Ingénierie linguistique et Fouille de textes

IBEKWE-SANJUAN Fidelia
ERSICOM
Université Jean – Moulin, Lyon 3
ibekwe@univ-lyon3.fr

SANJUAN Eric
IUT STID – LITA (EA 3097)
Université de Metz
eric.sanjuan@univ-metz.fr

Résumé.
Nous présentons le système TermWatch pour l'analyse des données textuelles qui allie ingénierie linguistique et techniques d'aggrégation et de visualisation. L'accent est mis sur la refonte des modules d'ingénierie linguistique dans TermWatch. Celle-ci a nécessité le choix d'outils performants pour permettre une amélioration dans l'extraction de termes et l'identification de relations sémantiques pour enrichir celles syntaxiques sur lesquelles s'appuient la classification. Les résultats obtenus sont très intuitifs et montrent la manière dont se structurent les thématiques de recherche dans un domaine donné ainsi que les interconnections entre des domaines de recherche jusqu'alors séparés.

Mots-clés.
Fouille de textes, Cartographie thématique, Ingénierie linguistique, Variations terminologiques.

1. Introduction

Les termes "knowledge discovery in databases" (KDD) et "data mining" (DM) désignent à l'utilisation des techniques d'analyse des données pour traiter des données structurées résidant dans des mémoires électroniques. Selon Fayyad (1997), KDD a pour objet d'extraire des informations utiles des bases de données et la fouille de données (FD) n'est qu'une étape dans ce processus itératif et intéractif[1]. Plus précisément, un bon système de KDD doit découvrir des connaissances implicites, potentiellement utiles et jusqu'alors inconnues à partir de cette masse de données. Selon cette définition, les buts du KDD apparaissent comme étant très larges et recouvrent de nombreuses études dans des domaines variés qui traitent de l'analyse des données : gestion des bases de données, reconnaissance des formes, intelligence artificielle, techniques de visualization des donnes, réseaux de neurones, classification et aggrégation des données, techniques d'apprentissage, etc.

Le défi posé par ce nouveau paradigm de recherche étant le traitement des données massives, toujours croissantes qui ont nettement surpassé des capacités de traitement humain et face auxquelles il convient d'inventer ou d'adapter des méthodes existantes d'analyse des données. C'est un fait admis que tout système de KDD doit inclure l'ensemble du processus qui commence par l'entrepôt de données, la sélection et le néttoyage des données, l'acheminement vers une technique de fouille, l'application de la technique de fouille, le choix d'un modèle de données, et enfin la validation et l'interprétation des résultats. Au coeur des techniques de fouille se trouvent des méthodes d'aggrégagtion ou de classification automatique dont l'objectif est de réduire les données brutes en des aggrégats facilement visualisables et interprétables. Ainsi, la réduction des données est essentielle dans tout système de fouille.

La découverte des connaissances dans des textes (KDT) et la fouille de textes (FT) sont le pendant s'appliquant aux données non structurées. Par analogie avec la KDD, la KDT serait la découverte des informations non triviales, implicites, inconnues jusqu'alors et potentiellement utiles

---

1 *"KDD is concerned with extracting useful information from databases"* **and "***data mining is but a step in this iterative and interactive process".(Fayyad, 1997 : 1)*

(Feldman, 1998 ; Lent *et al.*, 1997). Selon Kodratoff (1999), la KDT est " *la science qui découvre des connaisances dans des textes, où "connaissances" est à prendre dans son acception en KDD, c'est-à-dire que la connaissance extraite doit reposer sur des bases réelles, et va modifier le comportement d'un agent humain ou méchanique*[2]". Dans le même ordre d'idée, Hearst (1999) affirme que l'objectif de la fouille de données est de "*découvrir ou dériver de nouvelles informations à partir des données, de trouver des modèles intéressants des données et de séparer le signal du bruit*[3]."Le processus de fouille de textes est alors similaire à celui de la FD sauf que les types de traitement effectués sont plus adaptés au traitement automatique des textes. C'est ainsi que de nombreuses études ancrées sur l'ingénierie linguistique peuvent accomplir différentes tâches de FT. Selon ce site Internet[4] de vulgarisation sur la FT, les étapes nécessaires pour effectuer une FT sont :

1. Recherche d'informations (à partir du WEB ou de bases de données)
2. Selection des documents pertinents
3. Nettoyage des données : vérification orthographique, retrait mots vides, lemmatisation,
4. Identification des mots pertinents : analyse statistique, analyse semantique, analyse syntaxique ou structurelle,
5. Filtrage : selection des mots les plus pertinents,
6. Truncation[5]
7. Application d'un algorithme de fouille de textes : clustering, règles d'association, classification, extraction d'information (concepts/abstracts)

Les étapes 1 et 2 sont des tâches classiques de recherche d'information. Les étapes 3-5 sont des tâches habituellement effectuées par des systèmes de traitement de l'information au sens large : systèmes visant l'indexation automatique, systèmes linguistiques cherchant à identifier des unités de textes significatives pour diverses application (des syntagmes nominaux par exemple, des relations sémantiques particulières). L'étape 4 marque l'intégration des outils linguistiques en amont des techniques de réduction des données (étape 7). Des tâches jadis accomplies par des systèmes resposant sur l'ingénierie linguistique peuvent désormais revêtir l'appellation de "fouille de textes". Un signe extérieur, bien visible de cette évolution est le vocabulaire employé par des éditeurs commerciaux de solutions de FT[6]. Ils incluent des tâches effectuées par des outils d'ingénierie linguistique dans la fouille : le résumé automatique (*automatic summarization/abstracting*), la catégorisation des textes ou des documents, la reconnaissance d'entités nommées et l'extraction d'information (IE), l'extraction de relations sémantiques ou conceptuelles entre unités textuelles. Ceci témoigne de l'entrée de l'ingénierie linguistique dans un champ habituellement dominé par l'application des techniques statistiques. Cette rentrée s'impose par la nature de l'objet "fouillé". Bien que la FT implique nécessairement l'application d'algorithmes de réduction des données, ceux-ci doivent intervenir dans la phase finale, après extraction ou identification, par des techniques linguistiques, des unités d'informations pertinentes sur laquelle portera la fouille.

