Terminologie et classification automatique des textes

Eric SanJuan[1], Fidelia Ibekwe-SanJuan[2]

[1]LAPCS – Université Claude Bernard – 50 av. Tony Garnier – 69366 Lyon Cedex 07 – France – eric.sanjuan@univ-lyon1.fr

[2]ERSICO – Université Jean Moulin – 4 cours Albert Thomas – 69008 Lyon – France – ibekwe@univ-lyon3.fr

Abstract

We present a linguistically-based method, CPCL (Classification by Preferential *Clustered* Link) for clustering text in order to detect research topics. After natural language processing which extracts terms from english texts, our approach identifies terminological variations between these terms. We use local grammars implemented in the INTEX linguistic tool (Max Silberztein, 1993) to extract terms. Programs written in Awk then identify the syntactic variation relations amongst terms. These variations describe operations of expansion, insertion and substitution that take place in terms. The variation relations are then represented in the form of a di-graph which form the basis of our classification method. Clustering is a two-step procedure. First connected components are computed basing on a subset of the variation relations. They are then clustered into classes using the second subset of variation relations. We defined properties that enable the user exploit and interpret classes obtained from this approach : class size, centrality, variation index and transformation index.

The clustering technique, though similar to hierarchical ascending classification, differs from it in that clustering is based not on atomic information units (terms) but on subsets of these units (the connected components). Our approach does not use a distance measure for clustering but rather computes an edge differenciation coefficient that indicates the proportion of each variation type. It also does not impose the number of classes to be obtained nor their sizes. This clustering algorithm has been described in Ibekwe-SanJuan (1998a) and Ibekwe-SanJuan & SanJuan, (1999). Here we will focus on the automating of the first linguistic process which deals with term extraction. This stage was not automated until now.

Résumé

Cet article présente une méthode de classification automatique de textes, CPCL (Classification by Prefrential *Clustered* Link) fondée sur des relations linguistiques. Le but est d'identifier des thèmes de recherche contenus dans le corpus. Après une analyse linguistique automatique des textes en anglais visant à extraire des unités terminologiques (termes), la méthode recherche des variations syntaxiques entre ces termes. Les termes sont extraits grâce à des grammaires locales définies et implantées dans le logiciel INTEX (Silberztein, 1993). Les relations de variations syntaxiques sont recherchées par des programmes écrits en AWK qui comparent les opérations d'expansion, d'insertion et de substitution entre les termes. Les relations de variations sont représentées sous forme de graphes orientés et constituent la base de l'algorithme de classification CPCL.

L'étape de classification opère en deux temps. On calcule d'abord les composantes connexes du graphe à partir d'un sous-ensemble de relations. Ensuite, on agglomère ces composantes connexes en classes avec le deuxième sous-ensemble de relations. La technique de classification, semblable à la classification ascendante hiérarchique à saut minimal, diffère de celle-ci dans la mesure où la classification porte non pas sur des unités atomiques de

compte (les termes) mais sur les composantes connexes du graphe (ensembles de termes). La méthode n'utilise pas une mesure de distance pour agglomérer les composantes connexes en classes mais va plutôt calculer un coefficient de différenciation d'arêtes du graphe selon le type de variation qu'elles représentent. Notre méthode n'impose ni la taille des classes obtenues, ni leur nombre. Nous avons défini des propriétés pour caractériser les classes obtenues par cette méthode : la taille, la centralité, l'indice de variation interne et externe, l'indice de transformation. Cette méthode de classification a été exposée dans Ibekwe-SanJuan (1998a) et Ibekwe-SanJuan & SanJuan (1999). Ici, nous nous intéressons plus particulièrement à l'automatisation de l'étape d'extraction des termes qui ne l'était pas jusqu'ici.

Mots-clés : Terminologie, variation terminologique, TALN, classification automatique de textes, théorie des graphe, veille scientifique et technique.

1. Introduction

Nous avons développé une méthode de classification automatique de textes, à partir de relations linguistiques de variations que nous détectons entre les unités textuelles pertinentes présentes dans ces mêmes textes. Ces unités sont des termes, reconnus largement comme des unités textuelles de choix car ils permettent d'accéder aux concepts ou objets manipulés par des chercheurs ou spécialistes dans leurs domaines de spécialité (Condamines, 1993 ; Bourigault 1994). L'objectif visé par cette méthode est d'identifier les thèmes (sujets) intéressants contenus dans un corpus et de montrer les relations qu'ils partagent.

De nombreuses méthodes d'analyses de données ont été appliquées aux textes (Lebart & Salem, 1994). Le choix de l'unité de compte varie selon la méthode et l'objectif visé : les segments répétés dans Alceste (Reinert, 1993), les mots-clés associés dans Leximappe et ses variantes (Callon et al. 1991). Dans ces méthodes, la classification est basée sur les critères d'occurrences ou de co-occurrences des unités de compte. Récemment, d'autres approches de l'analyse des données textuelles ont commencé à s'intéresser aux termes comme des unités de compte de la classification (Neuroweb : Lelu, 2001) ou SDOC (variante de la méthode des mots associés implantées par l'INIST) mais la classification s'appuie toujours sur des critères de co-occurrences. Une mesure de distance est habituellement appliquée sur une matrice de co-occurrence afin de générer les classes.

