ProxiDocs : un outil de cartographie et de catégorisation thématique de corpus

Thibault Roy[1], Pierre Beust[1]

[1]GREYC- ISLanD CNRS UMR 6072 – Université de Caen – 14032 Caen Cedex – France
thibault.roy@etu.info.unicaen.fr, pierre.beust@info.unicaen.fr

Abstract

This paper presents a software, called ProxiDocs, which constructs a representation (a map) of the thematic structure of a whole of textual documents. ProxiDocs allows its users to realize a thematic analysis of a corpora according to his needs. These maps show thematic similarities and differences between documents which belong to a same corpora. The hypothesis tested with ProxiDocs affects the instrumentation of the meaning's thematic dimension with statistical processing, like principal components analysis and hierarchical clustering.

Résumé

Cet article présente une application dédiée à la visualisation de propriétés thématiques d'un ensemble de documents électroniques. Cette application, nommée ProxiDocs, a pour but d'assister son utilisateur dans des tâches de veille technologique en lui donnant les moyens d'une analyse thématique de corpus par rapport à ses propres centres d'intérêts. ProxiDocs produit des cartes interactives mettant en évidence les proximités et les différences thématiques des documents composant le corpus donné, ainsi que la répartition des différents thèmes utilisés dans ces documents. Les hypothèses testées dans le cadre de ProxiDocs concernent donc l'instrumentation de la dimension thématique du sens par des traitements statistiques, tels l'analyse en composantes principales et la catégorisation hiérarchique ascendante.

Mots-clés : sémantique textuelle, logiciel anthropocentré, cartographie thématique de corpus, analyse en composantes principales, catégorisation hiérarchique ascendante.

1. Introduction

Cet article prend place dans une recherche en Informatique, et plus précisément en Traitement Automatique des Langues (TAL) sur la sémantique des documents textuels. Il décrit les principes et l'intérêt d'un outil logiciel que nous avons développé au sein de notre équipe. Cet outil, appelé ProxiDocs, vise à assister son utilisateur dans des tâches de veille technologique en lui donnant les moyens d'une analyse thématique de corpus. Dans bons nombres de situations (analyse de flux documentaires, recherche d'informations, extraction d'informations …) l'appréhension de la thématique des documents ainsi que l'homogénéité thématique d'une collection de documents est une première analyse importante et souvent délicate. ProxiDocs a pour objectif de fournir à son utilisateur une aide dans ce type d'analyse en lui permettant de construire, en fonction de ses propres centres d'intérêt, des cartes thématiques de documents. En ce sens, c'est un système anthropocentré tel que le définit (Thlivitis, 1998), c'est-à-dire qu'il n'a pas une exécution guidée par des ressources propres, valables pour tout utilisateur, mais au contraire que les traitements qu'il réalise sont personnalisés car intégralement conditionnés par les besoins et les choix de l'utilisateur.

Les cartes thématiques produites avec ProxiDocs sont interactives et révèlent les proximités et les différences qui structurent un corpus supposé homogène du fait de sa provenance. Elles

permettent de mettre cette homogénéité à l'épreuve des classes lexicales choisies et définies par l'utilisateur. ProxiDocs est un outil logiciel gratuit open source, disponible avec sa documentation sur le Web (cf. http://www.info.unicaen.fr/~troy/proxidocs). C'est un logiciel d'étude au sens de (Nicolle, 1996), c'est-à-dire qu'il est conçu dans le but de vérifier des hypothèses sur les langues en les expérimentant sur du matériau textuel attesté. Les hypothèses testées dans le cadre de ProxiDocs concernent l'instrumentation de la dimension thématique du sens par des traitements statistiques. De même que d'autres outils développés au sein de notre équipe, tels que ThemeEditor (Beust, 2002) ou encore Anadia (Nicolle *et al.*, 2002), ProxiDocs fait partie d'une plate-forme de logiciels d'étude en constante évolution pour l'analyse linguistique informatisée des corpus de documents électroniques.

Nous donnerons, dans la première partie de cet article, les principes des méthodes mises en œuvre pour la construction des cartes thématiques et nous discuterons comparativement leur intérêt. La deuxième partie sera dédiée aux améliorations apportées au logiciel par l'implémentation d'une méthode de catégorisation automatique des documents d'une carte. Enfin, nous conclurons en abordant les évolutions que nous souhaitons encore apporter à l'application ProxiDocs ainsi que les pistes de recherche que le logiciel nous ouvre.

