Extraire semi-automatiquement des connaissances dans la littérature biomédicale

Jessica Pinaire, Jérôme Azé, Sandra Bringay, Paul Landais

Extraire semi-automatiquement des connaissances dans la littérature biomédicale

Jessica Pinaire[1,3], Jérôme Azé[1], Sandra Bringay[1,4], Paul Landais[2,3]

[1] LIRMM, UMR 5506, Université de Montpellier, France
prenom.nom@lirmm.fr

[2] ÉQUIPE D'ACCUEIL 24-15, Institut Universitaire de Recherche Clinique, Université de Montpellier, Montpellier, France
paul.landais@inserm.fr

[3] CHU, Département d'information médicale, Nîmes, France
paul.landais@chu-nimes.fr

[4] AMIS, Université Paul Valéry, Montpellier, France
sandra.bringay@univ-montp3.fr

Résumé :
La littérature biomédicale résume les connaissances scientifiques actuelles. La quantité de données disponibles est trop importante pour pouvoir être analysée manuellement. Il devient crucial de construire des outils d'analyse automatisés pour supporter les activités de recherche bibliographique. Dans ce contexte, nous proposons une méthode semi-automatique, supportée par des visualisations, pour explorer les bases bibliographiques, que nous appliquons à la thématique des trajectoires de patients.

Nous avons utilisé et évalué plusieurs modèles statistiques et leurs représentations visuelles pour caractériser le contenu des articles et donner au chercheur une vision d'ensemble du contenu de la base.

L'approche a permis d'identifier 81 articles dans la littérature parmi 11 490 articles, qui ont été étudiés lors d'une revue systématique manuelle.

Nous proposons une approche semi-automatique efficace pour l'exploration de la littérature biomédicale.

Mots-clés : Fouille de textes, Visualisation de connaissances, Recherche d'informations.

1 Introduction et motivations

Notre travail se situe dans le contexte de l'aide à la réalisation d'études bibliographiques. Notre intérêt pour le domaine d'application biomédical est motivé par le fait que ce domaine a connu la plus forte croissance de tous les domaines scientifiques en terme de volume de publications. En février 2016, le moteur de recherche PubMed indexe plus de 25 millions de citations. Si plusieurs communautés (*e.g.* recherche d'informations, fouille de textes) se sont penchées sur le défi de l'extraction automatisée d'informations dans la littérature (Fleuren & Alkema, 2015; Huang & Lu, 2016; Vazquez *et al.*, 2011; Song, 2014; Geifman *et al.*, 2015), à ce jour, il n'existe pas, à notre connaissance, d'outils permettant aux chercheurs d'explorer facilement ces grands volumes de données.

Dans cet article, nous proposons de combiner plusieurs outils de fouille de textes pour aider le chercheur à visualiser le contenu de très nombreux articles en mettant en évidence les thèmes abordés, à l'aide de mots saillants et de réseaux de mots. L'objectif de ces visualisations est d'aider l'expert à formuler une question de recherche précise qui lui permettra dans un deuxième temps de filtrer le corpus et d'aboutir à un nombre de documents manuellement exploitables pour réaliser une revue systématique classique. Cette méthode a été appliquée avec succès pour une étude sur le thème des trajectoires de patients.

La section 2 motive ces travaux et décrit un rapide état de l'art. La section 3 décrit la stratégie employée et les outils mis en œuvre. La section 4 détaille la partie applicative sur le corpus de textes des trajectoires. Finalement, nous concluons et donnons quelques perspectives.

2 Motivations et bref état de l'art

Un travail de recherche doit nécessairement débuter par une analyse des bases bibliographiques spécialisées, afin de positionner ces travaux par rapport à l'état des connaissances scientifiques reflété via les publications.

Dans ce contexte, une revue systématique correspond à une recherche bibliographique basée sur une question clairement formulée. Dans un premier temps, une telle revue utilise des méthodes systématisées, répétitives et explicites pour identifier, sélectionner et évaluer de façon critique des articles de recherche répondant à cette question. Dans un deuxième temps, la revue permet de recueillir et d'analyser les données extraites des publications filtrées. Une méta-analyse peut alors être utilisée en complément d'une revue pour analyser et résumer les résultats des études. Cette méta-analyse se base sur des techniques statistiques. On trouve plusieurs méthodes de revues systématiques comme la méthode PRISMA [1] (Moher *et al.*, 2009) ou Cochrane [2] (Higgins *et al.*, 2008). Si ces méthodes s'avèrent très efficaces quand la question de recherche est clairement formulée, elles ne donnent que peu de directives quand le besoin informationnel du chercheur est plus vague.

