Indexation conceptuelle par propagation. Application à un corpus d'articles scientifiques liés au cancer

Nicolas Fiorini[*] – Sylvie Ranwez[*] – Vincent Ranwez[**] – Jacky Montmain[*]

[*] *Centre de recherche LGI2P de l'école des mines d'Alès, Parc Scientifique Georges Besse, F-30 035 Nîmes cedex 1*
prenom.nom@mines-ales.fr
[**] *Montpellier SupAgro, UMR AGAP, F-34060 Montpellier*
vincent.ranwez@supagro.inra.fr

RÉSUMÉ. *Si la recherche d'information conceptuelle a montré son efficacité dans différents contextes, elle nécessite de disposer de corpus de ressources indexées avec des concepts issus d'une ontologie de domaine. Or le processus d'indexation est souvent lourd et fastidieux et des solutions doivent être imaginées pour assister les experts dans cette tâche. Nous avons étendu notre méthode de propagation d'indexations vectorielles au cas de l'indexation conceptuelle, ce qui nous permet de suggérer à l'utilisateur une indexation conceptuelle pour une nouvelle ressource, lorsque celle-ci est positionnée sur une carte sémantique. Pour cela nous maximisons une fonction objectif mesurant la similarité entre l'indexation proposée et celles des documents présents dans le voisinage du document à annoter. Cette méthode est appliquée à l'indexation de publications scientifiques dans le domaine du cancer.*

ABSTRACT. *Concept-based information retrieval is known to be a powerful and reliable process. However, the need of a semantically annotated corpus and its respective data structure – e.g. a domain ontology – can be problematic. The conception and enlargement of a semantic index is a tedious task, which needs to be addressed. We previously suggested an annotation propagation approach in a vector space representation of the corpus to help users enriching a corpus. In this paper, we propose an extension of this process for semantic indexations. Starting from a map showing the documents of the corpus, a user will just have to place a new resource on this map to obtain a first annotation of this resource. This annotation is obtained by optimizing an objective function, which assesses the semantic similarity between the annotation suggested for this new resource and those of documents found in its vicinity. Here, we illustrate this strategy on tumor-related scientific papers.*

MOTS-CLÉS : *Indexation par propagation, indexation conceptuelle, similarité sémantique, recherche d'information conceptuelle, interaction homme-machine.*

KEYWORDS: *propagation algorithm, conceptual indexing, semantic similarity, conceptual information retrieval, man-machine interaction.*

1. Introduction

La société numérique bouleverse notre rapport à l'information et nos organisations sociales, économiques, éducatives ou politiques. Derrière l'attrait que représente un volume de ressources numériques accessibles en constante progression, se cache la réalité d'une ouverture à un monde inexplorable et incontrôlable en termes de quantité et de fiabilité des informations délivrées. Les enjeux de la recherche d'information (RI) n'en sont que plus stratégiques notamment dans les secteurs industriels et académiques, en particulier pour la veille technologique et l'innovation industrielle… Une majorité de systèmes de recherche d'information (SRI) se base sur des techniques de traitement automatique de la langue (NLP pour *Natural Language Processing*) et sur une indexation (à la fois des ressources et des requêtes pour les retrouver) sous forme de sac de termes (mots-clés), éventuellement pondérés. Cependant ces approches trouvent parfois leurs limites dues en partie aux ambigüités qui peuvent exister sur les termes de l'index utilisés, des relations entre ces termes ou, plus simplement, quand les ressources considérées ne sont pas textuelles (*e.g.* des gènes ou des images). Dans ces cas-là, la recherche d'information conceptuelle a souvent montré son efficacité (Haav et al., 2001; Sy et al., 2011). Le SRI dispose alors d'une base de ressources préalablement indexées avec des concepts issus d'une ontologie de domaine munie d'une mesure de similarité sémantique. L'appariement entre requête et ressource repose sur l'évaluation de la similarité sémantique entre le groupe de concepts de la requête et celui de l'indexation des ressources et exploite, pour ce faire, les relations de spécialisation et de généralisation de l'ontologie. La principale limite de cette approche est inhérente à sa définition : il faut disposer d'un corpus indexé avec des concepts. L'indexation est malheureusement un processus souvent fastidieux, coûteux en temps et qui nécessite un fort degré d'expertise du domaine.

Pour faciliter le processus d'indexation, nous avons proposé il y a quelques temps, une méthode d'*indexation par propagation* qui a été éprouvée dans le domaine de la musique et mise en œuvre pour indexer des titres musicaux puis, dans un autre contexte, pour indexer des photographies numériques (Crampes et al., 2009, 2006). L'objectif de cette approche était de déduire de nouvelles indexations à partir d'indexations existantes. Le principe était le suivant. Disposant d'une visualisation synthétique et interactive d'un échantillon représentatif de ressources indexées (le *support*), l'utilisateur est amené à *déposer* un nouvel élément à indexer à proximité des ressources qui lui semblent les plus semblables. Une indexation est alors proposée automatiquement pour ce nouvel élément en fonction des indexations de ces ressources voisines. Dans ces travaux, la représentation vectorielle du corpus associait à chaque élément du support ses coordonnées selon les dimensions d'analyse retenues. L'indexation d'un nouvel élément placé sur la carte résultait de la moyenne des indexations des k plus proches voisins pondérée par les distances avec ceux-ci. Les indexations ainsi inférées se sont avérées satisfaisantes et nous avons montré que cette méthode permettait de renseigner toutes les dimensions de l'indexation en un temps acceptable (le vecteur contenait 23 dimensions et plus de vingt

titres musicaux étaient indexés en moyenne en moins de cinq minutes par les DJ experts). Nous proposons aujourd'hui d'étendre cette approche à l'indexation conceptuelle. Dans la représentation vectorielle d'un corpus, les dimensions de l'espace sont considérées indépendantes. Ce n'est évidemment plus le cas pour une indexation conceptuelle où l'ontologie permet justement d'indiquer les liens entre les différents concepts utilisés pour l'indexation. La fusion des indexations des *k* plus proches voisins ne peut plus se ramener au calcul d'une simple moyenne arithmétique dans un espace conceptuel. Nous proposons, dans cet article, d'utiliser la notion de similarité sémantique pour définir l'indexation conceptuelle proposée pour un nouveau document à partir de l'indexation de ses voisins sur la carte sémantique. Nous détaillons une heuristique gloutonne qui nous permet d'effectuer cette tâche en un temps raisonnable.

