Term clusters labelling by consensus

Nicolas Turenne

Nommage de classes de termes par consensus

Nicolas Turenne

UMR INRA-INAPG – Biométrie et Intelligence Artificielle (BIA)
16 rue Claude Bernard, 75231 Paris cedex 5
turenne@inapg.inra.fr

Abstract

Textual databases become the biggest amount of data available in any domain. We have to know how to extract information and meaning of such textual data to solve a user need. A possible way could be term clustering (i.e. ; association fields) but, since a significant number of previous studies showed that it was impossible to get an absolute confidence in the efficiency of automatic processing, this task requires a careful evaluation of its results. We have considered in our work the consensus as a kind of fusion of given data (set of general knowledge classes) on the one hand and automatically obtained data on the other. We have used this model to drive an evaluation of the clustering results. We try to show that a consensus model may induce decision making to optimally label clusters.

Résumé

Les bases de données textuelles constituent la plus grande quantité de données disponibles dans n'importe quel domaine. Pour satisfaire les besoins d'un utilisateur nous devons savoir comment extraire l'information et le sens de ces données textuelles. Une solution possible peut être l'extraction de classes de termes (i.e. champs associatifs) mais, depuis qu'un nombre important d'études antérieures ont montré qu'il était impossible d'avoir une confiance absolue dans l'efficacité d'un tel processus automatique, cette opération demande une évaluation minutieuse de ses résultats. Nous avons considéré dans notre travail le consensus comme une sorte de fusion de données fixes (ensemble de classes de connaissances générales) d'un côté et de données obtenues automatiquement d'un autre côté. Nous avons utilisé ce modèle pour conduire une évaluation de résultats de classification automatique non supervisée. Nous essayons de montrer que le modèle du consensus peut induire une aide à la décision pour nommer des classes de termes.

Mots-clés : Modèle de consensus; Evaluation de classe; Fusion de données; Classification non supervisée; Traitement statistique de corpus; Fouille de textes.

1. Introduction

De nos jours, d'énormes bases de données textuelles sous format électronique sont disponibles forçant un utilisateur à spécifier ses besoins. Il est possible d'utiliser une collection de textes comme support d'extraction spécifique de connaissances dans une tâche de traitement du langage naturel. Ainsi une telle tâche peut être utile pour aider un utilisateur à extraire de l'information (Grefenstette, 1994). La fouille de textes se préoccupe du traitement de corpus dans la perspective de mieux comprendre la description du contenu d'un ensemble de textes et le sens des unités textuelles (Yarowsky, 1992)(Zernik, 1991).

Classiquement pour qualifier une classe de termes, on se sert d'un terme provenant de la classe elle-même et servant de prototype. Une autre voie est l'utilisation d'une hiérarchie de référence à laquelle on compare, à l'aide d'une mesure quantitative, les classes obtenues par classification automatique (Agarwal, 1995)(Turenne, 2000). L'objectif de cet article est de pouvoir évaluer des classes de termes en les nommant grâce à une ressource sémantique de sens commun externe et définie a priori. Pour décider de l'attribution de telle ou telle étiquette on projette les classes sur le thesaurus afin d'atteindre une correspondance à la majorité.

Nous appelons ce processus le modèle du consensus, généralement utilisé dans la classification non supervisée pour la construction des arbres (Leclerc, 1996). La méthodologie du consensus consiste à agréger plusieurs objets en un objet unique du même type. Il existe plusieurs stratégies et d'heuristiques de consensus. Les plus connues sont : le consensus strict, le consensus par quota et le consensus de la médiane. Le modèle du consensus est proche d'un modèle de vote. Traditionnellement un votant exprime un vote par rapport à un choix qui lui est soumis. Dans le cas de l'analyse des textes, un votant sera un document ou une classe de termes; l'objet est un terme qui peut être contenu dans le votant. L'ensemble des opinions produit les valeurs relatives à l'objet soumis aux votants. On regroupe les réponses de façon à sélectionner celle qui dépasse un certain seuil. Ainsi on extrait une opinion à la majorité, qui déterminera l'état de l'objet ou de sa représentation. On peut illustrer le modèle de consensus par un exemple (EX) concernant la prise de décision entre 2 orthographes d'un terme comme e_1=« événement » et e_2=« évènement » sachant qu'ils sont écrits dans n documents. Le terme e_1 apparaît n_1 fois dans n alors que e_2 apparaît n_2 fois, comme $n_1 > T$ (T étant un seuil acceptabilité) alors e_1 est retenu comme orthographe valide.
Dans la section 2, nous donnons le cadre de notre analyse de consensus. Dans la section 3, nous donnons un exemple de notre modèle pour l'indexation de pages d'Internet, et ensuite nous présentons l'application du modèle pour l'étiquetage sémantique. Le modèle est essentiellement utilisé pour valider la couverture des termes d'une classe par une étiquette générale et significative. Finalement, dans la section 4 nous présentons d'autres approches d'extraction d'information à partir de textes et leur moyen de qualifier les résultats.