C'est justement dans ce contexte que nous nous positionnons. Notre objectif est de proposer une méthode semi-automatique, facilitant le processus de recherche bibliographique en permettant une exploration sans *a priori* des articles. Notre objectif est d'aider le chercheur à identifier les thèmes d'intérêts de sa communauté, de se focaliser sur certains et de formuler une série de questions de recherche plus précises, qui pourront ensuite être utilisées en entrée d'une revue automatique.

Nous prenons ici comme cas d'étude la thématique des "trajectoires de patients". Ce champ d'étude a donné lieu à un nombre croissant de publications scientifiques dans de nombreuses revues médicales au cours des dix dernières années.

Nous utilisons le moteur de recherche PubMed développé par le NCBI [3] et considéré comme l'une des références dans les domaines de spécialisation de la biologie et de la médecine. Il donne accès à la base de données bibliographique MEDLINE qui contenait en février 2016 plus de 25 millions de citations publiées depuis 1950 dans environ 5 000 revues biomédicales [4]. Des travaux ont démontré que MEDLINE possède la couverture la plus complète des références bibliographiques biomédicales (Kelly & St Pierre-Hansen, 2008). PubMed est largement utilisé par la communauté scientifique biomédicale. En 2015, il y a eu, en moyenne, près de 8 millions de requêtes par jour [5].

1. Preferred Reporting Items for Systematic Reviews and Meta-Analyses
`http://www.prisma-statement.org/`
2. `http://community.cochrane.org/cochrane-reviews`
3. National Center for Biotechnology Information
4. `http://www.ncbi.nlm.nih.gov/pubmed`
5. `https://www.nlm.nih.gov/services/pubmed_searches.html`

Pour notre cas d'étude, si l'on interroge pubmed avec les mots clés "health", "patient(s)", "trajectories", "trajectory", "path", "pathway(s)", sur une période de presque 15 ans, nous obtenons un corpus de 11 490 documents, qu'il est impossible d'exploiter manuellement. Cet exemple montre que développer des méthodes pour un accès intelligent aux données contenues dans l'ensemble des publications scientifiques biomédicales reste un véritable défi.

Dans le domaine de la recherche bibliographique, les méthodes de fouilles de textes ont été utilisées pour faciliter le travail du chercheur en ciblant mieux la recherche documentaire et en réduisant le temps de travail (Cohen *et al.*, 2006). Dans le cadre de review systématiques, par exemple, il existe quatre tâches (Thomas *et al.*, 2011) pour lesquelles les techniques de fouille de textes sont généralement employées : 1) la reconnaissance automatique de termes dans les textes (Frantzi *et al.*, 2000) ; 2) la classification supervisée de documents dans des thèmes spécifiques (Joachims, 1998; Mo *et al.*, 2015; Sebastiani, 2002; Frunza *et al.*, 2011) ; 3) la classification non supervisée de documents qui regroupe les documents dans des thèmes. Chaque groupe correspond au thème partagé par l'ensemble des documents du groupe et par aucun autre document de la collection (Blei *et al.*, 2003; Reinert, 1983; Bada, 2014) ; 4) le résumé fait en sélectionnant des phrases à partir de chaque document basé sur l'importance de ses termes qui sont combinés avec des techniques de classement (Bollegala *et al.*, 2010).

D'autres auteurs utilisent la fouille de textes pour d'autres finalités. Par exemple, dans (Lin *et al.*, 2008), les auteurs créent des bases de données de correspondances reliant les auteurs avec les abréviations de leurs noms et réalisent une analyse des co-auteurs. Dans (Leitner & Valencia, 2008), ils annotent les abstracts de deux façons, d'abord le gène ou la protéine étudiée, puis les interactions de protéines et/ou les fonctions du gène. *In fine*, ils catégorisent les documents suivant ces annotations.

Si ces méthodes sont intéressantes, elles ne permettent pas d'explorer de manière globale et sans *a priori*, c'est-à-dire sans question de recherche précisément définie, les grands corpus d'articles. Notre méthode permet non seulement une telle exploration visant à filtrer les documents pertinents pour une revue plus systématique mais également d'accompagner cette exploration par des visualisations qui facilitent l'interprétation du chercheur. Dans la section suivante 3, nous décrivons les différentes étapes de cette méthode.

3 Analyse bibliographique semi-automatisée par fouille de textes

Notre méthode semi-automatique, basée sur des techniques de fouille de textes, aide : 1) à interpréter de gros volumes de données générés par la littérature biomédicale ; 2) à explorer le corpus par raffinements successifs jusqu'à la formulation d'une question de recherche. Nous détaillons dans cette section les trois étapes de la méthode (voir figure 1).

