île Rousse 2005
Journée sur les systèmes d'information élaborée

MESURE QUALITATIVE DE LA SIMILARITE DE CLASSES PRODUITES PAR DES METHODES D'AGREGATION DIFFERENTES

Fidelia Ibekwe-SanJuan
ibekwe@univ-lyon3.fr

Université de Lyon 3 - 4, cours Albert Thomas - 69008 Lyon

Mots clés : méthodes d'agrégations, comparaison de classes, relations linguistiques, associations lexicale, indice de similarité

Résumé.
La comparaison de classifications produites par des algorithmes d'agrégation différents est un problème complexe. A notre connaissance, il n'existe pas de protocole ou de métrique directs pour comparer de telles classifications. Il en résulte une quasi-impossibilité de discuter des apports respectifs des méthodes existantes, dès lors que les métriques habituelles (Rand Index) ne peuvent pas s'appliquer. Dans cette communication, nous proposons des critères qualitatifs de comparaison de classes issues de méthodes d'agrégation différentes. Nous avons pris deux situations extrêmes dans lesquelles il subsiste d'importantes différences dans les deux méthodes, l'une est basée sur la méthode des mots associés et l'autre sur les relations linguistiques. Des différences existent à plusieurs niveaux dans ces deux méthodes : au niveau des entrées à l'algorithme de classification, au niveau de la taille des matrices, au niveau de la taille des classes produites et enfin au niveau des paramètres de classification. Le but étant d'arriver à classer en tête, par un indice de relation fondé sur l'association lexicale des éléments des classes, les classes des deux méthodes dont les contenus seraient les plus similaires. La similarité être comprise ici dans son acception linguistique. Les opérations d'associations lexicales qui permettent de rapprocher deux éléments de deux classes donnent lieu, sur le plan sémantique à des relations d'équivalence (correspondance lexicale exacte), de synonymie (via une ressource externe telle WordNet), d'hyperonymie ou d'hyponymie et enfin à des relations d'association. Les résultats expérimentaux montrent la pertinence de la démarche.

1. Introduction

La comparaison de classifications produites par des algorithmes d'agrégation différents est un problème complexe. Il est étudié dès les années '80 par Milligan & Cooper, (1986) puis par Jain & Dubes, (1988). Ce problème se complique lorsqu'il existe d'importantes différences à plusieurs niveaux dans les algorithmes concernés : métrique utilisée (distance ou similarité), paramètres fixés (taille et nombre de classes, seuils d'occurrences et de co-occurrences, nombre d'itérations pour les méthodes de classification automatique). Une des exigences de base soulignée dans des études visant à comparer les classes produites par des algorithmes différents (Jain & Dubes, 1988 ; Zamir & Etzioni, 1998 ; Yeung *et al.*, 2001) est que ces algorithmes doivent produire le même nombre de classes, et si possible de taille équivalente. Dans Zamir & Etzioni (1998), pour comparer des méthodes d'agrégation différentes sur la même collection de textes, les auteurs choisissent un nombre constant de documents dans chaque classe. Lorsque cette exigence de taille équivalente est vérifiée, il devient possible de calculer un degré de similarité entre les classifications produites par des méthodes différentes et une classification de référence (*gold standard*). Cette démarche s'inscrit dans une approche dite de 'critère externe' (external criteria). L'indice couramment utilisé est le 'Rand Index' (Rand, 1971) ou sa version améliorée "the adjusted Rand index" (Hubert & Arabie, 1985). Or, l'utilisation du 'Rand Index' suppose l'existence préalable de la classification de référence. Cela suppose que des experts se mettent d'accord sur les classes à former à partir d'un corpus et sur le contenu de chaque classe. Cette situation est rarement atteignable pour plusieurs raisons : l'indisponibilité des experts, le coût en temps et en argent, le niveau variable de maîtrise du corpus par les experts, la difficulté de parvenir à un consensus. S'ensuit en aval de la constitution de la classification de référence, l'impossibilité de mesurer le silence, *a fortiori* sur des gros corpus. Par ailleurs, Pantel & Lin (2002), observent que le Rand Index peut donner lieu à des résultats contre-intuitifs. En l'absence d'une classification de référence, il est usuel dans la pratique, de recourir à une expertise humaine (Schiffrin & Börner, 2004) pour valider les résultats des méthodes d'agrégation. Cette pratique a pour inconvénients majeurs d'être subjective, arbitraire, difficilement modélisable et reproductible d'une expertise à une autre.

Les approches courantes de comparaison des résultats de méthodes d'agrégation ne permettent pas la comparaison lorsque les résultats sont différents. Notamment, les approches numériques de comparaison (Rand index) sont directement inapplicables lorsque deux méthodes prennent en entrée des unités d'information différentes (termes extraits du corpus, mots-clés issus d'un thésaurus ou mots isolés), adoptent une représentation différente (matrices de co-occurrences vs graphes de relations) qui sont de tailles très différentes et produisent des classes de tailles également très différentes.

Il est nécessaire de rechercher d'autres modalités de comparaison qui ne soient ni entièrement humaines, ni subordonnées à l'existence d'une classification de référence. Dans cette perspective, notre objectif ne sera pas tant d'évaluer les résultats (dire quelle méthode est meilleure que l'autre) mais de déterminer le degré de similitude dans le contenu des classifications proposées par différentes méthodes. Il s'agit donc plus de la comparaison des vocabulaires des deux classifications que de la comparaison « formelle » des résultats, au sens mathématique. Un usager qui lance plusieurs méthodes sur un même corpus va en effet pouvoir déduire quelque chose du contenu des différentes classifications, même en présence d'espaces mathématiques différents. Notre proposition est de l'aider dans cette tâche de comparaison en l'accélérant. Pour cela, nous nous appuyons sur des indices linguistiques inhérents aux unités de compte. Par conséquent, cette méthode ne peut s'appliquer qu'aux cas de classification portant sur des unités textuelles (termes, mots-clés).

Afin de tester l'applicabilité de notre méthode de comparaison, nous l'appliquerons aux classes issues de deux méthodes qui prennent en entrée des unités textuelles : la méthode des mots-associés (Callon et al., 1983), et plus spécifiquement la version implantée dans le logiciel SDOC sur la station STANALYST® (Polanco *et al.*, 2001), et la méthode d'agrégation par relations linguistiques TermWatch (Ibekwe-SanJuan, 1998 ; Ibekwe-SanJuan et SanJuan, 2004). Les deux méthodes ont été appliquées à un même corpus de 3355 titres et résumés des articles sur la recherche d'information en anglais (désormais corpus IR). Le corpus a été collecté sur la base PASCAL. SDOC a produit des classes à partir des mots-clés ayant servi à indexer les articles. TermWatch a produit des classes à partir des termes extraits des titres et résumés du corpus.

