Cartographie d'un corpus de domaine médical

Thibault Roy (1), Aurélie Névéol (2, 3)

(1) Laboratoire GREYC UMR 6072 CNRS
Université de Caen / Basse-Normandie
Boulevard Maréchal Juin 14032 Caen Cedex
thibault.roy@info.unicaen.fr

(2) Equipe CISMeF
aneveol@insa-rouen.fr

(3) NLM, 3600 Rockville Pike, Bethesda, MD 20894
Etats-Unis

RÉSUMÉ. Cet article présente les premiers résultats d'une expérience menée dans le cadre de la cartographie d'un corpus de documents médicaux. Notre objectif est de proposer à des experts dans le domaine de la santé (médecins, documentalistes, etc.) des vues globales sur des corpus situés dans ce domaine. Afin de fournir de telles vues, nous utilisons la plate-forme ProxiDocs de cartographie et de catégorisation de corpus permettant de prendre en considération les particularités du domaine médical. Les cartes ainsi construites à partir du corpus d'étude permettent de visualiser des proximités et des regroupements entre documents du corpus.

MOTS-CLÉS : Cartographie de Corpus, Catégorisation de Textes, Terminologies Médicales, Indexation

1 Introduction

Cet article présente les résultats d'une expérience récente menée dans le cadre de la cartographie d'un corpus de documents du domaine médical. Après avoir présenté la problématique générale de la représentation visuelle d'ensembles documentaires, nous précisons nos objectifs à travers la cartographie d'un corpus de documents médicaux. La deuxième partie détaille le corpus d'étude ainsi que la méthode de cartographie exploitée. La troisième partie présente la carte construite à partir du corpus d'étude et une analyse détaillée de cette carte. Enfin, nous concluons sur les résultats obtenus.

2 Cadre de travail

Le nombre de documents électroniques textuels produits et échangés chaque jour ne cesse de croître. Afin d'isoler les principales informations contenues dans des ensembles de documents, il peut être intéressant d'en proposer des représentations globales. Depuis quelques années, des outils d'analyse textuelle exploitent une technique de visualisation particulière appelée cartographie. À la manière d'une carte routière mettant en évidence des villes et des routes les reliant, une carte d'un ensemble de données textuelles met en évidence des proximités sémantiques et des liens entre entités textuelles, tels des mots, des textes, etc. Depuis 2001, les métamoteurs de recherche cartographiques KartOO (http://www.kartoo.com) et MapStan (http://www.mapstan.net) sont disponibles sur l'Internet. De nombreux logiciels dédiés à l'analyse de données textuelles proposent également des résultats d'analyses

sous forme de cartes. Parmi ces logiciels, nous pouvons citer Hyperbase d'Etienne Brunet, BI de Michel Kerbaol ou encore Lexico3 de l'équipe CLA2T de Paris III.
Les documents scientifiques dans le domaine de la santé ne sont pas épargnés par l'essor du numérique. Plusieurs projets se donnent alors pour objectif de guider les utilisateurs dans leur recherche d'information en santé. Ainsi, la fondation Suisse HON (Health On the Net – http://www.hon.ch) propose un portail vers une information de santé de qualité dans plusieurs langues européennes. La base documentaire MEDLINE® (http://www.pubmed.gov) recense l'ensemble des publications scientifiques dans le domaine de la santé depuis plusieurs décennies. Depuis 1995, le Catalogue et Index des Sites Médicaux Francophones (CISMeF – http://www.cismef.org) recense des ressources de santé institutionnelles à l'usage des professionnels de santé, des étudiants en médecine et du grand public. Afin de retrouver des informations pertinentes dans de tels ensembles documentaires, les méthodes traditionnelles consistent à interroger des bases documentaires à l'aide de mots-clés fournis aux moteurs de recherche dédiés. Notre objectif est de proposer à des experts de la santé (médecins, documentalistes, etc.) des vues globales sur des ensembles de documents médicaux. De telles vues doivent permettre de localiser les principales informations contenues dans l'ensemble documentaire, mais aussi des similarités et des différences entre documents de l'ensemble. Pour ce faire, nous proposons d'utiliser la plate-forme ProxiDocs de cartographie et de catégorisation de corpus [ROY 05] à laquelle des informations liées au domaine de la santé ont été intégrées.

