Combinaisons d'approches statistiques et sémantiques appliquées aux bibliothèques numériques scientifiques pour la promotion de la recherche pluridisciplinaire

Combinations of statistical and semantic approaches applied to scientific digital libraries for the promotion of multidisciplinary research

Fabrice Muhlenbach[1], Hussein T. Al-Natsheh[2,3]

[1]Université de Lyon, UJM-Saint-Etienne, CNRS, Lab. Hubert Curien UMR 5516
18 rue du Professeur Benoît Lauras, F-42023 Saint Etienne, France
fabrice.muhlenbach@univ-st-etienne.fr

[2]Université de Lyon, Lyon 2, ERIC EA 3083
5 avenue Pierre Mendès-France, F-69676 Bron Cedex, France
h.natsheh@ciapple.com

RÉSUMÉ. La connaissance dans tous les domaines scientifiques est maintenant disponible dans les bibliothèques numériques. Le problème est que les articles appartenant à différentes communautés de recherche n'emploient pas le même vocabulaire pour parler du même sujet. L'accès aux documents pertinents avec des outils de recherche d'information, des moteurs de recherche ou des systèmes de recommandation d'articles scientifiques échouera si ces méthodes ne permettent pas d'intégrer cette variabilité linguistique. Dans ce travail, nous présentons des stratégies d'utilisation des technologies d'intelligence artificielle pour parvenir à étendre la recherche documentaire de manière appropriée afin d'apporter de la diversité dans les résultats recommandés et favoriser ainsi une recherche pluridisciplinaire.

ABSTRACT. The knowledge of all science domains is now available on digital libraries. The problem is that the papers belonging to different research communities do not use the same vocabulary to talk about the same subject. Access to relevant documents with information retrieval tools, search engines or research-paper recommender systems will fail if these methods do not consider this linguistic variability. In this work, we present strategies for using artificial intelligence technologies to successfully expand the literature search to bring diversity to the recommended results, thereby promoting multidisciplinary research.

MOTS-CLÉS. système de recommandation, bibliothèque numérique, recherche pluridisciplinaire, similarité sémantique, plongement lexical, épistémologie.

KEYWORDS. recommender system, digital libraries, multidisciplinary research, semantic similarity, word embedding, epistemology.

1. Introduction

Les réflexions et travaux présentés dans cet article font suite à une constatation de problèmes rencontrés par les auteurs, alors qu'ils n'étaient encore qu'étudiants, quand ils se sont confrontés à un contexte pluridisciplinaire. Le premier avait effectué des études de sciences cognitives, cette discipline ayant pour objet la « nature de l'esprit » Simon (1986) à travers des facettes aussi variées que les neurosciences, la linguistique computationnelle, l'anthropologie cognitive, la psychologie cognitive, la philosophie de la cognition ou l'intelligence artificielle. Lors des séminaires où intervenaient des spécialistes de chacun des domaines, il s'était rendu compte qu'il était particulièrement difficile pour les chercheurs de disciplines différentes de parvenir à se faire comprendre de leurs collègues, alors que ces derniers, malgré la diversité de leurs approches, avaient *in fine* le même objet d'étude. Le second auteur, en effectuant son stage de master au CERN de Genève – le lieu où était né le Web ! –, avait entendu parler autour

de lui de “multivariate analysis” par les spécialistes en physique quantique sans savoir exactement de quelle technique statistique il s’agissait. Il avait découvert avec surprise que cela correspondait au domaine qu’il connaissait sous le nom d’apprentissage automatique : tout comme Monsieur Jourdain du *Bourgeois Gentilhomme* de Molière et la prose, il faisait de « l’analyse multivariée » sans le savoir !

Utiliser des termes ou expressions différents pour exprimer une même notion (synonymie), ou employer le même terme ou la même expression avec des sens différents (polysémie), voilà autant de difficultés qui peuvent compliquer les recherches menées au sein d’une discipline scientifique donnée, mais qui rendent d’autant plus périlleuses celles menées dans un contexte pluridisciplinaire. Nous considérons que c’est pourtant à la frontière des disciplines, ou quand plusieurs disciplines échangent leurs savoirs communs, que se produisent les grandes découvertes et avancées scientifiques.

À l’heure de la révolution numérique, les connaissances scientifiques se retrouvent dans des documents (articles, ouvrages, rapports...) directement accessibles sur Internet dans des bibliothèques numériques, qu’elles soient pluridisciplinaires et couvrent l’étendue de la connaissance humaine ou spécialisées sur une thématique donnée (une période historique, une certaine culture...), qu’elles soient en libre accès (comme le projet *Gallica*) ou les produits d’éditeurs scientifiques disponibles par abonnement ou paiement à l’article (comme *ScienceDirect* d’Elsevier ou *SpringerLink* de Springer).