2. La méthodologie de comparaison proposée

Les critères de comparaison que nous retenons concernent les relations lexico-sémantiques pouvant exister entre le contenu des classes à comparer. Deux types d'indices seront examinés : la correspondance lexicale de surface et les relations sémantiques induites par des phénomènes d'association lexicale. Ces deux indicateurs s'appliquent au contenu mais aussi aux libellés des classes. En effet, le libellé attribué à une classe par une méthode d'agrégation est généralement l'unité la plus représentative au regard d'un critère donné. Ce critère peut être le nombre de liens internes, dans ce cas, le libellé est censé représenter le contenu de la classe. Si le critère choisi est le nombre de liens externes, le libellé caractérise davantage la classe par rapport aux classes avoisinantes. Ainsi, le degré de similitude des libellés est un indice intéressant, à côté du degré de similitude du contenu des classes produites par deux méthodes différentes.

2.1 Correspondance lexicale exacte

Le premier test intuitif de similarité est de mesurer le nombre d'unités communes à deux classes des deux méthodes différentes. Pour tenir compte des écarts importants entre la taille des classes produites par les deux méthodes, nous normalisons ce nombre en le divisant par la taille de chacune des classes. Pour tenir compte des phénomènes de variations morphologiques dans la forme des termes, deux types d'appariement sont envisagés :

- *correspondance lexicale stricte* : les termes ou libellés des deux classes c_i et c_j de deux méthodes correspondent, modulo des variations flexionnelles (nombre, genre et temps). Pour éviter que les phénomènes de flexion gênent cette reconnaissance, il faut prendre en compte les versions lemmatisées des mots[1]. On note LSC le nombre de libellés de classes communs stricts aux deux classifications et T_{ij} la proportion de termes communs aux classes c_i et c_j.

Le nombre de libellés/termes communs est la relation sémantique la plus forte (avec la synonymie), il convient donc de l'« avantager » dans le calcul par rapport à d'autres types de relations qui induisent une proximité sémantique plus faible.

- *correspondance lexicale partielle ou inclusion lexicale* : les éléments d'un terme/libellé de la classe c*i* de la méthode A sont inclus dans un terme/libellé de la classe c*j* de la méthode B ou vice versa. La correspondance lexicale partielle permet de prendre en compte des variations dans la stratégie d'extraction des unités de compte. A titre d'exemple, dans le système Termwatch, nous privilégions l'extraction de termes complexes du domaine qui peuvent être des séquences assez longues (*valued propositional truth value logic lvpl*). Ceci résulte en un faible taux de correspondance exacte entre nos termes et ceux d'une méthode privilégions l'extraction d'unités plus simples, voire des mots-isolés. La détermination de la relation induite entre termes lorsqu'il y a une correspondance partielle est traitée dans la section suivante car cet indice est combiné avec d'autres indices linguistiques qui permettent de déterminer la relation sémantique exacte ainsi que le degré de proximité sémantique induite entre deux termes.

2.2 Association lexicale et proximité sémantique

La recherche de proximité sémantique entre termes constitue le deuxième niveau de comparaison. Nous distinguons deux cas de figure : cas où la détection des termes en relation fait appel à l'usage de ressource sémantique externe (2.2.1) et les cas où la proximité sémantique se déduit directement du type d'opération lexicale rencontrée entre deux termes issus de deux classes différentes (2.2.2 – 2.2.3).

2.2.1 Relations de synonymies entre termes de même longueur (substitution forte)

Elle nécessite le recours à une ressource sémantique externe tel le thésaurus ou une base sémantique générique. Dans l'idéal, on privilégiera l'usage de ressources spécialisées du domaine mais l'on se heurte vite à deux obstacles majeurs : les ressources sémantiques n'existent pas pour tous les domaines et lorsqu'elles existent, elles peuvent ne pas être accessibles. Devant cet écueil, l'alternative consiste à

[1]Ceci est possible dans le cadre d'extraction automatique des termes. Les lemmes sont indiqués dans le dictionnaire utilisé par l'extracteur de termes.

utiliser des ressources sémantiques générales. De telles ressources ont été construites pour plusieurs langues : anglais (WordNet, Fellbaum, 1999) et EuroWordNet (Vossen, 1998) pour les autres langues européennes. WordNet définit pour chaque mot de la langue anglaise ses classes d'équivalences sémantiques appelées "synsets". Chaque "synset" regroupe tous les mots ayant le même sens que le mot considéré. Ainsi, les mots dans un même synset peuvent être considérés comme des synonymes.
WordNet établit également d'autres types de relations sémantiques entre mots : des relations hiérarchiques, (hyperonymie/hyponymie ou générique/spécifique) et des relations d'association. Des travaux résumés dans Pedersen *et al.*, (2004) proposent différents indices pour calculer la similarité sémantique entre deux mots dans WordNet. Le problème qui se pose pour nous est d'étendre les calculs de similarité « mot - mot » à « terme - terme ».
Pour identifier des termes synonymes, nous procédons de la manière suivante : étant donné deux termes de même longueur qui ne diffère que par un mot (des substitutions), nous vérifions si les deux mots différents appartiennent au même 'synset' de WordNet. Dans le cas positif, une relation de synonymie est établie entre les deux termes. Ainsi, le terme «*impact study* » de la classe « *user service* » produite par la méthode des mots associés (SDOC) est lié au terme «*impact survey*» produite par une des classes de la méthode d'association lexicale (TermWatch). Nous appelons cette relation la « *substitution forte* » qui correspond sur le plan sémantique à une relation de synonymie. Pour calculer la proportion de termes synonymes entre classes issues de deux méthodes différentes, nous utiliserons une notations analogue à celle de la correspondance lexical : ainsi on notera T_{SYNij} le nombre de relations de synonymie entre classes c*i* et c*j* de deux méthodes différentes. Par la suite, on accordera la même importance aux relations d'équivalence lexicale et de synonymie parce qu'elles induisent la plus grande proximité sémantique entre termes. Ce calcul, étendu aux libellés est noté L_{SYN}.

2.2.2 Relations d'hyperonymie/hyponymie et d'association entre termes

Il s'agit ici des cas d'inclusion lexicale ne nécessitant pas le recours à une ressource sémantique extérieure. Dans ce qui suit, nous recherchons des relations sémantiques entre termes qui partagent au moins un élément lexical en commun (même mot) ou dont les mots proviennent d'une même famille morphologique.
L'inclusion lexicale exprime le fait que les éléments d'un terme/libellé de la classe c*i* de la méthode A sont inclus dans un terme/libellé de la classe c*j* de la méthode B. A titre d'exemple, le terme "*approximation operators*" est lexicalement inclu dans le terme "*rough set approximation operators*". Selon la fonction grammaticale des éléments communs (centre ou modifieur[2]), la relation sémantique engendrée est différente. Deux cas se présentent :
i- *Inclusion lexicale avec centre commun* (*expansion gauche ou insertion*) *: relation d'hyperonymie/hyponymie*
Dans ce cas, le terme inclus est le hyperonyme (générique) et le terme englobant est l'hyponyme (spécifique). Ces exemples illustrent cette hypothèse :

citation analysis --> TS web-based citation analysis
automatic translation --> TS automatic query translation

Le premier est un cas d'expansion gauche et le deuxième un cas d'insertion. Ce type de relation, notée T_{Mij} vient en seconde position (après les termes communs et synonymes) par degré de proximité sémantique induite. De manière analogue, on notera L_{MAB} le nombre de ces relations entre libellés.