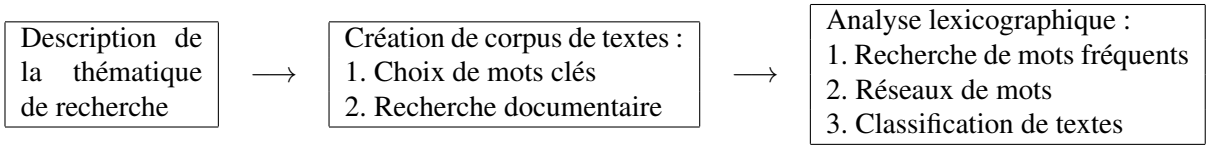

FIGURE 1 – Stratégie d'analyse

3.1 Étape 1 : Définition de la thématique de recherche

Afin de débuter l'exploration du corpus, nous demandons à l'expert de définir sa thématique de recherche en langage naturel en explicitant son besoin informationnel. Cela nous permet de définir des questions sans *a priori* qui intègrent des aspects thématiques *e.g. quelles communautés travaillent sur ma thématique ? Quelles maladies sont concernées par cette thématique ? Quels sont les sujets abordés par les articles ?*

Dans un deuxième temps, l'expert doit choisir une liste de mots clés pour représenter cette thématique qui sera utilisée à l'étape 2 pour générer le corpus.

3.2 Étape 2 : Création du corpus de textes

À partir du premier ensemble de mots clés, nous utilisons l'API de Pubmed pour rechercher les documents. Un corpus de textes, constitué de l'union du titre et de l'abstract est ensuite créé. Afin de ne retenir que les autres termes, les mots clés utilisés pour la sélection sont supprimés du corpus précédemment créé. Ensuite, une première exploration du corpus par Latent Dirichlet Allocation (LDA) (Blei *et al.*, 2003), à l'aide du package LDAviz (Sievert & Shirley, 2014) dans R, permet de repérer des abréviations (*e.g.* ppci, rr, hr) peu répandues dans le langage courant, que nous écartons du corpus.

Enfin, les prétraitements suivants sont appliqués : a) lemmatisation du texte : les verbes sont ramenés à l'infinitif, les noms au singulier et les adjectifs au masculin singulier ; b) enrichissement du dictionnaire : le logiciel détecte des termes non reconnus. Pour ne pas perdre trop d'informations, ces termes non reconnus ont été récupérés, lemmatisés par TreeTagger [6] et réinjectés dans le dictionnaire après vérification manuelle et ajout de termes médicaux spécifiques ou sigles bien connus comme AMI (Acute Myocardial Infarction). Dans la suite, les analyses sont réalisées avec les formes pleines (noms, adjectifs, adverbes et verbes).

3.3 Étape 3 : Analyse lexicographique

Nous analysons le corpus précédent avec le logiciel IRaMuteQ [7]. Il s'agit d'une Interface de R pour les Analyses Multidimensionnelles de Textes et de Questionnaires (Ratinaud & Déjean, 2009). Il permet de faire des analyses statistiques sur des corpus de textes (Ratinaud & Marchand, 2012).

Nuage de mots : C'est une représentation synthétique de la distribution des termes. Les mots les plus fréquents sont au centre et la taille de police varie proportionnellement au nombre d'occurrences.

Analyse de similarités : C'est une technique, reposant sur la théorie des graphes, classiquement utilisée pour décrire des représentations sociales, sur la base de questionnaires d'enquête (Flament, 1981). L'objectif de l'analyse de similarités (ADS) est d'étudier la proximité et les relations entre les éléments d'un ensemble, sous forme d'"arbres maximum". Pour chaque corpus, nous avons choisi la représentation en arbre *graphopt* décrite dans (Csardi, 2015) et l'algorithme de (Brandes, 2001) pour décrire les communautés au sens du plus court chemin, ce qui met en évidence les mots les plus souvent associés dans une même phrase ou un texte.

6. http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

7. http://www.iramuteq.org/

Classification de textes : La classification de Reinert (Reinert, 1983) est une classification hiérarchique descendante (CDH) s'effectuant en plusieurs étapes. Elle propose une approche globale du corpus. Après partitionnement de celui-ci, elle identifie des classes statistiquement indépendantes de mots. Ces classes sont interprétables grâce à leurs profils, qui sont caractérisés par des mots spécifiques corrélés entre eux. La CDH résume cela par un dendrogramme.

4 Application au corpus Trajectoire

Le CHU de Nîmes s'intéresse aux trajectoires de patients. Dans cette section, nous allons appliquer notre méthode en collaboration avec un expert de cette thématique. L'expert est un professeur de médecine spécialisé dans l'analyse des bases médico-économiques.