La richesse en diversité de ces bibliothèques numériques est néanmoins mise à mal par la porte d’accès aux documents qu’elles recèlent. Alors que dans une bibliothèque physique pourvue de livres et revues imprimés sur papier, des accidents heureux peuvent se produire quand un chercheur tombe sur un document dont le titre ou la couverture attirent son attention alors qu’il s’aventurait dans les rayons à la recherche d’autre chose, ces découvertes réalisées sur le mode de la sérendipité n’ont pas l’occasion de se produire dans le monde numérique. L’accès à l’information se fait en effet au moyen de moteurs de recherche issus des techniques produites en recherche d’information, et les résultats sont présentés à travers des systèmes de recommandation, or les algorithmes de ces méthodes ne visent que la précision ou pertinence des résultats.

L’article présenté ici détaille des stratégies élaborées par les auteurs pour utiliser les technologies d’intelligence artificielle non pour obtenir des documents très spécialisés correspondant parfaitement aux mots-clés saisis par les utilisateurs des bibliothèques numériques et limités à leurs disciplines d’origine, mais au contraire à étendre la recherche documentaire de manière appropriée afin d’apporter de la diversité dans les résultats recommandés et favoriser ainsi une recherche pluridisciplinaire.

Ces stratégies sont hybrides dans la mesure où elles combinent deux angles d’attaque. D’une part, puisqu’elles emploient comme matériel de base des textes scientifiques rédigés par des êtres humains, elles intègrent des méthodes développées en traitement des langues naturelles mettant en avant le sens des mots, des phrases ou des documents et seront désignées par la suite en étant accompagnées de l’épithète « sémantique ». D’autre part, comme la masse des documents employés dans les bibliothèques numériques est vraiment conséquente, ces stratégies d’accès aux documents tirent également bénéfice des approches récentes concernant l’apprentissage profond où les aspects liés au sens émergent de représentations vectorielles issues de caractéristiques statistiques des termes au sein des phrases dans l’ensemble des documents, aussi utiliserons-nous l’adjectif « statistique » pour qualifier ces approches.

2. Avant-propos épistémologique

Suivant l'usage en cours dans la tradition française Brenner (2003) ; Brenner et Gayon (2009), nous entendons par « épistémologie » la discipline qui prend la connaissance scientifique pour objet, qui fait partie de la philosophie des sciences, et qui s'intéresse aux questions portant sur la production des connaissances scientifiques, leur validation, leur organisation, leur évolution, etc.

De nombreuses questions épistémologiques se posent sur la manière dont peuvent se produire les connaissances scientifiques. Typiquement, l'activité d'un chercheur – lorsqu'il n'effectue pas les autres tâches nécessaires à sa fonction (enseignement, administration, recherche de crédits, évaluation du travail de ses pairs...) – prend l'un des deux chemins suivants : l'*exploration* ou l'*exploitation* Berger-Tal et al. (2014).

Dans la phase d'exploitation, le chercheur se base sur ses connaissances existantes pour produire de nouvelles connaissances en suivant un procédé de découverte propre à sa discipline : un mathématicien s'attaque à la démonstration d'un théorème, un archéologue va sur le terrain pour réaliser des fouilles, un psychologue fait passer des expériences à des sujets humains, etc.

Dans la phase exploratoire, le chercheur remet à jour ses connaissances en discutant avec les collègues de son laboratoire, en lisant des ouvrages ou des articles scientifiques, en assistant à des séminaires ou en participant à des conférences. La recherche exploratoire est une phase essentielle de l'activité de chercheur. Selon Gary Marchionini, « la recherche exploratoire fait de nous tous des pionniers et des aventuriers d'un nouveau monde riche d'informations à découvrir, avec des pièges et des coûts nouveaux » Marchionini (2006).

Au cours de ces deux phases interviennent des processus qui ne suivent pas le seul déroulement d'une démarche de travail méthodique : intuition, créativité et imagination ont aussi leurs places Langley et al. (1987), de même que la capacité à établir des analogies entre disciplines, voire à savoir tirer opportunément les conséquences de l'apparition d'un événement s'étant produit dans un contexte fortuit mais ayant une véritable valeur de découverte scientifique, et ceci de façon inattendue (*sérendipité*).

À la suite de l'arrivée de la micro-informatique dans les années 1980, du *web* dans les années 1990 et de l'informatique nomade dans les années 2000, une révolution numérique a pu naître et faire émerger une véritable « *e*-science » où la disponibilité de vastes ensembles de données à manipuler et d'outils informatiques capables de les traiter constituent un nouveau paradigme de recherche pour les chercheurs afin de les aider dans leurs activités Hey et al. (2009).

Les bibliothèques numériques, et plus particulièrement les bibliothèques numériques scientifiques, représentent un des constituants clés de ce nouveau paradigme de recherche. Par « bibliothèque numérique », nous désignons une bibliothèque dans laquelle une proportion conséquente de ses ressources est disponible dans un format lisible par une machine (à la différence d'un format papier ou sur microfiches) et accessible au moyen d'un ordinateur. Le contenu numérique d'une telle bibliothèque peut être détenu localement ou accessible à distance via des réseaux informatiques Kresh (2007).