Nous présentons une approche différente qui permet de visualiser les associations entre les thèmes de recherche non par le biais d'un critère de 'présence-absence' mais par celui des transformations linguistiques entre les unités représentant les concepts ou objets du domaine étudié. Outre le fait de fonder la classification sur des relations linguistiques - de variations terminologiques - la méthode CPCL utilise le modèle de graphe pour représenter celles-ci. Ce graphe, orienté et coloré va permettre de former les composantes connexes du graphe et ensuite d'agglomérer ces composantes connexes en classes qui représentent les thématiques du corpus.

Cette méthode avait été testée sur un premier corpus de biotechnologie végétale mais la première étape qui consiste à extraire les termes à partir des textes n'avait pas été automatisée. Elle l'est aujourd'hui grâce à l'implantation de grammaires locales dans le moteur linguistique INTEX développé par (Silberztein, 1993). Après une description de la phase d'extraction de termes mise en œuvre dans INTEX (§1), les secton §2 et § 3 décrivent succintement les relations de variation identifiées et la méthode de classification, CPCL. La section §4 sera consacrée aux caractéristiques des classes qu'il produit. La dernière section (§5) montrera l'application de cette méthode à un corpus de textes scientifiques sur la panification.

2. Extraction des termes à partir des textes

De nombreuses études en extraction terminologique souligne l'importance des termes comme signe linguistique qui accompagne l'apparition d'un nouveau concept dans un domaine. A ce jour, dans la plupart d'applications visant à extraire automatiquement des termes à partir d'un corpus, la validation des unités extraites passe nécessairement par une expertise humaine à cause de l'ambiguïté morpho-syntaxique inhérente à la composition d'un terme (confusion avec un SN ordinaire dépendant du texte). Pour extraire des termes à partir d'un corpus, nous utilisons la boîte à outils linguistique INTEX (Silberztein, 1993). INTEX est un environnement linguistique permettant d'analyser morphologiquement et syntaxiquement un texte afin d'y faire divers traitements tel que la recherche de séquences de formes, de lexèmes, de catégories morphologique. Il permet aussi de faire l'analyse syntaxique ou encore une étude stylistique du corpus. Notre corpus de travail est constitué de 564 titres et résumés bibliographiques en anglais, extraits des bases scientifiques *Food science and technology* et *Food science and technology abstracts*[1] sur une période de 10 ans (1988-98). Ces résumés d'articles scientifiques traitent des conditions de conservation et d'amélioration de la pâte boulangère pour la fabrication du pain. Ce corpus fait environ 72000 mots.

2.1. Définition de grammaires locales dans INTEX

Notre stratégie d'extraction des candidats termes peut se résumer ainsi : nous privilégions la détection de termes complexes qui peuvent indiquer les associations établies par les auteurs des articles entre concepts atomiques du domaine. La décomposition des séquences de syntagmes nominaux complexes, appelées SN maximaux, ne descendra pas forcément au niveau atomique. INTEX permet de définir des automates à états finis qui décrivent plusieurs propriétés linguistiques d'une unité textuelle. L'extraction des termes va consister à appliquer, de façon itérative, des automates de complexité décroissante au corpus. Nous avons défini tout d'abord la structure d'un syntagme nominal minimal (SN-min) : à savoir une unité composée de déterminants, de modifieurs mais obligatoirement d'un nom centre :

SNmin : (<D>+<E>)(<NB>+<E>)(<ADV>+<E>)<A> <N><N>** *(1)*

où + = opération de disjonction; E = chaîne vide; D = déterminant ; NB = nombre; ADV = adverbe ; A = adjectif ; N = nom;* = opérateur de Kleene.

Cet automate reconnaît des SN tels que :

a hydrophilic powdered lecithin *(1a)*

traditional sour dough starter cultures *(1b)*

Un SNmin est donc une structure nominale composée sans élément joncteur (préposition). Une fois un automate défini dans INTEX, il peut être appelé par d'autres automates. Nous avons ensuite défini une structure de SN maximale (SNmax) comme un environnement pouvant imbriquer plusieurs SNmin :

SNmax : SNmin (of +prép1+cc) (SNmin cc) SNmin (<E> +((of +prép1) SNmin (<E>+ (of+prép1+cc) (SNmin cc)* SNmin (E + SNmin))))* *(2)*

où : *prép1* = une classe de prépositions fréquentes dans des termes (from, to, by, for, with, in); *cc* = la coordination n-aire (and, or, la virgule) pouvant relier plusieurs membres.

[1] Corpus fourni par l'équipe URI de INIST (Institut de l'information scientifique et technique).