4.1 Étape 1 : Définition de la thématique de recherche

La table 1 décrit les questions sans *a priori* formulées par le chercheur.

Questions sans *a priori*
Q1 : Existe-t-il des études sur les trajectoires de patients ?
Q2 : Quels sont les thèmes abordés dans ces études ? (la prise en charge, le traitement, les coûts,...)
Q3 : Pour quelles pathologies sont étudiées les trajectoires ?

TABLE 1 – Questions sans *a priori*

À partir de ces questions, l'expert choisi des mots clés décrivant la thématique qu'il souhaite explorer. Nous retenons les termes suivants pour décrire la notion de trajectoire : "trajectoire", "parcours" et "chemin".

4.2 Étape 2 : Création du corpus de texte

Dans PubMed, nous avons recherché les documents qui traitent des trajectoires ou parcours de soins de patients. Nous avons effectué une recherche selon le thème et contraintes résumés dans le tableau 2. Ainsi, la requête : **C1 + T1 + C2 + C3**, sélectionne les articles du domaine médical qui traitent des trajectoires, écrits entre le 1^{er} janvier 2000 et le 31 octobre 2015, en anglais. Il a résulté de la recherche documentaire un total de 11 490 articles.

Thème et contraintes	**Mots clés**
C1 : contexte medical	"health", "patient(s)"
T1 : Trajectoire	"trajectories", "trajectory", "path", "pathway(s)"
C2 : dates	January 1st 2000 to October 31th 2015
C3 : langue	English

TABLE 2 – Mots clés utilisés dans la recherche documentaire

4.3 Étape 3 : Analyse Lexicographique

Nuage de mots : Les termes les plus saillants de la figure 2 sont "care", "study", "cancer", "cell", "treatment", "disease". Les termes "study" et "care" nous apportent des éléments sur le contenu du corpus : ces articles traitent de la notion de soins. Ensuite, avec les termes, "treatment", "disease", il est avant tout question de parcours de soins du patient. Nous pouvons alors répondre positivement à la question **Q1** : il existe bien des études sur les trajectoires de patients.

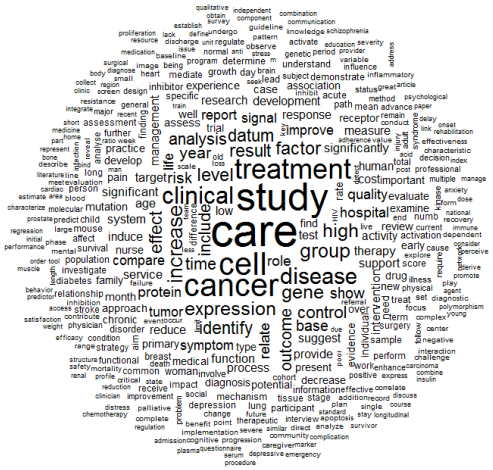

FIGURE 2 – Nuage de mots du corpus Trajectoire

Cette première analyse fait ressortir les mots importants caractéristiques du corpus. Toutefois, au delà de l'intérêt visuel, le nuage n'apporte pas beaucoup d'informations. Avec l'ADS, nous allons explorer en détail ce corpus, en rapprochant ces termes d'autres termes pour en déduire leur signification et définir un contexte.

Analyse de similarités : Cette présentation en arborescence met en évidence des réseaux de termes fortement co-occurents et permet ainsi de cerner les thématiques les plus souvent abordées au travers des articles. La figure 3 montre que l'étude est fortement liée au soin, à la maladie et plus spécifiquement au cancer. Dans la représentation des communautés, on voit apparaître celle des maladies étudiées, causant des dysfonctionnements sévères, chroniques, du cœur, des reins ou encore des poumons, que l'on soit jeune, âgé, homme ou femme. Ces deux dernières remarques répondent à la question **Q3**.
Notons que la notion de cancer est représentée dans une feuille différente de celle des maladies. Cette observation permet à l'expert de se poser la question suivante : *pour quelles raisons cette pathologie est-elle particularisée dans le contexte des trajectoires ?* Cette feuille contient le mot "treatment", qui est une notion intrinsèque à celle de trajectoire car cohérent avec le travail réalisé par les chercheurs de cette discipline ces dernières années pour organiser les soins de cette pathologie lourde.