S'inscrivant dans cette philosophie de rencontre entre ingénierie linguistique et techniques d'analyse de données, nous avons mis au point un système d'analyse des données textuelles qui accomplit certaines tâches de FT. Ce système vise à cartographier les sujets contenus dans un corpus de textes. De ce fait, ses résultats peuvent être qualifiés de FT dans la mesure où la carte obtenue

---

2"*the science that discovers knowledge in texts, where «knowledge is taken with the meaning used in KDD», that is the knowledge extracted has to be grounded in real world, and will modify the behavior of a human or mechanical agent.*" (Kodratoff, 1999).

3"*goal of data mining is to discover or derive new information from data, finding patterns across datasets, and/or separating signal from noise*" (Hearst, 1999).

4(http://www.inf.ufrgs.br/~wives/english/textmining.html), visité le 8 mars 2004.

5 Nous n'avons pas pu déterminer avec certitude à quoi se réfère cette notion.

6 Voir par exemples les sites de NetOwl (http://www.netowl.com/products.html), SAS TextMiner (www.sas.com/technologies/analytics/ datamining/textminer/), Temis (www.temis-group.com/).

révèle la disposition spatiale des sujets de recherche. Ce système TermWatch inclut en amont tout une série de traitements relevant de l'ingénierie linguistique et visant à sélectionner les untés de traitement. Ce sont les termes du domaine, unités porteuses de sens dans un discours. Après identification des termes, TermWatch recherche des relations syntaxiques et sémantiques entre ces termes. Pour cela, il s'appuie sur des outils linguistiques développés dans la communauté d'ingénierie linguistique. Ce sont ces relations qui vont constituer la base pour la technique d'aggégation que nous avons développée (Ibekwe-SanJuan, 1998). Ainsi, TermWatch ne fait pas appel à la notion d'occurrence ou de co-occurrence. Ses résultats sont des « clusters » représentant des sujets de recherche dont on peut suivre l'évolution dans le temps. Ces « clusters » sont présentés sous forme d'un réseau que l'on peut visualiser et explorer via logiciel de visualisation de graphes, AiSee® (http://www.aisee.com). Des aspects de cette méthodologie ont été publiés ailleurs (Ibekwe-SanJuan et SanJuan, 2004, 2003). Dans cet article, l'accent est mis sur les développements récents apportés à la chaîne de traitements linguistiques, notamment l'identification de nouvelles relations sémantiques qui viendront enrichir les bases du clustering et par conséquence, les résultats du système. Le travail d'ingénierie linguistique a été achevé pour l'anglais.

Dans la suite de cet article nous décrivons le processus d'ingénierie linguistique permettant d'id les unités terminologiques du corpus et les relations utilisées pour la classification. Ensuite nous décrivons brièvement la technique d'aggrégation et son application à un corpus de textes sur la recherche d'information. La dernière section est consacrée à l'analyse des résultats.

Nous avons constitué un corpus répondant à un objectif de FT. En sélectionnant 16 revues scientifiques réputées de langue anglaise publiant des articles dans le domaine de la recherche d'Information (Information Retrieval – IR), nous avons extrait les titres et résumés pour la période 1997-2003 à partir de la base de données PASCAL (INIST-CNRS). Cela constituait 3355 textes courts totalisant 455 000 mots. Notez que les résumés sont bien d'auteurs et non des résumés documentaires.

2. Ingénierie linguistique

C'est la première phase de tout le procesus menant à la cartographie des thèmes. Etant donné que l'objectif de notre système est de cartographier les sujets de recherche dans le corpus et non de bâtir une représentation linguistique de chaque phrase, il n'est pas nécessaire de mettre en oeuvre une analyse linguistique poussée, allant par exemple jusqu'à une représentation syntaxique ou sémantique des énoncés. Nous nous limitons ainsi à une analyse de surface, partielle des énoncés, se focalisant essentiellement sur les syntagmes nominaux qui renferment la majorité des termes. La chaîne de traitement linguistique fait appel à deux outils libres, distribués par la communauté d'ingénierie linguistique : un détecteur d'acronymes développé par des chercheurs à l'University de Nevada (USA) et un étiquêteur grammatical (LTPOS) développé à l'Université d'Edinburgh. LTPOS nous fournit les étiquettes morphologiques sur lesquelles nous projetons nos règles d'extraction des termes.

2.1 Recherche d'acronymes

Avant tout traitement, nous recherchons à identifier les candidats terms acronymes (les abréviations des termes sous forme développée). L'identification des acronymes est basée sur un algorithme développé à l'Unversité de Nevada. Ce programme a identifié 711 acronymes de notre corpus. Après élimination des doublons, nous avons retenu 655 acronymes dont sept étaient incomplets et 17 étaient erronés (le programme a choisi le mauvais candidat terme pour l'acronyme). A titre d'exemple, il a identifié la séquence: *"of the SQL92 query language*" comme forme développée de l'acronyme "*OQL"*. Une recherche dans le texte a montré que la bonne forme est "*object query language*". En revanche, il a correctement identifié des acronymes tels que "*NCSTRL*" pour "*Networked Computer Science Technical Reference Library*".
La précision de ce programme est très satisfaisante puisque 631 sur 655 (96%) variantes d'acronym étaient correctement identifiées. Les acronymes indiquent une équivalence entre termes : l'acronyme

et sa forme développée renvoient au même concept. Leur identification à ce stade initial évite la dispersion des différentes formes d'un même concept dans les étapes ultérieures de traitement.