2. La cartographie thématique de corpus

2.1. *Définitions et intérêts*

À la manière du logiciel Umap (cf. http://www.umap.com), l'outil que nous proposons va chercher à mettre en évidence des similarités et des différences entre les documents d'un corpus homogène. Plus précisément, ProxiDocs va extraire les tendances thématiques des documents d'un tel corpus via une représentation graphique que nous appellerons une carte thématique. Cette représentation nous permettra de mettre en évidence les thèmes abordés dans ce corpus, mais également leur répartition et leur cohésion au sein des documents de cet ensemble. Chaque document sera représenté par un point sur la carte, la position de ce point dépendra des thèmes abordés dans le document représenté selon la propriété suivante :

> *Si deux documents abordent des thèmes similaires, alors nous supposerons que les points les représentant seront proches sur la carte.*

À l'inverse, si deux documents n'abordent pas les mêmes thèmes, alors nous pouvons faire l'hypothèse que les points les représentant seront éloignés sur la carte. Les intérêts de ProxiDocs sont alors multiples, par exemple le regroupement des documents thématiquement similaires peut aider à la création de logiciels permettant une « lecture rapide » d'un grand nombre de documents. Dans une tâche de recherche documentaire sur Internet, la cartographie thématique appliquée aux pages retournés par un moteur de recherche permettrait d'organiser les liens obtenus et de proposer des pages moins redondantes, aux thématiques différentes et se complétant.

2.2. *La construction des cartes thématiques*

2.2.1. *Les thèmes utilisés*

De la même manière que (Pichon et Sébillot, 1999), nous entendrons ici par « thèmes », les sujets abordés dans un texte. Traiter la thématique d'un texte revient donc pour nous à mettre en évidence les principaux sujets abordés dans ce dernier. Les documents électroniques que nous analysons ne sont jamais isolés et font toujours parti d'un ensemble de documents homogènes. C'est par exemple le cas des documents fournis par un moteur de recherche en réponse à une requête. Afin de construire la carte thématique d'un corpus, il sera nécessaire

de prendre en entrée les thèmes à faire émerger des documents du corpus analysé. Ces thèmes sont choisis et définis par l'utilisateur de l'application, afin de ne représenter que les thèmes et les mots pertinents de son point de vue. Techniquement, l'utilisateur va saisir ses définitions de thèmes dans un fichier XML (W3C, 2001a), compatible avec le logiciel ThemeEditor (cf. http://www.info.unicaen.fr/~beust/ThemeEditor.html). Par exemple, un utilisateur pourra définir la liste des 18 lexies (mots ou mots composés) suivantes correspondant à une définition succincte du thème « Aviation » :

Aviation : avion, avions, appareil, appareils, Airbus, A320, vol, vols, pilote, pilotes, pilotage, aéronautique, passager, passagers, Boeing, Air France, décollage, décollages.

2.2.2. *La chaîne de traitement*

Les différentes étapes de calculs de ProxiDocs s'effectuent l'une après l'autre, d'une manière chaînée. Le schéma suivant résume alors le fonctionnement de notre application :

