simultanée d'électricité et de chaleur (eau chaude, vapeur ou eau surchauffée, par exemple).

d'un corpus homogène du domaine de la cogénération. Le résultat de ces filtrages permettra ainsi de filtrer et ordonner des documents pour les présenter à l'expert pour la validation.

2. Matériel

Pour sélectionner les documents les plus pertinents pour la tâche d'acquisition terminologique, nous utilisons une collection de textes et des mots-clés du domaine de la cogénération. Les programmes de filtrages ont été écrits en `Perl 5`.

2.1. Mots-clés

Une première liste de mots-clés nous a été fournie par les consultants en information d'EDF. Elle a été complétée et triée. Les mots-clés dont nous disposons n'ont pas tous le même statut. Nous en avons distingué deux types par rapport au domaine de la cogénération :

- mots-clés que nous avons appelés « transversaux » (ou bien « généraux »). Ces mots clés appartiennent de manière non-ambiguë au domaine de la cogénération. La présence d'un de ces mots dans un texte est l'indice que le texte aborde le domaine. Ces mots-clés comportent le terme *cogénération* ainsi que ses variantes et équivalents (*co génération, production combinée d'électricité et de chaleur, etc.*)
- mots-clés que nous avons appelés « spécifiques ». Ils sont spécifiques dans le sens où ils servent à préciser, parmi les documents abordant le domaine étudié, des points de vue différents. Nous avons distingué les points de vue suivants : économie, écologie, règlementation et technique.

Nous supposons que ces mots-clés sont discriminants et permettent de délimiter le domaine de la cogénération et les facettes distinguées de ce domaine.

2.2. Collection de textes

La collection de textes dont nous disposons a été collectée à la suite de plusieurs recherches de documents effectuées sur le Web. Ces recherches ont été faites (1) avec des moteurs de recherche généraux (AltaVista, Google et Lycos), (2) avec le moteur de recherche du site de la Communauté européenne, et (3) sur les sites ATEE (Association Technique Energie Environnement) et Energie plus [3]. Nous avons ainsi collecté 336 documents concernant la cogénération. Ces documents sont essentiellement commerciaux et juridiques. C'est pourquoi nous avons décidé d'enrichir cette collection de textes avec des documents éventuellement plus spécialisés provenant des bases de données utilisées par EDF et des actes des conférences sur la cogénération. Ce choix est tout à fait légitime, car (1) ces sources de données sont « contrôlées » et donc fiables et (2) le métamoteur SAFIR sera destiné également à interroger des bases de données d'EDF.

Les recherches dans les bases Galaxie, Prodec, Elsevier et le portail de cogénération d'EDF nous ont fourni 38 documents (aux formats `HTML`, `PDF`, `DOC`), tous convertis au format `HTML`. Les documents provenant des actes des conférences sur la cogénération (52 textes) ont été digitalisés, corrigés orthographiquement et également convertis au format `HTML`.

Nous disposons finalement de 426 documents sur la cogénération. Nous considérons que cette collection de documents est suffisamment représentative et variée pour l'utiliser comme source

3. ATEE et Energie plus sont des sites consacrés aux énergies alternatives, dont la cogénération.

pour l'acquisition terminologique[4].

3. Méthode

Lors de la recherche de documents sur le Web, nous avions fait une série de filtrages basés sur le paramétrage des moteurs de recherche (langue des documents), le contrôle des URL[5] (élimination des doublons d'URL) et une analyse très superficielle du contenu des documents (contrôle du nombre minimal de mots-clés par page). Ces filtrages ont permis d'enlever une partie considérable du bruit. Mais d'autres filtrages, avec cette fois une analyse plus poussée du contenu des documents, sont nécessaires pour avoir un corpus homogène. Car nous considérons que disposer de documents caractéristiques du domaine ne suffit pas pour constituer un corpus pour l'acquisition terminologique. En effet les documents que nous avons réunis doivent être filtrés et évalués quant à leur pertinence par rapport au domaine de la cogénération. Nous faisons l'hypothèse que plus un corpus est homogène, plus les candidats termes dont les fréquences sont élevées feront office de bon indicateur quant à leur pertinence et à leur centralité par rapport au domaine.