Après avoir présenté l'indexation par propagation dans ses grandes lignes, nous discuterons des limites de son application à l'indexation conceptuelle dans la section 2, en distinguant trois points clés à la base de notre réflexion. Ces trois axes serviront de fil conducteur à la présentation de notre approche dans la section 3. La section 4 présentera les résultats que nous avons obtenus sur un corpus de documents scientifiques liés au cancer et indexés par des concepts issus du MeSH (*Medical Subject Headings*). Enfin, nous conclurons en discutant les résultats, les limites de l'approche et les perspectives ouvertes par ces travaux.

2. Indexation par propagation : état de l'art, principe et mise en œuvre

L'indexation de documents est une étape indispensable avant tout processus de RI. Il en découle que la qualité des résultats fournis par un SRI est directement impactée par l'index. Aussi, (Baziz et al., 2005) ont montré que l'utilisation d'annotations sémantiques pouvait mener à une amélioration qualitative de tels systèmes. La nécessité d'associer des métadonnées à des documents numériques de plus en plus nombreux a conduit à imaginer des mécanismes pour déduire de nouvelles indexations à partir d'indexations existantes (Lazaridis et al., 2013). Ainsi l'auteur de (Marchiori, 1998) propose de propager des métadonnées, en utilisant les hyperliens contenus dans les pages Web comme support de cette propagation. Les métadonnées d'une page Web sont propagées à celles qui sont en lien avec elle, avec un poids d'affaiblissement. S'inspirant de ce principe (Abrouk et al., 2005) propose d'utiliser les co-citations entre articles scientifiques pour propager leurs indexations. A partir de l'union des annotations des documents co-cités, les annotations les plus pertinentes au regard de plusieurs critères de représentativité, sont proposées pour indexer un nouveau document. Cependant ces approches en ignorant le contenu des pages et en se focalisant sur des liens de référencement, peuvent propager des erreurs (*e.g.* si une page ou un document est cité comme contre-exemple). Les autres solutions proposées dans la littérature concernant la propagation d'indexation terminologique ou sémantique ont principalement été trouvées dans le domaine de l'image mais ne correspondent pas exactement à notre contexte. Généralement, elles utilisent

conjointement l'analyse d'image (analyse de contenu) et des indexations existantes d'autres images pour en déduire, par des approches statistique, des termes possible pour indexer une nouvelle image (*e.g.* basées sur l'entropie maximale (Jeon & Manmatha, 2004), des *n-grams,* des *inférences bayésiennes* (Zhang et al., 2001) ou les *SVM – Support Vector Machine*). (Lazaridis et al., 2013) se distingue de ces approches d'une part parce que les ressources indexées ne sont pas exclusivement des images, et parce que leur indexation est conceptuelle. Cependant, leur version actuelle ne considère qu'un seul concept et se ramène à un problème de classification. L'extension de notre méthode d'indexation par propagation au niveau conceptuel semble répondre à cette demande à la fois d'une indexation rapide et relativement complète. Toujours dans le domaine de l'indexation multimédia, (Pastorello et al., 2008) propose une propagation d'indexations sémantiques qui repose sur les technologies du Web sémantique (triplets RDF). Les indexations propagées concernent des données multimédia qui, étant impliquées dans différents processus, peuvent être soumises à une ou plusieurs transformations. Leur objectif n'est pas de proposer une indexation nouvelle, mais plutôt de les compléter à partir des axiomes présents dans l'ontologie et qui peuvent également provenir de la caractérisation du processus de transformation. Le terme de propagation n'est donc pas employé dans le même sens que nous l'entendons ici.

2.1. *Principe de l'indexation par propagation*

Les travaux cités ci-dessus font tous état de la fastidiosité de la tâche d'indexation, et du niveau d'expertise de domaine qu'elle requiert (Shevade et al., 2003). Il ne s'agit donc pas d'automatiser complètement ce processus ni d'évincer l'homme de cette tâche à forte valeur cognitive, mais plutôt de gérer systématiquement pour lui la connaissance du domaine modélisé par l'ontologie. L'objectif est donc, plus modestement, de l'accompagner et d'accélérer le processus d'indexation au travers d'un outil convivial. Nous lui proposons de positionner l'élément à indexer à proximité de ressources qu'il estime lui être *proches* afin d'exploiter leurs indexations pour caractériser semi-automatiquement la nouvelle ressource. Cette méthode suppose donc à la fois une interface sémantiquement riche et interactive et des algorithmes de propagation efficaces. L'automatisation de la prise en compte des relations de spécialisation et de généralisation de l'ontologie, sur lesquelles repose le calcul des similarités sémantiques, réduit le degré d'expertise nécessaire de l'opérateur humain. En effet, le placement de l'élément à annoter sur la carte sémantique ne nécessite qu'une connaissance locale du corpus pour identifier un voisinage pertinent et non pas la maitrise globale du modèle de connaissances et du corpus, *e.g.* il suffit de bien connaitre le genre policier (connaissance experte propre et locale), pour être à même d'y classer un roman d'Agatha Christie sans pour autant connaître l'organisation globale de la bibliothèque. Etre expert d'un domaine ne suppose pas que l'on maitrise l'artéfact mis en place pour faciliter le recueil des connaissances. Par ailleurs, déléguer à l'outil informatique le calcul des similarités entre groupes de concepts, inférer des relations de subsomption ou de spécialisation permet de limiter l'impact de la subjectivité et de l'imprécision inhérente à un juge-

ment d'expert. En effet, on peut considérer que l'ontologie et la mesure de similarité sémantique qui lui est associée sont, elles aussi empreintes de subjectivité, mais l'ontologie reste par définition un modèle issu d'un consensus entre les experts d'un domaine. Ainsi, si le calcul de voisinage reste affecté par la subjectivité et l'imprécision du modèle, le traitement informatique gèrera ses effets de façon homogène pour chaque processus d'indexation.