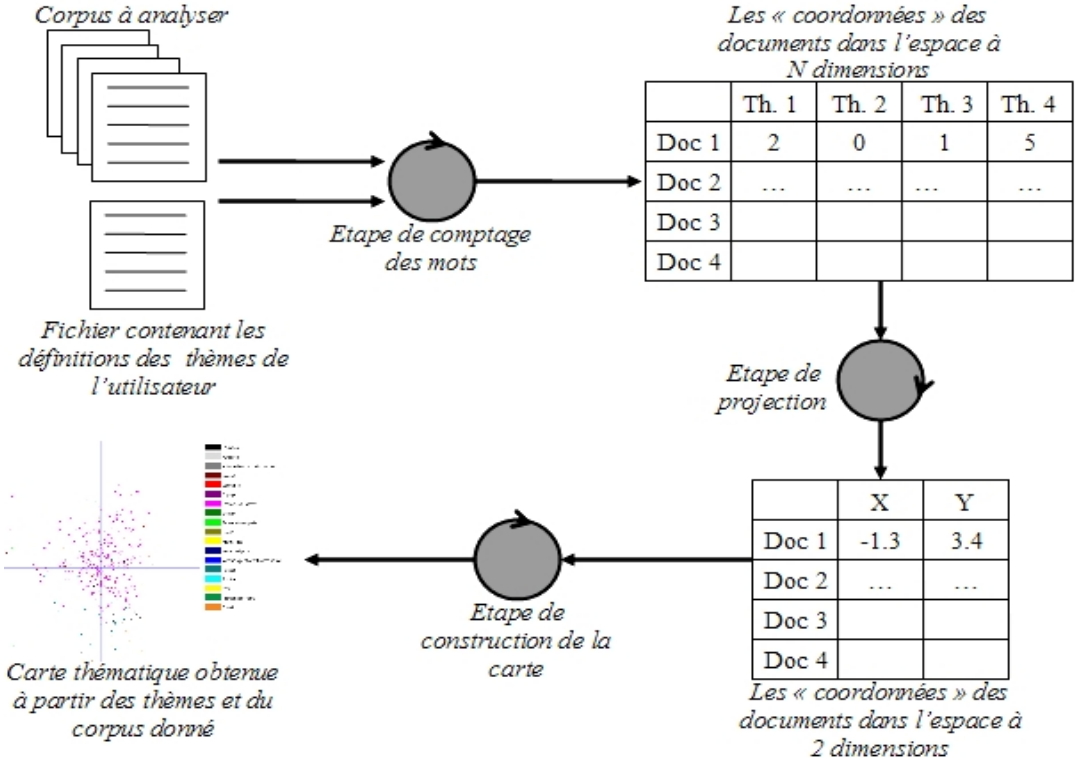

Figure 1. La chaîne de traitement de l'application ProxiDocs

Les étapes de comptage des mots et de projection sont paramétrables par l'utilisateur (les différentes possibilités seront détaillées par la suite). L'un des buts de ProxiDocs est de permettre de comparer ces possibilités. Les cartes thématiques produites en sortie de l'application sont représentées dans le format SVG (W3C, 2001b), ceci afin de permettre à l'utilisateur d'effectuer des zooms et des déplacements sur ses cartes, et de garantir la portabilité des cartes. Ce dernier pourra également visualiser les documents figurant sur la carte en cliquant sur les points les représentant. Chacun de ces points se verra attribuer une couleur correspondant au thème majoritaire dans le document qu'il représente.

2.2.3. *L'étape de comptage*

Cette étape de comptage est nécessaire pour établir une représentation vectorielle des documents du corpus étudié. Il faut compter, pour chaque document, le nombre des occurrences de mots de chaque thème. Chaque document se voit donc attribuer un vecteur contenant un nombre de données égal au nombre de thèmes utilisés. Soit *N*, un tel nombre, le vecteur décrira donc un point possédant *N* coordonnées. Soit *D* un document où figure 3 mots du thème « Bourse », 0 du thème « Météo » et 2 du thème « Sport », le vecteur obtenu sera alors : *vecteur_D* = (3, 0, 2).

La méthode précédente est appelée « absolue », du fait qu'elle ne fait pas intervenir la taille des textes lors du comptage. Une seconde méthode, appelée « relative », va chercher à déterminer pour chaque document du corpus, les proportions des mots de chaque thème par rapport au nombre de mots du document. Ainsi, pour un texte de 300 mots contenant 35 mots appartenant au thème « Sport », la méthode relative considérera que ce thème occupe (35/300*100=) 11,6% du texte. Cette seconde méthode sera particulièrement importante lorsque la taille des documents constituant le corpus varie significativement.

Une difficulté rencontrée lors du comptage des mots est qu'un mot peut appartenir à plusieurs thèmes. Ce serait par exemple le cas du mot « avocat », que l'on pourrait aussi bien affecter au thème des aliments qu'au thème de la justice. Dans un tel cas, nous avons choisi d'attribuer le mot au thème le plus représenté dans le texte. Ceci revient à prolonger le plus possible les isotopies génériques (Rastier, 1987) du texte (c'est-à-dire favoriser la redondance thématique).

2.2.4. *Les méthodes de projection*

Après application de l'une des deux méthodes de comptage précédentes, les documents sont représentés par des points possédant *N* coordonnées. La carte étant un support en deux dimensions, il va falloir projeter ces points sur ce support. Afin de réaliser une telle projection, nous nous sommes inspirés des principes généraux suivants :

i. Déterminer les deux composantes (ou axes) « principales », c'est-à-dire les deux composantes les plus caractéristiques de l'espace de départ à N dimensions,

ii. Projeter l'ensemble des points de l'espace de départ à N dimensions sur les deux axes obtenus précédemment.