2.2. Extraction de multitermes

Cette étape a subi une refonte dans la mesure où nous avons changé d'outil linguistique permettant d'effectuer cette extraction. Nous utilisions jsuqu'ici le système INTEX (Silbertzein, 1993) mais ce système étant basé sur une approche non-déterministe de l'analyse linguistique, il ne choisit pas une étiquette pour chacun des mots analysés. Ainsi, nous nous retrouvions avec beaucoup d'ambiguïtés ('nom/verbe' très abondantes en anglais), ce qui avait pour conséquence de nuire à la précision du résultat de l'extraction. Nous avons alors opté pour un étiquêteur grammatical, déterministe par définition puisqu'il choisit une étiquette pour chaque mot. Nous avons choisi le LTPOS développé à l'Université d'Edinburgh. Celui-ci associe à chaque mot d'un texte sa partie du discours (nom, verbe, préposition, adjectif, adverbe, etc) et est couplé à un détecteur de syntagmes (LTChunker) qui balise des structures nominales simples (sans attachement prépositionnel). Le LTPOS a été entraîné sur un très large corpus en anglais et atteint un niveau de performance honorable (entre 96-98% d'étiquettes correctes selon ses auteurs). Nous donnons quelques exemples de séquences étiquetées dans laquelle les structures nominales simples ont été délimitées :

1. [[the_DT US_NNP Digital_NNP Millennium_NNP Copyright_NNP Act_NNP]]
2. [[binary-independent_JJ and_CC inverse-document-frequency_JJ weightings_NNS]]
3. [[automatic_JJ and_CC manual_JJ indexing_VBG techniques_NNS]]

Le symbole "[[ ]]" délimite les syntagmes nominaux simples (SN) reconnus par LTChunker. LTChunker ne reconnaît pas les structures nominales complexes qui peuvent également être des termes. Afin d'extraire ces séquences, nous avons écrit une dizaine de règles qui au regard du contexte des étiquettes produites par LTChunker va ou non reconnaître les structures avoisinantes au SN. A titre d'exemple, la règle suivante :

4. <modifier> <N>* of <modifier> <N>* (<prep1>|<verb>) <modifier> <N>*
où :
< > indique des symboles non terminaux
<modifier> = un déterminant et/ou un adjectif portant soit l'étiquette DT | JJ
<N>= un nom avec une de ces étiquettes NN | NNP | NNPS | NNS
<prep1> = toutes les prépositions sauf "*of*
* = opérateur de Kleene (la boucle)

stipule que si l'on trouve un SN suivi par la préposition "*of*" suivie d'un autre SN qui à son tour est suivie d'une préposition qui n'est pas le "of" et qui se termine par une structure en SN, alors extraire les sous-séquences suivantes :

4a. - <modifier> <N>* of <modifier> <N>*
4b. - <modifier> <N>*

Cette régle exprime tout simplement le rôle prépondérant joué par la préposition "*of*" dans la composition des termes anglais et exprime l'interdiction de segementer la séquence à cette endroit. Appliquée à la sséquence suivante, elle produit les candidats termes indiqués ci-après :

[[the_DT meaningful_JJ processing_NN]] of_IN [[information_NN]] in_IN [[relation_NN]] to_TO [[two_CD systems_NNS]] of_IN [[information_NN processing_NN]] :

a- [[the_DT meaningful_JJ processing_NN]] of_IN [[information_NN]]
b- [[relation_NN]]

c- [[two_CD systems_NNS]] of_IN [[information_NN processing_NN]]

A l'aide de ces règles, 53200 termes candidats ont été extraits du corpus IR. Il est fréquent que les termes extraits n'aient qu'une unique occurrence dans tout le corpus.

2.3. Identification de relations syntaxiques et sémantiques

Les termes d'un même domaine partagent plusieurs relations qui peuvent être de nature morphologique, structurelles ou sémantiques. Les relations morphologiques peuvent concerner les phénomènes suivants : l'usage des formes abrégées d'un même terme (*WWW / world wide web*), la composition (*online web access / on line web access / on-line web access*), la variation orthographique, (*b-cell chronic lymphocytic leukaemia ; b-cell chronic lymphocytic leukemia*), la morphologie dérivationnelle (*tumor promoter ; tumor promotion*). La variation structurelle ou syntaxique concerne des opérations qui affectent la longueur, la structure d'un terme ou les éléments lexicaux dans le terme. Ainsi, l'inversion par l'usage de prépositions (*information retrieval ↔ retrieval of information*), l'expansion d'un même par l'ajout de nouveaux éléments ("*access to information → equal access to information*) ou la substitution d'un élément dans le terme (*bibliographic hypertext system ↔ bibliographic retrieval system*). Les relations sémantiques peuvent être réparties en deux catégories larges : relations hiérarchiques (hyperonyme / hyponyme, meronyme) et relations transversales (co-hyponymie, synonymie, l'association). Etant donné que les acronymes ont déjà été identifiés en amont, nous allons nous intéresser ici aux deux autres catégories de relations : syntaxiques et sémantiques.

2.3.1 Relations syntaxiques

C'est sur cette catégorie de relations que s'appuie la classification dans TermWatch. Il s'agit ici d'en rappeler brièvement les principes. Les relations syntaxiques que sont subdivisées en deux catégories COMP et CLAS.

- *Variations de modification (COMP).* Le premier ensemble de variations, appelé COMP, comprend deux expansions et une substitution de modifieurs. Celles-ci sont :

  1- l'expansion gauche (Exp_G) qui permet d'identifier comme variantes ces deux termes : *academic library → uk academic library*.

  2- l'insertion (Ins) qui concerne l'ajout de nouveaux éléments à l'intérieur d'un terme et à une même position : *academic library→ academic biology library*.

  3- la substitution de modifieur (Sub_M) qui consiste à changer un unique élément nominal modifieur : "*academic library users ↔ public library users* .

- *Variations de centre (CLAS)*

  Ce deuxième sous-ensemble de relations appelé CLAS concerne les opérations d'ajout ou de substitution qui entraînent le changement de l'élément centre dans un terme. On retrouve les deux types d'opérations précédemment cités. Dans les deux premières relations de variation qui suivent, notées Exp_D et Exp_GD, l'ancien centre devient modifieur dans le nouveau terme :

  4- l'expansion droite (Exp_D) : *academic library → future of academic library*

  5- l'expansion gauche-droite (Exp_GD) : *academic library → canadian academic library privilege*

  6- la substitution de centre (Sub_C) : *directors of academic library ↔ future of academic library.*

Ce sont actuellement ces six relations qui sont considérées pour la classification. Cependant, elles rencontrent une limite liée à la nature même de leur définition. Etant donné qu'elles se fondent toutes sur une base lexicale – l'inclusion lexicale, il arrive que certaines variantes ne permettent pas de dégager une thématique dans le domaine. Alors que les relations d'expansion (L-Exp, Ins, Exp-D, Exp-GD) génèrent grosso modo des relations de type hiérarchiques (voir figure 1 ci-après), la signification des relations engendrée par les substitutions n'est pas toujours apparente.