L'automate de SNmax reconnaît des séquences telles que :

a lengthy process, development of traditional bread flavour *(3)*

a guide to the formulation and processing of traditional yeast-containing doughnuts *(4)*

Des automates intermédiaires vont décomposer les SNmax en fonction de la présence de certains marqueurs (préposition suivie de déterminant, coordination, présence de deux prépositions dans la séquence). Ces critères sont détaillés dans (Ibekwe-SanJuan 2001).

Les deux séquences de SNmax seront décomposées de la sorte :

a lengthy process (3a)

development of traditional bread flavour *(3b)*

a guide *(4a)*

*formulation** *(4b)*

*processing of traditional yeast-containing doughnuts** *(4c)*

(4b) et (4c) sont des découpages erronés dû au fait que nous ne traitons pas encore la coordination. Il faudrait pouvoir reconnaître les membres coordonnés et les éléments communs, à restituer en cas de découpage. Le découpage s'arrête lorsqu'aucune séquence ne contient ni une coordination, ni deux prépositions, ni une suite 'préposition + déterminant'.

2.2. Filtrage de candidats termes extraits

Les unités extraites par nos automates sont des termes candidats qu'il faut valider pour ne retenir que ceux susceptibles de correspondre à de réels termes du domaine. De ce corpus, nous avons extrait 10 324 candidats termes. Cette liste a été soumise à un spécialiste du domaine de l'agroalimentaire[2] pour validation.

Après filtrage, 5268 candidats plausibles ont été retenus. Ce sont ces termes qui seront soumis aux phases ultérieures du traitement.

3. Identification des relations de variations

Après l'extraction de termes candidats dans INTEX et leur validation, la liste retenue est soumise à des programmes de recherche de variations syntaxiques. Les types de variation recherchés ainsi que leur intérêt sur le plan terminologique a été décrit dans (Ibekwe-SanJuan 1998b). Nous les rappelons brièvement ici. De nombreuses études ont déjà souligné l'intérêt des phénomènes de variations terminologiques pour plusieurs applications : extraction et mise à jour terminologique, indexation automatique et recherche d'informations. Les variations sont des transformations morphologiques, syntaxiques (Jacquemin, 1994 ; Daille 1994) ou sémantiques (Hamon, 1998) qui se produisent au sein des termes pour en modifier la forme ou la structure. Par exemple, au niveau syntaxique, entre "*bread baking"* et *"bread dough baking"*, entre *"bread dough"* et *"quality of french bread dough"*. Le fait de mettre en évidence ces variations peut permettre de suivre l'évolution terminologique dans un domaine, et ainsi les associations entre les concepts de ce domaine.

Nous avons choisi de nous intéresser à un sous-ensemble de variations syntaxiques qui affectent la structure et le nombre d'éléments dans un terme. Il s'agit de phénomènes d'expansions, d'insertions et de substitutions. Sur l'axe grammatical, on peut les répartir en

[2] Ingénieur documentaliste de l'INIST qui indexe régulièrement les articles dans ce domaine.

deux catégories : des variations affectant les éléments modifieurs dans un terme et des variations affectant l'élément centre dans un terme. Dans un syntagme nominal (SN), l'élément centre est le dernier nom dans une structure composée : *baking* dans "*bread baking"* ou le dernier nom avant la préposition dans une structure syntagmatique : *quality* dans "*quality of french bread dough".* Les autres éléments du terme sont donc ses modifieurs : "*bread"* et "*french bread dough"* respectivement dans les deux exemples précédents. Syntaxiquement, il n'y a qu'un élément centre dans un terme tandis qu'il peut y avoir plusieurs éléments modifieurs. Donc les possibilités de variation sont multiples.

3.1. Variations affectant les éléments modifieurs dans un terme

Elles appartiennent à deux catégories : les expansions de modifieurs et la substitution de modifieurs.

Les expansions décrivent deux sortes de transformation : expansion gauche et insertion. L'**expansion gauche** (Exp_G) concerne l'ajout d'un modifieur à gauche d'un terme. Par exemple, entre "*bread baking*" et "*microwave bread baking*". **L'insertion** (Ins) est tout simplement l'ajout d'un ou plusieurs éléments entre les bornes gauche et droite du terme mais dans une seule position. Ainsi, "*bread characteristics*" et "*bread dough quality characteristics*".

La **substitution de modifieur** (Sub_M) concerne le changement d'un élément modifieur dans un terme : "*bread dough leavening*" et "*composite dough leavening*".

3.2. Variations affectant l'élément centre dans un terme

Elles appartiennent également à deux catégories : les expansions et la substitution de centre.

L'expansion droite (Exp_D) dénote l'ajout d'un nouvel élément centre dans un terme : "*bread baking*" et "*bread baking trial*".

L'**expansion gauche_droite** (Exp_GD) réunit les deux cas élémentaires décrits précédemment. C'est par exemple la relation entre "*bread dough*" et "*frozen bread dough preparation*".