Nous avons tout d'abord développé deux méthodes simples se basant sur ces principes. La première d'entre elles, appelée « méthode des deux plus grandes distances », consiste à prendre les deux axes passant par les deux couples de points les plus éloignés[1] dans l'espace de départ.

La seconde de ces méthodes simples, appelée « méthode du produit scalaire », consiste à prendre comme premier axe principal, celui reliant les deux points les plus éloignés de notre espace de départ. Le second axe sera celui passant par deux points de ce même espace et dont le produit scalaire avec le premier axe sera le plus proche de 0. Une fois les deux principaux axes de l'espace de départ déterminés par l'une des deux méthodes précédentes, il ne restera plus qu'à projeter tous les points à *N* coordonnées sur ces axes.

[1] La distance euclidienne sera utilisée pour calculer la distance entre deux points de l'espace à *N* dimensions. Ce calcul de distance pourrait être paramétrable par l'utilisateur, les méthodes de calcul de distance que nous pourrions proposer seraient, entre autres, la distance du *Khi-2* et la distance hamiltonienne.

La dernière méthode que nous avons développée est l'analyse en composantes principales (ACP) (Bouroche et Saporta, 1980). Cette méthode étant bien connue dans le domaine de l'analyse de données, nous ne la détaillerons pas ici mais nous nous limiterons à rappeler une idée de son application à notre problématique. L'idée majeure de l'ACP est que si les thèmes sont corrélés entre eux alors ils sont partiellement redondants. Prenons un exemple simple pour illustrer cette idée. Supposons que l'utilisateur ait défini les thèmes « Equitation » et « Jeux » parmi l'ensemble de ses thèmes. Supposons de plus, que son corpus soit principalement composé d'articles concernant les courses hippiques. Il est évident dans ce cas que ces deux thèmes sont fortement corrélés vis-à-vis du corpus. Ils peuvent donc être regroupés en un seul groupe de thèmes. L'ACP va généraliser ce processus à l'ensemble des thèmes, pour ne retenir que les deux principaux groupes de thèmes. Ces deux groupes sont appelés les composantes principales, la projection des documents se fera alors sur ces deux axes.

2.3. *Interprétation des résultats*

Nous allons présenter les cartes produites avec les méthodes du produit scalaire et de l'ACP. Le corpus étudié est constitué de 594 articles économiques du journal « Le Monde » dans la période 1987-1989, totalisant 500 000 mots. Le fichier de thèmes ici utilisé est défini par un utilisateur et il décrit 17 thèmes de son choix. La méthode de comptage relative est utilisée. La carte suivante sera obtenue par la méthode du produit scalaire :

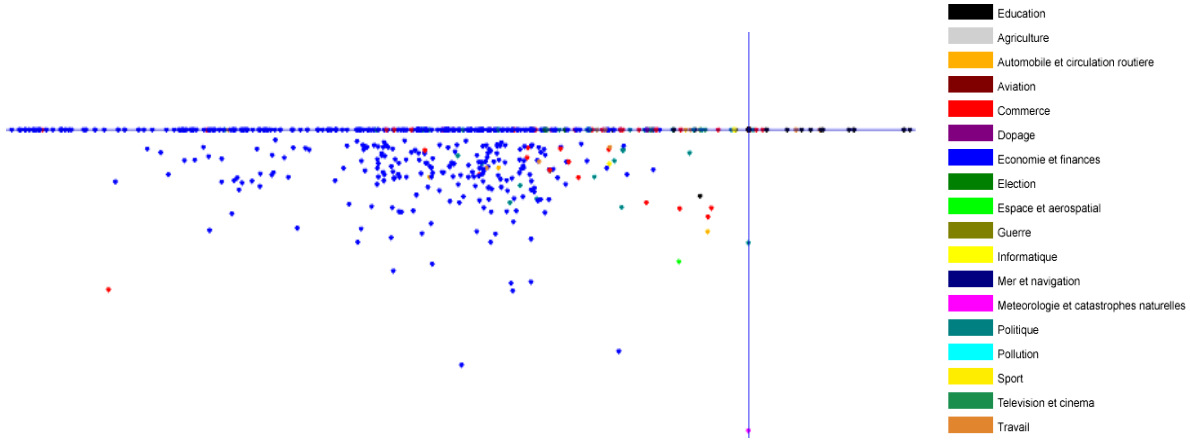

Figure 2. La carte obtenue avec la méthode du produit scalaire. Chaque point a la couleur du thème majoritaire dans le document qu'il représente. En cliquant sur l'un de ces points, on affiche le contenu du document représenté.
(version en couleur disponible sur http://www.info.unicaen.fr/~troy/proxidocs/cartes/fig2.html)