2.1.1. *Création du support de l'indexation (carte sémantique)*

Un échantillon représentatif de ressources préalablement indexées est utilisé comme état initial. On l'appelle le *support* d'indexation. En fonction de la nature de l'indexation (lexicale, vectorielle, conceptuelle), différentes mesures peuvent être utilisées pour estimer une proximité entre les ressources indexées. Une projection sur un espace à deux dimensions conservant au mieux les distances de l'espace métrique de départ peut alors être proposée à l'utilisateur. Une telle carte, peut être obtenue par une implémentation de l'algorithme des ressorts (Crampes et al., 2006). Ici, nous avons choisi de générer cette carte à l'aide d'une projection MDS – Multi-Dimensional Scaling –, facile d'obtention à partir d'une matrice de dissimilarités. Ainsi l'utilisateur dispose d'une représentation où la proximité physique des éléments sur la carte sémantique ainsi produite traduit une proximité sémantique de ces éléments dans l'espace d'indexation.

2.1.2. *Propagation de l'indexation*

La phase d'indexation peut alors commencer. Positionnant sur cette carte de nouvelles ressources à indexer, l'expert indique graphiquement les ressources indexées qu'il estime être les plus proches sémantiquement de celles qu'il doit indexer. Ces dernières se voient attribuer automatiquement une indexation déduite de celles de leurs k plus proches voisins. Différents algorithmes peuvent assurer cette propagation en fonction de différentes stratégies. L'utilisateur peut retoucher ou préciser cette indexation afin de la rendre plus pertinente.

Les principaux atouts de cette méthode sont les suivants : i) puisqu'il reste au cœur du processus, l'opérateur humain et son expertise confèrent une grande qualité à l'indexation produite ; ii) l'indexation propagée est relativement complète et peut prendre en compte différents points de vue sur la ressource indexée ; iii) le processus est rapide et convivial, même s'il nécessite une bonne appropriation par l'utilisateur du support d'indexation. Les résultats que nous avions obtenus dans le cas d'une indexation vectorielle où chaque dimension est indépendante des autres étaient très positifs (Crampes et al., 2009, 2006), l'attrait pour la méthode s'expliquant en partie par l'analogie entre l'interactivité avec la carte et les regroupement de disques que peuvent faire les DJ sur une table de mixage. Mais on peut raisonnablement penser que ce type d'analogie est plus général (photos présentées sur un pêle-mêle, regroupement d'articles scientifique dans des classeurs thématiques, etc.)

2.2. *Extension de l'approche par propagation à une indexation conceptuelle*

Si la recherche d'information conceptuelle a montré son utilité et son efficacité pour pallier les problèmes de synonymie et de polysémie, et lever de nombreuses ambiguïtés (Sy et al., 2012), sa pertinence est fortement dépendante de la qualité de l'indexation sous-jacente. Or cette indexation préalable reste difficile à automatiser. Le but d'une indexation par propagation est de faciliter cette tâche en proposant un ensemble de concepts pour caractériser une ressource, à partir de ressources déjà indexées. Le principe de propagation décrit plus haut n'est pas directement transposable pour une indexation conceptuelle et trois problématiques requièrent une attention particulière :

i) Créer le support le mieux adapté au contexte de l'utilisateur. Il est généralement impossible de représenter de manière lisible la totalité d'un corpus sur une carte sémantique, il faut donc réduire l'espace des ressources déjà indexées à un échantillon représentatif. Cet échantillon doit être à la fois proche du contexte de l'utilisateur et très explicite pour éviter les ambiguïtés lors de l'indexation.

ii) Déterminer le voisinage à prendre en compte et les règles de propagation à appliquer lors de l'indexation d'une nouvelle ressource. La prise en compte de documents trop « éloignés » risquerait de réduire la spécificité de l'indexation proposée. L'indexation résultant de la propagation doit refléter au mieux le contenu sémantique du document tout en restant concise (éviter les concepts redondants).

iii) Choisir une ou des mesures de similarité sémantique pertinentes à la fois pour créer la carte sémantique présentant l'échantillon évoqué en i) et pour appliquer les règles de propagation des indexations évoquées en ii).

Ces trois axes serviront de fil conducteur dans la section suivante, où le protocole complet de l'indexation par propagation est détaillé, et illustré par l'indexation de publications scientifiques liées au cancer en utilisant des concepts du MeSH.

3. Indexation conceptuelle par propagation

Dans le cadre de l'accompagnement de plusieurs communautés scientifiques à la gestion de leur collectif, nous avons développé plusieurs plateformes collaboratives. L'une d'entre elles, à l'initiative de l'ITMO[1] *cancer*, est dédiée aux acteurs de la recherche française impliqués dans la lutte contre le cancer. Une base d'environ 38 000 articles scientifiques publiés par les membres de cette communauté servira de cas d'étude pour la présentation de notre approche. Le contexte est donc le suivant : l'utilisateur (expert en cancérologie) dispose d'un ensemble de documents indexés par des concepts non-pondérés issus du MeSH. Cette indexation est celle

[1] Institut Thématique Multi-Organismes de l'AvieSan (Alliance pour les sciences de la vie et de la santé – http://www.aviesan.fr/)

proposée par la base de données PubMed du NCBI[2]. De cet ensemble est extrait un échantillon de référence dont il va se servir pour indexer de nouveaux documents. Ceux-ci peuvent être des compte-rendus de réunion, des rapports de recherche internes, ou encore des publications scientifiques non référencées dans PubMed (*e.g.* publication dans des colloques francophones).

