N° d'ordre NNT : 2019LYSE2062

THESE de DOCTORAT DE L'UNIVERSITÉ DE LYON

Opérée au sein de

L'UNIVERSITÉ LUMIÈRE LYON 2

École Doctorale : ED 512 Informatique et Mathématiques

Discipline : Informatique

Soutenue publiquement le 15 février 2019, par :

Hussein AL NATSHEH

Text Mining Approaches for Semantic Similarity Exploration and Metadata Enrichment of Scientific Digital Libraries.

Devant le jury composé de :

Nathalie AUSSENAC-GILLES, Directrice de Recherche, Université Toulouse 3, Présidente

Juliette DIBIE, Professeure des universités, AGROPARISTECH, Rapporteure

Gilles VENTURINI, Professeur des universités, Université de Tours, Rapporteur

Sabine LOUDCHER, Professeure des universités, Université Lumière Lyon 2, Examinatrice

Djamel Abdelkader ZIGHED, Professeur des universités, Université Lumière Lyon 2, Directeur de thèse

Fabrice MULHENBACH, Maître de conférences, Université Jean Monnet Saint-Etienne, Co-Directeur de thèse

Contrat de diffusion

N° d'ordre NNT : 2019LYSE2047

THESE de DOCTORAT DE L'UNIVERSITÉ DE LYON

Opérée au sein de

L'UNIVERSITÉ LUMIÈRE LYON 2

École Doctorale : ED 512 Informatique et Mathématiques

Discipline : Informatique

Soutenue publiquement le 15 février 2019, par :

Hussein AL NATSHEH

Text Mining Approaches for Semantic Similarity Exploration and Metadata Enrichment of Scientific Digital Libraries.

Devant le jury composé de :

Nathalie AUSSENAC-GILLES, Directrice de Recherche, Université Toulouse 3, Présidente

Juliette DIBIE, Professeure des universités, AGROPARISTECH, Rapporteure

Gilles VENTURINI, Professeur des universités, Université de Tours, Rapporteur

Sabine LOUDCHER, Professeure des universités, Université Lumière Lyon 2, Examinatrice

Djamel Abdelkader ZIGHED, Professeur des universités, Université Lumière Lyon 2, Directeur de thèse

Fabrice MULHENBACH, Maître de conférences, Université Jean Monnet Saint-Etienne, Co-Directeur de thèse

Contrat de diffusion

Résumé long

Pour les scientifiques et chercheurs, s'assurer que la connaissance est accessible pour pouvoir être réutilisée et développée est un point crucial. De plus, la façon dont nous stockons et gérons les articles scientifiques et leurs métadonnées dans les bibliothèques numériques détermine la quantité d'articles pertinents que nous pouvons découvrir et auxquels nous pouvons accéder en fonction de la signification réelle d'une requête de recherche. Cependant, sommes-nous en mesure d'explorer tous les documents scientifiques sémantiquement pertinents avec les systèmes existants de recherche d'information au moyen de mots-clés ? Il s'agit là de la question essentielle abordée dans cette thèse.

L'objectif principal de nos travaux est d'élargir ou développer le spectre des connaissances des chercheurs travaillant dans un domaine interdisciplinaire lorsqu'ils utilisent les systèmes de recherche d'information des bibliothèques numériques multidisciplinaires. Le problème se pose cependant lorsque de tels chercheurs utilisent des mots-clés de recherche dépendant de la communauté dont ils sont issus alors que d'autres termes scientifiques sont attribués à des concepts pertinents lorsqu'ils sont utilisés dans des communautés de recherche différentes.

Afin de proposer une solution à cette tâche d'exploration sémantique dans des bibliothèques numériques multidisciplinaires, nous avons appliqué plusieurs approches de fouille de texte. Tout d'abord, nous avons étudié la représentation sémantique des mots, des phrases, des paragraphes et des documents pour une meilleure estimation de la similarité sémantique. Ensuite, nous avons utilisé les informations sémantiques des mots dans des bases de données lexicales et des graphes de connaissance afin d'améliorer notre approche sémantique.