2. Cadre général

Définition du consensus: soient N objets qui ont à décider d'une valeur binaire (0,1). Chaque objet possède sa propre valeur initiale:
- Si tous les objets ont la valeur initiale 0, alors ils décident 0,
- Si tous les objets ont la valeur initiale 1, alors ils décident 1,

Les deux conditions précédentes sont en général affaiblies en renforçant la prémisse « tous les objets ont la valeur initiale n », ainsi :
- Tous les objets décident d'une même valeur.

Nous développons la méthode de consensus dérivée de la méthode des quotas assurant la déduction d'une décision par quota sans systématiquement obtenir la majorité absolue. On décrit dans la suite comment déterminer des valeurs de choix (0 ou 1). Soit un ensemble d'objets E à classer et un ensemble S de votants. Soient $e \in E$, et $s_k \in S$:

$$S=\{s_1, s_2, \ldots, s_n\} . \quad (1)$$

$\Re$(a,b) est la relation de vote de b par rapport à a. I est l'ensemble des votes tels que :

$$i_k = e \Re s_k . \quad (2)$$

Soit $p_I(e)$ est l'ensemble de projection de e dans l'espace I :

$$p_I(e) = \{i \mid i \in I\} . \quad (3)$$

Cette relation $\Re$ est définie comme une fonction d'appartenance.

$$i_k = \begin{cases} 1 \Rightarrow e \subset s_k, & s_k \ vote \ \ pour \ \ e \\ 0 \Rightarrow e \not\subset s_k, & s_k \ vote \ \ contre \ \ e \end{cases} \quad (4)$$

Dans notre exemple EX (§1) s_k est un document k et E={"événement", "événement"}. On définit la fonction F qui rassemble les opinions positives pointant vers l'objet e (Fig. 1) telle que:

$$F(e)= \sum_{k=1}^{card(S)} i_k \qquad (5)$$

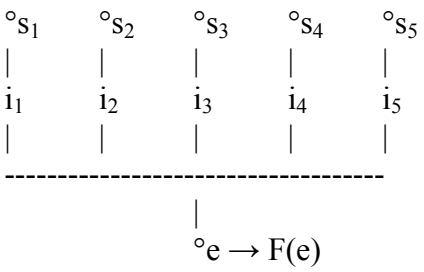

Fig. 1. *Vote pour e de chaque individu de l'espace S.*

Pour plusieurs objets $\{e_1,...,e_n\}$=E à soumettre on aboutit à l'ensemble $\Im$ des valeurs calculées par F (Fig. 2) :

$$\Im=\{F(e_1),\ldots,F(e_n)\} \qquad (6)$$

Dans notre exemple EX , $\Im=\{n_1, n_2\}$.

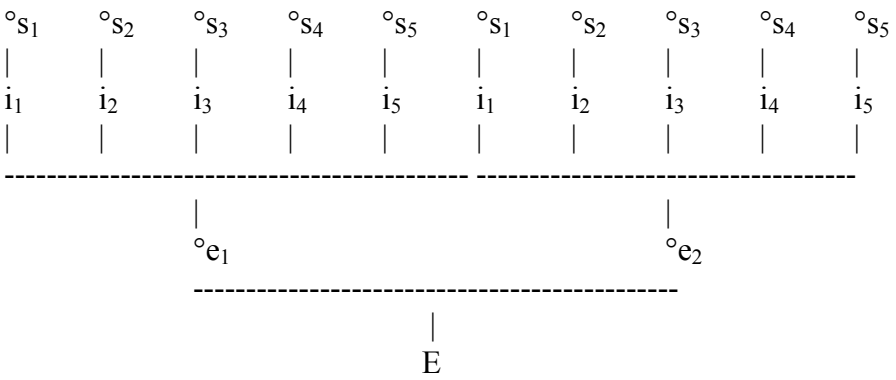

Fig. 2. *Vote pour 2 objets e_1 et e_2 pour chaque individu s_i de l'espace S.*

On peut aussi définir la fonction inverse F^{-1} telle qu'elle calcule le score d'appartenance pour tous les objets e_j liés à (avec i_k) un individu donné s_k :

$$F^{-1}(s_k)= \sum_{j=1}^{card(E)} i_k(e_j) \qquad (7)$$

Ainsi $\Im$ est l'ensemble suivant:

$$\Im=\{F^{-1}(s_1),\ldots,F^{-1}(s_n)\} \ . \qquad (8)$$

Nous définissons un seuil d'acceptation T concernant les valeurs à conserver. Ce seuil dépend de l'objectif à développer un consensus acceptable et interprétable. Ce seuil est borné par :

$$\min(\Im) < T < \max(\Im) \ . \qquad (9)$$

La limite inférieure assure une décision correcte du critère de consensus.