Nous décrivons dans cette section les traitements automatiques appliqués pour obtenir un corpus homogène quant au domaine de la cogénération. L'application de ces traitements aux données ne permet en aucun cas de construire un corpus homogène d'une manière complètement automatique. Le résultat de ces traitements est perçu comme des propositions qui devront être validées ou invalidées manuellement.

3.1. Sélection des documents les plus représentatifs du domaine

Après une première analyse des documents collectés, dont une majeure partie provient du Web, nous avons jugé nécessaire de faire les traitements suivants pour sélectionner les documents les plus représentatifs du domaine :

- détection des documents doublons ($D_a = D_b$) et des documents inclusions ($D_a \subset D_b$),
- détection des pages de liens (nous faisons une distinction entre les pages de navigation composées essentiellement de liens hypertextes et les pages informatives intéressantes en soi pour leur contenu textuel),
- établissement d'un ordre de pertinence des documents par rapport au domaine de la cogénération pour n'en sélectionner que les plus importants et pour faciliter leur sélection manuelle.

3.1.1. Détection de documents doublons et inclusions

Lors des recherches et du rapatriement des documents du Web, nous avons systématiquement éliminé les doublons des URL grâce à un système de mémorisation des URL relevées. Cependant, et nous l'avons remarqué très rapidement, il restait des doublons et inclusions de documents qui provenaient d'adresses différentes. En effet, un même document peut être publié sur des sites différents. De plus chaque site peut découper et répartir les documents de manière différente. Suite à l'enrichissement avec des documents venus d'autres sources, le risque d'avoir des doublons augmente. Il est donc nécessaire de faire une analyse plus poussée du contenu des documents afin de détecter les doublons et les inclusions.

4. Si par la suite il s'avère que le corpus n'est pas assez riche et représentatif, il devra être complété.
5. Uniform Source Locator, adresse d'une page Web.

Il existe déjà de nombreux travaux en détection de documents similaires et de doublons. Le travail de thèse de (Nauer, 2001) avait pour but la fouille de données bibliographiques multibases à destination des scientifiques. A côté des difficultés que présente l'intégration des références bibliographiques provenant de sources différentes (format, normalisation de l'indexation, etc.) Emmanuel Nauer se pose également la question des doublons de documents. Ce sont les couples auteur - label (« *label* » correspond à la clé dans le format de description des références bibliographiques en BibTex) de chaque document qui sont utilisés pour la détection des documents doublons. Une autre expérience est celle du site ResearchIndex [6] consacré à la diffusion de la littérature scientifique. Dans le souci d'offrir une information plus complète ResearchIndex propose de consulter des documents similaires. Le calcul de similarité est basé sur le pourcentage de phrases proches entre les documents. (Dachelet, 1990) mentionne un autre principe pour mesurer la similitude entre des documents, basé sur les liens de citation bibliographique. Ces liens témoignent de la similitude du sujet abordé. Il existe également des travaux de recherche de similitude en morphologie. Par exemple (Jacquemin, 1997) propose une mesure vectorielle de distance entre deux suffixes qui permet de trouver des suffixes allomorphes ou des variantes de suffixes (par exemple, « *-tion* », « *-ation* »).

Dans la tâche de constitution d'un corpus de spécialité exploitable, nous voulons repérer les documents doublons et inclusions afin de les éliminer, car nous considérons qu'ils faussent le corpus et par là-même le résultat de l'acquisition terminologique.

Nous appliquons pour cela la technique de recherche documentaire qui est utilisée d'habitude pour calculer la proximité entre la requête et les documents-réponses, où les documents sont représentés par un certain nombre de descripteurs pondérés et comparés à la requête composée de mots-clés également pondérés. Dans notre cas, nous utilisons cette technique pour calculer la proximité entre deux documents.

Trois moments sont à distinguer : (1) recherche de descripteurs dans les documents, (2) attribution de poids à ces descripteurs, (3) calcul de proximité entre les documents.