En outre, la thèse présente quelques implémentations de cas d'utilisation du modèle que nous avons proposé. Enfin, plusieurs évaluations expérimentales ont été menées afin de valider l'efficacité de notre approche. Les résultats de l'approche hybride, basée à la fois sur une représentation sémantique de petits textes et sur l'information sémantique des mots extraits de bases de données lexicales, ont été très encourageants. Nous pensons que nos nouvelles approches basées sur les techniques de fouille de texte permettent d'obtenir en pratique les résultats escomptés en ce qui concerne la limitation de l'exploration sémantique dans les systèmes classiques de recherche d'information des bibliothèques numériques.

L'avantage de notre approche est qu'elle s'applique aux grandes bibliothèques numériques multidisciplinaires. En ce sens, nous utilisons les informations trouvées

dans les métadonnées de ces bibliothèques afin de les enrichir de balises sémantiques supplémentaires.

Par conséquent, les métadonnées améliorées et enrichies permettent aux chercheurs de récupérer des documents plus pertinents d'un point de vue sémantique qui seraient autrement restés inexplorés sans cet enrichissement. Nous pensons que notre étude et les approches que nous proposons fourniront des solutions pratiques à l'accès aux connaissances et contribueront aux communautés de recherche et aux domaines de la fouille de texte et de la gestion des données dans les bibliothèques numériques.

Les bibliothèques numériques deviennent de plus en plus interdisciplinaires et leur nombre de documents continue d'augmenter rapidement. Récemment, de nouveaux concepts tels que l'accès ouvert et le savoir ouvert sont apparus, où nous avons commencé à assister à de nombreuses bibliothèques numériques à accès ouvert. Cependant, la gestion et la récupération des connaissances stockées dans ces bibliothèques numériques posent de nombreux défis. Par exemple, les bibliothèques numériques ne sont pas centralisées et il n'existe pas de schéma de données uniformisé à l'échelle mondiale, ni parmi les éditeurs ni entre les différentes disciplines scientifiques.

En tant que texte généré par l'homme, la connaissance est représentée dans un langage naturel qui n'est pas encore compris par la machine, contrairement aux langages de programmation. Un autre défi posé par l'utilisation de texte pour représenter le langage naturel est le fait qu'il n'y a pas de correspondance parfaite entre les mots et les pensées. Une pensée peut être exprimée de nombreuses manières en utilisant différentes formes syntaxiques possibles et différents mots. Dans le même temps, un mot peut avoir plusieurs significations. Les systèmes classiques de recherche d'informations, tels que les moteurs de recherche par correspondance de texte classiques, sont sémantiquement insuffisants en raison de ces défis ultérieurs.

Nous aborderons ensuite certains des problèmes rencontrés par les bibliothèques numériques multidisciplinaires dans la gestion des connaissances et la recherche d'informations. La principale motivation de cette thèse est de résoudre ces problèmes clés en utilisant des techniques d'exploration de données et de gestion des connaissances.

- Utilisation de métadonnées, mots-clés, balises et ontologies :

Le concept de bibliothèques numériques à accès ouvert a été adopté par de nombreuses organisations pour répondre à la demande croissante de chercheurs en sciences à accès libre. Nombre de ces initiatives en libre accès regroupent les publications de nombreux éditeurs. Cependant, il n'y a pas de métadonnées standard pour différents éditeurs, même pour ceux en libre accès. Mapper différentes

métadonnées de l'éditeur sur le schéma de données combiné est un problème ouvert qui doit être traité par les bibliothèques numériques multi-sources. Par exemple, la catégorisation des sujets scientifiques diffère, que ce soit dans «Web of Science» ou «Scopus». La dénomination et la numérotation des catégories et sous-catégories principales sont différentes.

Il existe différentes manières de fournir des informations complémentaires, des données non structurées aux données structurées. La structuration est généralement fournie par une métadonnée qui utilise des balises ou des mots-clés choisis par les auteurs d'articles scientifiques. Lors de la soumission de leurs articles scientifiques, les auteurs sont généralement limités à un nombre maximal de balises et de mots-clés par document. Dans de nombreux cas, les travaux de recherche concernent de nombreuses disciplines scientifiques et contiennent de nombreuses variantes de nommage de sujets. qui tous ensemble nécessitent un plus grand nombre de mots-clés.