ii- Inclusion lexicale avec modifieurs communs (*expansion droite ou gauche-droite*) *: relation d'association*
Lorsque la partie commune concerne les éléments modifieurs du terme, la relation engendrée est l'association (*VOIR AUSSI*). Ces exemples illustrent cette hypothèse :

african American VA african American households
authentic reasoning VA authentic reasoning expert systems
clustering algorithm VA robust hierarchical clustering algorithm ROCK

Les deux premiers exemples sont des cas d'expansion droite. Le troisième est un cas d'expansion

[2]Le modifieur s'oppose au centre du terme dans la mesure où il est le qualificatif du terme. Un terme peut avoir plusieurs modifieurs.

gauche-droite. On notera T_{Cij} de relations d'association entre termes de deux classes différentes et L_{cAB} le nombre de ces relations entre libellés.
Pour tenir compte de la signification respective des deux types de relations – hyperonymie/hyponymie (L_{MAB} , T_{Mij}) d'une part et association (L_{cAB} , T_{Cij}) d'autre part, nous pénaliserons cette dernière dans le mode de calcul puisqu'elle introduit une plus grande distance sémantique entre termes que la relation hyperonymie/hyponymie.

2.2.3 Association lexicale sans relations explicite entre termes de même longueur

Il nous reste à considérer un dernier cas : le cas où deux termes de même longueur partagent des éléments communs sans être dans une des relations sémantiques définies ci-dessus. Il renvoie à la situation suivante : deux termes d'une même longueur ne diffèrent que par deux mots qui ne sont pas des synonymes (ne sont pas dans le même synset de WordNet). Nous appelons cette catégorie de variantes les « *substitutions faibles* » car il s'agit d'une substitution lexicale sans lien sémantique explicite. En fonction du rôle grammatical de l'élément différent, la proximité sémantique sera plus ou moins faible. Si la substitution porte sur un élément modifieur (*academic library* ↔ *public library*), la proximité sémantique entre les deux termes est plus forte que si la substitution concerne l'élément centre (*british library* ↔ *british musuem*). Dans les deux cas, on peut parler d'une « association » au sens général : association de propriétés ou d'objets pour la substitution de modifieurs et association de concepts pour la substitution de centres. La relation induite par la substitution de modifieurs créée plus de distance sémantique entre termes que l'hyperonymie/hyponymie mais moins celle créée par la substitution de centre. Donc, la substitution de modifieurs est placée à mi-chemin entre l'hyperonymie/hyponymie et substitution de centre. C'est la raison pour laquelle le calcul de son indice le pénalise moins fortement que celui calculant l'inclusion lexicale avec modifieurs communs.

i- *Le nombre de substitution de modifieurs* est noté $T_{SUB\text{-}Mij}$ tandis que le nombre de ces variantes entre libellées de classifications différentes est noté : $L_{SUB\text{-}MAB}$.
i- *Le nombre de substitution de centre* est calculé de manière analogue au calcul du nombre d'association impliquant un changement un changement de centre, donc la même notation est applicable aux termes et aux libellés ce qui donne respectivement : $T_{SUB\text{-}Cij}$ et $L_{SUB\text{-}CAB}$.

2.3 Pondération globale

Ces indices ont été élaborés en tenant compte de la signification de chaque type de relation sémantique entre termes de deux classifications différentes. Les relations sémantiques les plus fortes ont été avantagées tandis que les relations moins fortes ont été pénalisées. Ainsi, les indices de termes/libellés équivalents et synonymes seront les plus élevés, suivi des indices calculant la proportion de relations d'hyperonymie/hyponymie, ensuite l'indice de relation d'association sans changement de centre, et enfin celui de relations d'association avec changement de centre. En additionnant les résultats des calculs de ces différents indices, pondérés en fonction de la proximité sémantique qu'ils induisent et en normalisant le résultat pas le nombre total de termes dans les classes considérées on obtient un indice global de relation (REL_{ij}) entre termes de deux classes produite par deux méthodes différentes.
REL_{ij} :

$$\frac{T_{ij}+T_{SYN_{ij}}+\frac{T_{M_{ij}}}{2}+\frac{T_{SUBMij}}{3}+\frac{T_{Cij}}{4}+\frac{T_{SUBCij}}{5}}{T_{ci}+T_{cj}}$$

De même en combinant les indices pour les libellées, on obtient un indice général REL_{AB} de proximité entre libellés de deux classifications différentes.
REL_{AB} :

$$\frac{L_{AB} + L_{SYN_{AB}} + \frac{L_{M_{AB}}}{2} + \frac{L_{SUBM_{AB}}}{3} + \frac{L_{C_{AB}}}{4} + \frac{L_{SUBC_{AB}}}{5}}{L_A + L_B}$$

où T_{ci} est le nombre de termes dans la classe *ci* et T_{cj} le nombre des termes dans la classe *cj*. L_A et L_B donnent ces mêmes informations pour les libellés des deux classifications.
Observons qu'en introduisant au dénominateur de ces formules le nombre total de termes dans une classe, nous tenons compte de la proportion des termes non-reliés dans ces deux classes. Il en va de même pour les libellés.

3. Application aux classes produites par deux méthodes différentes

Nous avons calculé les différentes relations décrites ci-dessus entre classes produites par TermWatch (Ibekwe-SanJuan, 1998 ; Ibekwe-SanJuan & SanJuan, 2004) et celles de SDOC (Polanco *et al.*, 2001). Nous décrivons brièvement les principes d'agrégation de chacune de ces méthodes avant de présenter les résultats de la comparaison.

3.1. SDOC : la méthode des mots-associés

SDOC est fondé sur l'analyse de la matrice de co-occurrence des mots-clés employés pour décrire les documents. Pour construire les réseaux des associations de mots, la première étape consiste à calculer le nombre d'occurrences m_i de chaque mot-clé *i* dans l'ensemble d'articles et le nombre de co-occurrences m_{ij} de chaque paire de mots-clés m_i et m_j. On calcule donc un coefficient d'association normalisé :

$$E_{ij} = m_{ij}^2 / m_i * m_j$$

c'est-à-dire la co-occurrence au carré des mots-clés *i* et *j*, divisée par le produit de leurs fréquences respectives. Il vaut 0 si les mots-clés *i* et *j* n'apparaissent jamais simultanément et 1 dans le cas inverse. Dans ce cas, on a l'égalité : $m_{ij} = m_i = m_j$. Par ailleurs, ce coefficient est analogue aux indices bien connus de Dice, de Jaccard et de Salton. SDOC s'appuie sur l'algorithme de classification ascendante hiérarchique (CAH) à simple lien ou du saut minimal (single linkage). Si *x, y, z* sont trois objets, et si les objets *x* et *y* sont regroupés en un seul éléments noté *h*, on peut définir la distance de ce groupement à *z* par la plus petite distance des divers éléments de *h* à *z* : $d(h,z) = \text{Min } \{d(x,z), d(y,z)\}$. Cette distance s'appelle le simple lien ou le saut minimal.