Cette carte présente les informations d'une manière particulière : les différents points sont agglomérés autour de l'axe des abscisses, et aucun point ne figure au dessus de cet axe. Ce phénomène peut s'expliquer par une relative « inefficacité » de l'axe des ordonnées choisi, puisqu'un grand nombre de points ont un projeté de valeur nul sur cet axe. Grâce à la coloration des points, nous observons tout de même une forte présence des documents abordant le thème « Economie et Finances », ceci attestant bien l'homogénéité du corpus, mais nous ne distinguons pas de groupes de documents au thème majoritaire différent.

La carte suivante a été réalisée avec les mêmes entrées que la carte précédente, la méthode utilisée est l'analyse en composantes principales :

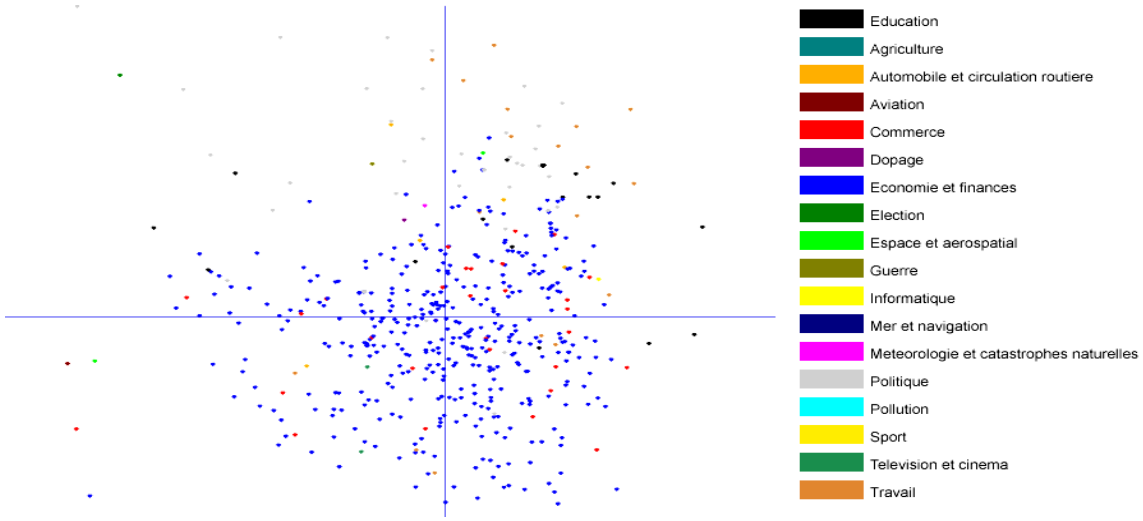

Figure 3. La carte obtenue avec la méthode de l'analyse en composantes principales (version en couleur disponible sur http://www.info.unicaen.fr/~troy/proxidocs/cartes/fig3.html)

Nous voyons que la carte met différemment en évidence les informations : les points ne sont pas agglomérés autour des axes et se répartissent mieux sur la carte. Bien évidemment, les documents de thèmes majoritaires « Economie et Finances » sont tout aussi nombreux que précédemment, mais cette fois-ci, ils sont présents dans une large zone centrée sur l'origine du repère. Des documents de thème majoritaire « Commerce » se mélangent à ces documents dans cette zone, il est donc possible de déduire un certain lien entre ces deux thèmes dans les documents du corpus. Dans la partie située en haut et à droite de la carte, des documents de thèmes majoritaires « Travail » et « Education » se mélangent. Là encore, ces deux thèmes semblent liés vis-à-vis des documents du corpus. La partie en haut et à gauche de la carte contient des documents de thème majoritaire « Politique », étant donné que ces derniers sont légèrement éloignés des autres points, nous pouvons en déduire qu'ils ne sont pas liés avec les autres documents. La carte ne laisse pas réellement paraître des groupes distincts de points, ceci attestant encore une fois l'homogénéité du corpus par rapport aux thèmes donnés.