Les types d'heuristiques ne sont pas limités. Par exemple nous pouvons fixer un critère à la majorité comme suit.
Ce critère est lié à la médiane:

$$T=(\max(\Im) - \min(\Im))/2 . \quad (\mathbf{10})$$

Ce critère est lié aux valeurs maximales :

$$T= \max(\Im) . \quad (\mathbf{11})$$

Ou bien celui-ci est lié aux fréquences relatives cumulées, par exemple on fixera T à 65% des plus fortes fréquences.

Ou bien celui-ci est lié au quota fixé par l'amplitude des votes :

$$T=\frac{\max(F(e_j))-\min(F(e_j))}{\max(F(e_j))+\min(F(e_j))}.\max(F(e_j)) \quad (\mathbf{12})$$

Finalement la règle du consensus par quota s'énonce comme suit :

$$e \subset C \text{ si } F(e) >= T \text{ , ou } s \subset C \text{ si } F^{-1}(s) >= T . \quad (\mathbf{13})$$

Dans le cas d'un quota de fréquences cumulées :

$$e_k \subset C \text{ si } \frac{\sum_{j=1}^{j=k} F(e_j)}{\sum_{j=1}^{card(E)} F(e_j)} \geq T \quad \text{avec} \quad F(e_{i-1}) \geq F(e_i) \ \ \forall i, \text{ ou } s \subset C \text{ si } \frac{\sum_{j=1}^{j=k} F^{-1}(s_j)}{\sum_{j=1}^{card(S)} F^{-1}(s_j)} \geq T \quad (\mathbf{14})$$

$$\text{avec} \quad F^{-1}(s_{i-1}) \geq F^{-1}(s_i) \forall i$$

C représente la solution de la décision de sélection des objets par consensus, c'est-à-dire l'ensemble des objets e retenus par vote favorable et vérifiant un seuil d'acceptabilité ; T est fixé empiriquement par rapport à la tâche désirée.

Dans notre exemple EX, en utilisant un quota à l'amplitude $T=\frac{n_1-n_2}{n_1+n_2}.n_1$ avec $n_1>n_2$, $n_1>T$ et $n_2 < T$ si $n_2 < n_1/2$ et donc e1 est choisi.

3. Utilisation du consensus

Nous présentons 2 utilisations possibles du modèle de consensus dans le cadre de la recherche d'information ou de la fouille de textes. Le consensus a pour but de résoudre un problème d'aide à la décision en tenant compte de sa nature sémantique.

3.1 Consensus pour l'indexation

Cet exemple illustratif est lié aux pages d'Internet. Contrairement à l'exemple précédent une page Web n'aura pas le statut de votant mais ce sont les requêtes qui ont permis d'accéder cette page Web grâce aux moteurs de recherche. Dans notre cas de figure, la page est ma page Web personnelle. Il y bien sûr une part d'arbitraire dans le choix d'un seuil qui est adapté aux données traitées. Dans notre cas on choisit un quota de fréquences cumulées à 65% pour lequel nous savons que les termes neurone, information…sont pas ou très peu significatifs.
Voici les caractéristiques du modèle de consensus :

Taille de la base des votants	:	N = 400 requêtes en français
Objet	:	e = un mot de la page Web donnée
Votant	:	s = une requête d'Internet
Résultat d'un vote	:	i = 1 ou 0.
Seuil	:	T = 3 (quota des fréquences relatives cumulées à 65%).

E={"mining", "text", …} ; F={58, 45,…} ; C={"mining", "text",…, "apprentissage"}.
Les requêtes forment un corpus d'individus votant (environ 400 requêtes provenant de moteurs comme AltaVista, Google, Yahoo). Chaque mot est soumis à l'ensemble des requêtes pour obtenir les valeurs F de ces mots (Table 2.). Au-delà d'un seuil (environ 3) les mots (colorés en gris) sont pertinents par rapport à la page Web, à l'exception de quelques mots considérés comme appartenant à une liste d'arrêt (i.e. *de*, *du*). Ces mots identifiés par consensus parmi les mots interrogés de la page pourraient faire partie d'une liste typique qui résumerait le contenu de la page Web.