Lors de l'agrégation SDOC adapte les principes de l'agrégation par lien simple afin de produire des classes de taille homogènes, non emboîtantes. Il ne produit donc une hiérarchie de classes. Du fait que les paramètres de la classification implique la fixation d'une taille pour les classes (souvent 10), les classes sont toutes de même niveau. Il s'agit plutôt d'une partition des liens entre mots-clés par taille homogène, en commençant par les liens les plus forts.

Du fait de cette coupure des liens, il existe à la fois des associations internes aux classes (intra-classes) et des associations externes (inter-classes). Les classes sont disposées sur un diagramme stratégique reflétant leur centralité et densité. SDOC permet aussi de visualiser le contenu des classes sous forme de graphes. Il est accessible via la station STANALYST®, développée par l'Unité de Recherche et Innovation de l'INIST/CNRS. Un module d'indexation automatique des textes a été développé, ce qui permet à la plateforme de travailler directement sur des unités textuelles extraites du corpus en faisant une extraction terminologique. Dans cette expérience, nous avons utilisés les mots-clés issus du lexique PASCAL pour la classification. Il s'agit donc d'un vocabulaire contrôlé, non issu du corpus.

3.2. TermWatch : méthode d'agrégation par relations sémantiques

Termwatch dispose d'une chaîne de traitement linguistique permettant d'extraire automatiquement, à partir des textes, les termes. Le système classifie directement les termes, automatiquement extraits du corpus, en fonction de leurs seuls liens lexico-sémantiques, donc sans utiliser de notion d'occurrence ou de co-occurrence d'unités textuelles dans les textes. Les relations utilisées pour l'agrégation sont distinguées selon leur signification et le but recherche par l'agrégation.

- Une première catégorie, appelée COMP comprend des relations sémantiques fortes (expansion gauche, substitution forte filtrée par WordNet),
- Une deuxième catégorie, appelée CLAS comprend les relations qui induisent plus de distance sémantique entre termes (substitution faible, insertion, expansions droite, expansions gauche-droite).

Ces relations sont identiques à celles utilisées pour la comparaison de conteu des classes (§2.2.1 – 2.2.3). La technique de classification implémentée dans TermWatch, nommée CPCL (Classification by Preferential Clustered Link) est une variante de la classification ascendante hiérarchique (CAH) adaptée spécialement à la réduction d'un graphe de termes reliés par un nombre illimité de relations linguistiques. L'agrégation se déroule en deux étapes.

L'étape 1 est celle de la réduction du graphe initial pour former des connexes au sens de la théorie des graphe. Ces composantes sont formées avec les relations de COMP.

L'étape 2 agrège ces composantes en « clusters » (classes) à l'aide des relations de CLAS. La technique d'agrégation s'effectue selon le principe de l'agrégation par lien simple. Pour ce faire, il est nécessaire de calculer 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 $\{1,\ldots,k\}$, $N_r(i,j) = 0$;

ii) $d(i,j) = 0$ si i = j ;

iii) $d(i,j) = 1 / \sum_{r=1}^{k} \frac{N_r(i,j)}{|R_r|}$ où $R_1 \ldots R_k$ 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.
L'ensemble du système fonctionne dans une approche de «classification non supervisée » et permet de réaliser une analyse thématique d'un corpus textuel. TermWatch n'a théoriquement pas de limite sur la taille initiale du graphe (sauf celle imposée par la performance des machines). Les résultats sont présentés sous forme d'un graphe que l'on peut visualiser et explorer avec le logiciel AiSee®. TermWatch est une plateforme développée conjointement par l'Université Lyon3 et le laboratoire LITA de l'Université de Metz.

Dans cette tâche de comparaison, les représentations graphiques des classes des deux méthodes n'ont pas été utilisées. Seule une comparaison de la similarité du contenu des classes est effectuée.

3.3 Convergences et divergences des deux méthodes

Les deux méthodes ont en commun le fait de considérer des multi-termes comme unités de compte et non des mots isolés. Les termes peuvent provenir d'une ressource sémantique externe, typiquement le thésaurus dans le cas de la méthode des mots associés ou être extraits directement des textes dans le cas de TermWatch. Un deuxième point commun entre les deux méthodes est la technique d'agrégation adoptée. Il s'agit de la classification ascendante hiérarchique (CAH) par saut minimal dans une approche de classification non supervisée. De plus, les deux méthodes étiquettent automatiquement leurs classes par le terme le plus actif, en terme de co-occurrence pour SDOC et en termes de relations linguistiques pour TermWatch. Les classes produites par les deux méthodes représentent des thématiques du corpus. Chaque classe est assimilable à un thème désigné par un libellé automatiquement extrait.
Cependant, de nombreuses différences existent : SDOC démarre la classification à partir d'une matrice de co-occurrences là où TermWatch démarre à partir d'un graphe avec un nombre de relations plus élevé. TermWatch fait intervenir l'algorithme de CAH seulement à la deuxième étape après un regroupement préalable des termes sur des bases linguistiques tandis que l'agrégation se fait au niveau des termes dès le calcul de l'indice d'association dans SDOC. SDOC requiert la fixation d'un seuil d'occurrence et de co-occurrence minimal ainsi que la taille des classes dans SDOC. Dans TermWatch, l'algorithme fonctionne sans fixation de seuil, de nombre de classes ou de la taille de celles-ci. L'utilisateur peut indiquer un seuil de lien minimal et le nombre d'itérations souhaités. Alternativement, il peut laisser l'algorithme converger.

En conséquence, la taille et le nombre de classes produites dans TermWatch sont très variables alors qu'elles ont tendance à être fixes dans SDOC. Dans la présente étude, étant donné que nous nous focalisons sur la comparaison du contenu des classes et non sur la qualité de la représentation graphique (disposition spatiale des classes), on n'analysera pas les liens externes mais uniquement le degré de similitude des classes produites par les deux méthodes du point de vue de leur contenu. Dans ce cas, une simple liste de classes avec leur contenu suffit pour effectuer cette comparaison.
Le tableau suivant donne les détails de la classification produite par chacune des méthodes sur le corpus IR.

Classification par SDOC		**Classification par TermWatch**	
Nombre total des mots-clés dans l'index	5168	Nombre total de termes extraits du corpus	44 665
Fréquence min. mots-clés	>= 2	Nombre d'itérations	1
Co-occurrence des mots clés	>= 2	Nombre de composantes	1595
Nombre de classes	148	Nombre de classes	674
Taille max. d'une classe	10	Taille max. d'une classe	135
Total mots-clés dans classes	1142	Total termes dans classes	5632

Tableau 1. Paramètres des classifications utilisés par les deux méthodes.