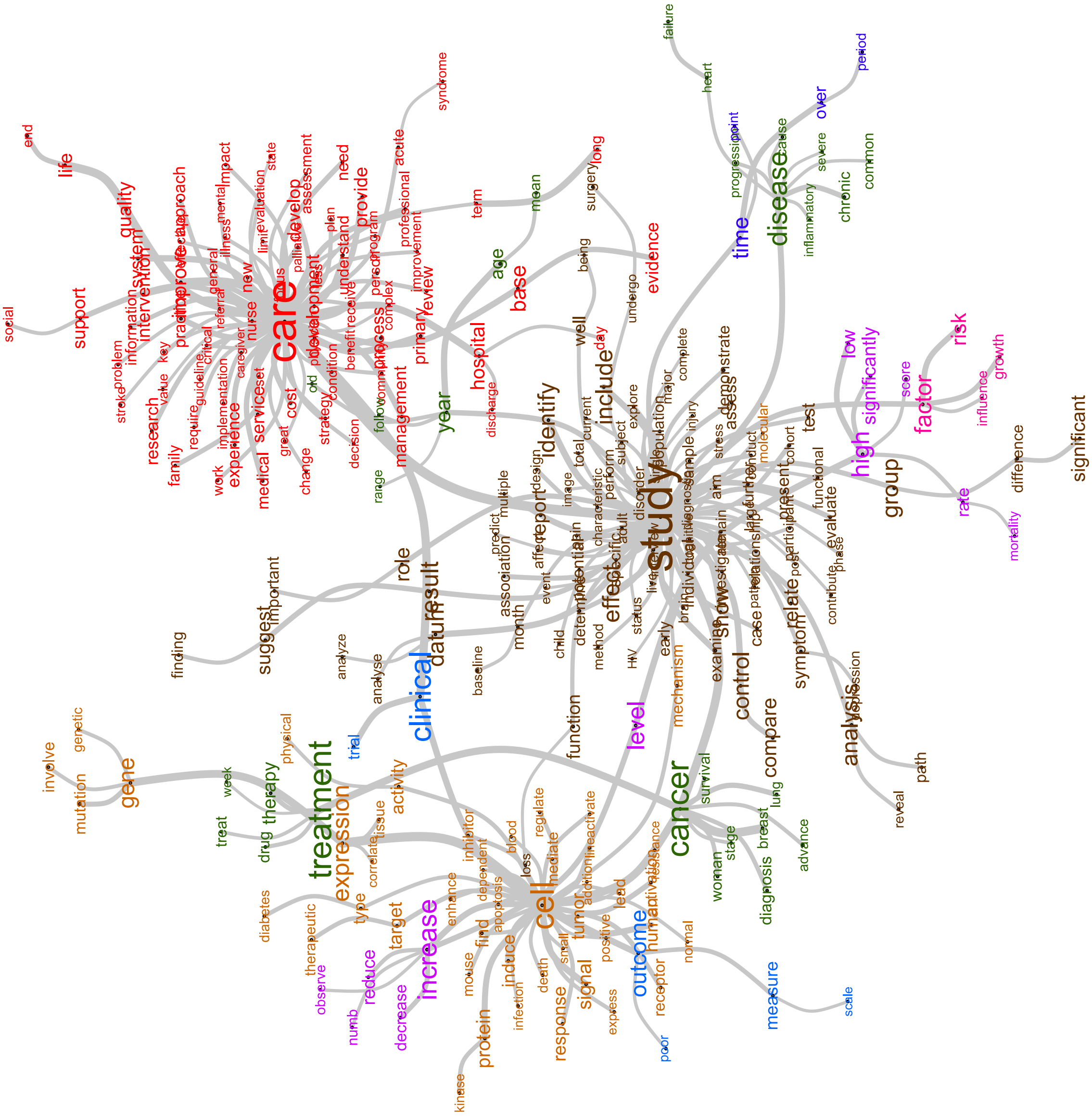

FIGURE 3 – ADS et communautés du corpus Trajectoire

Dans une autre branche la notion de cancer est liée à celle de génétique. L'utilisation du mot clé "pathway" fait ressortir tous les articles évoquant les voies de signalisation cellulaire et les gènes. Dans notre contexte applicatif, cette visualisation nous permet d'éliminer ce thème car il ne s'agit plus de la trajectoire du patient au sens série d'événements médicaux que nous souhaitons étudier mais d'interactions au niveau des cellules. De plus, ceci montre que plusieurs définitions de la notion de trajectoire existent et que le sens doit être affiné via un item à ajouter dans la grille de lecture de la revue systématique.

L'ADS a permis d'explorer en partie le contenu des articles, en identifiant des thèmes, caractérisés par des associations de termes. Afin de compléter ces analyses, nous allons maintenant chercher à savoir si ces thèmes sont suffisamment représentatifs du contenu du corpus pour y classer les articles. Nous nous focaliserons ensuite sur les articles qui n'auront pas trouvé leur place dans ce regroupement afin d'en apprendre davantage sur les articles à la marge des travaux courants.