Fréquence	Fréquence relative	Terme
58	22.39%	mining
45	17.37%	text
18	6.94%	textmining
16	6.17%	turenne
13	5.01%	data
6	2.31%	clustering
3	1.15%	statistique
3	1.15%	traitement
3	1.15%	de
3	1.15%	supervisé
3	1.15%	jérôme
3	1.15%	thil
3	1.15%	français
3	1.15%	filtrage
3	1.15%	du
3	1.15%	apprentissage
2	0.77%	langage
2	0.77%	dinformation
2	0.77%	web
2	0.77%	neurone

Table. 2. *Consensus sur les termes-index des requêtes.*

3.2 Consensus sur l'étiquetage

Ce dernier exemple met en valeur la fonction inverse F^{-1} à travers l'utilisation de différentes catégories d'un thésaurus. Notre objectif est de tester la cohésion lexicale d'un groupe de termes assimilés à une classe (Loukashevich, 2000)(Elman, 2000). Pour cela nous essayons d'étiqueter la classe de termes avec une catégorie d'un thesaurus de référence.

3.2.1 Structure du thésaurus

Le thésaurus est composé de 100000 termes répartis sous 1000 catégories sémantiques que nous appelons aussi champs sémantiques (groupes conceptuels ayant un code d'identification).
Le thésaurus que nous utilisons est disponible sous format électronique dans le logiciel MS-Word™. Il a été édité par Larousse en 1984, avec le titre : "thésaurus des idées"™. Ce thésaurus possède la même structure que le Roget en anglais. Ainsi les résultats peuvent être reproduit en anglais.

La structure du thésaurus est celle d'un polyarbre. C'est un graphe dont un nœud peut avoir plusieurs pères et plusieurs fils à trois niveaux. Le thésaurus contient autour de 120000 formes classées sous 873 catégories qui sont elles-mêmes réparties sous 26 catégories:
- niveau 0 : termes, environ 120000,
- niveau 1 : sous-catégories, 873 en tout,
- niveau 2 : catégories (thèmes), 26 en tout.

Après lemmatisation le nombre de formes est réduit à environ 100000. L'outil de lemmatisation utilise un dictionnaire de 200000 formes / 50000 lemmes dont certaines règles basiques permettent de lever certaines ambiguïtés (Turenne, 2000).

Partie 1 liste des catégories:
La première liste représente les sous-catégories. Sa structure est la suivante:
exemple : sous-catégorie numéro (873 en tout)

chimie	75
photographie	775

La seconde liste représente les catégories. Sa structure est la suivante :

exemple :	intervalle des sous-catégories	numéro de catégorie (26 en tout)
	773-791	10 (art)

Partie 2 corps des sous-catégories :
Cette partie présente les instances des sous-catégories qui sont les termes appartenant aux sous-catégories. Les termes mentionnés peuvent être listés dans une colonne, quelques informations peuvent apparaître entre parenthèses, un point, une virgule ou un tiret peut séparer quelques termes, et un terme peut être suivi d'un code pointant vers une autre sous-catégorie. Un terme peut être précédé par une locution (comme 'fam. :', 'ou', 'mus.-', ' -.mil. :'…) mais nous l'ignorons.
exemple :

775 (numéro de sous-catégorie)
1 appareil ; appareil-photo ; viseur
2 laboratoire ; studio ; bain ; atténuateur ; développeur ; fixateur ; révélateur ; déclencheur ; nitrate d'argent

Nous avons stocké cette table comme index dans une base de données relationnelle :

terme	étiquette syntaxique	code de sous-catégorie

Une seconde table relie les codes de sous-catégorie (parmi les 873) avec les codes des catégories (parmi les 26).

code de sous-catégorie	code de catégorie

3.2.2 Algorithme

Avant d'utiliser le modèle de consensus nous extrayons les classes de termes avec notre classifieur de termes Galex (Graph Analyzer for Lexicometry)(Turenne, 2000). Le corpus est une collection de documents « plein texte » se référant à un sujet. Un extracteur de termes découpe le corpus en groupes nominaux considérés comme des termes du domaine la méthode est basé sur une méthode des bornes [Turenne, 1999]. Ensuite le classifieur construit des groupes de termes ou classes de termes que nous voulons évaluer comme étant des champs sémantiques liés au domaine du corpus. Le système Galex utilise l'exploitation d'un graphe d'associations pour extraire des champs thématiques. Il repose sur l'utilisation d'un motif de graphe qui contraint l'association de termes avec d'autres termes vérifiant le réseau du motif. Un terme pôle Tp admettant une fréquence privilégiée est associé à d'autres termes T via deux autres termes pivots Tpi pour établir la similarité d'association. La similarité n'est pas basée sur un calcul mais sur l'appariement à un réseau précis. Ainsi les phrases suivantes {la coronarographie montre une lésion athéromateuse, la coronarographie montre un angor d'effort, la coronarographie montre une angine de poitrine, la coronarographie révèle un infarctus, la ventriculographie révèle un angor d'effort, la ventriculographie révèle une angine de poitrine} donnent la classe vérifiant le motif (T_p,T_{pi},T_{pi},T…T) : (T_p= coronarographie, T_{pi}=lésion athéromateuse, T_{pi}=angor d'effort, T=angine de poitrine, T=infarctus) grâce aux relations que ces termes entretiennent avec les verbes *montrer* et *révéler* [Turenne, 1999].
La technique d'évaluation utilise le modèle du consensus pour savoir si une étiquette sémantique prise dans un thésaurus statique et externe, choisi comme référence, peut couvrir les termes d'une classe. Chaque terme de la classe interroge les champs sémantiques considérés comme individu votant. Nous utilisons F^{-1} (eq. (6))). Cela signifie que pour un terme donné