3.1. *Création du support d'indexation*

Pour simplifier le rôle de l'utilisateur dans le processus d'indexation, une carte sémantique, que nous appelons aussi paysage de référence, doit lui être présentée, sur laquelle seront disposés les documents identifiés dans le corpus indexé. La première étape consiste donc à identifier ces ressources et à en calculer une projection dans un espace à deux dimensions pour obtenir cette carte sémantique. En effet, considérer l'ensemble du corpus lors de la propagation est inenvisageable. Non seulement cela augmenterait considérablement les temps de calcul, mais surtout cela submergerait l'utilisateur d'informations inutiles présentées sur une carte surchargée et donc illisible.

Rappelons que sur le support d'indexation, la proximité physique entre les documents traduit leur proximité sémantique. Pour obtenir une telle représentation graphique nous procédons en deux étapes : i) le calcul des distances sémantiques entre chaque couple de documents à positionner et ii) l'utilisation d'un outil de positionnement multidimentionnel (MDS) qui place les documents sur un espace 2D en préservant au mieux les distances originelles. Cette dernière étape est effectuée dans notre cas à l'aide de la librairie Java MDSJ[3].

Afin de garantir une efficacité algorithmique et visuelle, le support d'indexation présenté à l'utilisateur ne doit contenir qu'un nombre restreint de documents. Le corpus C est donc réduit à un nombre r de documents de référence $Cr \subset C$ (Figure 1, partie haute). Cette réduction n'impacte que très faiblement l'indexation finale, puisque la propagation est calculée à partir des documents les plus proches de celui qui sera annoté, les plus éloignés n'apportant que peu d'information. Cette réduction doit cependant être judicieuse afin de ne pas omettre les documents importants pour la future indexation et limiter la présence de documents inappropriés qui apporteraient essentiellement du bruit. L'idéal serait donc de constituer un corpus réduit contenant les r documents qui sont sémantiquement les plus proches de celui à annoter. Évidemment, l'identification exacte de ces documents est impossible puisque l'on ne dispose pas, encore, de l'indexation sémantique du document que l'on cherche à annoter.

Nous proposons d'adopter une stratégie en deux étapes. Un premier sous-ensemble de *r* documents est présenté à l'utilisateur sur la base des informations

---

[2] http://www.ncbi.nlm.nih.gov/pubmed

[3] Algorithmics Group. *MDSJ: Java Library for Multidimensional Scaling (Version 0.2).* http://www.inf.uni-konstanz.de/algo/software/mdsj/. University of Konstanz, 2009.

disponibles associées à l'article à annoter. Cet ensemble peut, par exemple, être constitué de documents cités par l'article à annoter, ceux publiés par les mêmes auteurs, ou encore ceux qui sont jugés proches sur la base d'une analyse lexicale. L'utilisateur peut alors indiquer la zone de cette première carte où il souhaite positionner son document et une première indexation I_1 lui est proposée sur cette base. Si l'indexation est jugée insuffisante par l'utilisateur, une recherche sémantique est utilisée pour trouver les documents proches de I_1 au sein du corpus annoté et une nouvelle carte est construite. Cette seconde carte peut être vue comme un zoom sur la zone choisie par l'utilisateur sur la première carte et permet d'affiner l'indexation.

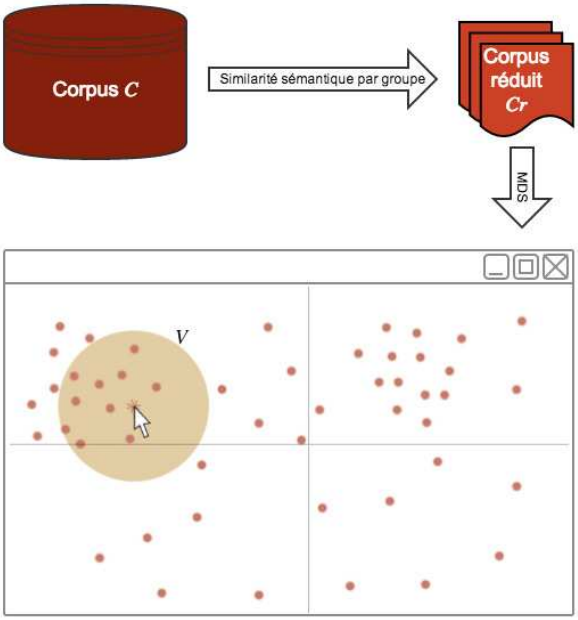

Figure 1. *Réduction du corpus C avec une mesure de similarité sémantique puis affichage de la carte sémantique. Un voisinage V est calculé pour chaque clic de l'utilisateur.*

Dans ce scénario, chaque clic à un endroit d'une carte permet de lancer le processus d'indexation par propagation que nous détaillons dans la section suivante. Le résultat de ces indexations pouvant être éventuellement affinée manuellement.

3.3. *Stratégie de propagation des indexations conceptuelles*

La première étape du processus est l'identification des documents de la carte qui seront pris en compte. Ces documents sont ceux qui sont dans le voisinage $V \subset$ Cr définit comme étant l'ensemble des k plus proches voisins de l'endroit où

l'utilisateur a cliqué. La seconde étape est la restriction des concepts de l'ontologie à un sous-ensemble I_0 qui définit l'espace des indexations qui sera exploré lors de la recherche de l'indexation optimale du nouveau document.