Comme nous pouvons le constater, les entrées des deux méthodes sont très différentes, la taille des matrices/graphe de départ également (5168 *vs* 44 665). La ratio des classes produites par les deux méthodes est de 1:6. Bien que les deux méthodes emploient la CAH, le fait de fixer la taille des classes dans SDOC et de la laisser libre dans TermWatch conduit à des écarts importants de taille des classes. La plus grande classe dans TermWatch a 135 termes là où la taille maximale des classes dans SDOC est fixée à 10 mots-clés. Par les métriques habituelles (Rand index), les deux résultats seront incomparables. Mais, étant donné que les deux portent sur des unités textuelles (mots-clés ou termes), le contenu des classes peut être comparé à l'aide des critères qualitatifs que nous avons exposés en section §2.

4. Similitude des classes SDOC et TermWatch (TW)

Les critères de comparaison énoncés en section §2 ont été appliqués aux classes produites par ces deux

méthodes. Les indices REL_{ij} et REL_{AB} ont été calculés pour chaque paire de classe SDOC et TW.

4.1 Similitude des libellés

Le calcul de similitude des libellés porte sur la totalité de la classification puisque il n'y a qu'un libellé par classe. Tout naturellement, l'indice de relation entre libellés (REL_{AB})classe en tête les libellés exactes (en cas d'exéquo, c'est d'ordre alphabétique qui prévaut). L'indice REL_{AB} a permis de mettre en relation 117 couples de libellés. Cela traduit un nombre important de relations strictes entre libellés.

Cependant, un libellé identique entre deux classes ne va pas forcément de pair avec un contenu lexical similaire. En effet, l'indice de relation calculé sur le contenu des classes (RELij) a parfois classé assez loin deux classes ayant un libellé identique. C'est le cas de deux classes étiquetées « f*uzzy logic* » par les deux méthodes qui arrivent à la 43ème position avec un indice de relation de 1.98 et de deux autres classes libellés « *vector space* » qui arrivent à la 118ème place avec un REL_{ij} de 1 (voir §4.2, tableau 3 ci-après). La figure 1 ci-dessous donne les contenus de ces deux dernières classes. Bien qu'elles aient le même libellé, on observe très peu de liens entre termes des deux classes. Les termes à l'origine des liens sont mis en gras.

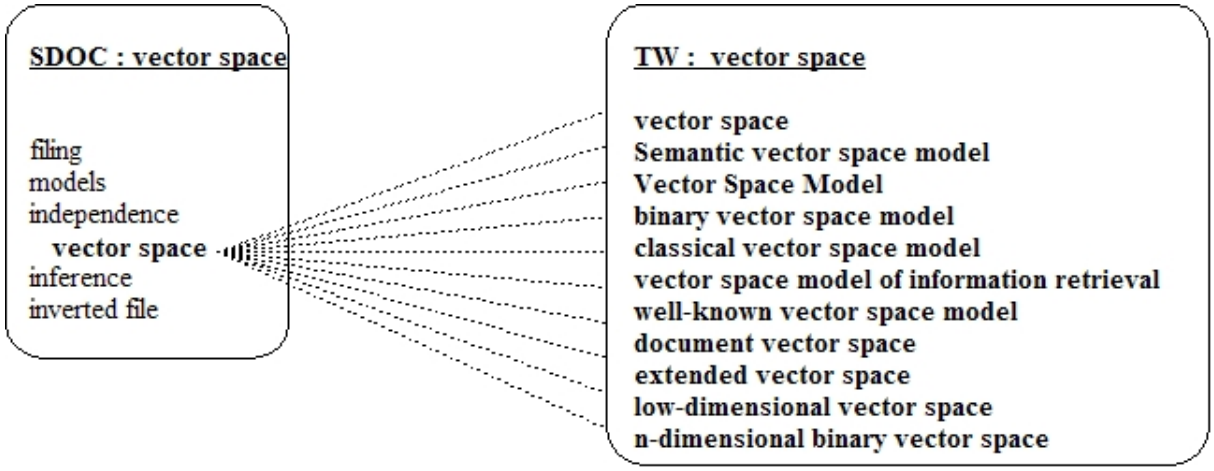

Figure 1. Liens entre termes de deux classes SDOC et TW ayant un libellé identique.

Il est intéressant de noter que tous les termes de la classe TW sont liés au seul mot-clé « *vector space* » de la classe SDOC. Il n'est en effet pas évident de lier thématiquement les autres mots-clés de la classe SDOC (*filing, models, independence*, ...) aux les termes de la classe TW. Ces derniers précisent les différents types de « *vector space model* » dont il est question dans le corpus. Ainsi, la faiblesse de l'indice de similarité attribuée à ces deux classes, en dépit de l'identité de leurs libellés, se justifie.

4.2 Similitude dans le contenu des classes

Notre approche de comparaison a permis de relier 751 couples de classes et 1639 couples de termes. Le tableau ci-après résume les associations lexicales trouvées et les relations sémantiques induites entre termes des deux classifications.

Type d'opération lexicale	***Relation sémantique***	***Nombre de liens***
Correspondance exacte	équivalence	51
Substitution forte	Synonymie	3
Insertion	Hyperonymie/hyponymie	26
Expansion gauche	Hyperonymie/hyponymie	1074
Expansion droite	Association (voir aussi)	483

Type d'opération lexicale	*Relation sémantique*	*Nombre de liens*
Expansion gauche-droite	Association (voir aussi)	785
Substitution faible	Association (voir aussi)	3

Tableau 2. Type d'opérations lexicales entre termes des classes produites par SDOC et par TW.

Les cas de correspondance exacte entre termes dans les résultats des deux méthodes sont infimes. Les relations d'inclusion lexicale sont plus nombreuses entre les mots-clés du lexique PASCAL et les termes du corpus. En effet, les premiers sont souvent des termes binaires (formés de deux mots) lexicalement imbriqués dans les termes plus longs extraits par TW. Le faible nombre de synonymes repérés s'explique par le fait que WordNet est une ressource générale. De ce fait, il arrive fréquemment de nombreux termes d'un corpus spécialisé en soient absents. Ce qui conduit à un taux assez faible de reconnaissance de relations sémantiques.

Le tableau ci-dessous montre le résultat du calcul de REL_{ij} pour les vingt premiers couples de classes de SDOC et TW. Pour vérifier la cohérence globale de ce classement, nous avons ajouté quatre classes prises au hasard plus bas dans le rang afin de comparer leurs indices avec les classes en tête du classement. Sur les 751 paires de classes reliées, les valeurs de l'indice REL_{ij} vont de 16.33 à 0.2. Nous avons confronté les valeurs REL_{ij} à la proportion des termes impliqués dans ces relations pour chaque paire de classes afin de tester une éventuelle corrélation entre le « nombre de termes liés » et « l'indice de relation ». L'indice de relations témoigne de la qualité des liens en termes de proximité sémantique induite.