(e) nous collectons toutes les valeurs du résultat du vote (i) pour une catégorie donnée (s) et cela pour toutes les catégories.
Quelques caractéristiques du modèle de consensus :

Taille de la base des votants : N = 1000 catégories (Larousse ou Roget)
Objet : e = un terme d'une classe de termes donnée
Votant : s = un groupe de termes d'une catégorie
Résultat d'un vote : i = 1 ou 0.
Seuil : T = max(ℑ) (somme maximale des occurrences d'une catégorie).

Nos expériences ont été faites essentiellement sur un corpus médical traitant des maladies coronariennes ayant été constitué par des experts en médecine.

Termes de la classeErreur ! Source du renvoi introuvable.	**Codes collectés (groupes conceptuels)**	**Décision**
Anomalie, spasme coronarien, thallium, pronostic, méthergin, femme, question, fonction gauche, segment, coronaire droit	210, 248, 328, 345, 326, 331, 391, 331	Coeur(331)

Table 3. *Etiquetage automatique d'un groupe de termes par consensus*.

Nous collectons le score total obtenu pour chaque groupe conceptuel du thesaurus (Table 3.). Le code conceptuel ayant la fréquence maximale est assignée comme étiquette sémantique pour qualifier la cohésion lexicale de la classe de termes (Turenne, 2000) (Table 4).

Algorithme (étape 1 : étiquetage par le 1er niveau)

```
1- Collecter tous les codes des termes inclus dans la classe.
2- Si un terme est un terme composé
        chercher le code du terme entier s'il existe
        Sinon collecter les codes du premier mot et du dernier mot du terme sauf si le dernier est un adjectif.
3- Si il existe, choisir le code le plus fréquent comme étiquette de la classe
        si plusieurs codes ont la même fréquence maximale
                choisir la plus petite
        sinon chercher les codes du terme pôle
              choisir la plus petite.
```

(Les heuristiques consistant à choisir le plus petit code par s'explique par le fait que le code moins pondéré est plus général)

Algorithme (étape 2 : étiquetage par le 2nd niveau et étiquetage global)

```
1- Assigner chaque code de 1er niveau affecté à une classe à un code du 2nd niveau inclu dans un intervalle de codes du 2nd niveau. Par exemple : code 248 est dans l'intervalle 230-267 lié à "matière" ; donc le noeud 2 pour la classe est "matière".
2- Sélectionner les thèmes du corpus qui sont prédominants :
   - Collecter les codes de toutes les classes,
   - Calculer la fréquence de chaque code,
   - Trier les codes par ordre croissant de fréquence.
3- Choisir les 3 premiers codes de la liste ayant une fréquence ≥ 3
```

3.2.3 Application

Classe	Catégorie (niv. 1)	Catégorie (niv. 2)
star, planet, solar system, satellite, earth, moon, comet	World	Matter in general
human rights, council of Europe, minorities, convention, parliamentary assembly, committee of ministers, countries	Council	Voluntary action

Table 4. Exemples d'étiquetage avec quelques classes (manuellement créées).

Classe	Catégorie (niv. 1)	Catégorie (niv. 2)
Substance de contraste, évolution, oblique antérieur, technique, test, traitement, ventricule, ventriculographie	Médecine	Médecine
Seule incidence, diamètre de l'obstruction, possible de dilater, angioplastie	Dimension	Dimensions
valve aortique, coronaire, coronarien, aorte, circulation	Coeur et vaisseaux	Corps
tension artérielle, douleur, fraction d'éjection, ventriculaire, artère, altération	Coeur et vaisseaux	Corps
technique du cathéter cardiaque, lésion, examen, angor, cathéter	Méthode	Corps

Table 5. *Exemples d'étiquetage automatique avec quelques classes (obtenues automatiquement).*

Le second processus de généralisation induit la sélection de codes relatifs à chaque classe, pour les classer par ordre décroissant de façon à conserver les plus fréquents (Table 6.). Les 3 codes les plus fréquents sont retenus pour qualifier l'ensemble des classes et donc le corpus.