La **substitution de centr**e (Sub_C) dénote le changement de l'élément centre : "*bread dough leavening*" et "*bread dough elasticity*".

Plus de 3000 termes ont été ainsi mis en relation dans le corpus. Les termes techniques, loin d'être figés, varient. Ils partagent des relations linguistiques entre eux. Ce sont ces relations que nous utilisons pour la classification.

3.3. Génération d'un graphe de variantes

La mise en relation des termes conduit à la génération d'un graphe de variantes où les sommets sont les termes et les arêtes représentent chaque type de relation de variation décrite précédemment. Les arêtes sont orientées pour les relations d'expansions (Ins, Exp_G, Exp_GD). Elles sont non-orientées pour les substitutions (Sub_M, Sub_C). Cette distinction dénote le fait que les relations d'expansion génèrent un ordre entre les termes. C'est une relation anti-symétrique : *bread baking* < *bread baking trial* < *bread roll baking trial*. La relation de substitution génère une symétrie entre les termes : *bread dough leavening* σ *composite dough leavening*. Remarquez que cette relation est transitive sur des termes de longueur 2 et génère des sous-graphes complets (des cliques). La figure 1 ci-dessous donne un aperçu d'une portion du graphe autour des variantes de "*bread dough*".

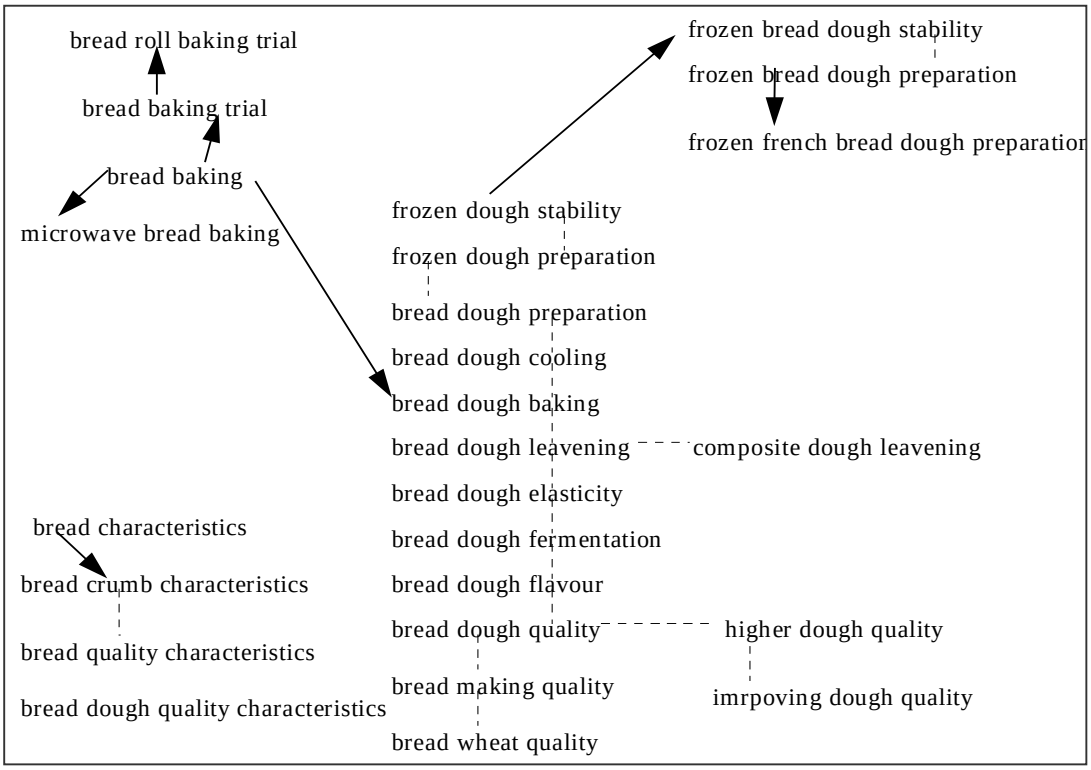

Figure 1. Un sous-graphe de variantes de "bread dough".

Un tel graphe, outre ses propriétés formelles, donne à voir des informations sur les associations des concepts/objets du domaine. Ainsi, on peut visualiser les modifieurs (propriétés) partagés par un élément centre (concept) : l'ensemble de termes autour de "*trial*" ou de "*characteristics*" par exemple. A l'inverse, on voit les concepts qui partagent les mêmes proriétés : celles de "*bread dough*". On voit également à quel endroit un changement de propriété intervient : entre "*bread dough quality*" et "*bread making quality*" par exemple. Pour le spécialiste du domaine, ce graphe de variantes constitue un résultat intermédiaire, riche de renseignements.

4. La classification automatique des termes variants

4.1. Paramètres de classification

Ils sont au nombre de deux : calcul d'un coefficient de différenciation d'arêtes, partage en deux catégories des relations de variation.