Rang	RELij	SDOC Libellé	Nb_Kw	TW libellé	Nb_T	Termes liés	Ratio
1	16.33	information retrieval	10	Information retrieval	135	51	0.35
2	6.67	search engine	9	Sophisticated search engines	54	20	0.32
3	5.2	Case study	4	Illustrative case studies	31	16	0.46
4	5	language processing	9	efficient data structure	23	19	0.59
5	4.87	information	8	scientific information environment	19	17	0.63
6	4.67	methodology	8	exploratory data analysis	15	16	0.7
7	4.33	user interface	10	new user interfaces	26	14	0.39
8	4	chemistry	7	scientific information environment	19	13	0.5
9	3.87	decision theory	4	collaborative decision making	23	14	0.52
10	3.33	internet resource	10	three-part research project	16	11	0.42
11	3.33	bibliometric analysis	10	parallel citation analysis	14	7	0.29
12	3.33	algorithm	10	efficient clustering algorithm	23	10	0.3
13	3.2	classification	10	subject classification	19	**24**	**0.83**
14	3	stress	10	three-part research project	16	9	0.35
15	3	set theory	7	generalized rough sets	13	7	0.35
16	3	information processing	7	temporal information processing	10	7	0.41
17	2.87	adaptation	8	Query expansion	15	7	0.3
18	2.73	neural networks	8	problem of query optimization	6	10	0.71
19	2.73	knowledge management	10	Information management	13	7	0.3
20	2.67	knowledge representation	5	fuzzy knowledge representation	10	6	0.4
101	1.27	Segmentation	10	Word segmentation	18	6	0.21
118	1	Vector space	6	Vector space	11	4	0.24
750		advantage	6	isolated computer systems	5	1	0.09
751		Advantage	6	visual programming environments	14	1	0.05

Tableau 3. Indice de relation des premiers couples de classes SDOC et TermWatch (TW).

Les deux premières colonnes donnent respectivement le rang et l'indice REL_{ij} pour deux classes. La colonne « Nb_Kw » indique le nombre total de mots-clés dans la classe SDOC, « Nb_T » donne l'information équivalente pour les termes des classes TW ; la colonne « Termes liés » est le nombre de termes impliqués dans les liens ; la colonne « Ratio » est la proportion de ce nombre sur le nombre total des termes dans les deux classes.
Comme on peut le constater, l'indice de relation (REL_{ij}) n'est pas corrélé avec la proportion de termes impliqués dans les liens. Si l'on considère uniquement les classes dans ce tableau, celles qui seraient les plus similaires par la proportion de termes relies seraient « *classification* » et « *subject classification* » (voir figure 3 ci-dessous) suivies par « *neural networks* » et « *poblem of query optimization* ». Or, comme les liens sont pondérés en fonction de la signification sémantique supposée par chaque type d'opération lexicale, le nombre de termes impliqués n'est pas forcément synonyme d'une plus grande similarité thématique.

Le classement par indice de relation REL_{ij} montre que les deux classes les plus similaires par leur contenu correspondent à la thématique générale du corpus (*information retrieval*). Il est intéressant de constater le lien établi entre la classe SDOC « *bibliometric analysis* » et celle de TW « *parallel citation analysis* » classées 11ème par REL_{ij}, dont les libellés ne sont pas lexicalement associés mais dont les contenus sont bel et bien liés. Les liens sont initiés par l'unique mot-clé « *citation analysis* » et les termes de la classe TW qui l'imbriquent. C'est ce que montre la figure 2 ci-dessous.

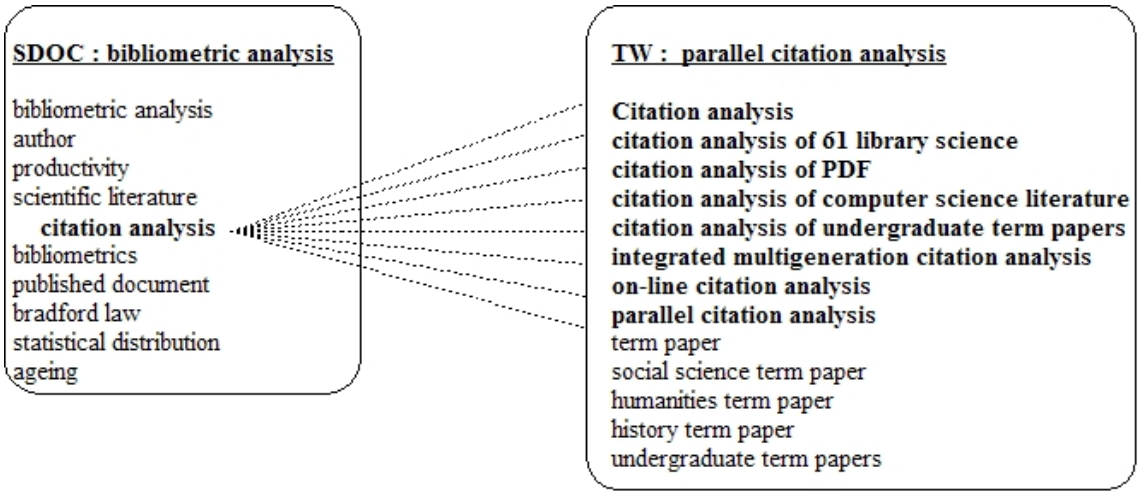

Figure 2. Liens entre la classe « *Bibliometric analysis* » de SDOC et « *parallel citation analysis* » de TW.

Ces liens sont motivés sémantiquement : l'analyse de la citation est utilisée dans les études bibliométriques. TermWatch, classifiant à partir des termes employés par les auteurs dans leurs articles, fait ressortir des concepts plus spécifiques de *citation analysis* : *on-line citation analysis, integrated multigeneration citation analysis*, *parallel citation analysis...* Cela tend à montrer que la comparaison par associations lexico-sémantiques peut permettre de rapprocher des concepts de niveaux génériques (lorsque la classification part d'une indexation humaine, issue d'un vocabulaire contrôlé) avec ses thématiques plus spécifiques (lorsque une autre classification part des termes employés par les auteurs eux-mêmes). Ici, la classe TW précise une des thématiques de la classe SDOC (*citation analysis*). Du fait que la classification SDOC dans cette expérience, s'est basée sur l'indexation manuelle issue d'un vocabulaire contrôlé, on peut trouver dans une même classe, des termes de niveaux hiérarchiques différents : *bibliometrics* est un terme générique de *bibliometric analysis* qui est le générique de « *Bradford law* » et de « *citation analysis* », etc.

Dans de nombreux cas, les liens s'expliquent par un seul mot-clé de la classe SDOC, lié à plusieurs termes de la classe TW.

Dans l'exemple suivant, les liens entre termes des deux classes « *classification* » (SDOC) et « *subject classification* » (TW) ont une structure plus imbriquée. La majorité des mots-clés dans la classe SDOC sont impliqués dans des liens alors que la totalité des termes de la classe TW sont concernés. Or ces deux classes n'arrivent qu'en 13ème position par l'indice de relation (voir tableau 3) alors que leurs contenus paraissent plus thématiquement homogènes que ceux des classes de la figure 2, classées 11ème par l'indice de relation.