Nous avons décidé de comparer les documents en les représentant par un vecteur de leurs formes non lemmatisées associées à leurs fréquences. Ce type d'index de chaque document a été établi avec un programme `Perl`. Nous considérons (1) que deux documents qui ont le même index (mêmes formes avec les mêmes fréquences) sont identiques, c'est à dire doublons ; (2) qu'un document est inclus dans un autre si son index est un sous-ensemble de l'index de l'autre.

Etant bien conscients du fait que les mêmes formes peuvent former des phrases pas toujours équivalentes, voire former des phrases avec des sens opposés, nous faisons l'hypothèse que ce phénomène n'apparaîtra pas entre des documents d'une taille importante.

Nous utilisons donc une méthode vectorielle, inspirée de la formule du cosinus, qui consiste à comparer deux vecteurs représentant deux documents D_a et D_b. La proximité entre ces deux documents correspond à la proximité de leurs vecteurs d'index, elle est obtenue en calculant leur indice de similarité selon la formule :
$sim(D_aD_b) = \frac{\sum_{i=0}^{n} F_{ai}F_{bi}}{\sqrt{\sum_{i=0}^{n}(F_{ai})^2 \sum_{i=0}^{n}(F_{bi})^2}} P_aP_b$,
où :

- F_{ai} et F_{bi} sont les fréquences d'un mot i dans les documents D_a et D_b,

6. ResearchIndex : `http://www.researchIndex.com/`.

- P_a est la proportion du nombre de formes communes entre D_a et D_b par rapport au nombre de formes dans le document D_a,
- P_b est la proportion du nombre de formes communes entre D_a et D_b par rapport au nombre de formes dans le document D_b.

Dans cette formule, inspirée de celle du calcul de cosinus, nous utilisons en plus les proportions P_a et P_b. Si la formule du cosinus sans ces proportions permettait de détecter les documents doublons, elle n'était pas efficace dans le cas des documents inclusions. La prise en compte de ces proportions permet de détecter également les documents inclusions. Il serait aussi intéressant de voir si le seul calcul des proportions P_a et P_b ne permettrait pas de repérer doublons et inclusions, en dehors de la formule du cosinus, ce calcul étant plus simple et plus rapide à réaliser que le calcul du cosinus.

Comme nous l'avons indiqué plus haut, nous avons observé dans la collection des documents les cas de figures suivants : (1) documents doublons, (2) documents inclusions, (3) documents différents. L'algorithme proposé permet de détecter les documents doublons et les inclusions.

Deux documents D_a et D_b sont supposés être doublons si : leur indice de similarité $sim(D_aD_b)$ est supérieur à 0,5, et les proportions P_a et P_b sont supérieures à 0,95.

Le document D_a est supposé être l'inclusion du document D_b si : leur indice de similarité $sim(D_aD_b)$ est compris entre 0,1 et 0,5, et la proportion P_a est supérieure à 0,95.

Les seuils ont été ainsi fixés suite aux observations faites sur les résultats intermédiaires. Le résultat de ce traitement (propositions de documents doublons et inclusions) doit être validé manuellement.

3.1.2. Détection des pages de liens

Un autre problème à affronter lorsqu'on travaille avec les données du Web est celui des types de documents. Un premier aperçu des documents a montré la présence de pages de texte et de pages de liens. Les pages de liens servent essentiellement à la navigation, leur fonction principale est de pointer vers d'autres pages du Web. La fonction principale des pages de texte est informative. Les pages de liens correspondent aux cas de figures suivants :

- page de type « *annuaire* » : la page est une liste de liens vers les pages d'autres sites,
- page de type « *sommaire* » (externe) : la page est une liste de liens vers d'autres pages du même site,
- page de type « *sommaire* » (interne) : la page comporte des liens (ancres) vers des paragraphes de la même page.