Pour déterminer I_0 on peut considérer que seuls les concepts présents dans l'indexation I_{dj} d'au moins un document d_j du voisinage V sont des concepts potentiellement pertinents pour l'indexation du nouveau document. L'indexation d'un document étant vue dans notre cas comme un ensemble de concepts, on peut donc définir I_0 de manière plus formelle par :

$$I_0 = \cup_{d_j \in V} I_{d_j} \qquad [1]$$

A partir de I_0, le but est d'identifier l'indexation I^* qui synthétise au mieux la sémantique de la zone indiquée par l'utilisateur et constitue donc une indexation sémantique raisonnable de son nouveau document. Pour ce faire nous proposons de chercher l'indexation médiane des documents de V, i.e. l'indexation dont la somme des similarités aux documents de V est maximale :

$$I^* = \underset{I \subset I_0}{\arg\max}\{f(I)\}, f(I) = \sum_{d_j \in V} sim(I, I_{d_j}) \qquad [2]$$

Algorithme 1 *Heuristique de recherche de I^* dans l'espace défini par I_0*

```
I*← I0
score_ I* ←f(I*)
optimum_local ← faux
tant que optimum_local = faux faire
    meilleur_score ←  score_ I*
    meilleure_indexation ←  I*
    pour chaque c ∈ I* faire
        I ← I*\{c}
        score← f(I)
        si score > meilleur_score alors
            meilleur_score ←score
            meilleure_indexation ← I
        fin si
    fin pour
    si meilleur_score > score_ I* alors
        I*← meilleure_indexation
        score_ I* ← meilleur_score
    sinon
        optimum_local ←vrai
    fin si
fin tant que
```

La solution et la difficulté de ce problème sont donc intimement liées à la mesure $sim(I,I_{dj})$ utilisée pour mesurer la similarité entre une indexation possible I et l'indexation I_{dj} d'un document du voisinage. Ce problème d'optimisation est polynomial pour certaines mesures de similarité triviales (*e.g.* si $\mathrm{sim}\left(I, I_{d_j}\right) = \left|I_{d_j} \cap I\right|$, alors la totalité de l'ensemble I_0 constitue toujours la solution maximale I^*). Cependant, ce problème d'optimisation est probablement NP-difficile dans le cas général (et pour la majorité des similarités réalistes). En effet, sa résolution exacte demande a priori dans le cas général de tester l'ensemble des $O(2^{|I_0|})$ indexations possibles, pour identifier celle maximisant la fonction objectif f(I). Nous proposons donc une heuristique gloutonne qui permet de résoudre de manière approchée ce problème en un temps polynomial.

Le principe de l'heuristique que nous utilisons pour trouver une approximation de I^* part de l'indexation la plus complète I_0. Puis à chaque étape un concept est supprimé tant que cela améliore la fonction objectif. Le concept supprimé à une étape est celui qui permet d'obtenir la meilleure amélioration. L'algorithme 1 détaille cette stratégie. Bien qu'elle soit plus longue en temps de calculs, cette solution a été préférée à l'approche inverse consistant à partir d'une indexation vide et à l'enrichir d'un concept à chaque étape car elle conduit à une meilleure optimisation. Ceci peut s'expliquer par le fait que cette stratégie permet d'avoir une vision globale de la complémentarité des concepts disponibles lors du choix des concepts à retirer.

3.4. *Variantes de l'approche*

Plusieurs variantes ou modulations de l'algorithme sont intéressantes à étudier. Tout d'abord, partant du principe que notre approche se base sur une structure de données, une ontologie de domaine, il peut être bon d'en profiter, en particulier lors de la définition de I_0. Par exemple, au lieu d'être la simple union des indexations des documents du voisinage V, I_0 pourrait contenir tout concept présent dans la sous-ontologie, *e.g.* telle que définie par (Ranwez et al., 2012), de cette union (cf. équation [3]). Cela revient à élargir l'espace de recherche en prenant en compte non seulement des concepts présents dans les indexations des documents du voisinage, mais également certains de leurs ascendants et descendants communs. En tirant partie de la structure de l'ontologie, on pourrait donc retrouver dans la suggestion d'indexations des concepts qui ne sont pas directement présents dans les indexations des documents dans V.

$$I_0 = \mathrm{concepts}\left(\mathrm{sous\_ontologie}\left(\bigcup_{d_j \in V} I_{d_j}\right)\right) \qquad [3]$$

Nous avons aussi observé durant nos tests que la condition d'arrêt était trop stricte. En effet, parfois, f(I) fournit un score légèrement moins bon qu'à l'étape précédente et l'algorithme s'arrête. Or, une très faible réduction du score peut être tolérée dans le sens où la concision du résultat final est également primordiale : fournir une indexation de 75 concepts à l'utilisateur est moins satisfaisant que de

fournir une indexation ayant un score comparable (baisse minime de la fonction objectif) avec seulement 8 concepts. Nous avons donc ajouté un paramètre autorisant une baisse du score. Dans nos tests, cette baisse est autorisée jusqu'à 1%, ainsi si deux indexations ont le même score à 1% près, la plus concise est préférée par notre approche. Cette tolérance est une manière simple de prendre en compte le fait qu'une indexation doit être aussi concise que possible du moment que cette concision n'affecte pas sa qualité.