Dans certains cas, l'éditeur a fourni un nombre réduit de mots-clés standard ou de noms de catégories fournis par les sociétés scientifiques, comme dans le système de classification ACM Computing ou l'IEEE Taxonomie.

Cela devient un problème critique lorsque le système de récupération d'informations utilise la liste de mots-clés susmentionnée pour filtrer et classer les documents. Si le chercheur interrogateur utilise une seule variante ou un mot clé associé qui n'est pas mentionné dans l'ensemble de mots-clés de la publication, cette publication sera exclue des résultats. Par exemple, si le chercheur utilise le mot clé "Moteurs de recherche" alors que la publication est uniquement étiquetée avec cette liste de mots clés: ["recherche d'informations,", "exploration de texte,", "linguistique informatique" et " “systèmes de recommandation basés sur le contenu”], la publication n'apparaîtra probablement pas dans les résultats de la recherche.

Nous devons également faire face à cette limitation dans les cas où diverses disciplines scientifiques ont tendance à utiliser des termes différents pour décrire le même sujet. Par exemple, en informatique, nous utilisons le terme «machine learning», tandis que le même concept est appelé «analyse multivariée de données» par la communauté de la physique des hautes énergies. Même au sein de la même communauté scientifique, le même sujet ou concept est exprimé sous différents termes ou noms au fil du temps (\ "Data mining" et "data science", par exemple).

En informatique, l'ontologie est une sorte de représentation de connaissances sans schéma. Contrairement aux bases de données documentaires et relationnelles, un réseau de faits connectés organisés selon un triple format (sujet, prédicat, objet) est utilisé pour représenter le savoir de manière sémantique. Développer et utiliser ce graphe de connaissances dans des bases de données plus difficiles et coûteuses que

dans des bases de données traditionnelles. Un exemple typique de graphe de connaissances en accès libre est DBpedia, qui a été créé à l'aide de RDF (Resource Description Framework), en tant que standard Web sémantique. DBpedia a été extrait des infoboxes de l'encyclopédie numérique Wikipedia.

Traiter avec un graphe de connaissances ou des données liées implique de nombreuses normes pour la définition du vocabulaire, des conventions de nom de concept, de la désignation des liens sémantiques, de la désambiguïsation des noms d'entités, du stockage, de la gestion des données en libre accès et bien sûr du langage et des méthodes de requête. Une ontologie bien connue et largement utilisée est schema.org. Le principal défi des données liées réside dans le fait qu'elles sont incomplètes.

La collecte de toutes les connaissances de l'homme (multilingue, multi-domaines) dans un graphe de connaissances centralisé est un travail infini en soi, en particulier si l'on prend en compte la mise à jour de son contenu au fil du temps. Même si le processus était en grande partie automatisé, la curation et l'annotation manuelle des données resteraient nécessaires, ce qui est coûteux et difficile à gérer. Ainsi, l'utilisation de données non structurées reste nécessaire en plus de l'utilisation de telles sources de connaissances structurées. Dans ce contexte, les techniques d'extraction de texte jouent un rôle important dans le traitement de diverses sources de texte brut ou de données textuelles sans aucun type d'annotation sémantique.

- Promouvoir la recherche interdisciplinaire :

Récemment, la communauté scientifique a été témoin de nouveaux domaines scientifiques interdisciplinaires émergents. Dans de nombreux cas, l'informatique devient un domaine transversal qui a été utilisé dans de nombreuses autres disciplines, facilitant les découvertes scientifiques ainsi que la gestion du stockage et de l'informatique pour des expériences scientifiques.
(haute énergie ou des études génétiques en biologie). La bioinformatique est un exemple de ce type de domaine transversal: elle est issue des deux domaines de l'informatique et de la biologie.

domaines interdisciplinaires ont de meilleures chances d'invention et de l'innovation car ils sont positionnés dans les frontières du domaine scientifique typique. De nombreuses idées et modèles inventifs peuvent être hérités d'autres domaines apportant une solution réussie. Rester au cœur de la discipline scientifique a généralement beaucoup moins de chances de trouver une découverte décisive par rapport aux études de recherche se situant à la frontière plus étroite de la discipline et chevauchant parfois avec d'autres disciplines et communautés de recherche.