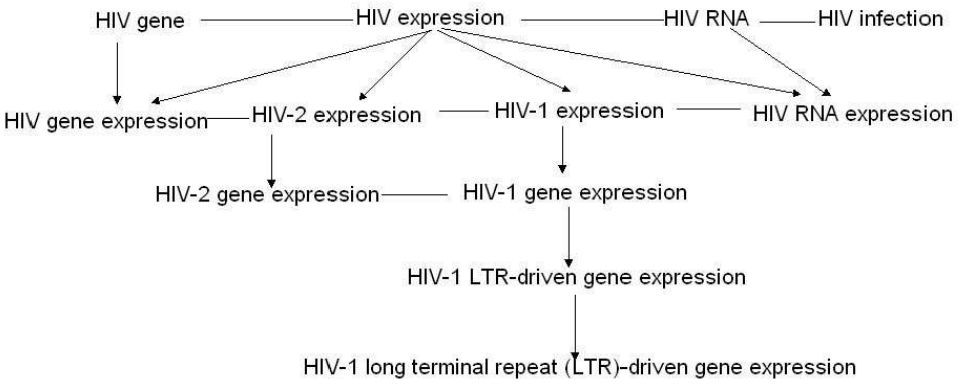

Figure 1. Structuration des termes avec les relations syntaxiques[7].

Parce que les variations syntaxiques ne permettent pas d'expliciter la relation exacte entre deux termes et parce que certaines de ces opératuions de variations peuvent engendrer du bruit, il est devenu nécessiare d'intégrer des relations sémantiques précises entre termes qui viendront enrichir celles syntaxiques déjà implantées. Cela nous fournit par la même occasion un moyen de filtrer des relations syntaxiques bruyantes telle que la substitution.

2.3.2 Relations sémantiques

Elles peuvent être identifiées par deux manières : en s'appuyant sur des traces linguistiques qui signalent ces relations en corpus (approche dite de « marqueurs ») ou en se référant à une ressource extérieure qui a déjà encodé ces relations sémantiques entre termes du domaine. Souvent les deux types d'approches sont combinées pour une identification optimale des relations sémantiques. Séguéla & Aussenac-Gilles (1999) proposent un environnement d'acquisition et de validations de différentes relations sémantiques.
Dansr l'approche de marqueurs, nous nous sommes penchés essentiellement sur des marqueurs de relations hyperonyme / hyponyme (Hearst, 1992, Morin & Jacquemin, 2003) et de synonymie (Suarez & Cabré, 2002).

- *Marqueurs d'hyperonymie / hyponymie*

Il s'agit de marques discursives laissées par les auteurs des textes dans leur discours et qui signalent les relations hiérarchiques de type générique / spécifique (« est-un » au niveau conceptuel). Les plus cités de ces marqueurs sont :

H1 : *such* NP_0 *as* NP_1 (<cc>)+ NP_2 …, (<cc>)* NP_n
H2 : NP_0 *such as* NP_1 (<cc>)+ NP_2 …, (<cc>)* NP_n
H3 : NP_0 (<cc> | NP_1) *like* (NP_2 | <cc>)+ NP_n
H4 : NP_0 (<cc> | NP_1) (*particularly* | *especially* | *in particular*) (NP_2 | <cc>)* NP_n
H5 : NP_0 (<cc> | NP_1) *including* (<cc> | (NP_2 | <cc>*)) NP_n
H6 : NP_0 ([;| :]) NP_1 [parenthesis] (,) NP_2* (<cc>) NP_n
H7 : NP_0 (<cc> | NP_1) [,] (<&>) *other* NP_2 …, (<cc>)* NP_n
H8 : NP1 (, | namely) NP_2 …, (<cc>)* NP_n

où NP est soit un syntagme nominal simple soit complexe avec un attachement prépositionnel, “cc” is est un élément de coordinating (*and, or, virgule*), “*” est l'opérateur de Kleene. Le marqueur H2 détectera la séquence suivante :

7 Ces exemples proviennent d'un autre corpus auquel nous avons appliqué notre méthode, il s'agit du corpus GENIA.

[HYPER] challenging requirements in non-traditional applications such as [HYPO] geographic information systems ([HYPO] GISs), [HYPO] computer-aided design ([HYPO] CAD), and [HYPO] multimedia databases.

Le symbole [HYPER] est inséré avant le candidat terme hyperonyme et le symbole [HYPO] devant chaque candidat terme hyperonyme. Ceci résulte en cinq paires de termes partageant une relation hyperonyme/hyponyme. Chaque paire de termes hyponymes est reliée en tant que "co-hyponymes" d'un même terme. On peut schématiser ces relations par la figure suivante :

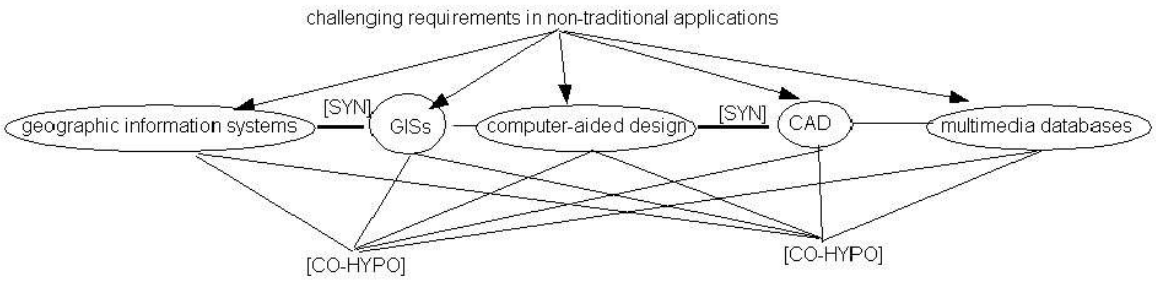

Figure 2. Relations d'hyperonyme/hyponyme acquises par marqueurs.