3 Présentation du corpus et de la méthode de cartographie

3.1 Corpus d'étude

Pour cette étude, nous avons travaillé avec un corpus de 70 ressources[1] extraites aléatoirement du catalogue CISMeF dans le cadre de différentes campagnes d'évaluation de systèmes d'indexation automatique. Chaque ressource du corpus de travail comporte une indexation à l'aide de descripteurs du thésaurus MeSH® (Medical Subject Headings). Cette indexation se présente sous la forme d'une liste pondérée de mots-clés ou de paires mot-clé/qualificatif issus du MeSH. La pondération « majeur » dénote les thèmes traités en profondeur dans la ressource, et la pondération « mineur » signale les thèmes traités plus succinctement.

3.2 Méthode de construction des cartes

La catégorisation du corpus en spécialités médicales est effectuée grâce à un outil bibliométrique [DAR 05] utilisant récursivement l'algorithme de catégorisation décrit dans [NEV 04]. Cet algorithme est fondé sur l'indexation MeSH des ressources, et exploite les liens sémantiques existant entre les mots-clés MeSH et les spécialités médicales d'une part, les qualificatifs MeSH et les spécialités médicales d'autre part. Ainsi, chaque descripteur MeSH attribué à une ressource permet de catégoriser la ressource sous la (les) spécialité(s) médicale(s) auxquelles renvoie le descripteur. Par exemple, une ressource indexée avec le mot-clé <*diabète*> relève de la spécialité « endocrinologie ». Le score attribué à « endocrinologie » sera de 100 si <*diabète*> est un thème majeur pour la ressource et de 1 si c'est un thème mineur. À partir du classement des spécialités établi avec la méthode précédente sur le corpus d'étude, nous obtenons une information globale sur cet ensemble.
Les spécialités ainsi classées servent de point de départ à la cartographie de l'ensemble documentaire. Le premier traitement réalisé consiste à attribuer une structure vectorielle à chaque ressource : une ressource est représenté par un vecteur de nombres réels compris dans un espace de dimension égale au nombre de spécialités où chaque coordonnée du vecteur est le score de la spécialité correspondante, l'ordre des coordonnées dans le vecteur étant similaire au classement global des spécialités. Les scores des spécialités de chaque ressource sont déterminés à l'aide de la méthode précédente, mais en ne prenant cette fois-ci en

[1] Afin de rendre compte de la multiplicité des documents électroniques que cela soit du point de vue de leurs formats ou des usages auxquels ils sont destinés, nous utiliserons le terme de « ressource ».

considération que la ressource et non la globalité du corpus. De cette manière, une ressource est représentée de la façon suivante :

$$Vecteur_{Res} = (Score_{Virology}(Res), Score_{Infectiology}(Res), Score_{Virology}(Res), etc.)$$

Si des spécialités apparaissant dans l'ensemble ne sont pas présentes dans la ressource, des valeurs nulles sont placées aux coordonnées correspondantes dans le vecteur. Ce processus est alors répété pour chaque ressource de l'ensemble étudié. Ainsi, un espace de grande dimension où les ressources prennent place a pu être construit.
Dans notre étude, 78 spécialités (sur 126) ont été utilisées pour catégoriser l'ensemble des ressources à l'aide de la méthode précédente ; l'espace des ressources possède donc 78 dimensions. Afin de visualiser graphiquement les documents prenant place dans un tel espace, nous avons choisi d'en réaliser une projection vers un espace à deux dimensions. La plate-forme ProxiDocs permet de réaliser cette opération selon différentes méthodes statistiques (cf. [ROY 04] pour plus de détails sur ces méthodes). Dans cette étude nous avons choisi d'utiliser la méthode de projection de Sammon pour les résultats satisfaisants qu'elle donne dans la projection d'espaces de grande dimension [SAM 69]. Des Analyses en Composantes Principales [BOU 80] et des Analyses Factorielles des Correspondances [BEN 80] ont également réalisées à l'aide de la plate-forme dans cette étude, mais les résultats obtenus se sont révélés moins pertinents.

Entrée : un espace de ressources à n dimensions (n > 2)
Sortie : un espace à deux dimensions où les ressources prennent place

1. Placer chaque ressource aléatoirement dans l'espace d'arrivée à deux dimensions (le placement aléatoire se fait dans $[0,1]^2$).
2. Pour chacune des ressources de l'ensemble : tester si les distances euclidiennes dans l'espace de départ à n dimensions entre la ressource courante et les autres ressources sont respectées dans l'espace d'arrivée à 2 dimensions (une faible constante près fixée empiriquement).
3. Si ce n'est pas le cas, les autres ressources peuvent effectuer un déplacement minimal (valeur du déplacement minimal donnée en entrée de l'algorithme) dans l'espace à 2 dimensions afin de « tendre » vers une situation où les distances entre chacune des ressources sont respectées dans l'espace à 2 dimensions.
4. Reprendre à l'étape 2. jusqu'à ce que les distances entre chaque ressource soient respectées entre l'espace de départ à n dimensions et l'espace d'arrivée à 2 dimensions.