Le modèle de l'invention consistant à résoudre certains problèmes fondamentaux dans une discipline donnée pourrait être emprunté dans un autre contexte d'une autre discipline et fonctionner avec succès comme solution inventive dans ce domaine. Cette idée de modèles inventifs communs a été introduite dans un nouveau domaine émergent de l'innovation systématique appelé TRIZ, qui désigne la théorie de la résolution inventive de problèmes.

Un exemple de cas interdisciplinaire utile peut être vu avec certains modèles mathématiques empruntés de la physique à l'exploration de données. Entropie et inertie des modèles mathématiques prises à partir du domaine de l'énergie sont utilisés dans les modèles d'apprentissage de la machine comme des arbres de décision et le regroupement.

Un autre type de fertilisation croisée interdisciplinaire pourrait être provoqué par la mise en commun de concepts entre deux domaines scientifiques. Par exemple, entre l'intelligence artificielle (IA) et la philosophie, il existe de nombreux concepts communs tels que «action», «conscience», «épistémologie».

Aujourd'hui, étant donné que la science est très large et spécifique, les scientifiques et les chercheurs sont très rarement exposés à plus d'une ou deux disciplines. Au Moyen Age, cependant, c'était possible de trouver de nombreux exemples de polymathes qui ont travaillé dans de nombreuses disciplines différentes. Quelques exemples de polymathes sont al-Khwarizmi, qui travaillait en mathématiques, en astronomie et en géographie. Leonardo da Vinci qui a travaillé dans de nombreux domaines, dont l'invention, la peinture, la sculpture, l'architecture, les sciences, la musique, les mathématiques, l'ingénierie, la littérature, l'anatomie, la géologie, l'astronomie, la botanique, l'écriture, l'histoire et la cartographie.
Ces polymathes ont pu avoir de nombreuses découvertes exposées à plusieurs domaines en même temps.

De nos jours, nous ne disposons que d'un nombre très limité de polymathes du fait de la spécialisation - et même de l'hyperspécialisation - des scientifiques, ainsi que de l'énorme quantité de connaissances dans chaque domaine scientifique. Cependant, Herbert A. Simon (1916-1961), qui était considéré comme un chercheur transdisciplinaire très spécial, est un bon exemple de l'époque moderne: il transcende les frontières disciplinaires dans une demi-douzaine de domaines (traitement de l'information, prise de décision, problème). -solving, théorie de l'organisation et systèmes complexes), il a formulé des modèles en psychologie pour des applications en intelligence artificielle, mais ces modèles ont également des conséquences sur le plan économique (la théorie de Simon sur la rationalité liée a conduit à un prix Nobel d'économie en 1978).

- Traitement des sources hétérogènes de métadonnées :

Un éditeur peut, dans une certaine mesure, unifier son schéma de données de publications par type. Par exemple, l'éditeur pourrait définir et nommer les champs obligatoires que tous les auteurs doivent respecter. Même si le schéma évolue au fil des années pour faire face aux nouvelles demandes, le schéma des anciennes publications gérées par l'éditeur peut être mis à jour pour préserver la cohérence. Les types de publication peuvent toutefois varier même pour un seul éditeur. Il existe par exemple des articles de conférence, des articles de journaux, des affiches, des critiques, des livres ou des chapitres de livres.

Le problème se pose lorsqu'il s'agit d'une bibliothèque numérique multidisciplinaire qui a généralement des sources d'éditeurs différents. Nous pouvons trouver ici beaucoup de différences dans le schéma en termes de catégorisation, de nommage, de hiérarchie, de types de données, etc. Ainsi, lorsqu'une bibliothèque numérique multi-sources regroupe une nouvelle publication, elle relève ce défi et finit généralement par avoir un schéma combiné plus grand afin de compter tous les nouveaux schémas différents. Cela entraîne un problème de duplication de données qui ajoute davantage de complexité au maintien de ce schéma combiné pour les systèmes de récupération d'informations.

- Augmentation de la taille du corpus scientifique :

Le volume et l'échelle des données posent de nombreux problèmes techniques. Chaque jour, le nombre de publications, d'auteurs et d'affiliations augmente rapidement. Les technologies Big Data tentent de résoudre ce problème en plus de la diversité des formes de données, de l'analyse des données en continu, de l'incertitude et de la qualité des données. Les systèmes de recherche d'informations doivent donc également tolérer ces problèmes de Big Data. L'introduction d'un système de récupération d'informations activé par la sémantique entraîne généralement des calculs lourds qui ne peuvent pas nécessairement être mis à l'échelle.