3.2. *Choix des mesures de similarité pertinentes*

Dans toutes les méthodes de propagation, les mesures de similarité entre concepts ou groupe de concepts jouent un rôle capital, ce qui a parfois conduit à en définir de nouvelles (Shevade et al., 2003). Une étude comparative d'un grand nombre de mesures de similarité sémantiques paramétrées montre qu'il existe une telle diversité de mesures sémantiques qu'il semble plus judicieux de sélectionner la plus appropriée à un contexte applicatif donné plutôt que de tenter d'en proposer une nouvelle (Harispe et al., 2013a). Pour faciliter ce choix, les auteurs de cette étude ont proposés un cadre unificateur pour définir ces similarités sur lequel ils se sont appuyé pour développer une librairie JAVA dans laquelle elles sont implémentées : la SML[4] (Harispe et al., 2013b). C'est cette librairie qui a été utilisée pour les développements de notre application.

Le critère optimisé dans notre approche repose sur une mesure de similarité sémantique entre paires de concepts. Cette mesure n'étant pas censée répondre à certaines particularités (*e.g.* la prise en compte de la spécificité des concepts comparés), notre choix s'est porté sur la mesure de Lin (Lin, 1998). Celle-ci utilise dans son calcul une mesure d'IC (Information Content), pour laquelle différentes formules sont disponibles. Notre choix s'est porté sur l'IC intrinsèque de Seco (Seco et al., 2004) qui permet de disposer d'une méthode générale applicable dès que l'on dispose d'une ontologie de domaine indépendamment de tout corpus. Enfin, pour comparer des groupes de concepts, nous avons fait le choix d'utiliser une méthode indirecte (par agrégation), BMA – *Best Match Average*, (Schlicker et al., 2006), une moyenne des meilleures paires de concepts pour les deux groupes comparés. Une telle méthode a l'avantage d'être rapide puisqu'elle s'appuie sur les distances entre paires de concepts (qui peuvent être pré-calculées) et ne nécessite donc pas de revenir à la globalité de l'ontologie.

4. Indexation conceptuelle d'articles scientifiques : évaluation et discussion

L'application de ces travaux concerne-l'indexation de documents scientifiques ou administratifs liés au cancer (50 dans notre cas).

---

[4]*Semantic Measures Library* – http://www.semantic-measures-library.org/sml/

4.1. *Protocole de test*

Nous avons réalisé nos tests selon les deux variantes sur le même jeu de données de 50 documents à annoter. Bien sûr, à chaque itération (indexation d'un document), ce document était temporairement supprimé du corpus. Nous identifierons les résultats obtenus à partir de I_0 définit selon l'équation [1] par "variante (1)" et ceux obtenus à partir de I_0 définit selon l'équation [3] par "variante (2)". Nous avons aussi mis en place quelques solutions moins élaborées pour les comparer aux résultats obtenus. Par exemple, on peut choisir d'annoter le nouveau document en reprenant à l'identique l'indexation du document le plus proche du clic sur la carte sémantique. Il est aussi possible de considérer I_0 — dans ses deux variantes — comme indexation proposée sans plus de traitement. La qualité de l'indexation proposée est évaluée selon trois critères : i) la similarité sémantique, que l'on nommera *score final,* entre l'indexation proposée et l'indexation manuelle issue de PubMed, que l'on qualifie de référence ; ii) la concision de l'indexation proposée, fournir trop de concepts est un handicap pour la validation finale attendue de l'utilisateur et constitue également un indicateur de redondance; iii) le temps de calcul nécessaire, qui est également un facteur clé pour favoriser l'interaction homme machine via un système réactif.

Nous avons évalué deux protocoles. Le premier est celui du cas d'utilisation énoncé dans la partie 3 : une carte des publications de tous les co-auteurs du document à annoter est présentée à l'utilisateur. Ce dernier clique sur la carte, permettant au système d'optimiser une première fois le critère. Si l'indexation proposée à partir des simples publications des co-auteurs n'est pas satisfaisante (dans l'évaluation, cette indexation sera toujours considérée comme insatisfaisante), celle-ci est utilisée pour récupérer les 50 documents les plus pertinents du corpus entier via la similarité sémantique définie dans la section 3.2. Ils sont alors utilisés pour définir une nouvelle carte et l'utilisateur est sollicité une seconde fois pour positionner son document sur cette nouvelle carte. Le second protocole se focalise sur l'évaluation du processus de propagation de l'indexation, qui est l'apport principal de cet article, et s'affranchit du problème de l'identification des documents pertinents pour cette tâche (qui constitue un problème relativement indépendant). Pour se faire, connaissant l'indexation PubMed du document à annoter, le programme récupère les 50 documents du corpus qui sont les plus proches sémantiquement du document à annoter (à l'exclusion du document lui-même). Nous utilisons pour cela la même méthode que pour le premier protocole, la différence étant dans le fait que le groupe de concepts utilisé pour récupérer Cr est l'indexation connue du document. Le programme retourne donc les vrais 50 documents les plus proches d'après la mesure Lin BMA.

4.2. *Résultats et discussion*

Nous notons qu'en terme de temps de calcul, les phases primaires sont les plus longues, à savoir la construction de la carte sémantique, se décomposant en (i) la récupération des documents pertinents avec Lin BMA, (ii) la construction d'une

matrice de dissimilarités et (iii) le calcul du MDS. Le temps que prend l'indexation – représenté dans le tableau par la différence entre le temps pour déterminer I_0 et celui du résultat final – est de l'ordre de quelques secondes.