Les pages de liens constituent en soi une source d'information importante. Pour (Amitay and Oberlander, 1997) les liens hypertextes sont étudiés d'un point de vue cognitif (disposition, organisation des pages HTML), de même que d'un point de vue structurant et pratique. Dans le projet TypWeb les liens hypertextes d'un site ont été relevés et quantifiés afin de discriminer les sites marchands des sites personnels (Beaudouin et al., 2001). Dans une autre perspective, le travail de (Adamic and Adar, 2000) vise à identifier les communautés d'individus en étudiant les liens hypertextes insérés dans les pages personnelles d'universitaires. Un des objectifs de ce travail est de mettre en relation des individus dont les centres d'intérêts sont similaires.

Pour la tâche d'acquisition terminologique à partir de ressources textuelles, nous nous intéressons aux pages de texte et aux pages de type sommaire interne. Les autres types de pages de liens (« *annuaire* », « *sommaire externe* ») peuvent également être exploitées mais dans une optique différente, par exemple pour compléter le corpus avec de nouveaux documents si cela s'avère nécessaire. Nous avons donc besoin de détecter les pages de liens pour les éliminer ensuite. Pour cela nous étudions le code HTML des documents et relevons :

- C_t, nombre total de caractères dans le texte (y compris dans l'hypertexte),
- C_l, nombre total de caractères dans les liens externes et internes (les balises <a href> pointant sur les pages d'autres sites ou du même site).

Pour faire une proposition quant au type d'une page (page de liens ou page de texte), nous calculons la proportion de l'hypertexte C_l par rapport à la totalité du texte C_t. Nous faisons intervenir plutôt le nombre de caractères que le nombre de liens pour obtenir une estimation plus pertinente dans cette caractérisation de pages HTML.

L'application de ce calcul permet de faire une première proposition des pages quant à leur caractère hypertextuel. Formellement, plus la valeur obtenue est importante, plus il y a de chances que la page en question soit de type « annuaire » ou « sommaire ». Ici encore il ne s'agit que d'une aide permettant une validation manuelle plus aisée et rapide.

3.1.3. Etablissement de l'ordre de pertinence des documents

La tâche de l'évaluation de la pertinence des documents par rapport au domaine de la cogénération constitue une autre étape dans l'homogénéisation de la collection des documents. Cette tâche aussi conduit aux travaux d'évaluation de la pertinence des réponses par rapport à la requête. Mais dans ce cas précis, c'est la proximité entre un document donné et le modèle du domaine qui est calculée. Nous représentons le domaine par un certain nombre de mots-clés, dont le poids est 1. Nous faisons ainsi l'hypothèse que la liste des mots clés du domaine pondérés à 1 représente le document « idéal » de ce domaine. Plus un document réel se rapproche de ce document « idéal », plus il sera pertinent. Comme dans le cas de la détection de documents doublons (sous-section 3.1.1), cette tâche se décompose en trois étapes : (1) recherche des descripteurs du domaine et des documents, (2) attribution de poids aux descripteurs de chaque document, (3) évaluation de la pertinence des documents par rapport au domaine.

Nous prenons comme descripteurs les mots-clés « transversaux » qui définissent le domaine de la cogénération ainsi que les noms communs du texte, en faisant ainsi l'hypothèse que ce sont les noms communs et les groupes nominaux qui représentent le mieux le contenu d'un document.

Pour la pondération des descripteurs dans les documents, nous avons choisi une des formules relevée dans (Salton and Buckley, 1987, p. 17) :
$P_i = 0.5 + 0.5\frac{f_i}{f_{max}}$ Elle est destinée à calculer la fréquence normalisée des descripteurs, en tenant compte de la fréquence f_i d'un descripteur i dans un document et de la fréquence f_{max} du descripteur le plus présent dans ce document. Nous avons préféré de ne pas utiliser les algorithmes qui diminuent le poids des termes distribués dans plusieurs documents de la collection, comme par exemple la pondération TF-IDF très répandue dans le milieu de la recherche d'information. Ceci parce que nous cherchons à rapprocher les documents et non à les différencier.

L'évaluation de la pertinence des documents par rapport au domaine est obtenue en appliquant

la formule : $sim_{Dom,Doc} = \frac{\sum_{i=0}^{n} P_i}{N_d}$ où P_i est le poids d'un mot-clé i d'un document et N_d est le nombre total des mots-clés du domaine. Le domaine Dom, vu comme le document « idéal », est ainsi représenté par la somme des poids de ses descripteurs (le poids de chacun étant égal à 1). Le document Doc est également représenté par la somme des poids pondérés de ses descripteurs.