Nos contributions pourraient être résumées dans les articles suivants:

- Proposition d'une similarité sémantique au niveau de la phrase à usage général utilisant un ensemble de fonctionnalités extensible par paire. L'approche proposée a montré des résultats d'évaluation cohérents lorsqu'il s'appliquait dans quelques cas d'utilisation, un dans les styles d'écriture et un autre dans le couplage sémantique des phrases de contenu abstrait à papier.
- Une nouvelle approche sémantique pour un système de recommandation basé sur le contenu pour un sujet scientifique interdisciplinaire. L'approche a également été utilisée pour l'expansion de corpus, ainsi que pour la modélisation

de sujets scientifiques et l'utilisation de mots-clés au fil du temps sur un corpus de texte multidisciplinaire.

- Un nouveau modèle évolutif pour l'enrichissement en métadonnées et le marquage sémantique automatique de bibliothèques numériques multidisciplinaires par rapport aux approches classiques de modélisation par sujet et aux techniques de multi-étiquetage.
- Modèle de désambiguïsation des noms d'auteurs utilisant l'apprentissage semi-supervisé avec une contribution Open Source sur le système de désambiguïsation des noms d'auteurs actuellement utilisé dans la bibliothèque numérique du CERN.
- Analyse de la diversité et des imprévus dans les systèmes de recommandation de papier scientifique sémantique et le balisage thématique sémantique et son rôle dans la promotion de la recherche transdisciplinaire.
- Solutions logicielles à source ouverte pour l'estimation de similarité sémantique de phrases et le balisage sémantique automatique de métadonnées, ainsi que des expériences de cas d'utilisation reproductibles dans une bibliothèque numérique multidisciplinaire.

Dans les corpus de textes scientifiques, certains articles issus de communautés de chercheurs différentes peuvent ne pas être décrits par les mêmes mots-clés alors qu'ils partagent la même thématique. Ce phénomène cause des problèmes dans la recherche d'information, ces articles étant mal indexés, et limite les échanges potentiellement fructueux entre disciplines scientifiques.

Notre modèle permet d'attribuer automatiquement une étiquette thématique aux articles au moyen d'un apprentissage des représentations sémantiques d'articles du corpus déjà étiquetés. Passant bien à l'échelle, cette méthode a pu être testée sur une bibliothèque numérique d'articles scientifiques comportant des millions de documents. Nous utilisons un réseau sémantique de synonymes pour extraire davantage d'articles sémantiquement similaires et nous les fusionnons avec ceux obtenus par un modèle de classement thématique. Cette méthode combinée présente de meilleurs taux de rappel que les versions utilisant soit le réseau sémantique seul, soit la seule représentation sémantique des textes.

L'activité des chercheurs a été bouleversée par un accès toujours plus important aux bibliothèques numériques en ligne. La recherche d'information dans ces bibliothèques numériques se fait le plus souvent au moyen de mots-clés entrés dans des moteurs de recherche. Néanmoins, l'appariement entre les mots-clés entrés et ceux utilisés pour décrire les documents scientifiques pertinents présents dans ces bibliothèques

numériques peut s'avérer limité si la terminologie employée n'est pas la même dans les deux cas. Tout chercheur appartient à une communauté avec laquelle il partage des connaissances et un vocabulaire communs.

Cependant, lorsque celui-ci souhaite étendre l'exploration bibliographique au-delà de sa communauté d'appartenance afin de recueillir des éléments d'information qui le conduisent à de nouvelles connaissances, il convient de lever plusieurs verrous scientifiques et techniques induits par la grande taille des bibliothèques numériques, l'hétérogénéité des données et la complexité du langage naturel. Les chercheurs qui travaillent dans un contexte pluri- et trans-disciplinaire doivent pouvoir accéder aux documents qui les intéressent sans pour autant être bloqués par la barrière d'un cloisonnement disciplinaire induit par une méconnaissance du vocabulaire employé par d'autres disciplines scientifiques. Le plus souvent, les réseaux sémantiques sont une bonne réponse aux problèmes de variations linguistiques en retrouvant des synonymes ou des champs lexicaux communs.