Classification de textes : À la suite de cette classification, 80% des articles ont été répartis dans onze classes disjointes (voir figure 4). Nous avons ensuite réalisé une deuxième classification sur le sous-corpus constitué des 3 160 articles non classés lors de la première analyse. Nous identifions cinq classes constituées de 99% des articles (voir figure 5). Seulement trois articles n'ont pu être classés.

À la suite de ces deux classifications, nous pouvons répondre à la question **Q2** en listant les thèmes étudiés dans les articles concernant les trajectoires de patients. Le premier thème est la maladie avec, par exemple, les troubles métaboliques comme le diabète et les complications cardiovasculaires. Certains articles concernent le ressenti du patient, ses angoisses et le vécu de sa maladie. Dans le parcours du patient, il y a le soutien par son environnement proche, la famille mais aussi les dispositifs mis en place comme l'intervention d'une infirmière à domicile. D'autres articles traitent de la fin de vie, des soins palliatifs et des procédés mis en place pour gérer cette dernière étape de la maladie. Un autre thème évoqué est la recherche clinique, avec la constitution de cohortes, la collecte de données, les méthodes employées dans ces différentes études. Ensuite, il y a l'organisation de l'hôpital, ses différents services, le personnel dont il dispose pour soigner les patients et les coûts associés. D'autres articles sont axés sur les réglementations et recommandations sanitaires régies par les guides de bonnes pratiques.

Avec cette méthode, il est relativement simple de repérer les thèmes qui nous intéressent plus spécifiquement, puis, soit de les explorer en pratiquant de nouvelles analyses de fouille textuelles, soit de les analyser avec des revues systématiques. Il est également aisé d'écarter tous les articles hors sujet, comme ceux traitant de la génétique, dont la thématique avait déjà été repérée par l'ADS. En d'autres termes, ces trois représentations nous donnent une vue d'ensemble des thèmes abordés dans ce corpus de textes, en organisant les articles dans ces catégories, ce qui permet à l'expert d'effectuer un premier tri en sélectionnant uniquement les articles d'intérêt pour son étude.

5 Discussions et Conclusions

Notre stratégie, sous forme de raffinements successifs, a permis de mettre en exergue les mots importants sous forme de nuages, puis de les connecter à d'autres mots avec l'ADS, mettant ainsi, en évidence un univers lexical. La classification des articles sans *a priori*, met en évidence les divers thèmes qui recouvrent l'ensemble de ces articles.

FIGURE 4 – Résultats d'une classification des articles pour le corpus Trajectoire

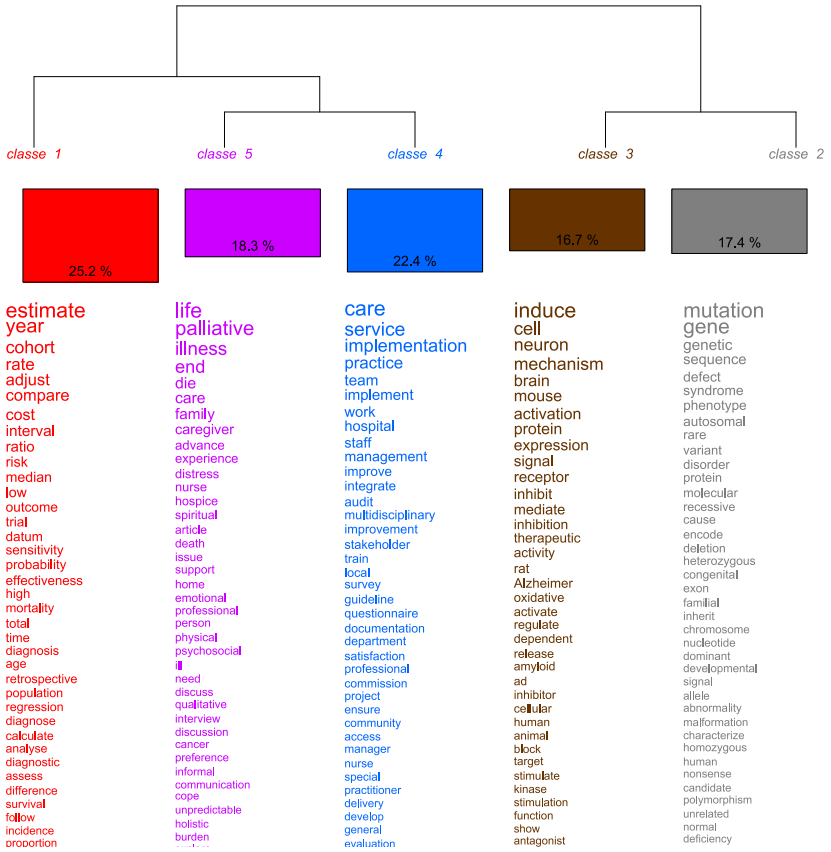

FIGURE 5 – Résultats de la deuxième classification des articles pour le corpus Trajectoire