Là aussi, une validation manuelle est nécessaire. C'est par rapport à cette validation que nous pourrons définir l'efficacité des calculs appliqués et sélectionner les documents pertinents par rapport au domaine de la cogénération. Ce sont ces documents qui seront utilisés dans la tâche d'acquisition terminologique.

4. Résultat et discussion

Un ensemble de 426 documents a été soumis aux filtrages automatiques définis dans la section 3. Ces filtrages ont permis (1) de préparer et de faciliter la sélection manuelle des documents et (2) d'homogénéiser ainsi le corpus.

Doublons et inclusions. Ce filtrage a montré que, parmi les 426 documents, 35 étaient impliqués dans la relation de doublon ou d'inclusion. Des 19 propositions de doublons toutes ont été validées exactes, un des documents doublons étant gardé à chaque fois. Dans la majorité des cas, la relation de doublon se retrouve entre deux documents, dans quatre cas entre trois et dans un cas entre sept documents (il s'agit d'un annuaire des projets et des études sur la production d'électricité).

Nous avons eu 8 propositions d'inclusions. Les cas d'inclusion ont été plus délicats à résoudre. Deux choix étaient possibles :

- garder les documents incluants et donc plus grands et hétérogènes,
- garder les documents inclus et donc plus courts et homogènes.

Nous avons préféré garder les documents inclus, car nous pensons que si les documents courts sont retrouvés séparément, ils sont plus pertinents du point de vue de leur thématique et de leur homogénéité. Les documents incluants sont souvent des documents polythématiques et donc hétérogènes en soi.

Liens hypertextes. La validation manuelle des pages que le système automatique de filtrage a caractérisé comme pages de liens hypertextes a permis de rejeter 36 documents. Nous n'avons pas pu déterminer de seuil permettant de distinguer de manière sûre les pages de liens et les pages de texte. Il serait utile d'affiner la formule en introduisant par exemple un facteur de normalisation pour rendre les valeurs plus comparables. Il serait aussi intéressant de ne prendre en compte que le comptage brut de liens internes et externes et de comparer ensuite les résultats des deux approches. Les propositions de caractérisation obtenues avec ce traitement nous ont permis de consulter les documents d'une manière plus systématique et d'aller plus vite dans la validation de cette caractérisation.

Pertinence par rapport au domaine. Les 355 documents gardés après ces deux filtrages ont été classés selon leur pertinence par rapport au domaine. Le calcul de cette pertinence est basé sur la représentation des documents et du domaine par les descripteurs pondérés. Les documents ont été proposés à l'expert pour la validation dans l'ordre décroissant de cette pertinence. Notre attente quant à la validation de l'expert portait sur plusieurs points : (1) sélectionner les documents pertinents, (2) évaluer l'efficacité de la formule appliquée, (3) définir le seuil de

pertinence qui permettrait (dans l'avenir) de discriminer automatiquement les documents pertinents.

Nous avons pris comme descripteurs les noms communs et les mots-clés « transversaux » de la cogénération. Mais d'autres mots, comme par exemple les adjectifs ou les verbes, peuvent être également pris en compte. Les documents ont été étiquetés et lemmatisés avec le logiciel Cordial 6. Le traitement était fait sur les formes lemmatisées. Puisque nous cherchons à comparer le contenu des documents, nous considérons qu'il est tout à fait légitime de s'abstraire des formes fléchies (de surface) et de travailler sur les lemmes.