Dans le domaine scientifique, toutefois, cette approche n'est pas suffisante car elle se heurte à la terminologie propre au jargon scientifique et technique qui, par nature, est très spécifique, et qui a la particularité d'évoluer très rapidement. Une autre solution pourrait être apportée par le plongement lexical . Cette technique permet de retrouver des termes liés par une proximité au sein d'un même document et, de là, de déduire une proximité sémantique. Cette approche présente malgré tout les problèmes de ne pas donner d'information sur le nombre de termes dont il faut tenir compte pour être encore considéré comme sémantiquement proche du terme initial et de ne pas trop bien fonctionner quand il s'agit d'un concept composé de plusieurs termes plutôt que d'un seul et unique terme.

Dans cet article, nous proposons une solution combinant deux sources d'information sémantique~: la première est issue de l'ensemble de synonymes déduits d'un réseau sémantique, la seconde provient de la représentation sémantique d'une projection vectorielle des articles.

Dans ce travail, nous avons étudié trois méthodes permettant d'attribuer sémantiquement des étiquettes de sujets scientifiques aux articles d'un corpus. Ces étiquettes sont issues d'une taxonomie de la collection “Web of Science”. Or les bibliothèques numériques multidisciplinaires combinent des corpus provenant de nombreux éditeurs scientifiques utilisant chacun leur propre taxonomie. Ce phénomène freine l'accès de certains articles à des chercheurs d'autres disciplines par leur emploi d'une terminologie et d'une taxonomie différentes.

En enrichissant la bibliothèque numérique avec plus de balises obtenues à travers la méthode d'auto-étiquetage thématique de textes scientifiques que nous proposons, la

taxonomie et les balises étendront l'exploration de la recherche à davantage d'articles sémantiquement pertinents.

L'approche combine deux sources d'information sémantique (synonymes issus d'un réseau sémantique et résultats de la représentation sémantique d'une projection vectorielle). Notre étude expérimentale montre une amélioration significative en terme de rappel par rapport aux résultats obtenus en utilisant seulement les synonymes de sujets extraits des réseaux sémantiques. Ajoutons que lorsqu'une requête est menée sur un mode exploratoire dans une bibliothèque numérique scientifique, il est difficile de connaître directement les termes exacts de la thématique des documents recherchés. La requête sera donc le plus souvent une périphrase composée de plusieurs termes, situation où la méthode retourne les meilleurs résultats.

Dans cette thèse, nous avons étudié le problème de la limitation, dans les moteurs de recherche, de la récupération de documents sémantiquement pertinents au-delà de la correspondance des mots clés. Une approche sémantique permettant d'explorer de tels documents est nécessaire pour élargir l'accès aux connaissances des chercheurs utilisant les bibliothèques numériques. Le besoin d'une telle solution est principalement requis pour la recherche interdisciplinaire où différents domaines scientifiques ont tendance à utiliser différentes terminologies pour décrire le sujet interdisciplinaire. Pour résoudre ce problème, nous avons fourni les contributions suivantes:

Tout d'abord, un modèle de représentation sémantique de phrases basé sur des traits par paires. Le modèle est capable d'utiliser à la fois des fonctionnalités linguistiques et des fonctionnalités d'intégration de mots et de phrases non supervisées. Le modèle a bien fonctionné non seulement dans un repère de similarité de texte sémantique, mais également dans quelques cas d'utilisation. Le premier cas d'utilisation était la capacité d'identifier des phrases et des documents similaires écrits dans des styles différents, tandis que l'autre cas était de mettre en évidence (de coupler) les phrases du résumé papier à leurs phrases sémantiquement similaires dans le contenu papier.