		Score final moyen	*Taille moyenne de l'indexation*	*Temps moyen de calcul (ms)*
Protocole cas d'utilisation	Variante (1)	0.732	6.76	8842
	Variante (2)	0.730	6.76	17285
	I_0 (1)	0.697	82.28	7204
	I_0 (2)	0.698	144.21	7896
	Plus proche	0.699	17.92	7192
Protocole référence similarité	Variante (1)	0.843	6.71	10929
	Variante (2)	0.843	6.70	12490
	I_0 (1)	0.787	61.72	10148
	I_0 (2)	0.775	108.98	8602
	Plus proche	0.792	17.90	10142

Tableau 1. *Résultats, respectivement en colonne : le score moyen (la similarité sémantique entre l'indexation proposée et celle de PubMed), la taille de l'indexation proposée et le temps de calcul. En ligne, les différents protocoles testés.*

Nous avons testé si les différences entre les stratégies possibles sont ou non statistiquement significatives. Tout d'abord, l'indexation produite en utilisant le protocole complet est significativement moins bonne que les indexations produites dans l'autre cas qui s'appuie sur une carte « idéale » (p-valeur = 2.5×10^{-15}). Cela montre clairement l'importance de disposer d'une carte sémantique adaptée au document considérée, et la marge de progrès associée à cette étape qui n'est pas le cœur de cet article.

L'utilisation de la sous-ontologie (variante (2)) ne permet pas d'obtenir une indexation significativement meilleure (p-valeur = 0.5921 et 0.5093 dans les deux protocoles respectifs). Ce résultat peu attendu (le choix de la sous-ontologie ouvrait les portes à des concepts plus généraux) peut s'expliquer par le fait que, bien souvent, on peut obtenir une concision satisfaisante tout en gardant une spécificité convenable. La spécificité, appréciée par le calcul d'IC dans les mesures, influe directement sur les choix durant la réduction. Aussi, les scores finaux obtenus par la variante (1) sont systématiquement meilleurs que ceux obtenus par des approches plus naïves, telles que l'indexation du plus proche voisin (p-valeur = 4.98×10^{-5} et 8.443×10^{-6}) ou I_0 (p-valeur = 8.4×10^{-4} et 8.93×10^{-8}).

Pour finir, en plus du résultat de la variante (1) dans les deux méthodes, le plus proche voisin déterminé dans le premier protocole est lui aussi moins bon que celui

obtenu par le second (p-valeur = 1.36×10^{-9}). Ceci explique certainement la différence de qualité entre les deux approches : si le voisinage est approximativement déterminé, l'indexation est biaisée, impliquant ainsi un score final moins bon que pour le protocole référence. Le critère semble donc réaliser son objectif mais repose sur une étape clé : la construction d'un support fiable. La difficulté du protocole complet repose dans l'établissement de la carte, décisive dans l'indexation. On voit que, malgré des scores plus faibles que les résultats optimaux, nous proposons une indexation plutôt bonne (en moyenne 0.732) que l'utilisateur pourra affiner.

4.3. *Limites et perspectives*

La réduction préalable du corpus est une étape primordiale afin de proposer une carte sémantique lisible à l'utilisateur. Cette étape s'appuie actuellement sur une recherche sémantique qui s'avère être un facteur limitant en terme de temps de calcul. Dans le cadre d'un corpus de 38 000 documents, la recherche se fait rapidement (de l'ordre de quelques secondes), mais en l'état actuel un tel protocole passera difficilement à l'échelle. Il faut cependant garder à l'esprit que cette première implémentation ne constituait qu'un prototype pour valider la pertinence de l'approche. Nous travaillons actuellement à accélérer cette étape en optimisant le calcul des mesures de similarité sémantique par groupe (*groupwise semantic similarity*) proposée dans la SML à notre cas particulier.

Actuellement, seul le voisinage le plus proche est considéré dans le critère. Or, nous aimerions aussi utiliser les autres documents présents dans le support d'indexation comme repoussoir. L'indexation finale devant être à la fois proche de V et distante de $C_r \setminus V$. Une telle règle permettrait d'améliorer la spécificité de l'indexation. Deux implémentations sont envisageables, soit sous la forme d'une extension du critère, soit sous la forme d'un filtre une fois l'algorithme de réduction achevé.

Lorsque la carte sémantique réalisée est imprécise, l'indexation qui en résulte est approximative. Ce biais apparaît lors de la première proposition de carte à l'utilisateur. Seuls les documents des co-auteurs sont actuellement sélectionnés durant cette phase. Cette approche un peu simpliste trouve ses limites dans le cas où des auteurs (*e.g.* jeunes doctorants) ont peu publié, il est alors difficile de retourner suffisamment de documents pertinents uniquement sur cette base. Dans tous les cas, cette approche se prive par construction de documents potentiellement très pertinents rédigées par d'autres personnes. Nous envisageons actuellement deux solutions complémentaires pour améliorer cette étape : (i) identifier les cas problématiques pour pouvoir les notifier à l'utilisateur et (ii) limiter les situations problématiques en enrichissant la carte proposée. Cependant quelle que soit la stratégie retenue, la qualité de la carte est intrinsèquement dépendante de l'adéquation entre le corpus dont on dispose et le document à annoter. Il est donc inévitable que dans certains cas l'utilisateur ait à positionner son nouveau document dans une zone très peu dense. Il serait possible de détecter automatiquement ces cas pour éviter de

proposer une indexation inadaptée[5]. Étant donné que nous voulions une approche générique, nous n'avons pas tenu compte du type d'entité traité. Or, dans le cas d'étude choisi dans cet article le corpus était composé de documents textuels. L'enrichissement de la méthode avec des outils de NLP tels que le NCBO Annotator (Jonquet et al., 2009) ou OpenCalais[6] permettrait certainement de remédier aux cartes imprécises lors de la première étape.

Outre les limites liées au mode calculatoire, la principale limite de cette approche, déjà rencontrée lors de nos précédents travaux, concerne l'usage. Une trop grande flexibilité de l'indexation (trop de critères à renseigner et donc des affinages difficiles à appréhender) et les limites liées à la visualisation sont les principaux freins. Nos efforts actuels se concentrent donc sur la visualisation à la fois de la carte sémantique de référence (support) et des documents indexés au cours du processus (doit-on conserver tous les documents déjà traités par l'utilisateur ?). Notons que l'opérateur humain reste au centre du processus et que c'est lui qui "dirige" l'indexation. Pour cela, une connaissance au moins partielle de l'ontologie du domaine est nécessaire et un environnement adapté doit être pensé. Des tests sont envisagés conjointement avec des spécialistes ergonomes et des utilisateurs en situation réelle pour améliorer les modes de visualisation et d'interaction.