Il existe de nombreuses autres façons de pondérer les descripteurs, elles sont par exemple mentionnées dans (Robertson and Jones, 1997; Salton and Buckley, 1987; Takao et al., 2000). A part l'algorithme décrit dans 3.1.3 nous en avons également appliqués d'autres (poids « combined weight » proposé par (Robertson and Jones, 1997), d'autres poids mentionnés dans (Salton and Buckley, 1987)). Pour l'établissement de l'ordre de pertinence des documents par rapport au domaine, il existe également d'autres algorithmes du modèle vectoriel ou bien d'autres modèles (modèle de Jaccard ou de Dice par exemple) qu'il pourrait être intéressant d'appliquer. En se fondant sur le résultat de la validation des documents par l'expert, il serait intéressant de faire une comparaison et une évaluation des pondérations différentes obtenues avec des calculs statistiques.

Il ressort de la validation de l'expert que les documents placés en tête du classement sont principalement dits pertinents. Tandis que les documents hors sujet sont plutôt condensés à la fin du classement. Mais il n'y a pas de frontière stricte qui délimiterait les documents pertinents du reste. Ce qui nous conduit à faire une autre constatation concernant les descripteurs utilisés, car ils tiennent une place très importante dans la représentation du domaine. Ainsi nous avons à tort considéré par exemple les termes « *cycle combiné* » et « *pile à combustible* » comme synonymes absolus de « *cogénération* », ce qui invalide les documents placés à la 24^{eme} et à la 32^{eme} positions. Pour arriver à un jeu de descripteurs performants, plusieurs jeux de descripteurs devraient être testés. Les documents hors sujet apparaissent à partir de la position 82 d'une manière assez ponctuelle, et assez massivement à partir de la position 253 (valeur de proximité du document avec le domaine est située autour de 0,0146). Ce qui montre que les premiers 71 % du nombre total des documents étaient globalement bien classés avec l'algorithme appliqué. D'autre part, le résultat de cette validation confirme notre hypothèse de présence des documents hétérogènes, qui traitent de plusieurs thématiques (facettes) de la cogénération à la fois.

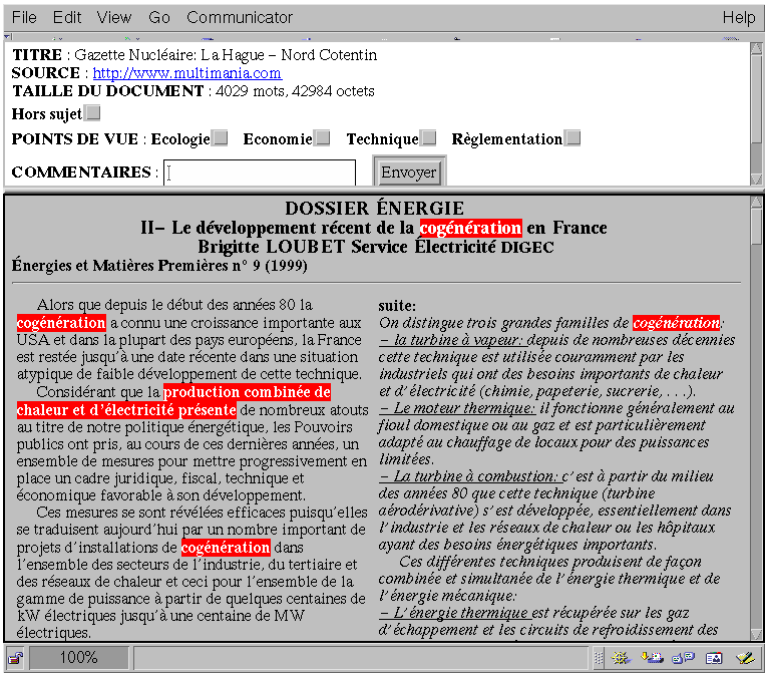
Interface de validation. Les 355 documents ainsi filtrés et classés ont été soumis à la validation de l'expert.

La question de l'ergonomie de l'interface de validation des documents occupe une place importante. Cette interface devra comporter l'information nécessaire à la validation sans l'encombrer pour autant.

Nous avons opté pour une interface simple réalisée sous forme de page `HTML`. Sur la page d'accueil nous décrivons la tâche de validation à réaliser et donnons quelques instructions quant à cette tâche. La page d'accueil permet d'accéder à la liste des documents à valider et à la liste des documents déjà validés (avec les traces de la validation de l'expert). Dans les deux cas, les documents sont rangés dans l'ordre de leur pertinence par rapport au domaine tel que calculé automatiquement. Les deux listes permettent d'accéder à la page de validation de chaque document. Ainsi l'expert peut revenir en arrière s'il change d'avis sur un des documents déjà validés.