Après quelques tests avec différents corpus et fichiers de thèmes, nous sommes en mesure de déterminer dans quelles types d'études utiliser telle ou telle méthode de projection. Sans surprise, l'ACP nous donne des cartes qui apparaissent très lisibles. Cependant, pour en extraire le plus d'informations possible une étape essentielle reste à mener : celle de l'interprétation des axes. Elle incombe à l'utilisateur (destinataire de la carte) par une analyse du contenu des documents. Les méthodes des plus grandes distances et du produit scalaire se révèlent quant à elles assez limitées pour projeter les points d'espaces de grandes dimensions. Nous obtenons cependant des résultats intéressants par ces méthodes en utilisant des espaces de « petite » taille, par exemple de 3 ou 4 dimensions.

3. Une méthode de catégorisation appliquée à la thématique de corpus

3.1. *Intérêt d'une méthode de catégorisation et intégration à la cartographie thématique*

Les cartes obtenues précédemment nous permettent de déduire des informations sur les relations thématiques existant entre les documents du corpus analysé. Malheureusement, si le corpus étudié est très homogène et les thèmes définis par l'utilisateur sont très corrélés entre eux, alors la carte obtenue ne contiendra qu'un seul groupe de points, sans mettre en évidence plusieurs groupes distincts. L'analyse d'une telle carte peut alors se révéler assez délicate.

L'intérêt d'intégrer une méthode de catégorisation est d'automatiser l'analyse manuelle de la carte en mettant en évidence sur cette carte des groupes de documents thématiquement similaires. Ainsi, l'interprétation de la carte sera alors plus simple, mais également plus fine, puisque pour chaque groupe de documents construit par cette méthode, il sera possible d'obtenir le document le plus représentatif de ce groupe.

La méthode que nous utilisons est appelée « catégorisation hiérarchique ascendante », son fonctionnement peut se résumer par les deux étapes suivantes (Bouroche et Saporta, 1980) :

i. *Parmi les n entités à classer, chercher les deux entités les plus proches. Ces deux entités sont ensuite agrégées en un nouveau groupe.*
ii. *Calculer les distances entre le nouveau groupe et les entités restantes. La configuration est alors identique à celle de l'étape i., hormis que l'on a seulement n-1 entités à classer.*

Et ainsi de suite, on cherche de nouveau les deux entités ou groupes les plus proches, que l'on agrége et ceci jusqu'à ce qu'à obtenir le nombre de groupes arbitrairement déterminé par l'utilisateur. L'application de cette méthode à notre problématique est alors assez simple, puisque nous assimilons les documents aux points d'un espace. Calculer la proximité de deux documents revient donc à calculer la distance euclidienne entre les points les représentant. Ainsi, la distance entre deux groupes de documents s'obtiendra en déterminant le centre de gravité de chacun de ces groupes, puis en calculant la distance euclidienne entre ces deux points. Le document supposé le plus représentatif d'un groupe sera alors celui étant le plus proche du centre de gravité de ce groupe.

3.2. *Intégration à la cartographie thématique*

L'étape de catégorisation va prendre en entrée les points de l'espace à deux dimensions caractérisant la carte. Cette étape de calcul va donc chercher à former des groupes de points, puis pour chacun de ces groupes, déterminer son document « caractéristique ». Nous allons présenter deux nouvelles cartes montrant les résultats de la catégorisation. Ces deux cartes seront construites à partir des mêmes entrées que les cartes précédentes, la méthode de projection utilisée sera l'ACP. Nous choisissons de faire figurer dans les deux cartes suivantes 15 groupes de documents. La première carte sera la suivante :