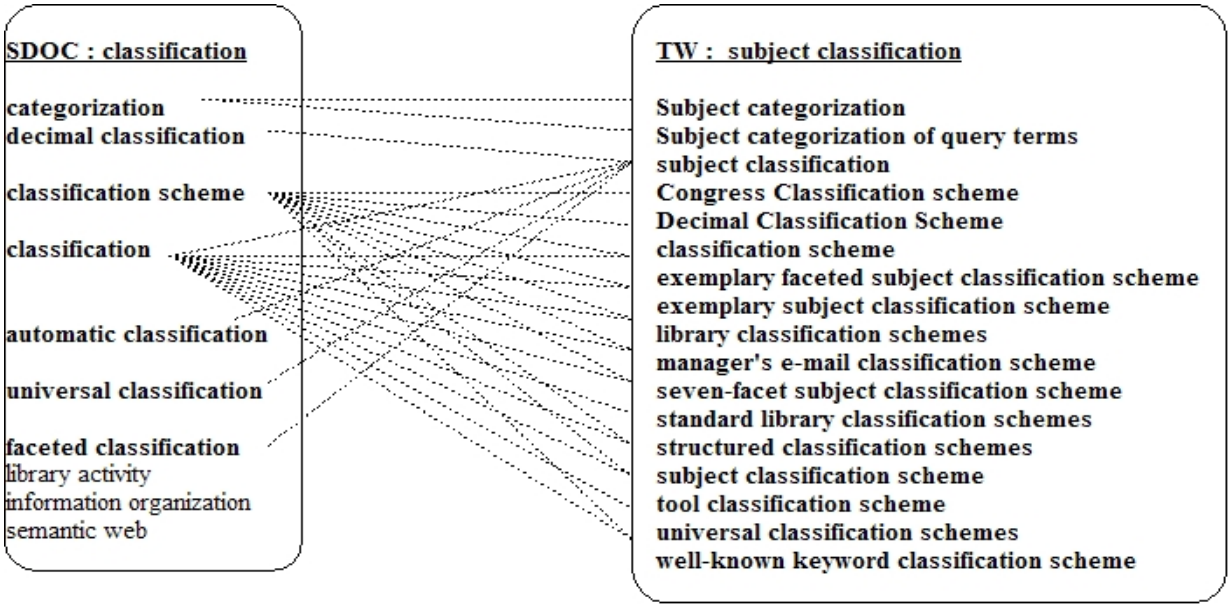

Figure 3. Liens entre la classe «*classification* » de SDOC et « *subject classification* » de TW.

Dans certains cas, REL_{ij} et « proportion de termes liés » sont corrélés, notamment pour des classes en bas du classement par REL_{ij}. Les deux paires de classes respectivement à la 101ème et 118ème positions par REL_{ij} ont des ratios de termes liés assez faibles, ce qui tendrait à confirmer leur classement selon le REL_{ij}. Le libellé identique des deux classes étiquetées « *vector space* » par les deux méthodes (SDOC et TW) aurait pu suggérer que ces deux classes seraient très proches, donc auront un REL_{ij} élevé mais l'examen de leurs contenus respectifs montre un éloignement thématique important (voir figure 1 ci-dessus). Les deux classes en dernière position par REL_{ij} (respectivement 750ème et 751ème) ont des ratios de termes très faibles. La position de ces paires de classes par l'indice de REL_{ij} est donc cohérente.

Une faiblesse théorique de l'approche de comparaison proposée vient de ce qu'elle s'appuie essentiellement sur des liens lexicaux, même si dans certains cas, nous avons la possibilité de détecter des synonymes via WordNet. Certains liens lexicaux, lorsqu'ils impliquent un mot isolé avec des termes plus longs peuvent être des liens accidentels. Ainsi, la plupart de liens entre deux classes peuvent être initiés par un mot isolé[3]. A titre d'exemple, la classe « *language processing* » de SDOC et la classe « *efficient data structure* » de TermWatch sont en rang 4 avec un indice de similarité élevé. Cet indice s'explique par de nombreux liens entre un mot-clé formé d'un uniterme « *structure* » et les termes contenant ce mot dans la classe de TermWatch. Nous représentons ces liens dans la figure 4 ci-dessous.
Bien que le mode de calcul de l'indice REL_{ij} pénalise ce type de liens car il induit une proximité sémantique faible, il suffit que leur nombre soit élevé (19 dans l'exemple ci-dessous) ou qu'il n'existe

[3]C'est souvent le cas si la méthode de classification prend en entrée un vecteur de mots isolés ou si l'entrée vient d'un vocabulaire contrôlé, comme c'est le cas ici. L'indexation contrôlée prône souvent des unités très courtes : unitermes ou termes binaires (formés de deux mots).

que ce type de liens entre termes de plusieurs classes pour que ces classes soient considérées comme très similaires. Lorsqu'on observe le contenu des deux classes, on s'aperçoit que la classe TW est homogène thématiquement, elle traite essentiellement de la structure de données. La classe SDOC a un contenu plus générique et rassemble des mots-clés issus de plusieurs niveaux hiérarchiques : « *language processing* » peut être considéré comme le générique de « *semantic analysis* » et de « *morphological analysis* » alors que « *linguistics* » et « *language processing* » peuvent être considérés comme des termes associés. Cette classe traite de différents aspects du traitement automatique des langues. De ce fait, le rang de ces deux classes par l'indice de similarité (4ème) peut paraître trop élevé au regard de leurs contenus respectifs.