L'interface de validation d'un document est présentée dans la figure ci-dessous. L'information suivante, obtenue le plus souvent grâce au repérage automatique, est disponible : titre du document, source du document (site Internet, base de données), taille du document en nombre de mots et d'octets.

Ce que nous demandons à l'expert, c'est d'indiquer son avis sur la pertinence du document par rapport au domaine de la cogénération. Ainsi si le document n'est pas pertinent la case « *Hors sujet* » est à cocher. Nous lui demandons également de sélectionner un des points de vue sur la cogénération : écologie, économie, technique ou réglementation. Le choix des points de vue n'est pas exclusif car nous considérons qu'un document peut être polythématique (et cela correspond à l'état réel des choses). La boîte « *Commentaires* » est prévue pour accueillir toute remarque de l'expert.



5. Perspectives et conclusion

Nous avons présenté les filtrages automatiques mis en oeuvre pour assister et aider la sélection manuelle des documents les plus pertinents par rapport à un domaine de spécialité. Pour certains de ces filtrages une liste des mots-clés du domaine est utilisée. Grâce aux filtrages effectués nous détectons les documents doublons et inclusions, afin de les éliminer, car nous considérons que leur présence fausserait le corpus et le résultat de l'acquisition terminologique. Nous détectons les pages de liens, également dans le but de les éliminer, et ne gardons que les pages de texte jugées plus informatives. Nous effectuons également un calcul de pertinence des documents

restant par rapport au domaine de la cogénération. D'autres traitements automatiques peuvent être appliqués pour obtenir un corpus plus propre et plus homogène.

Nous nous sommes rendu compte par exemple que, malgré le paramétrage des moteurs de recherche quant à la langue de recherche, le corpus contenait des textes écrits en anglais, en néerlandais et en espagnol. Il serait donc intéressant d'appliquer un outil de reconnaissance de la langue du document, qui soit capable de repérer non seulement les documents écrits entièrement dans une langue donnée, mais aussi ceux comportant des passages dans des langues différentes, car ceux-ci fausserait les traitements linguistiques ultérieurs (étiquetage morpho-syntaxique, acquisition terminologique et des relations, etc.).

Le point suivant dans notre méthode de constitution de corpus est le regroupement des documents en sous-corpus homogènes du point de vue de leurs thématiques. La tâche de regroupement des documents en sous-corpus homogènes fait référence aux travaux existants en classification et catégorisation des documents. Selon la distinction souvent faite entre ces deux types de regroupement, la catégorisation correspond à un regroupement supervisé selon des catégories préétablies. Tandis que la classification correspond à un regroupement de documents non supervisé, les classes obtenues étant interprétées a posteriori. Comme nous disposons déjà des mots-clés « spécifiques » qui caractérisent les points de vue sur le domaine de la cogénération, nous pensons catégoriser les documents selon ces listes de mots-clés. Nous pensons effectuer la création de sous-corpus thématiques d'après le même principe que celui utilisé pour le calcul de pertinence des documents par rapport au domaine de la cogénération. Mais cette fois c'est la liste des mots-clés « spécifiques » d'un point de vue donné qui sera utilisée pour représenter le document « idéal » et définir la similarité de ce document « idéal » avec d'autres documents du corpus validés par l'expert. Comme nous l'avons dit plus haut, nous supposons qu'un document peut être multicatégoriel et appartenir à plus d'un point de vue, avec un degré de pertinence plus ou moins élevé. La difficulté reste la même : choisir la bonne approche pour ce calcul et avoir une idée du seuil de pertinence efficace. Pour cela aussi nous utiliserons le résultat de la validation de l'expert. On pourra également comparer le résulat de ces catégorisations avec des classements proposés par d'autres outils, comme par exemple ceux développés par (Grivel and François, 1995; Ferret, 1998; de Chalendar, 2001) que nous avons testés.