Afin de mettre en évidence des regroupements entre les ressources ainsi projetées, nous avons choisi d'appliquer une Catégorisation Hiérarchique Ascendante (CHA) [BOU 80]. Son fonctionnement dans le cadre de notre étude peut se résumer par les deux étapes suivantes :

Entrée : un espace de ressources prenant place dans un espace à 2 dimensions
Sortie : un ensemble de n groupes de ressources, le nombre n étant choisi empiriquement par l'utilisateur

1. Parmi les entités à catégoriser, chercher les deux entités les plus proches (c'est-à-dire, dont la distance euclidienne est la plus petite) dans l'espace à deux dimensions. Ces deux entités sont ensuite agrégées en un nouveau groupe.
2. Calculer les distances entre le nouveau groupe et les entités restantes. La configuration est alors identique à celle de l'étape 1, hormis que l'on a seulement n-1 entités à classer.

Et ainsi de suite, on cherche de nouveau les deux entités les plus proches, que l'on agrége et ceci jusqu'à obtenir le nombre de groupes choisis par l'utilisateur.
Une fois les étapes de projection et de catégorisation réalisées, nous retournons des représentations graphiques de l'ensemble documentaire que nous appelons des cartes. Ces cartes permettent alors à

l'utilisateur de naviguer sur l'ensemble et de visualiser de façon interactive différentes informations, telles des regroupements entre ressources de l'ensemble. La partie suivante de cet article présente en détails une carte obtenue à partir du corpus d'étude.

4 Cartographie du corpus d'étude

Dans cette partie, nous présentons une carte construite selon la méthode détaillée précédemment à partir du corpus catégorisé en spécialités médicales. Cette carte met visuellement en évidence 12 groupes de ressources obtenus par CHA (nombre de groupes choisi empiriquement).

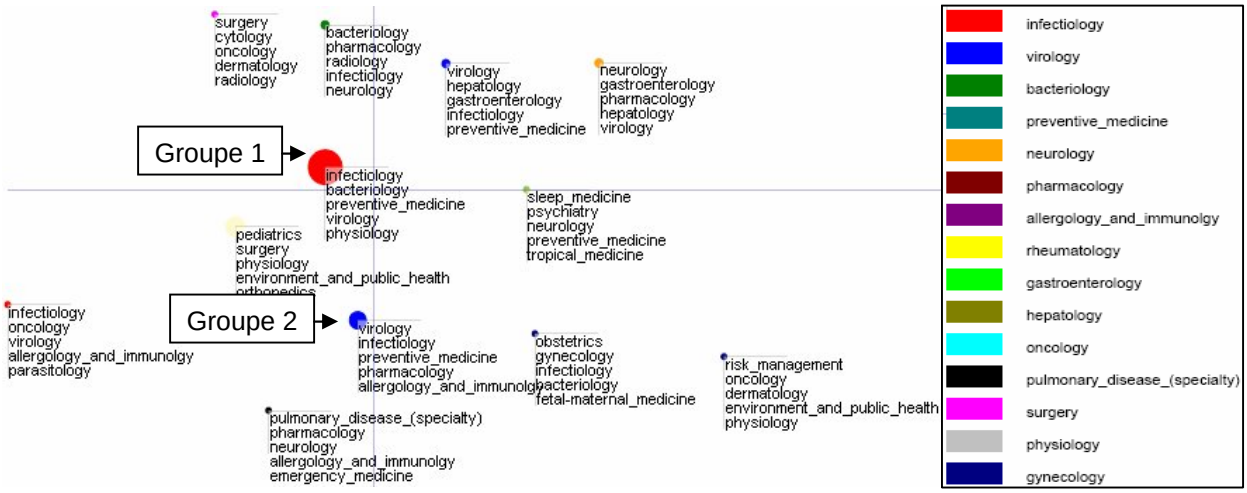

Figure 1 : Carte du corpus construite avec la plate-forme ProxiDocs. Version interactive de cette carte au format SVG : http://www.info.unicaen.fr/~troy/cismef/

Chaque groupe de ressources est représenté par un disque de taille proportionnelle à sa cardinalité. La couleur attribuée à chaque disque correspond à sa spécialité majoritaire, c'est-à-dire celle ayant le score le plus élevé dans les ressources du groupe. Une légende attribuant une couleur aux 15 spécialités majoritaires dans les groupes est disponible sur la partie droite de la figure. Chaque groupe est caractérisé par ses cinq spécialités de score le plus élevé. Chaque disque sur la carte est un lien hypertexte vers un rapport détaillé sur les propriétés du groupe. La carte révèle des regroupements entre ressources du corpus d'étude par rapport aux spécialités médicales.