- *Coefficient de différenciation d'arêtes*

A chaque arête du graphe on lui associe l'inverse du nombre d'arêtes issues d'un même type de variation. Ce coefficient permet de ne pas handicaper les variations qui sont faiblement représentées par rapport à celles qui sont sur-représentées. Ces différences découlent pour l'essentiel des propriétés formelles des différentes variations, ainsi la relation de substitution génère sur les termes à deux mots une relation d'équivalence et donc un grand nombre de cliques.

- *Détermination des rôles des relations de variation*

Avant l'étape de classification, un partage des relations de variation en deux catégories est opéré en fonction de l'importance que l'utilisateur leur accorde. Cela permet de disposer d'une première catégorie de relations (notée COMP) utilisée pour calculer des composantes connexes. Les relations choisies par défaut sont celles qui affectent les éléments modifieurs, à savoir Sub_M, Ins, Exp_G (cf. §2.1 supra). Cela permet de regrouper au préalable tous les paradigmes du corpus (les ensembles de termes partageant le même centre). Le deuxième sous-ensemble de relations (noté CLAS) sert à former des classes à partir des composantes connexes. Nous choisissons habituellement les relations affectant l'élément centre dans un terme, à savoir Exp_D, Exp_GD, Sub_C (cf §2.2 supra). Les relations de CLAS permettent ainsi d'agglomérer les paradigmes (concepts) associés.

4.2. L'algorithme de classification CPCL

Il comporte deux étapes distinctes : la recherche des composantes connexes à partir des relations de COMP et l'agglomération des composantes en classes à partir des relations de CLAS.

- *Etape 1. Recherche des composantes connexes avec les relations de COMP*

Cette première étape est une application des algorithmes usuels de recherche de composantes connexes. Partant du sous-ensemble COMP de relations, l'algorithme va générer les composantes connexes du graphe qui sont des ensembles maximaux de sommets tels qu'entre deux sommets, il existe un chemin dans COMP permettant de les relier. Cette étape ne fait pas intervenir le coefficient de différenciation d'arêtes défini précédemment. La taille des composantes obtenues est très variable.

- *Etape 2. Agglomération des composantes connexes en classes avec les relations de CLAS*

Les composantes connexes générées à l'étape précédente sont agglomérées en classes à l'aide des relations dans CLAS. C'est dans cette étape que l'algorithme fait intervenir le coefficient de différenciation d'arêtes. Remarquez qu'à ce stade, les sommets ne sont plus les termes (les unités de base) mais des ensembles de termes (composantes connexes) reliés par les relations de COMP. Ainsi, il peut y avoir plusieurs arêtes entre les sommets du graphe quotient obtenu à ce stade. Ce graphe est réduit en sommant la valeur des liens entre deux sommets du graphe. On applique ensuite à ce graphe réduit une classification ascendante hiérarchique par saut minimal. Ainsi deux composantes sont agglomérées si le lien qui les unit est plus fort que le lien qui unit l'une d'entre elles à tout autre composante. A chaque itération, l'algorithme va rechercher une composante à l'extérieur d'une classe qui partage le lien le plus fort avec une composante de cette classe. La procédure est réitérée autant de fois que l'utilisateur le désire ou jusqu'à ce que la partition du graphe initial aboutisse à la fermeture transitive de celui-ci.

Plus formellement, l'algorithme de classification peut être décrit ainsi:

Soit $G(V,E)$ un graphe (i.e. un ensemble V non vide de sommets et une collection E de paires de sommets dans V appelées arêtes), soit $\{COMP, CLAS\}$ une partition (les couleurs) de E et soit p une fonction de E dans un ensemble fini de rationnels.

De ce graphe nous dérivons un graphe réduit $dG(dV, dE)$ tel que : dV est l'ensemble de toutes les composantes connexes de $G(V, COMP)$ et dE est l'ensemble de paires de composantes $\{X, Y\}$ pour lesquelles il existe une arête $\{x,y\}$ dans $CLAS$ avec x dans X et y dans Y. Nous définissons aussi une fonction dp sur dE par :

$$dp(X,Y) = \Sigma\{p(\{x,y\}): \{x,y\} \in CLAS, x \in X, y \in Y\}$$

On applique alors à *dV* la méthode de classification hiérarchique par saut minimal en utilisant l'indice de dissimilarité *d* défini par:

$$d(X,Y) = 1/dp(X,Y) \text{ si } \{X,Y\} \in dE \text{ et } d(X,Y) = +\infty \text{ sinon.}$$

4.3. Implantation de la méthode de classification

La méthode CPCL a été implantée en langage AWK ce qui permet son insertion dans un shell script sur tout système d'exploitation. Les graphes sont représentés sous forme de listes d'adjacence auxquelles on accède rapidement grâce à des tables de hachage. L'étape ultime de classification hiérarchique se fait non pas par un balayage d'un tableau de dissimilarités, mais par un parcourt en largeur des arêtes d'un graphe ce qui, dans ce cas, réduit considérablement la complexité de la classification.