Nombre d'occurrences	Code de catégorie	Nom de la catégorie (1er-niveau)
31	*383*	*Maladie*
23	*331*	*Coeur*
21	*391*	*Médecine*
11	792	Travail
10	185	Période

Table 6. *Catégories les plus fréquentes automatiquement affectées à l'ensemble des classes.*

Concernant le corpus médical, un aperçu général du processus d'affectation montre les résultats suivants : 21 classes ont été étiquetées par une catégorie médicale et 43 ont été étiquetées par une catégorie d'état ; finalement 85% ont été "correctement" classées dans un thème relatif au contenu sémantique du corpus. Ce dernier résultat relève de l'appréciation humaine et souligne par là-même le problème difficile de l'évaluation à cheval entre un point de vue personnel et une heuristique robuste et automatique.

Finalement, nous testons notre algorithme d'étiquetage sur un corpus avec des thèmes mélangés. Pour cela nous construisons un corpus à partir de textes de l'Encyclopédie Universalis concernant d'une part l'aéronautique, et d'autre part l'histoire de la Russie. La taille du corpus est assez petite, environ 70000 mots. Le traitement du corpus conduit à une discrimination des sujets à travers l'extraction des classes de termes obtenues par notre classifieur Galex (Turenne, 2000). Des 61 classes extraites, 27 sont relatives à l'histoire de la Russie, 19 sont relatives au domaine de l'aéronautique et 15 restent ambiguës. En utilisant le même thesaurus

décrit 3.2.1, de l'ensemble des classes environ 75% peuvent être affectées pertinemment à leur thème respectif.
En ce que concerne le nommage de l'intégralité des classes (Table 5), c'est-à-dire l'affectation d'étiquettes du thésaurus pour l'ensemble des classes, on arrive à une précision de plus de 95% en considérant les 3 étiquettes les plus probables. Ce résultats est valable pour des corpus de type scientifique, technique ou encyclopédique. En ce qui concerne des corpus de messages électroniques la précision n'est plus très fiable.

4. Travaux antérieurs

Traditionnellement un problème d'aide à la décision d'un groupe est établi dans un environnement où il y a une question à résoudre, un ensemble d'options possibles, et un ensemble d'individus (experts, juges,...) qui présentent leur opinion ou préférence à travers l'ensemble des options possibles. Habituellement l'ensemble des individus admet initialement des opinions opposées. Ainsi, un processus de recherche de consensus est nécessaire pour atteindre un consensus général sur les options sélectionnées. Le consensus, de manière classique, signifie un accord complet et unanime des opinions individuelles (i.e. consensus maximum). Dans la pratique, ce consensus est difficile à mettre en oeuvre. Le processus est donc vu comme un processus dynamique dans lequel un « modérateur », via un échange d'information, essaie de persuader les individus de mettre à jour leur opinion. A chaque étape le degré du consensus existant et la distance par rapport à un consensus idéal sont mesurés (Herrera et al, 1995).
Nous présentons maintenant quelques systèmes automatiques qui visent à qualifier la cohésion d'une classe, dont la plupart sont dédiés au traitement du langage naturel étant donné la spécificité des données.
(Grefenstette, 1996) exploitent des thesaurus pour tester la cohésion lexicale des paires de mots (mot cible / mot contextuel le plus significatif) avec un corpus de 4 Mo et 3000 mots cibles ayant une fréquence supérieure à 10. Trois thesaurus ont été utilisés : Roget, MacQuarie et Webster. Nous rappelons que le Roget admet 30000 termes classes dans 1000 catégories et, Webster est un dictionnaire de définitions formatées en listes de mots nettoyées par une liste d'arrêt (434 mots vides). MacQuarie possède la même structure que le Roget. Les résultats visent à comparer la cohésion des mots les plus significatifs soit par une méthode syntaxique (relation nom-adjectif, nom-verbe…), ou par une méthode de cooccurrence basée sur une fenêtre de 10 mots (à gauche et à droite), les mots les plus significatifs étant déterminés par un coefficient de Jaccard. Les expériences montrent : premièrement qu'une paire appartient à la même catégorie dans moins de 50% des cas et ce pour les mots cibles les pus fréquents, deuxièmement, il semble difficile de distinguer la méthode syntaxique de la méthode de cooccurrence. Comme la probabilité que 2 termes appartiennent à la même catégorie est inférieure à 1%, l'utilisation d'un thesaurus semble utile mais pas assez suffisante à elle seule pour déterminer la sémantique d'une paire.
(Agarwal, 1995) développe une classification conceptuelle hiérarchique. Il utilise également le thesaurus Wordnet pour éliminer les termes d'une classe qui n'appartiendraient pas à la même classe Wordnet des autres termes de la classe. En matière d'évaluation un expert crée une classification manuelle de référence. Une fonction des paramètres de précision et de rappel (F-mesure) permet d'estimer la qualité des classes établies automatiquement par rapport à la classification de référence. Ce qui rend notre approche originale par rapport à celle-ci est l'utilisation d'une hiérarchie de référence de sens commun qui fait apparaître la polysémie. En effet dans le thesaurus un terme peut apparaître plusieurs fois dans un certain nombre de classes. En utilisant une technique basée sur le rappel et la précision on obtiendrait des scores peu discriminants. Beaucoup de classes de référence proches de la classe observée obtiendraient