Les termes "*geographic information systems*" et "*GISs*" sont dans une relation de synonymie ([SYN]) puisqu'il s'agit de variantes acronymes, la même chose vaut pour "*computer-aided design*" et "*CAD*". Ces relations se superposent à celle de co-hyponymie ([CO-HYPO]) que partagent chacun de ces hyponymes avec les quatre autres. Nous avons écrit des règles pour extraire des variantes hyperonyme/hyponyme directement du corpus. Au total, ces patrons ont permis d'identifier 306 séquences contenant des termes variants. Une vérification manuelle de ces termes montre que cette approche atteint 77% de précision sur ce corpus, ce qui est une performance honorable pour une méthode peu coûteuse.

- *Marqueurs de synonymie*

De la même manière que pour les relation hyperonym/hyponyme, les auteurs laissent également des traces entre des termes synonymiques dans leur écrit. Les marqueurs les plus cités (Suarez & Cabré, 2002) sont :

S1 : NP_1 (also | equally | now) *called* (now, also equally) NP_2
S2 : NP_1 (also | equally | now) *known as* NP_2
S3 : NP_1 (also | equally | now) *named* NP_2
S4 : NP_1 (also | equally | now) *viewed as* NP_2
S5 : NP_1 (also | equally | now) *termed as* NP_2
S6 : NP_1 *termed* (for | as) NP_2
S7 : NP_1 *referred to as* NP_2 (<cc> | NP_3)
S8 : NP_1 (<have> (also | equally | now) |<be>) defined as NP_2

S9 : NP^1 <be> no other than NP^2

où "(<MOT>)" est tout chaîne de caratères contenue entre parenthèses. Nous avons écrit des règles pour gérer les différents cas de figure (la possibilité d'avoir des adverbes comme "*now, equally*" avant ou après la forme verbale). Une recherche de ces patrons dans le corpus ont permis d'idientifier 107 séquences contenant des termes synonymiques. Un exemple est fourni ci-après. Le symbole [SYN] est placé devant les termes synonymiques qui sont soulignés.

A new approach for supporting reactive capability is described in the context of an [SYN] advanced object-oriented database system ***called*** *[SYN] ADOME-II.*

On acquiert la relation de synonymie entre "*advanced object-oriented database system*" et "*ADOME-II*". Les synonymes acquis par ces marqueurs ont un bon taux de précision dans le corpus IR, autour de 91%.
Notez que ces relations ont été acquises uniquement grâce aux marqueurs qu'on peut relever avec des patrons morpho-syntaxiques de surface, sans recours à une ressource extérieure. A l'inverse des relations syntaxiques, les termes mis en relation ici n'ont pas besoin de partager des éléments en commun et la relation acquise est explicitée grâce au type de marqueur utilisé pour les trouver.

- *Utilisation d'une ressource extérieure : taxonomie lexicale de WordNet*

Il est néanmoins nécessaire de compléter l'approche de marqueurs par une ressource séantique extérieure. En effet, la première ne permet pas d'identifier des termes sémantiquement proches mais qui ne sont pas explicitement reliés par les marqueurs retenus. Les ressources les plus utilisées pour l'acquisition de relations sémantiques sont des dictionnaires des domaines concernés, une taxonomie ou une ontologie. WordNet (Fellbaum, 1998) est une taxonomie lexicale qui structure les mots de la langue générale en "classes d'équivalence" appelées "synsets". Deux mots sont dans un même synset s'ils partagent un sens en commun. Un mot polysémique sera donc dans plusieurs synsets, avec un maximum de six synsets correspondants à six sens différents. Outre la relation transversale de synonymie, WordNet contient également des relations hiérarchiques de type hyperonyme/hyponyme puisque les synsets sont structurés. WordNet vient avec des programmes Perl qui permettent de calculer le degré de similarité entre deux mots selon divers indices (Purandare & Pedersen, 2004). Notre objectif est de déterminer les termes proches sémantiquement. Or, WordNet ne permet que de trouver les mots proches. Il nous faut ensuite rechercher où ces mots apparaissent dans notre liste de termes afin de les associer. A titire d'exemple, partant de la similarité entre "*Internet*" et "*Network*"donnée par WordNet, nous avons associé les deux termes "*Internet literacy » et « Network literacy ».* Le tableau ci-dessous donne d'autres exemples de termes proches associés par la même méthode.

Coeff. Similarité	***Terme1***	***Terme2***
29590099.43	Lookup tool	Search tool
29590099.43	Internet literarcy	Network literacy
29590099.43	Partition method	Segmentation method
29590099.43	Theme relationship	Topic relationship
29590099.43	Scholar use	Student use

Table 1. Les termes proches calculés à partir de WordNet.

La première colonne donne l'indice de similarité de deux termes. Actuellement, cette recherche a été faite pour les termes binaires en relation de substitution. Cela permet de sélectionner parmi les variantes syntaxiques de ce type, uniquement des termes proches du point de vue sémantique. Nous allons l'étendre prochainement aux autres types de variations et l'intégrer dans l'algorithme de classification. Cependant, les limites de ce type d'approche réside dans le degré de couverture des mots du corpus par la ressource extérieure et dans la multiplicité de sens (ambiguïté) retenue. Sur un corpus très technique, il est à craindre que cette couverture soit insuffisante.

La figure 3 ci-après résume la chaîne de traitements linguistiques opérés sur les textes avant classification et cartographie.