Nous avons présenté dans cet article les traitements automatiques mis en oeuvre pour assister et aider la tâche de sélection manuelle des documents pertinents d'un domaine et pour constituer ainsi un corpus de spécialité pour l'acquisition terminologique.

Remerciements

Nous remercions Jean-François Perrot de nous avoir lancé le défi d'appliquer les modèles statistiques à nos données ; les étudiantes Magali Antic et Blandine Jeannin qui nous ont aidées dans l'enrichissement du corpus ; Monique Slodzian, Pierre Zweigenbaum, Yannick Toussaint, Marie Pasquier, Jean-David Sta et tant d'autres pour le soutien et les conseils de lecture ; Clément Boré, expert EDF en cogénération, qui a bien voulu effectuer la validation des textes du corpus.

Références

Adamic L. A. and Adar E. (2000). Frequency of friendship predictors. Technical report, Xerox Palo Alto Research Center. Disponible à `http://www.parc.xerox.com/istl/groups/iea/papers/web10/`. Visité le 13/09/2001.

Amitay E. and Oberlander J. (1997). Conventionsays... In *The Eight ACM International Hypertext Conference Hypertext'97*. Disponible à `http://www.mri.mq.edu.au/~einat/`. Visité le 13/09/2001.

Bachimont B. (2000). Engagement sémantique et engagement ontologique : conception et réalisation d'ontologies en ingénierie des connaissances. In Charlet J., Zacklad M., Kassel G., and Bourigault D. editors, *Ingénierie des connaissances, évolution récentes et nouveaux défis*, chapter 19, pages 305–323. Eyrolles, Collection technique et scientifique des Télécommunications.

Beaudouin V., Fleury S., Habert B., Illouz G., Licoppe C., and Pasquier M. (2001). TyPWeb : décrire la Toile pour mieux comprendre les parcours. In *CIUST'01 (Colloque International sur les Usages et les Services des Télécommunications) – e-Usages*, pages 492–503, Paris. France Télécom R&D, ENST–Paris, IREST, ADERA.

Bourigault D. and Slodzian M. (1999). Pour une terminologie textuelle. Nantes. Tutoriel à TIA 1999.

Dachelet R. (1990). Etat de l'art de la recherche en informatique documentaire. la représentation des documents et l'accès à l'information. Technical report, INRIA Rocquencourt.

de Chalendar G. (2001). *SVETLAN', un système de structuration du lexique guidé par la détermination automatique du contexte thématique*. PhD thesis, Université Paris XI, Orsay.

Ferret O. (1998). Une segmentation thématique fondée sur la cohésion lexicale. In *TALN*, pages 32–41, Paris.

Grabar N. and Berland S. (2001). Construire un corpus web pour l'acquisition terminologique. In *Terminologie et intelligence artificielle*, pages 44–54, Nancy.

Grivel L. and François C. (1995). *Une station de travail pour classer, cartographier et analyser l'information bibliographique dans une perspective de veille scientifique et technique*, pages 81–113. Presses universitaires de Rennes.

Habert B., Illouz G., Lafon P., Fleury S., Folch H., Heiden S., and Prévost S. (2000). Profilage de textes : cadre de travail et expérience. In Rajman M. editor, *5èmes Journées d'Analyse des Données Textuelles (JADT 2000)*, Lausanne.

Jacquemin C. (1997). Guessing morphology from terms and corpora. In *ACM SIGIR*.

Nauer E. (2001). *Principes de conception de systèmes hypertextes pour la fouille de données bibliographiques multibases*. Thèse en informatique, Université Henri Poincaré Nancy 1.

Robertson S. E. and Jones K. S. (1997). Simple, proven approach to text retrieval. Technical report, Department of information science of City university, Computer laboratory of university of Cambridge.

Salton G. and Buckley C. (1987). Term weighting approaches in automatic text retrieval. Technical report, Department of computer science of Cornell university.

Takao S., Ogata J., and Ariki Y. (2000). Study on new term weighting method and new vector space model based on word space in spoken document retrieval. In *RIAO*, pages 116–131.