des scores très voisins ce qui poserait des difficultés de prise de décision pour le nommage final. En revanche l'approche par calcul de mesure (précision/rappel) s'avère intéressante quand la hiérarchie est spécialisée (i.e. peu ou pas polysémique). On peut noter toutefois que les facteurs de précision et de rappel peuvent varier car ils dépendent sensiblement de la hiérarchie de référence qui elle-même dépend du point de vue de l'expert.
(Tishby, 1999)(Feldman, 1997) ont mis au point manuellement une hiérarchie de concepts liée à un thème et d'après un corpus de documents de ce thème. Dans leur étude les distributions de termes sont comparées pour un nœud donné de la hiérarchie, à une entropie relative calculée. Le système *KDT* (Knowledge Discovery from Texts) est basé sur cette approche pour assurer une navigation dans les documents en comparant les distributions de mots. Le système *FACT* (Tishby, 1999) est un système de la même famille mais basé sur les cooccurrences. Il gère en plus un langage de requête avec une interface graphique.
Le système *Syndikate* (Hahn, 1997) propose de choisir le concept le plus significatif à partir d'une ontologie (i.e. subsumant) en fonction d'un processus de raisonnement terminologique. Ce système exploite : premièrement, une connaissance qualitative des propriétés linguistique présentes dans les textes libres, et deuxièmement la configuration structurale dans la base de connaissances d'un domaine. Un réseau de 345 concepts et 347 relations sont ainsi définis et décrits dans un formalisme appelé logique de description. Les logiques de description sont basées sur des prédicats composés de concepts primitifs (entités abstraites) et de relations primitives (rôles). A partir de ce réseau on peut dériver d'autres concepts et relations, et finalement, des objets et relations concrets sont rattachés au réseau. Pour un objet donné dans un texte (i.e. terme), quelques hypothèses sont générées, à partir de concepts identifiés dans son contexte textuel, pour trouver le plus proche concept généralisant (subsumant).
(Valtchev et al., 2001) a étudié la fusion de réseaux de concepts grâce à l'Analyse Formelle de Concept (AFC ou treillis de Galois). L'AFC est une discipline qui étudie les structures hiérarchiques induites par une relation binaire entre deux ensembles. La structure, construite de sous-ensembles fermés ordonnés par une inclusion d'ensemble, satisfait les propriétés d'un treillis. (Valtchev et al., 2001) fournit les bases d'une procédure d'assemblage efficace de treillis grâce à un filtrage du produit direct des treillis partiels qui retient les concepts du treillis global et leur relation de précédence. Cette approche semble intéressante puisque notre algorithme est proche d'une structure de type treillis, mais ils basent leur algorithme sur l'intersection d'attributs. La différence avec notre approche intervient au niveau des attributs. Malgré cela, notre approche de fusion par intersection de classe est très similaire.

5. Conclusion et perspectives

Dans notre article, nous avons présenté un modèle de consensus pour aborder le problème de l'évaluation d'une extraction de classes de termes. L'extraction utilisée est notre méthode de classification non supervisée qui est basée sur l'extraction de graphes. L'évaluation est fondée sur le nommage des classes de termes.
Une première tâche, l'indexation de page Web, a été présentée comme illustration de notre modèle. La seconde, l'étiquetage sémantique, est plus centrale à notre article concernant l'évaluation de classes.
Le modèle de consensus est basé sur une liste d'individus votant donnant une réponse binaire sur l'appartenance d'un objet (un terme) dans le votant. L'individu votant peut être une requête, un document ou un groupe de termes. Nous collectons les réponses sous forme d'une liste de fréquences (fonction F(terme) ou son inverse F^{-1}(individu votant)). Un seuil à la majorité et des heuristiques assurent la prise de décision (un terme, ou un code sémantique d'un thésaurus).