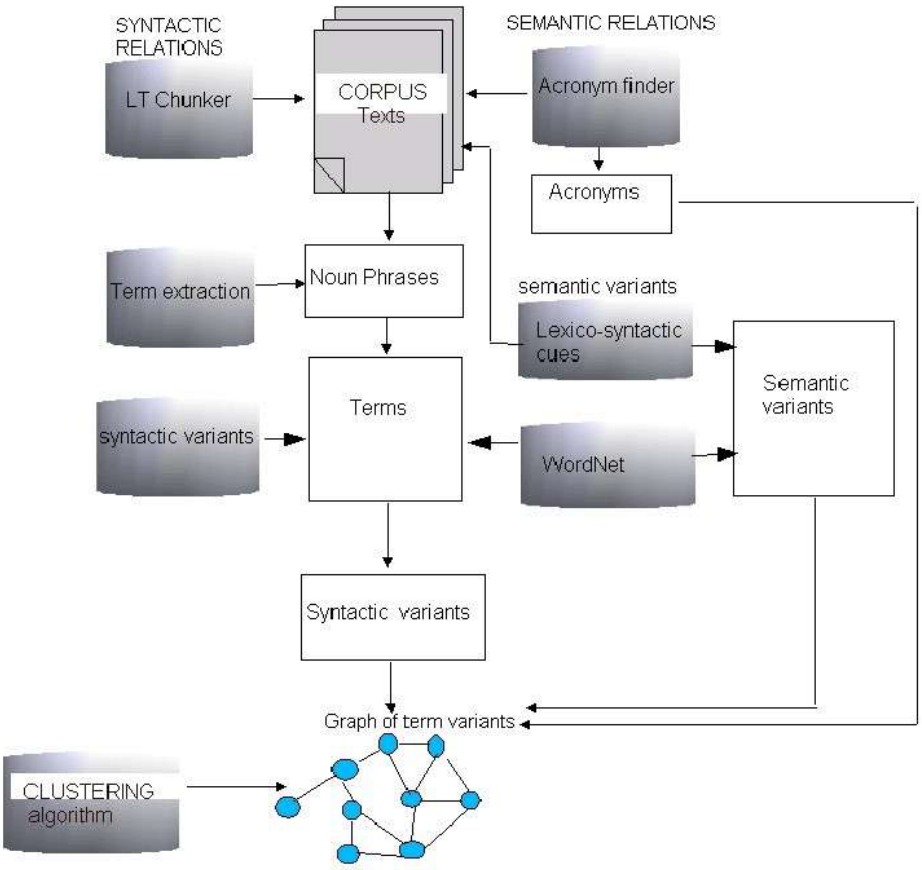

Figure 3. Le processus d'ingénierie linguistique mis en oeuvre dans TermWatch.

3. Réduction des graphes de termes variants.

Dans la présente expérience, les relations sémantiques n'ont pas été inclues dans la classification car elles sont encore en cours d'étude. Ainsi, nous montrons que les résultats obtenus avec les seules relations de variation syntaxiques (§2.3.1). Nous en rappelons succintement le principe ici.

3.1. Algorithme de classification non supervisée

L'algorithme Classification by Preferential Clustered Link (CPCL) implanté dans TermWatch procède en plusieurs étapes. Au prélable, l'utilisateur dissocie les types de variations en deux catégories selon qu'ils provoquent ou non un changement de centre (sujet) :

- COMP qui comprend les relations de modification (Exp_G, Ins, Sub_M),
- CLAS qui comprend les relations de changement de centre (Exp_D, Exp_GD, Sub_C).

On utilise la première catégorie COMP pour former des composantes connexes de termes, au sens de la théorie des graphe et la deuxième catégorie CLAS pour lier ces composantes et former des thématiques. Ces composantes représente le premier niveau de classification. Leur taille peut varier de 2 à 1 000 et leur regroupement nécessite la mise en oeuvre d'un algorithme spécifique qui ne néglige pas les plus petites composantes connexes.

Pour cela il utilise un indice de dissimilarité *d* qui à tout couple de composantes connexes, associe la somme des proportions de liens de variation des relations dans CLAS, entre ces deux composantes connexes. Plus formellement, *d* est une application dans [0,1] définie pour tout couple (*i, j*) de composantes connexes de la manière suivante :

i) $d(i,j) = 1$ si pour tout r dans $\{1,\ldots,k\}$, $N_r(i,j) = 0$ et $d(i,j) = 0$ si i = j ;

ii) $d(i,j) = 1/\sum_{r=1}^{k} \frac{N_r(i,j)}{|R_r|}$ où R_1 ... R_k désignent les relations dans CLAS et $N_r(i,j)$ est le nombre de liens dans R_r entre *i* and *j*.

Utiliser cette dissimilarité *d* pour agglomérer les composantes multi-termes en classes par classification ascendante hiérarchique revient à approcher *d* par une ultramétrique *u*, de choisir un niveau significatif du dendrogramme et de visualiser le graphe obtenu en agglomérant les composantes dans une même classe. Il est bien connu que la meilleure approximation inférieure serait l'ultramétrique associée à la classification par lien simple (CLS). Mais l'objectif n'étant pas la meilleure approximation numérique de *d* mais celle qui préserve la structure du réseau de composantes et évite l'effet de chaîne propre à la CLS. TermWatch utilise alors un critère d'agglomération local qui consiste à agglomérer deux classes seulement si la dissimilarité entre elles est plus faible qu'avec toute autre classe dans leur voisinage.

Cette ultramétrique particulière a la propriété de dégager des classifications avec un nombre de classes non triviales (non réduites à un singleton) bien plus important que la CLS, tout en partagent la majorité de ses propriétés mathématiques, dont l'unicité.

La technique de classification implémentée, la CPCL (Classification by Preferential Clustered Link) (Ibekwe-SanJuan, 1998) est une variante de la classification ascendante hiérarchique (CLS). La CPCL a été élaborée pour s'adapter spécialement à la réduction d'un graphe de termes reliés par un nombre illimité de relations linguistiques.

L'ensemble du système fonctionne dans une approche de «classification non supervisée » et permet de réaliser une analyse thématique de l'information traitée. TermWatch n'a théoriquement pas de limite sur la taille initiale du graphe (sauf celle imposée par la performance des machines). La figure 4 ci-après résume l'architecture du système.