Afin de construire cette carte, nous n'avons pris en considération que des spécialités médicales principales relevant d'un domaine médical (comme *infectiology*, *virology*, *neurology*, etc.). Les spécialités médicales dites transversales, c'est-à-dire des spécialités ne constituant pas un tout et étant applicables aux spécialités médicales principales, n'ont pas été prises en considération dans la cartographie (nous avons par exemple comme spécialités médicales transversales : *therapeutics*, *anatomy*, *diagnosis*, *economics*, *ethics*, *organization and administration*, etc.). La carte ne révèle donc que les domaines médicaux abordés dans le corpus et non les moyens de mettre en œuvre de tels domaines. Dans une étude préliminaire conservant l'ensemble des spécialités au même niveau, nous avions observe une influence significative des spécialités transversales sur la construction des cartes. La prise en considération de ces spécialités, globalement majoritaires dans l'ensemble des groupes, avait pour effet de masquer les thématiques dénotées par les spécialités principales. Il semblait donc plus pertinent de les étudier séparément.

La carte ainsi met en évidence la répartition des spécialités principales dans le corpus d'étude. Le groupe 1 sur la carte possède 36 ressources, les trois spécialités majoritaires dans ce groupe sont *infectiology*, *bacteriology* et *preventive medecine*. Un parcours rapide des ressources de ce groupe révèle qu'elles abordent des thématiques assez variées, certes liées aux spécialités principales, mais sans réelle lien entre les ressources. Au contraire le groupe 2 contenant 11 ressources et possédant comme spécialité majoritaire *virology*, *infectiology* et *preventive medecine* regroupe des ressources toutes étroitement liées au domaine de la virologie (par exemple, sur des ressources traitant du virus de la grippe et des différents vaccins

existants contre ce virus). Ce phénomène se retrouvent dans une très grande majorité des autres groupes de la carte : les ressources abordent des thématiques étroitement liées aux spécialités majoritaires du groupe les contenant.

5 Conclusion

Nous avons présenté dans cet article les premiers résultats d'une expérience dédiée à la cartographie d'un corpus du domaine médical. La carte ainsi construite a permis de visualiser des proximités et des différences entre documents. Cette carte a révélé une répartition des spécialités très différentes de celle obtenue lors de l'analyse globale du corpus. Ainsi, des groupes de documents de spécialités majoritaires particulièrement « enfouies » dans le classement global du corpus ont pu être mis en évidence. D'une certaine manière, des signaux faibles dans l'analyse globale du corpus sont ressortis à travers la carte, ceci à l'aide d'une prise en considération d'un niveau d'analyse de granularité différente : le groupe de documents.

6 Bibliographie

[BEN 80] Benzecri J.-P., *L'analyse des données - tome 2 : l'analyse des correspondances*, éditions Bordas, 1980.

[BOU 80] Bouroche J.-M., Saporta G., *L'Analyse des Données*, Paris : PUF, 1980.

[DAR 05] Darmoni S.J., Névéol A., Renard J.M., Gehanno J.F., Soualmia L.F., Dahamna B., Thirion B., "A MEDLINE Categorization Algorithm", *BMC*, sous presse, 2005.

[NEV 04] Névéol A., Soualmia L.F., Douyère M., Rogozan A., Thirion B. et Darmoni S.J., "Using CISMeF MeSH "Encapsulated" Terminology and a Categorization Algorithm for Health Resources", International *Journal of Medical Informatics* Vol. 73(1), 57-64, 2004.

[ROY 04] Roy T., Beust P., "ProxiDocs, un outil de cartographie et de catégorisation thématique de corpus", Actes des *Journées internationales de l'Analyse des Données Textuelles*, 978-987, 2004.

[ROY 05] Roy T., "Une plate-forme logicielle dédiée à la cartographie thématique de corpus", Actes de *TALN/RECITAL*, 545-554, 2005.

[SAM 69] Sammon J. W., "A Nonlinear Mapping for Data Structure Analysis", *IEEE Transactions on computers* C-18(5), 401-409, 1969.