5. Conclusion

Afin de favoriser la création de corpus de ressources indexées conceptuellement (à l'aide de sacs de concepts), nous proposons dans cet article une extension de notre méthode d'indexation par propagation. Disposant d'un ensemble de références de documents indexés, le système en propose à l'utilisateur une représentation graphique, dite carte sémantique, avec laquelle il interagit pour indexer de nouvelles ressources. Celles-ci se voient attribuer une indexation conceptuelle propagée à partir des indexations de références voisines de la zone désignée par l'utilisateur comme étant la position escomptée de la nouvelle ressource.

Cette approche est basée sur la construction de l'indexation médiane des indexations des documents présents dans la zone indiquée par l'utilisateur. Une approximation de cette indexation médiane est obtenue en un temps polynomial grâce à une heuristique gloutonne.

Les auteurs tiennent à remercier le programme ITMO Cancer de l'AvieSan qui a financé une partie de ces travaux.

---

[5] Des travaux sont en cours à ce sujet.
[6] http://www.opencalais.com/

12. Bibliographie

Abrouk, L., Gouaïch, A., & Raïssi, C. (2005). Annotation automatique de documents - Analyse des citations. In *INFORSID 2006* (pp. 483–497). Hammamet, Tunisie.

Baziz, M., Boughanem, M., & Aussenac-Gilles, N. (2005). A Conceptual Indexing Approach for the TREC Robust Task. In *TREC*.

Crampes, M., de Oliveira-Kumar, J., Ranwez, S., & Villerd, J. (2009). Visualizing Social Photos on a Hasse Diagram for Eliciting Relations and Indexing New Photos. *Ieee Transactions on Visualization and Computer Graphics*, *15*(6), 985–992.

Crampes, M., Ranwez, S., Velickovski, F., Mooney, C., & Mille, N. (2006). An integrated visual approach for music indexing and dynamic playlist composition. In U. Chandra & C. Griwodz (Eds.), *Proc. SPIE 6071, Multimedia Computing and Networking 2006* (Vol. 6071, p. 7103). San Jose, California.

Haav, H.-M., & Lubi, T.-L. (2001). A Survey of Concept-based Information Retrieval Tools on the Web. In *5th East-European Conference, ADBIS 2001* (Vol. 2, pp. 29–41). Vilnius, Lithuania.

Harispe, S., Janaqi, S., Ranwez, S., & Montmain, J. (2013a). From Theoretical Framework To Generic Semantic Measures Library. In Y. T. Demey & H. Panetto (Eds.), *On the Move to Meaningful Internet Systems: OTM 2013 Workshops*. Graz, Austria: Springer Berlin Heidelberg.

Harispe, S., Ranwez, S., Janaqi, S., & Montmain, J. (2013b). The semantic measures library and toolkit: fast computation of semantic similarity and relatedness using biomedical ontologies. *Bioinformatics (Oxford, England)*.

Jeon, J., & Manmatha, R. (2004). Using maximum entropy for automatic image annotation. *Proc. CVIR, Lecture Notes in COmputer Science*, *3115*, 24–32.

Jonquet C, Shah NH, Musen MA. (2009). The open biomedical annotator. Summit on Translat Bioinforma. 2009 Mar 1;2009:56-60.

Lazaridis, M., Axenopoulos, A., Rafailidis, D., & Daras, P. (2013). Multimedia search and retrieval using multimodal annotation propagation and indexing techniques. *Signal Processing: Image Communication*, *28*(4), 351–367.

Lin, D. (1998). An Information-Theoretic Definition of Similarity. In *Fifteenth International Conference on Machine Learning* (pp. 296–304). Morgan Kaufmann Publishers Inc.

Marchiori, M. (1998). The limits of Web metadata, and beyond. *Computer Networks and ISDN Systems*, *30*(1-7), 1–9.

Pastorello Jr., G. Z., Daltio, J., & Medeiros, C. B. (2008). Multimedia Semantic Annotation Propagation. In *2008 Tenth IEEE International Symposium on Multimedia* (pp. 509–514). IEEE.

Schlicker, A., Domingues, F. S., Rahnenführer, J., & Lengauer, T. (2006). A new measure for functional similarity of gene products based on Gene Ontology. *BMC bioinformatics*, *7*, 302.

Seco, N., Veale, T., & Hayes, J. (2004). An intrinsic information content metric for semantic similarity in WordNet. *Ecai 2004: 16Th European Conference on Artificial Intelligence, Proceedings*, *110*, 1089–1090.

Shevade, B., & Sundaram, H. (2003). Vidya: an experiential annotation system. In *Proceedings of the 2003 ACM SIGMM workshop on Experiential telepresence - ETP '03* (pp. 91–98). Berkeley, California, USA: ACM Press, New-York USA.

Sy, M. F., Ranwez, S., Montmain, J., & Ranwez, V. (2012). OBIRS-feedback, une méthode de reformulation utilisant une ontologie de domaine. In *COnférence en Recherche d'Information et Applications - CORIA 2012*. Bordeaux, France.

Sy, M.-F., Ranwez, S., Montmain, J., Regnault, A., Crampes, M., & Ranwez, V. (2011). User centered and ontology based information retrieval system for life sciences. *BMC Bioinformatics*, *13*(Suppl 1), S4.

Zhang, H. J., & SU, Z. (2001). Improving CBIR by Semantic Propagation and Cross Modality Query Expansion. In *International workshop on Multimedia Content-Based Indexing and Retrieval*. Brescia, Italy.