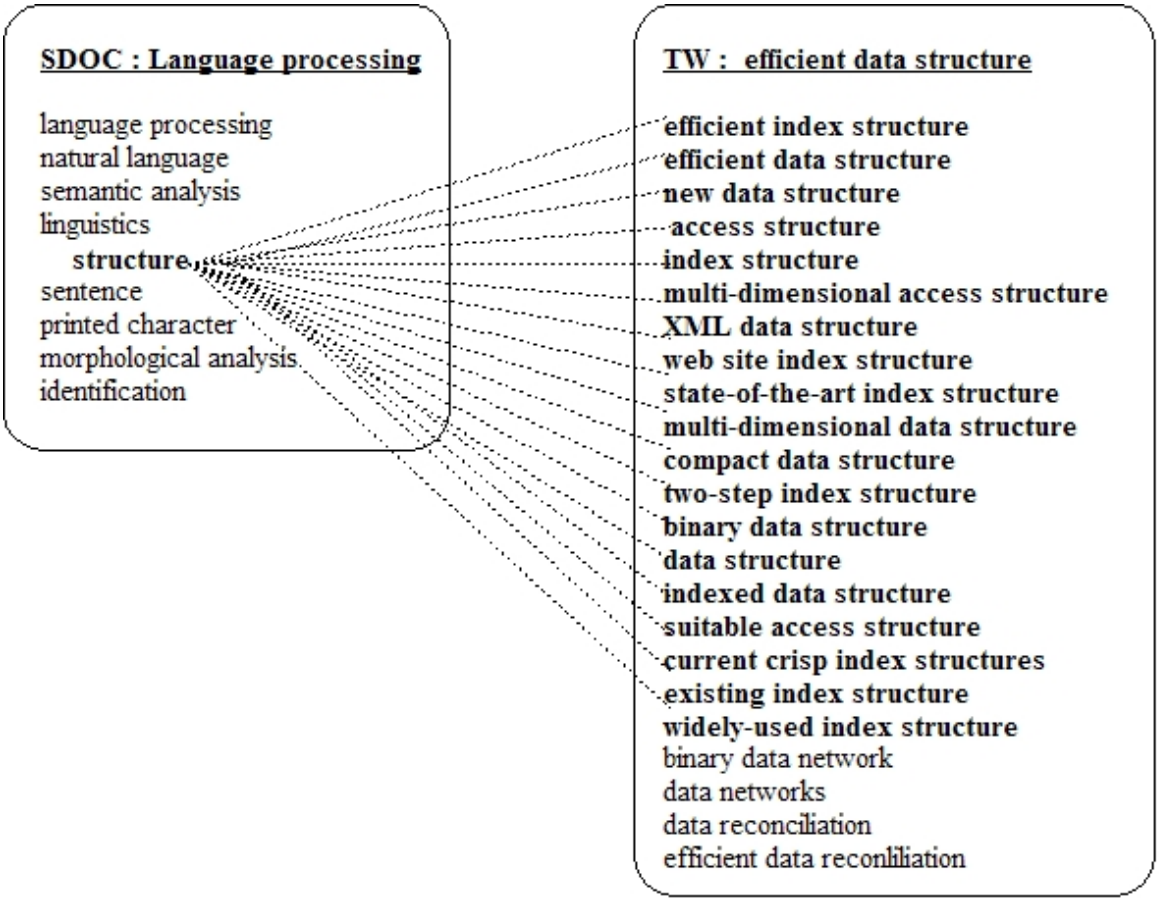

Figure 4. Termes à l'origine des liens entre une classe SDOC et une classe TermWatch.

Perspectives

Le but de cette étude est de proposer une piste alternative au problème épineux de comparaison des classifications issues de méthodes différentes, voire de deux classifications tout court, a fortiori lorsque les méthodes fournissent des résultats très différents. Les critères de comparaison que nous proposons sont d'ordre qualitatifs et mettent en jeu une analyse linguistique de surface. L'objectif est d'aboutir à une méthode automatisable qui en garantisse la reproductibilité et donc un gain de temps appréciable. Elle réduira par la même occasion le caractère arbitraire de l'évaluation humaine tout en conservant un de ses aspects les plus intéressants : son aspect qualitatif. En fonction des pondérations effectuées, cette méthode permettra de désigner les classes qui sont les plus proches du point de vue de leur contenu, de les classer par ordre décroissant de similarité.
Un des résultats que nous obtenons tient tout d'abord à un constat : classifier à partir des termes issus directement du corpus ou classifier à partir des mots-clés issus d'un vocabulaire externe conduit à des résultats assez différents. En effet, seules une centaine de paire de classes ont des indices au dessus de

1. Les écarts vont de 16 pour les deux classes les plus fortement reliées à 0.2. L'indice de relation est assez faible pour la majorité des classes, tendant à montrer d'importants écarts dans les deux vocabulaires utilisés par les deux méthodes.
Néanmoins, la méthode de comparaison que nous proposons permet de rapprocher des classes qui en apparence n'ont pas de libellés lexicalement associés, ce qui est un des buts recherchés. Il suffit que leurs contenus respectifs aient des relations lexicales. Cependant, certains liens peuvent être dus au hasard notamment lorsqu'il s'agit des unitermes. En outre, cette méthode de comparaison évite le piège de considérer comme très similaire deux classes ayant un libellé identique. Le poids du libellé doit être confirmé par d'autres liens à l'intérieur des deux classes. La méthode de comparaison a montré que pour des classes positionnées en bas du classement, il peut exister une corrélation entre « proportion de termes liés » et l'indice REL_{ij}
A l'avenir, pour réduire les écarts de vocabulaire, il serait intéressant de refaire cette comparaison à partir d'une même entrée (soit termes issus du corpus, soit mots-clés externes).

Remerciements
Nous remercions Eric SanJuan, Maître de Conférences, Université de Metz, pour avoir écrit les programmes informatiques nécessaires au calcul des indices de similarité.

Bibliographie

1. Callon M., Courtial J-P., Turner W., Bauin S. (1983). From translation to network : The co-word analysis. *Scientometrics*, 1983, 5(1).
2. Dobrynin, V., Patterson D., Rooney N. (2004). Contextual Document Clustering. Proceedings of the European Conference on Information Retrieval (ECIR'04), Sunderland, UK, April 5-7 2004, 167-180.
3. Fellbaum, C. *et al.* 1999. *Wordnet. An Electronic Lexical Database*. Cambridge, London: The MIT Press.
4. Hubert L,. Arabie P. (1985). Comparing partitions. *Journal of Classification*, 193-218.
5. Ibekwe-SanJuan, F., SanJuan, E. (2004a) Mining textual data through term variant clustering: the termwatch system. In Proceedings of Recherche d'Information assistée par ordinateur (RIAO'04). Avignon, 2004, 487-503.
6. Ibekwe-SanJuan F., Polanco X., SanJuan E. (2004). SDOC et TermWatch : deux méthodes complémentaires de cartographie de thèmes. Atelier « Fouille de données », dans Congrès sur l'Extraction et Gestion des Connaissance, EGC'04, Clermont-Ferrand, 20 janvier 2004, 16p.
7. 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 August, 1998, 170-174.
8. Yeung, K, Y., Ruzzo W, L. (2001). Details of the Adjusted Rand Index and clustering algorithms. Supplement to the paper "An experimental study on Principal Component Analysis for clustering gene expression data". *Bioinformatics,* 2001, 17, 763-774.
9. Jain, A.K., Dubes, R.C. (1988). Algorithms for Clustering Data. Prentice Hall, Englewood Cliffs, NJ, 1988.
10. Milligan, G.W., Cooper M.C. (1986). A study of the comparability of external criteria for hierarchical cluster analysis. *Multivariate Behavioural Research*, 1986, 21, 441-458.
11. Pantel P., Lin D. (2002). Document clustering with committees. Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'02, 8p.
12. Pedersen T., Patwardhan, Michelizzi (2004). WordNet::Similarity – Measuring the relatedness of concepts. Proceedings of the 5th Annual Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL-04), May, 3-5 2004, Boston, 4p.
13. Polanco, X., François C., Royauté J., Besagni D., Roche I. (2001). *STANALYST: An Integrated Environment for Clustering and Mapping Analysis on Science and Technology,* Proceedings of the 8th International Conference on Scientometrics and Informetrics, Sydney, Australia, July 16-20th 2001, Vol. 2, pp. 871-873.
14. Rand, W.M (1971). Objective criteria for the evaluation of clustering methods. *Journal of American Statistical Association*, 66, 846-850.
15. Schiffrin R., Börner K. (2004). Mapping knowledge domains. *Publication of the National Academy of Science* (PNAS), 2004, 101(1), 5183-5185.

16. Vossen, P. (ed.). 1998. *EuroWordNet: a Multilingual Database with Lexical Semantic Networks*. Dordrecht: Kluwer Academic Publishers.
17. Zamir, O. and Etzioni, O. (1998) Web document Clustering, A feasibility demonstration, *in* ACM SIGIR Conference on Research and Development in Information Retrieval 1998, 46-54.