L'ADS donne le contexte d'utilisation des mots. Cet outil d'analyse macroscopique offre une visualisation condensée et compartimentée des différents thèmes abordés par l'ensemble des articles. Dans cette étape, les articles sont considérés comme un ensemble à part entière. L'agencement des réseaux de mots permet de cerner l'importance d'un thème, de le placer par rapport aux autres dans un contexte déterminé, comme nous avons pu le faire avec le thème cancer. Cette représentation sous forme d'arbre, est particulièrement intéressante pour explorer non pas des articles mais un thème, sans se soucier de savoir s'il recouvre majoritairement un article.

La troisième étape de classification permet d'avoir un point de vue microscopique sur le corpus, celui des articles et de savoir si les thèmes sont représentatifs de ces articles. Ces derniers sont considérés comme des entités à part entière. Grâce à cette représentation, nous pouvons également mettre en évidence des thèmes qui recouvrent le corpus, mais aussi sélectionner les articles qui vont le plus nous intéresser, et *a contrario* identifier et retirer rapidement les documents hors sujet.

Cette approche a été efficace pour explorer le thème "Trajectoire" en collaboration avec un expert de la thématique qui a pu répondre à des questions sans *a priori* pour lesquelles il n'était pas possible de chercher une liste finie d'indicateurs dans les textes.

La recherche documentaire sur les trajectoires de patients, a montré que ce type d'étude

suscite un intérêt croissant dans la communauté biomédicale : que ce soit pour en apprendre davantage sur l'évolution d'une pathologie grâce au suivi du patient ou pour comparer les parcours de soins afin de mettre en place des stratégies par des procédures de soins facilitant le travail des personnels de santé tout en fournissant un cadre rassurant au patient et en réduisant les coûts. Nous retenons de cette étude que le concept de trajectoire est exploré plus particulièrement en cancérologie.

Cette première approche ouvre des perspectives intéressantes d'aide à l'analyse documentaire. À court terme, nous allons utiliser la méthode décrite dans cet article pour une analyse plus complète des trajectoires de patients, en croisant cette thématique avec le contexte d'étude des bases hospitalières de la tarification à l'activité mais également celui de l'infarctus du myocarde. Nous compléterons cette analyse avec une revue systématique en appliquant la méthode PRISMA. Nous utiliserons la grille de lecture associée à cette méthode et intégrerons de nouveaux indicateurs issus de l'exploration *a priori*, comme la définition du concept de trajectoire. À plus long terme, nous souhaitons améliorer la méthode proposée qui reste préliminaire. D'autres méthodes de fouille de textes peuvent être appliquées pour améliorer l'analyse bibliographique. Nous envisageons notamment la visualisation des termes d'intérêt correspondant aux items de la méthode PRISMA dans les publications pour aider l'analyse manuelle systématique (Abi-Haidar *et al.*, 2016) ou encore le résumé des thématiques qui consiste à prendre les points essentiels d'un texte pour en faire un paragraphe (Liu *et al.*, 2015).

Références

ABI-HAIDAR A., YANG B. & GANASCIA J.-G. (2016). Mapping the first world war using interactive streamgraphs. *Sociology and Anthropology*, **4**, 12–16.

BADA M. (2014). Mapping of biomedical text to concepts of lexicons, terminologies, and ontologies. *Methods in Molecular Biology (Clifton, N.J.)*, **1159**, 33–45.

BLEI D. M., NG A. Y. & JORDAN M. I. (2003). Latent Dirichlet Allocation. *J. Mach. Learn. Res.*, **3**, pp. 993–1022.

BOLLEGALA D., OKAZAKI N. & ISHIZUKA M. (2010). A bottom-up approach to sentence ordering for multi-document summarization. *Information Processing and Management*, **46**(1), 89–109.

BRANDES U. (2001). A Faster Algorithm for Betweenness Centrality. *Journal of Mathematical Sociology*, **25**(2), 163–177.

COHEN A. M., HERSH W. R., PETERSON K. & YEN P.-Y. (2006). Reducing Workload in Systematic Review Preparation Using Automated Citation Classification. *Journal of the American Medical Informatics Association*, **13**(2), 206–219.