Deuxièmement, nous avons proposé un nouveau pipeline pour élargir un corpus de sujets de recherche interdisciplinaires. Le pipeline recommande des articles sémantiquement pertinents qui n'utilisent pas nécessairement les mêmes terminologies du nom du sujet dans toutes les disciplines scientifiques connexes. Nous avons présenté un cas d'utilisation d'une bibliothèque numérique multidisciplinaire contenant des millions d'articles sur un sujet interdisciplinaire. Les experts en la matière ont évalué manuellement les articles recommandés par rapport à un autre système de recommandation dans lequel notre pipeline fonctionnait beaucoup mieux. Le modèle montre également des résultats d'évaluation corrélés à l'aide du modèle d'estimation de similarité sémantique de titre à titre que nous avons développé pour la similarité sémantique de phrases. Une analyse de la diversité des sous-thèmes a été menée sur

les résultats recommandés des deux modèles, notre approche ayant également fourni une meilleure diversité des recommandations.

Troisièmement, nous avons appliqué une approche hybride utilisant le pipeline du modèle précédent d'expansion de corpus et l'extension des requêtes sémantiques utilisant des bases de données lexicales pour enrichir les métadonnées d'une bibliothèque numérique multidisciplinaire. Nous avons fourni une étude de cas de 33 catégories de sujets scientifiques tirée de «Web of Science» et mené une expérience d'évaluation de la technique de modélisation par sujets. Nous avons également déterminé expérimentalement le système hybride le plus performant sur la base du degré de fusion entre les deux approches sémantiques. Les résultats ont également montré une bonne diversité et des résultats inattendus en utilisant les métadonnées enrichies.

Enfin, nous avons fourni un aperçu des niveaux de granularité du texte et des différentes approches d'exploration de texte. Nous avons conclu que les approches hybrides basées à la fois sur des caractéristiques sémantiques d'apprentissage automatique statistique et sur l'utilisation de réseaux sémantiques sont généralement meilleures pour résoudre les problèmes de similarité de texte sémantique. L'approche hybride est non seulement meilleure en précision, mais elle est également utile pour d'autres indicateurs d'évaluation ciblés comme la diversité et les imprévus.

Nous aimerions améliorer certains points dans les approches proposées. Tout d'abord, nous souhaitons enrichir le pipeline hybride d'enrichissement des métadonnées avec deux composants supplémentaires: l'incorporation de phrases dans les titres papier ainsi que l'utilisation de l'incorporation de mots pour développer les ensembles de synonymes dans le processus d'expansion de la requête. Nous pourrions également extraire des termes plus liés sémantiquement pour le développement de la requête, y compris, par exemple, les noms de catégorie du nom de sujet scientifique trouvé dans le réseau sémantique. En outre, pour notre modèle d'estimation de similarité sémantique de phrase proposé (SenSim), nous prévoyons de mener d'autres expériences sur tous les autres points de repère disponibles. Nous souhaitons également combiner d'autres fonctionnalités à nos ensembles de fonctionnalités par paire (\ ie l'incorporation de phrases et l'incorporation de mots dans des synonymes).

Deuxièmement, nous aimerions explorer et étudier les avantages de l'adoption des dernières avancées en matière de méthodes d'apprentissage en profondeur basées sur l'apprentissage en profondeur (publiées après nos contributions). Par exemple, il existe deux approches très intéressantes, l'une appelée ELMo et l'autre appelée BERT.

Troisièmement, nous souhaiterions fournir davantage d'études de cas et d'applications de la méthode que nous proposons pour «l'enrichissement en métadonnées à base sémantique» dans des domaines autres que les bibliothèques numériques

scientifiques. Par exemple, la même approche pourrait être utilisée dans d'autres systèmes de récupération d'informations d'articles de presse, d'encyclopédies électroniques (comme Wikipedia) et de systèmes de recommandation basés sur la sémantique de publicités présentant un contenu pertinent.

Enfin, nous pensons que le langage naturel est l'une des tâches les plus complexes de l'Intelligence Artificielle (IA). Sur la base de notre étude sur le terrain, nous pensons que nous sommes encore loin de résoudre ce problème. Cependant, l'application des avancées récentes de l'IA en matière d'extraction de texte pour résoudre les problèmes de récupération d'informations dans les bibliothèques numériques, y compris nos contributions, a donné des résultats encourageants. Nous pensons que les recherches dans cette direction devraient se poursuivre. Nous espérons avoir une validabilité humaine

Mots clés
Recherche d'information, similarité sémantique, enrichissement de métadonnées, fouille de texte, désambiguïsation d'entités nommées, gestion de la connaissance, bibliothèque numérique.