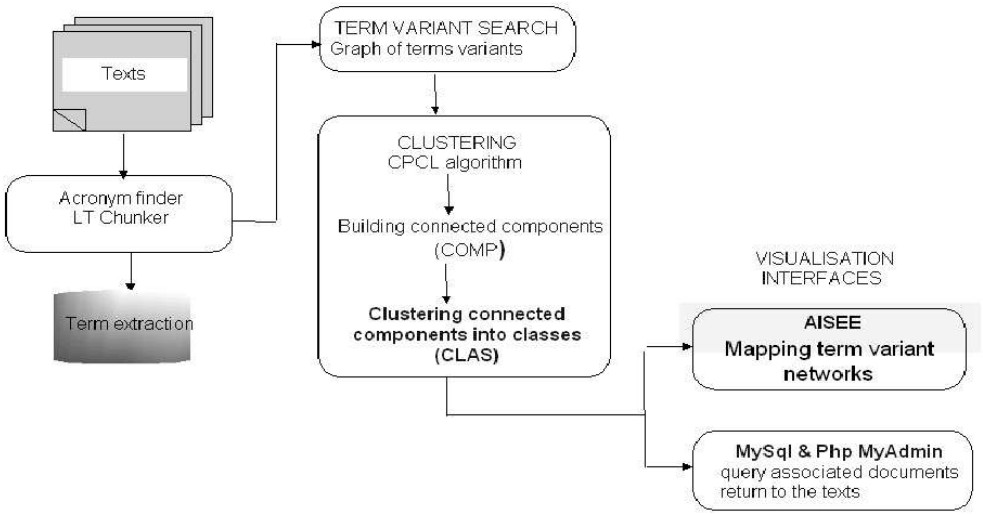

figure 4 : L'archtitecture du système TermWatch.

L'algorithme de classification de TermWatch décrit ci-dessus a été itéré deux fois sur ce corpus. A cette étape 6 849 termes regroupés en 4 252 composantes connexes ont été classées en 397 classes. Les classes sont automatiquement libellées par le terme ayant la plus forte activité de variation. Les libellés des classe permettent une première approximation de leur contenu. Les classes obtenues sont de tailles très variables. Le nombre de termes devient ainsi un bon indicateur de l'importance de la thématique tandis que le nombre de documents ayant au moins un terme dans la classe renseigne sur la transversalité de la thématique dans le corpus. En outre, pour chacun de ces thèmes il est possible d'extraire les documents ou les textes qui leur sont le plus associés et de faire ainsi correspondre les acteurs humains (auteurs) et institutionnels (laboratoires, entreprises, pays) qui sont responsables de leur développement.

3.2 Visualisation et analyse de la carte thématique

L'interface *AiSee* cherche une meilleure approximation planaire du graphe ainsi décrit avec les caractéristiques suivantes :

- les arêtes sont droites mais de longueur variable,

- le nombre de croisements est minimisé.

Il en résulte que les sommets de l'image obtenue peuvent montrer soit les classes soit les composantes connexes.
L'analyse de l'ensemble des résultats sur le corpus IR a déjà fait l'objet d'une publication (Ibekwe-SanJuan & SanJuan, 2004). Ici, nous nous focaliserons sur un sous réseau de thématique identifié : le sous-réseau construit autour de la thématique de la théorie floue (*Rough sets*) afin de comprendre ses liens avec de la thématique de recherche d'information (*information retrieval / IR*).

3.3. Lien entre les thématiques « *Rough Set* » et « *Information retrieval* »

Surprenant et sans-doute plus intéressant d'un point de vue fouille de textes est la révélation de la classe *Rough Set* liée à la classe *IR* qui est centrale dans la carte globale. Pour comprendre ce positionnement, il est nécessaire de retourner au contenu des textes. Ce retour est effectué au moyen d'une Base de données SQL interfacée sous Php MyAdmin. Le tableau suivant donne les titres des sept documents ayant le plus de termes dans la classe *Rough Set*.

Rang	***Titre***	***Année***	***Nombre***
1	Validation of authentic reasoning expert systems	1999	7
2	Double-faced rough sets and rough communication	2002	6
3	Canonical forms of fuzzy truthoods by meta-theory based upon modal logic	2001	6
4	On axiomatic characterizations of crisp approximation operators	2000	6
5	alpha -RST: a generalization of rough set theory	2000	5
6	Application of rough sets to information retrieval	1998	5
7	Parallel fuzzy inference based on level sets and generalized means	1997	5

Tableau 2. Titres des documents les plus fortement associé à la classe *Rough Set*

Sur ces sept documents, cinq traitent directement des *Rough Sets*. Il s'agit des documents 1, 2, 4, 5 et 6. Cela n'était pas a priori évident du fait que cette classe contient 232 termes et qu'elle a été constituée sans aucune information statistique sur la présence absence des termes dans les documents.

Les graphes générés par TermWatch sont codés de manière à ce que chaque sommet qui représente une classe puisse être éclaté en composantes de manière interactive avec l'interface AiSee. Celles-ci sont alors drapéees sur un même fond de couleur et positionnées en conséquence. La figure 5 ci-après montre le développement des classes *Rough Set* et *Information Retrieval*. On remarque ainsi que la classe Rough Set est constituée de thèmes liés à la théorie des ensembles et de ses extensions: R*ough Set*, *Fuzzy Set* and *Set Theory*. La mise en avant de cette classe par TermWatch est une bonne illustration du type de thématique émergeante que ce système peut traquer. En effet, la théorie des *Rough sets*, initialement conçue pour des applications telles que les télécommunications et les systèmes experts, comme l'illustre le titre des deux documents les plus fortement associés à cette classe, a récemment été appliqué à l'IR comme le signale le titre du sixième document. C'est d'ailleurs le terme *Rough Set Theory to Information Retrieval* extrait de ce document qui est à l'origine du lien entre les deux classes *Rough Set* et *IR*. D'autre part cette théorie admet désormais une extension "*Fuzzy*" introduite sous le nom de *alpha-RST*, nom qui est présent dans le titre du cinquième document à l'origine des variantes qui ont conduit à l'immersion dans une même classe des thèmes "*Fuzzy*" et "*Rough*". On retrouve ainsi l'association entre les termes *Rough Set* et *Set Theory* signalée par TermWatch.