CSARDI M. G. (2015). Package 'igraph'. *The Comprehensive R Archive Network. See http ://cran. r-project. org/web/packages/igraph/igraph. pdf.*

FLAMENT C. (1981). Similarity analysis : A technique for researches in social representations. *Cahiers de Psychologie Cognitive*, **1**(4), 375–395.

FLEUREN W. W. & ALKEMA W. (2015). Application of text mining in the biomedical domain. *Methods*, **74**, 97–106.

FRANTZI K., ANANIADOU S. & MIMA H. (2000). Automatic Recognition of Multi-Word Terms : the C-value/NC-value method. *International Journal on Digital Libraries*, **3**(2), 115–130.

FRUNZA O., INKPEN D., MATWIN S., KLEMENT W. & O'BLENIS P. (2011). Exploiting the systematic review protocol for classification of medical abstracts. *Artificial Intelligence in Medicine*, **51**(1), 17–25.

GEIFMAN N., BHATTACHARYA S. & BUTTE A. J. (2015). Immune modulators in disease : integrating knowledge from the biomedical literature and gene expression. *Journal of the American Medical Informatics Association*.

HIGGINS J. P., GREEN S. *et al.* (2008). *Cochrane handbook for systematic reviews of interventions*, volume 5. Wiley Online Library.

HUANG C.-C. & LU Z. (2016). Community challenges in biomedical text mining over 10 years : success, failure and the future. *Briefings in Bioinformatics.*, **17**(1), 132–44.

JOACHIMS T. (1998). Text Categorization with Support Vector Machines : Learning with Many Relevant. *Machine Learning : ECML-98*, **1398**, 137–142.

KELLY L. & ST PIERRE-HANSEN N. (2008). So many databases, such little clarity : Searching the literature for the topic aboriginal. *Canadian family physician Médecin de famille canadien*, **54**(11), 1572–1573.

LEITNER F. & VALENCIA A. (2008). A text-mining perspective on the requirements for electronically annotated abstracts. *FEBS Letters*, **582**(8), 1178–1181.

LIN J. M., BOHLAND J. W., ANDREWS P., BURNS G. A., ALLEN C. B. & MITRA P. P. (2008). An analysis of the abstracts presented at the annual meetings of the Society for Neuroscience from 2001 to 2006. *PLoS ONE*, **3**(4).

LIU F., FLANIGAN J., THOMSON S., SADEH N. & SMITH N. A. (2015). Toward abstractive summarization using semantic representations. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 1077–1086.

MO Y., KONTONATSIOS G. & ANANIADOU S. (2015). Supporting systematic reviews using LDA-based document representations. *Systematic Reviews*, **4**.

MOHER D., LIBERATI A., TETZLAFF J. & ALTMAN D. G. (2009). Preferred reporting items for systematic reviews and meta-analyses : the prisma statement. *British Medical Journal*, **339**.

RATINAUD P. & DÉJEAN F. (2009). IRaMuTeQ : implémentation de la méthode ALCESTE d'analyse de texte dans un logiciel libre ». In *MASHS2009*, p. 1–22.

RATINAUD P. & MARCHAND P. (2012). Application de la méthode ALCESTE à de « gros » corpus et stabilité des « mondes lexicaux » : analyse du « CableGate » avec IRaMuTeQ ». In *Actes des 11ème Journées internationales d'Analyse statistique des Données Textuelles*, p. 835–844.

REINERT A. (1983). Une méthode de classification descendante hiérarchique : application à l'analyse lexicale par contexte. *Les cahiers de l'analyse des données*, **VIII, (2)**, 187–198.

SEBASTIANI F. (2002). Machine Learning in Automated Text Categorization. *ACM Comput. Surv.*, **34**(1), 1–47.

SIEVERT C. & SHIRLEY K. (2014). LDAvis : A method for visualizing and interpreting topics. In *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*, p. 63–70, Baltimore, Maryland, USA : Association for Computational Linguistics.

SONG M. (2014). Takes : Two-step Approach for Knowledge Extraction in Biomedical Digital Libraries. *Journal of Information Science Theory and Practice*, **2**(1), 6–21.

THOMAS J., MCNAUGHT J. & ANANIADOU S. (2011). Applications of text mining within systematic reviews. *Research Synthesis Methods*, **2**(1), 1–14.

VAZQUEZ M., KRALLINGER M., LEITNER F. & VALENCIA A. (2011). Text Mining for Drugs and Chemical Compounds : Methods, Tools and Applications. *Molecular. Informatics.*, **30**(Issue 6-7), 506–519.