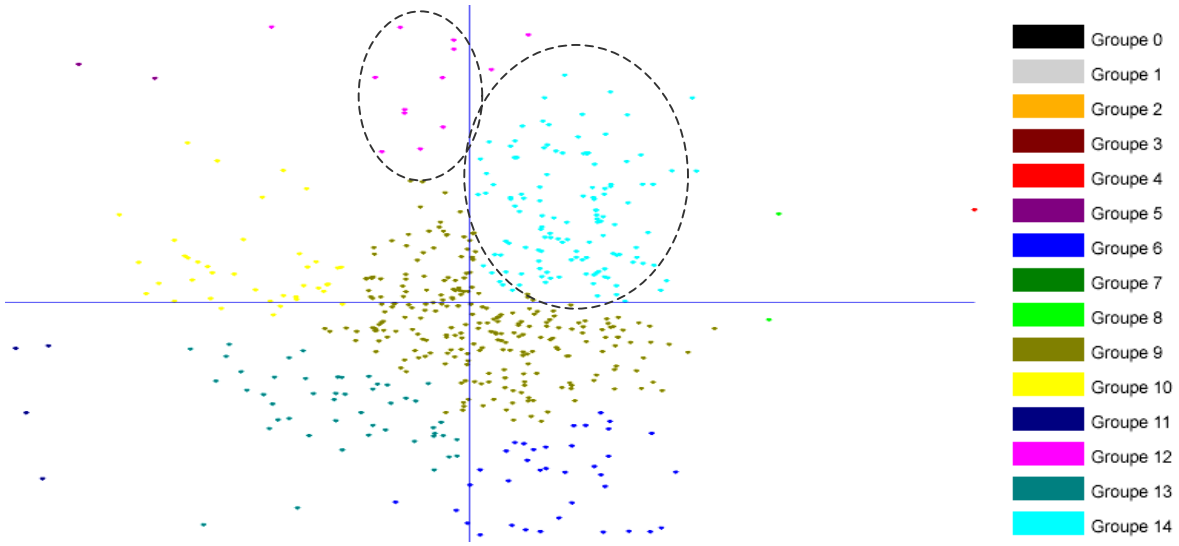

Figure 4. La carte thématique après catégorisation, chaque point est colorié suivant son groupe. Pour des raisons de lisibilité, nous avons entouré manuellement les groupes 12 et 14. (version en couleur disponible sur http://www.info.unicaen.fr/~troy/proxidocs/cartes/fig4.html)

La disposition des points sur la carte est identique à la carte de la figure 3., seule la coloration des points change. Au lieu de colorier chaque point en fonction de son thème majoritaire, les points sont colorés avec une couleur correspondant à leur groupe. De cette manière, nous voyons les groupes de documents aux thématiques similaires, ainsi que leur répartition sur la carte. En visualisant des documents appartenant à un même groupe, nous obtenons un « aperçu » de la thématique partagée par les documents de ce groupe. Ainsi, en consultant des documents du groupe 14, nous remarquons qu'ils abordent les thèmes de l'économie, de la bourse et du commerce. En légère opposition, des documents du groupe 12 abordent plus particulièrement des thèmes liés à la politique et aux actions de politiques économiques.

Afin de caractériser plus précisément chaque groupe, il est intéressant de considérer son document le plus représentatif. La carte suivante met en évidence cette propriété :

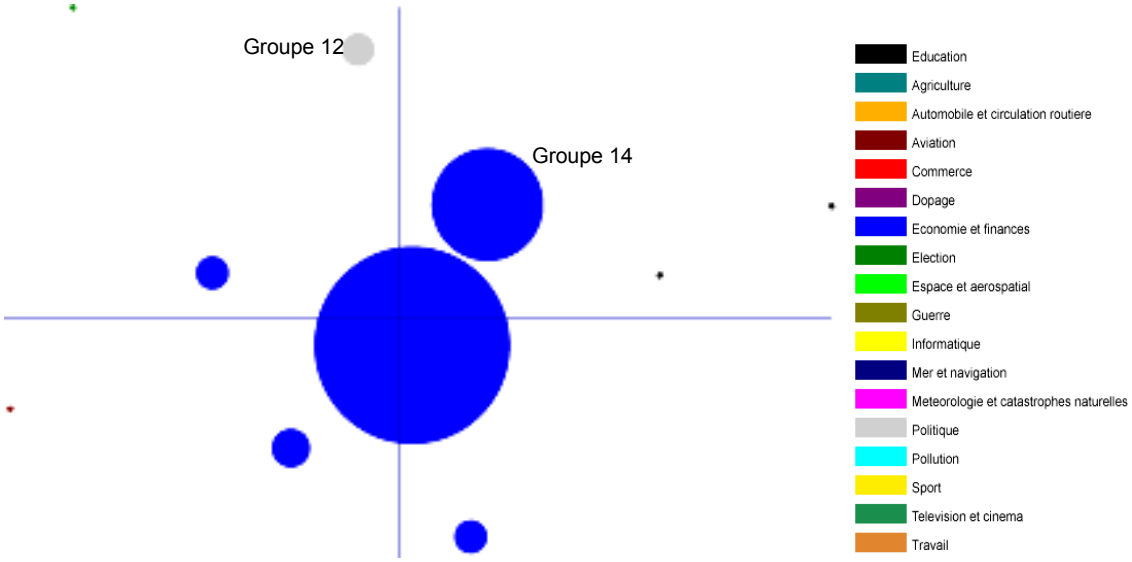

Figure 5. La carte illustrant les groupes de documents aux propriétés thématiques communes. Pour des raisons de lisibilités nous avons identifié les groupes 12 et 14. En cliquant sur un groupe, on obtient son document caractéristique.
(version en couleur disponible sur http://www.info.unicaen.fr/~troy/proxidocs/cartes/fig5.html)