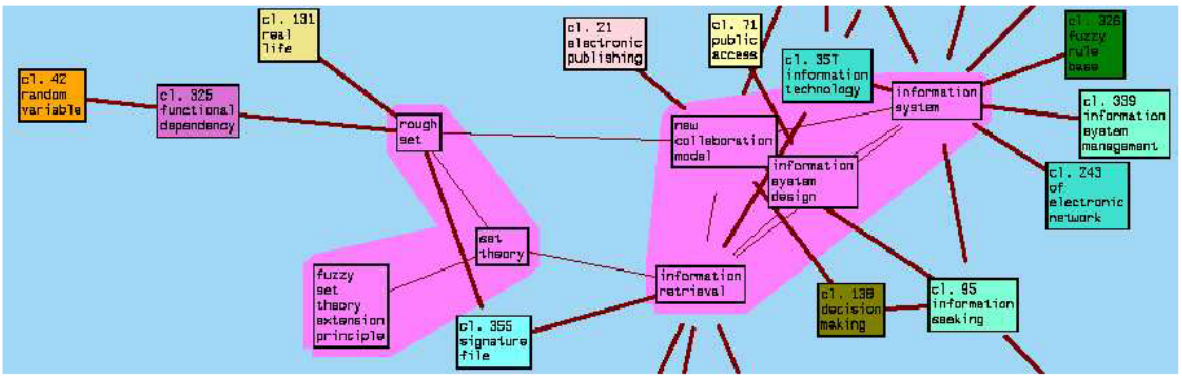

Figure 5 : Développement des classes *Rough Set* et *Information retrieval* (*IR*).

TermWatch a mis en évidence ce vers quoi les concepteurs de cette théorie voudraient tendre, à savoir des applications à la « vie réelle » telles que l'IR. Cet objectif influe directement sur la terminologie qu'ils emploient ce qui explique le positionnement différent de cette même classe par TermWatch. D'autres analyses notamment chronologiques sont en cours pour déterminer plus précisément l'évolution des thématiques sur la période couverte par le corpus. Pour suivre l'évolution des réseaux de termes variants, il est possible d'ajouter ou retirer les liens entre termes qui n'apparaissent qu'à une période donnée (intervalles de dates préalablement choisis par l'utilisateur).

4. Conclusion

TermWatch est un système d'analyse et de classification non supervisée de données textuelles fondé sur la linguistique. TermWatch a permis de dégager de manière automatique et sans ressources sémantiques extérieures, le sous-réseau de thématiques transversales dans le corpus et a permis d'identifier une thématique émergeante, la classe *Rough Set*, qu'il positionne en fonction de sa terminologie. Le caractère émergent de cette thématique a pu être vérifié par un retour au contenu des textes. Le système allie les récentes avancées en ingénierie linguistique, notamment en terminologie computationnelle, et en techniques de classification et de visualisation. La notion d'association entre termes est entièrement fondée sur les relations linguistiques que peuvent partager les termes. Cela représente une intéressante alternative à l'utilisation de la co-occurrence comme critère d'association, tout particulièrement lorsqu'il s'agit de faire émerger une information rare.

Références

Dowdall J., Rinaldi F., Ibekwe-SanJuan F., SanJuan E. *Complex structuring of term variants for Question Answering. Workshop on Multiword expressions : Analysis, Acquisition and Treatment.* In 41st Meeting of the Association for Computational Linguistics (ACL, 2003), Sapporo, Japan, 12 Juillet, 2003, 1-8p.

Fayyad U., (1997). Editorial. *Data Mining and Knowledge discovery*, 1 (1997), n° 1, 5-10.

Feldman, R., Fresko, M., Kinar, Y. *et al.* (1998). Text Mining at the term level. In Zytkow J.M. & Quafafou M. (Eds.), : *Principles of Datamining and knowledge discovery. Proceedings of the 2nd European symposium PKDD'98.* (pp. 65-73). Nantes, France. Berlin-Springer.

Hearst M.A. (1999). Untangling Text Data Mining. *Proceedings of the 37th Annual meeting of the Association for Computational Linguistics*, Maryland, June 20-26, 1999. [Invited paper].

Hearst M.A. (1992). Automatic acquisition of hyponyms from large text corpora. *Proceedings of the COLING'92*, Nantes, 539-545.

Ibekwe-SanJuan F., SanJuan E. (2004). Mining textual data through Term Variant clustering : the TermWatch system, In *RIAO-04 (Recherche d'Information assistée par ordinateur)*, Avignon, 26-28, Avril 2004, 487-503.

Ibekwe-SanJuan F., SanJuan E. (2003). TermWatch : cartographie de réseaux de termes. *5ème Conférence "Terminologie et Intelligence Artificielle " (TIA'03)*, Strasbourg, 31 Mars– 1 Avril 2003, 124-134.

Ibekwe-SanJuan, F. (1998). A linguistic and mathematical method for mapping thematic trends from texts. *Proceedings of the 13th European Conference on Artificial Intelligence* (ECAI'98), Brighton UK, 23-28 Août 1998, 170-174.

Kodratoff Y. (1999). Knowledge discovery in texts : A definition and applications, in Foundation of Intelligent systems, Ras & Skowron (eds.) *Lecture Notes in Artificial Intelligence*, n° 1609, Springer-Verlag, pp. 16-29, 1999.

Lent, B., Agrawal, R., & Ramakrishan, S. (1997). Discovering trends in Databases. *Proceedings of the 3rd International conference on knowledge discovery in databases (KDD'97),* 227-230.

Fellbaum C. (1998). WordNet : An electronic lexical database. Cambridge, M.A MIT press.

Morin E, Jacquemin C. (2003). Automatic acquisition and expansion of hypernym links. *Computer and the humanities.* Kluwer Academic press. 36p.

Pedersen T., Patwardhan, Michelizzi (2004). WordNet::Similarity – Measuring the relatedness of concepts. *A paraître dans « Proceedings of the 5th Annual Meeting of the North American Chafter of the Association for Computational Linguistics* (NAACL-04), Mai, 3-5 2004, Boston, 4p.

Séguéla P., Aussenac-Gilles N. (1999). Extraction de relations sémantiques entre termes et enrichissement de modèles du domaine. Actes de la conférence IC'99 - Plate-forme AFIA. Palaiseau (F), 14-18 Juin 1999. pp 79-88.

Silberztein M. (1993) Dictionnaire électronique et analyse automatique des textes. Le système INTEX. Masson, Paris.

Suarez M., Cabré T. (2002) Terminological variation in specialized texts : linguistic traces for automatic retrieval. *Proceedings of the VIII Ibero-American Symposium on Terminology (RITERM)*, 28-31 Octobre, 2002, 10p.