5. Les propriétés des classes

Les classes obtenues par la méthode CPCL peuvent être caractérisées par ces propriétés : taille, centralité, indices de variation (interne et externe) et indice de transformation.

Les deux premières, assez classiques, permettent d'appréhender l'importance d'un thème dans le corpus.

1. *Taille :* Une des spécificités de la méthode CPCL est qu'elle n'impose pas de taille maximale aux classes. Elle évite ainsi l'écueil d'une coupure artificielle des classes. De ce fait, les classes que nous obtenons sont de tailles très variables, ce qui peut indiquer l'importance d'un thème dans le corpus.
2. *Centralité* : Cette notion a ici la même acceptation que dans la méthode des mots-associés [Callon et al. 1991]. La centralité donne une indication de la position d'une classe vis-à-vis des autres classes. Elle reflète l'organisation des thèmes du corpus et permet de désigner comme thèmes centraux, les classes ayant une valeur élevée de liens externes et comme thèmes périphérique ou isolés, les classes à valeur faible (ou nulle) de liens externes.
3. Plus spécifique à la méthode CPCL, les classes sont décrites par leurs activités de variation interne et externe.
   - *L'indice de variation interne,* Int_i donne une indication de la force d'activité interne de variation dans une classe. C'est le rapport R_i/T_i entre la somme R_i des liens de variation interne à la classe i et le nombre total T_i de termes dans cette même classe.
   - *L'indice de variation externe,* Ext_j indique le degré d'activité de variation entre une classe et les autres classes. Il peut conforter l'information donnée par la centralité. Il

$$Ext_i = \frac{T^-_i}{T_i} \times \frac{T^+_i}{T}$$

   est formulé ainsi :

   où T^-_i est le nombre de termes de la classe i en relation avec des termes à l'extérieur de celle-ci ; T_i est le nombre total de termes dans la classe ; T^+_i est le nombre de termes à l'extérieur de la classe en relation avec un terme dans celle-ci et T est le nombre total des termes considérés.

4. Indice de transformation

Il mesure le degré de transformation d'une classe sur deux périodes différentes. Il sert à suivre l'évolution des thématiques dans le temps et est calculé par couple de périodes. Il s'agit du rapport $TRANS_{ij} = V_{ij} / (N_{ij}^2 + 1)$ qui emploie deux paramètres : le nombre N_{ij} de termes communs entre la classe i et la classe j et la somme V_{ij} des liens de variation entre la classe i et la classe j. Cet indice sera d'autant plus élevé que deux classes i et j à deux périodes différentes P_i et P_{i+1} ne possèdent pas de terme commun mais partagent beaucoup de liens de variations. Avec ces notations, le degré de transformation $TRANS_i$ de la classe i pour les périodes P_k et P_{k+1} est la moyenne des degrés de transformations $TRANS_{ij}$ par rapport aux classes j de P_{k+1} auxquelles i est liée.

6. Application à un corpus scientifique sur la panification

6.1. Affinage du filtrage des termes

Au terme de l'extraction des termes (cf. §1.2), nous avions retenu 5268 termes qui ont été soumis à la méthode de classification décrite en §3. Après des essais préliminaires, nous avons remarqué qu'une classe grossissait fortement (plus de 600 termes !) très tôt dans les phases de classification (dès la 2ème itération). Un examen attentif de son contenu a montré que cette agrégation était due à un phénomène de variation abondante provoqué par les termes binaires (de longueur 2). Ces termes étaient en relation avec beaucoup d'autres termes sans que la classe ainsi constituée soit réellement pertinente du point de vue des thèmes contenus. Nous avons donc procédé à un nouveau filtrage a posteriori du corpus afin d'éliminer certains termes binaires très généraux par rapport au domaine. Il s'agissait des termes ayant servi à élaborer l'équation de recherche pour la constitution du corpus. Ce sont essentiellement des termes ayant pour centre, les mots *dough*, *property*, *flour*, *bread*, *production*, *characteristic*, *quality*, *product*,… De ce deuxième filtrage, nous avons retenu 3651 termes qui feront l'objet de la classification.

6.2. Résultats de la classification

Le tableau ci-dessous montre le nombre de classes obtenues à différentes itérations.

Itérations	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Nb. classes	44	33	28	25	21	15	13	13
Nb. termes	938	996	1032	1093	1156	1196	1216	1216
Taille de la plus grande classe	160	218	309	336	916	1099	1143	1143

Tableau 1. Caractéristiques des classes obtenues à différentes itérations de l'algorithme

- *Remarques*

Le nombre total des termes impliqués par instance de classification n'augmente pas sensiblement à travers les 9 itérations (gain de 278 termes de la 2ème à la 9ème itération).