Dans l'étiquetage sémantique, dans un premier temps notre classifieur Galex extrait un ensemble de classes de termes à partir d'un corpus. Une stratégie d'étiquetage implémente un consensus pour décider quelle étiquette sémantique est la plus proche d'une classe de termes donnée. L'étiquette provient d'un ensemble de catégories définies par une hiérarchie dans un thésaurus. La projection d'un terme sur le thésaurus fournit un ensemble de catégories qui pointent vers ce terme. Ainsi une stratégie de comptage d'occurrences des codes recueillis permet de déduire la proportion maximale d'une catégorie qui sera désignée comme étiquette. Cette stratégie exploite aussi l'ensemble des classes pour désigner les 3 catégories couvrant cet ensemble. Une telle stratégie permet d'évaluer un ensemble de classes obtenu automatiquement en validant leur cohésion lexicale par un processus d'étiquetage sémantique.
Le consensus tente d'implémenter une sorte de fusion d'un simple niveau de hiérarchie avec un ensemble de classes de termes. Si cette fusion est satisfaisante cela signifie que les regroupements de termes sont suffisamment homogènes.
Les expériences révèlent quelques difficultés à obtenir un étiquetage efficace avec des corpus constitués de messages électroniques. Les syntagmes (i.e. groupes nominaux) ne sont pas référencés dans le thesaurus ou apparaissent dans la même classe bien que ne possédant pas d'usage courant en commun. Avec des textes encyclopédiques nous obtenons de meilleurs résultats avec un taux de succès d'étiquetage dépassant 70%. L'affectation d'étiquettes à l'ensemble des classes marche très bien avec un taux de réussite dépassant 95% quel que soit le thème du corpus traité.
L'intérêt du modèle de consensus repose sur sa simplicité de mise en oeuvre et sur son efficacité à informer grossièrement de la généralité d'une classe en fonction d'une référence. On espère utiliser un tel étiquetage sémantique dans un processus de filtrage automatique pour vérifier la capacité des catégories, fournies par les classes, à correspondre à une catégorie assignée à un nouveau document. Une autre perspective possible serait de compléter ce modèle par un réseau terminologique et d'implémenter un processus de subsomption logique de façon à extraire un concept généralisant significatif. Inversement on pourrait analyser si une étiquette sémantique obtenue par le modèle du consensus est réellement un subsumant dans le réseau terminologique et ce pour tous les composants de la classe.

Références

Agarwal R., Evaluation of Semantic Clusters, dans les actes de 33rd Annual Meeting of the Association for Computational Linguistic (ACL), Cambridge (USA), 1995.

Elman J., On the Generality of Thesaurally derived Lexical Links, dans les actes de *International Conference of Textual Statistical Data Analysis (JADT)* , Lausanne (Suisse), 2000.

Feldman R., Dagan I., Knowledge Discovery in Textual Databases (KDT), dans les actes de *1st International Conference on Knowledge Discovery (KDD),* Montreal (Canada), 1997.

Grefenstette G., Evaluation Techniques for Automatic Semantic Extraction: Comparing Syntactic and Window Based Approaches, in *Corpus processing for Lexical Acquisition* ed. B.Boguraev,J.Pustejovsky MIT 1996.

Grefenstette G., SEXTANT: Extracting Semantics from Raw Text: Implementation Details, Heuristics, *Integrated Computer-Aided Engineering*, vol.1 n°6 pp. 527-536 1994.

Hahn U., Schnattinger K., Knowledge Mining from Textual Sources, dans les actes de *International Conference on Information and Knowledge Management (CIKM)*, Las Vegas (USA), 1997.

Hearst M., Untangling Text Data Mining, dans les actes de *Association for Computational Linguistics Conference (ACL)*, University of Maryland (USA), 1999.

Herrera F., Herrera-Viedma E., Verdegay J.L., Basis for a Consensus Model in Group DecisionMaking with Linguistic Preferences, dans les actes du 3ième congès européen *Fuzzy and Intelligent Technologies* (EUFIT), Aachen (Allemagne), 1995.

Leclerc B., Consensus of classification: the Case of Trees, dans les actes de *Society of the Classification of North America Workshop*, Amherst (USA), 1996.

Loukashevich N., Dobrov B., Thesaurus as a Tool for Automatic Detection of Lexical Cohesion in Texts, dans les actes de *International Conference of Textual Statistical Data Analysis (JADT),* Lausanne (Suisse), 2000.

Tishby N., Pereira F. and Bialek W., The Information Bottleneck Method, dans les actes de *37th Annual Allerton Conference on Communication Control and Computing*, University of Illinois (USA) 1999.

Turenne N., Apprentissage statistique pour l'extraction de concepts à partir de textes. Application au filtrage d'informations textuelles, PhD thesis of the Université Louis-Pasteur, Strasbourg, 2000.

Valtchev P, Missaoui R, Lebrun P A partition-based approach towards constructing Galois (concept) lattices, to appear in *Discrete Mathematics*, 2001.

Yarowsky D., Word-Sense Disambiguation using Statistical Models of Roget's Categories trained on Large Corpora, dans les actes de *Computational Linguistics Conference (COLING)*, Nantes (France), 1992.

Zernik U. , Train 1 vs Train 2: Tagging Word Sense in a Corpus , in Zernik,U (ed.) *Lexical Acquisition: Exploiting on-Line Resources to Build a Lexicon*, Hillsdale, NJ: Lawrence Erlbaum Associates, 1991.