Sur cette carte, un disque représente un groupe de documents. La taille de ce disque est proportionnelle au nombre de documents constituant le groupe. Chaque disque est centré sur le centre de gravité du groupe qu'il représente et il se voit attribuer une couleur correspondant au thème majoritaire dans les documents de ce groupe. Les observations que nous avons déduites de la carte précédente se voient donc facilitées puisqu'il suffit de visualiser le document le plus représentatif d'un groupe pour obtenir des informations sur la thématique des autres documents du groupe. Ainsi, les observations réalisées précédemment se confirment, puisqu'en visualisant les documents le plus caractéristiques des groupes 12 et 14, nous obtenons respectivement un document abordant une décision politique en rapport au marché boursier, et un document abordant les opérations boursières de grandes entreprises.

4. Conclusions et perspectives

Le logiciel d'étude ProxiDocs que nous venons de présenter nous a permis de vérifier notre hypothèse de départ, à savoir que l'on peut mettre à profit des méthodes statistiques pour des analyses thématiques centrées autour des besoins d'un utilisateur. De cette manière, ProxiDocs permet de vérifier l'homogénéité des corpus recueillis, d'en extraire les principales ten-

dances thématiques par rapport aux attentes de l'utilisateur, ainsi que de mettre en évidence des groupes de documents aux propriétés thématiques similaires.

Plusieurs améliorations seront envisagées dans les prochaines versions de l'application actuellement en développement. Par exemple, il serait plus facile de pouvoir définir des thèmes en entrant simplement les lemmes des lexies plutôt que des formes fléchies (ce qui demanderait à intégrer une lemmatisation à la phase de comptage des occurrences de lexies dans les documents). D'autre part, en fonction de certains types de corpus et certaines définitions de thèmes, une projection sur un espace 3D serait peut-être plus intéressante que sur une carte 2D. Pour améliorer notre application et valider plus encore nos hypothèses, il va devenir incontournable de l'expérimenter avec un plus grand nombre d'utilisateurs et des corpus d'origines diverses.

Cette expérimentation et surtout son évaluation ne sera pas sans poser problèmes car il n'est pas simplement question ici de juger (par exemple en terme de rappel et de précision) si le logiciel donne des résultats satisfaisants ou non, mais il convient plutôt d'évaluer la façon dont plusieurs utilisateurs s'approprient un même outil chacun à leur façon et en fonction de leurs buts. En somme, ce n'est pas simplement le logiciel qu'il faut évaluer mais le couple outil-utilisateur. À travers une telle expérimentation, le besoin d'autres améliorations, voire d'autres outils logiciels complémentaires, se fera sûrement sentir nous persuadant ainsi à toujours mieux instrumentaliser la dimension intertextuelle de la sémantique des langues.

Références

Beust P. (2002). Un outil de coloriage de corpus pour la représentation de thèmes. In *Actes des JADT 2002*.

Bouroche J.M. et Saporta G. (1980). *L'analyse des données*. PUF, collection Que Sais-je ?.

Nicolle A. (1996). L'expérimentation et l'intelligence artificielle. *Intellectica*, vol. (22) : 9-19.

Nicolle A. *et al.* (2002). Un analogue de la mémoire pour un agent logiciel intéractif. *Cognito*, vol. (21) : 37-66.

Pichon R. et Sébillot P. (1999). Différencier les sens des mots à l'aide du thème et du contexte de leur occurrences : une expérience. In *Actes de TALN1999* : 279-288.

Rastier T. (1987). *Sémantique interprétative*. PUF.

Thlivitis T. (1998). Sémantique interprétative Intertextuelle : assistance anthropocentrée à la compréhension des textes. Thèse d'informatique de l'Université de Rennes I.

W3C (2001a). eXtend Markup Language (XML), http://www.w3.org/XML/.

W3C (2001b). Scalable Vector Graphics (SVG) 1.0 Specification, http://www.w3.org/TR/SVG/.