L'algorithme converge à la 9ème itération, les classes restent stables. Dès la 2ème itération, on remarque une assez grande classe (avec 160 termes) qui grossira pour représenter un tiers des termes à la 4ème itération et ¾ des termes à la sixième itération. L'explication de ce phénomène ne se réduit pas à l'utilisation d'un saut minimal et suggére que le vocabulaire de ce corpus est

fortement connexe. Il y a un grand réseau de variantes de termes et toute la difficulté va consister à choisir une itération qui offre une vision cohérente des classes. Au vu de ces données et après examen des résultats, nous avons choisi la 3ème itération qui offrait le meilleur rapport entre le nombre total de termes classés (regroupés avec des termes de centres différents) et le nombre maximal de termes dans une même classe (218 termes dans ce cas). Le tableau 2 ci-dessous résume la composition de certaines classes obtenues à la 3ème itération.

Classe	Nb. termes	Exemples de termes
1	13	bromate measurement, wheat dough surface stickiness
4	5	bread firmness, crumb firmness, white pan bread firmness, bread softener
13	53	sour dough fermentation quotient, cold storage stability, frozen bread dough stability, barley flour sour dough fermentation, yeast dough fermentation
15	4	dough pump, new dough pump, new dough pump from campbell technology inc
19	42	bread crumb texture, crust texture, crust hardness, crumb colour, excellent crust colour, loaf appearance, bread crumb structure, bread crumb elasticity
20	90	potassium bromate addition, low molecular weight glutenin subunit, stabilising agent, liquid batter, bread loaf volume, good loaf volume, flour composition
22	218	new baking technology, poor wholemeal flour baking performance, bread dough leavening, dough quality profile, high protein level, innovative microbial alpha amylase, cell wall degrading enzyme, grindamyl exel 16 bakery enzyme, polyglycerol ester, extending product shelf life, bread shelf life improvement, glutathione effects, honey effects
24	16	dough surface property, dough strength weakening, greater dough elasticity
29	19	arabic bread quality, bran enriched wheat bread quality, white pan bread quality
30	55	bakers' yeast strain, freeze tolerant yeast strain, lactobacillus brevis strain new yeast strain development, commercial bakers' yeast, torulaspora yeast
32	198	bread roll baking trial, practical baking test, bakery product quality, microwave baking, baking powder, raising capacity, crisp roll, nutritional wheat flour dough property, physicochemical wheat flour pentosan property, hard red winter wheat lama wheat, aqueous wheat flour dough phase, soft red winter wheat var

Tableau 3. Composition de quelques classes à la 3ème itération.

- *Remarques*

Le contenu des classes semblent cohérent même pour un non spécialiste. On peut identifier des groupes de classes autour d'une même thématique. Ainsi, les deux classes les plus grosses, 22 et 32, semblent traiter de la technologie boulangère (*new baking technology*), des enzymes, de la durée d'exposition du pain (*bread shelf life improvement*) et des tests de fabrication et de la qualité du produit. La classe 22 est par ailleurs liée à la classe 20 qui traite du volume du pain (*bread loaf volume*). La classe 30, liée à la classe 32 développe le thème de la levure (*yeast*). La qualité de la croûte (*crumb firmness, bread crumb texture*) est représentée par les classes 4 et 19. Cette dernière est liée à la classe 24 qui est formée autour du thème de la propriété du pâton (*dough strength weakening, greater dough elasticity*). Les problèmes de la congélation du pâton apparaît dans la classe 13, par ailleurs liée aux classes 24, 19, 22 et 32. La figure 2 ci-après montre ces liens externes entre les classes.

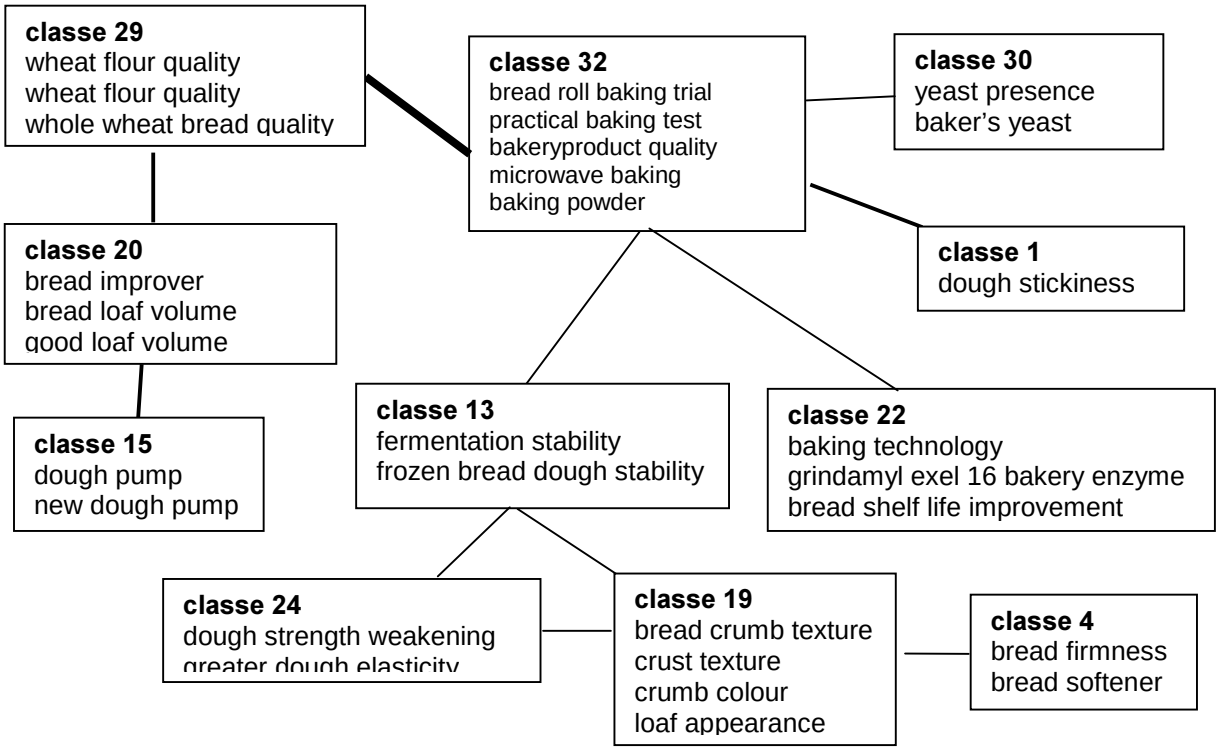

Figure 2. Quelques classes partageant des liens externes

7. Conclusion

Nous avons présenté une méthodologie pour la classification automatique des textes scientifiques et techniques qui exploite le plus possible, les relations linguistiques que les unités textuelles pertinentes partagent entre elles. L'ensemble des étapes de cette méthodologie a donné lieu à une implantation informatique qui est encore à l'étape de prototype. Nous avons expérimenté cette méthodologie sur un nouveau corpus et les résultats semblent encourageants. Nos efforts futurs doivent également porter sur le développement d'outil de visualisation des classes pour en faciliter leur interprétation.

Références

Bourigault D. (1994). LEXTER, un Logiciel d'Extraction Terminologique. Application à l'acquisition des Connaissances à partir de textes. Thèse de doctorat, Ecoles des Hautes Etudes en Sciences Sociales, Paris, 352p.

Callon M., Courtial J-P. Turner W. (1991). La méthode Leximappe : un outil pour l'analyse stratégique du développement scientifique et technique, in *Gestion de la recherche : nouveaux problèmes, nouveaux outils*, dir by VINCK, Boeck Editions, Bruxelles, pp. 207-277.

Condamines A., Amsili P. (1993) Terminology between language and knowledge. An example of terminological base. *Actes TKE'93, Terminology and Knowledge engineering*. Frankfurt, Indeks-Verlag, 316-323.

Daille B. (1994). Study and implementation of combined techniques for automatic extraction of terminology. The Balancing Act : Combining Symbolic and Statistical Approaches to Language, in Proceedings of the *"Workshop of the 32nd Annual Meeting of the ACL"*, Las Cruces, New Mexico, USA, 9p.

Hamon T., Nazarenko A. (1998). A step towards the detection of semantic variants of terms in technical documents. Proceedings of the joint conference of ACL-COLING'98, 10-14 Montréal, 498 - 504.

Ibekwe-SanJuan F. (1998a). 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, pp. 170-174.

Ibekwe-SanJuan F. (1998b). Terminological variation, a means of identifying research topics from texts. *Joint International Conference on Computational Linguistics (COLING-ACL'98)*, Montréal Québec, 10-14, août 1998, 564-570.

Ibekwe-SanJuan F., SanJuan E. (1999). L'analyse formelle de corpus terminologiques, in *Actes Société Francophone de Classification (SFC'99)*, Nancy, septembre 1999, 155 - 162.

Ibekwe-SanJuan F. (2001). Extraction terminologique avec INTEX. *Journées INTEX*, Bordeaux 10-11 juin 2001, 13p. *A paraître*.

Jacquemin C., Royauté J. (1994). Retrieving terms and their variants in a lexicalized unification-based framework. *ACM-SIGIR 94*, Dublin, juillet, pages 132-141.

Lebart L., Salem A. (1994) *Statistique textuelle*. Ed. Dunod, 342p.

Lelu A. (2001) Synthèse d'information en ligne : bilan du prototype NeuroWeb. *3ème Congrès du Chapitre français de l'International Society for Knowledge Organisation* (ISKO), Paris, 5-6 juillet 2001, 187-195.

Reinert M., (1990) ALCESTE : une méthodologie d'analyse des données textuelles et une application : Aurélia de G. de Nerval. Bulletin Méthodologie sociologique, n° 26.

Silberztein M. (1993) *INTEX*© manual, 2000-2001. ASSTRIL - LADL, 201p.