Les bibliothèques numériques ne sont pas seulement une nouvelle technologie, elles constituent un changement radical dans le domaine de la production de la connaissance. L'arrivée des bibliothèques numériques a suscité l'espoir de faire tomber les barrières existant d'une part entre le public – qui est le

plus souvent à l'origine de la production de la recherche en étant à la source des capitaux finançant les chercheurs et instituts – et le privé – le produit de la recherche se retrouve sous la forme de journaux, d'actes de conférences et de livres publiés auprès de grands groupes éditoriaux qui monnaient cet accès au savoir Goetz et al. (2016) –, et d'autre part entre les différentes communautés utilisatrices et productrices de tout ce savoir Van House (2003).

Nous considérons que le cœur du problème se situe au niveau de l'interface entre les utilisateurs (les chercheurs) et les bibliothèques numériques présentant toute cette richesse en informations et connaissances. En effet, tout comme il existe des systèmes de personnalisation au sein des moteurs de recherche les plus employés afin d'« aider » les internautes à travers le filtrage de la masse d'information du *web*, les amenant de façon malheureuse à un isolement intellectuel et culturel Pariser (2011), les utilisateurs des bibliothèques numériques scientifiques se verront le plus souvent restreints à une « bulle de filtres » les empêchant d'accéder au contenu des articles d'autres disciplines, faute de connaître le vocabulaire à employer et saisir les bons mots-clés de ces disciplines dans les systèmes d'interrogation de ces bibliothèques.

Dans la recherche scientifique, les chercheurs venant de disciplines différentes, n'étudiant pas les mêmes objets et n'ayant ni les mêmes approches ni les mêmes méthodologies, se sont forgés des outils conceptuels spécifiques décrits par un vocabulaire adapté, un « jargon » propre à leur discipline. Cette division disciplinaire, professionnelle et sociale en disciplines et sous-disciplines aboutit à une hyper-spécialisation, que ce soit dans le domaine des sciences appliquées, des sciences naturelles ou des sciences sociales Lahire (2012). Cette tendance à produire des subdivisions disciplinaires vise à favoriser la professionnalisation de certains pans de la recherche en permettant l'établissement d'experts dans leurs domaines respectifs mais, dans le même temps, elle appauvrit aussi la recherche menée à la marge, aux frontières des disciplines, ou partagée par différents domaines de recherche. Il en découle que deux disciplines différentes, mais ayant des éléments d'étude communs, auront du mal à échanger les découvertes effectuées dans leurs domaines respectifs en raison des habitudes spécifiques de désigner les concepts sous un vocabulaire particulier qui n'est pas partagé par les autres disciplines.

Autrefois, les savants ne se limitaient pas à une seule discipline scientifique : ils étaient aussi bien philosophes que physiciens, mathématiciens, astronomes, inventeurs, auteurs ou artistes, et leurs connaissances couvraient aussi bien les sciences (exactes et naturelles ou humaines et sociales) que les arts. Ce fut le cas notamment du perse Al-Khwârizmî au IX^{e} siècle, de Léonard de Vinci dans la deuxième moitié du XVI^{e} siècle, des français René Descartes et Blaise Pascal au XVII^{e} siècle ou de l'allemand Gottfried Leibniz au début du XVIII^{e} siècle. De nos jours, trouver un esprit aussi « polymathe » n'est pas chose aisée, mais l'exemple d'Herbert A. Simon (1916–2001) est une belle illustration du succès représenté par la capacité à faire sauter les divisions disciplinaires : pionnier de l'intelligence artificielle, il avait su adapter ses découvertes faites en psychologie cognitive dans le domaine de l'économie, recevant la consécration suprême de la part des économistes en 1978 Simon (1996).

Les chercheurs d'aujourd'hui ont la chance de disposer de bibliothèques numériques scientifiques pluridisciplinaires potentiellement capables de les aider à réaliser de fructueuses passerelles entre disciplines scientifiques. Pourtant, sans interface d'interrogation adaptée capable de faire émerger des documents associés à un sujet donné mais désigné par un vocabulaire différent dans une autre discipline, l'utilisateur sera restreint aux documents de son seul domaine scientifique. Il faut ainsi veiller à doter les systèmes d'interrogation, de recherche d'information et de recommandation des bibliothèques numériques

de contre-mesures permettant de promouvoir la diversité pluridisciplinaire. Sans cela, un chercheur utilisant une bibliothèque numérique scientifique n'aura accès qu'aux sources d'information de sa propre discipline. Il en résultera que les documents trouvés ne feront que confirmer ou affirmer le point de vue initial du chercheur utilisant cette bibliothèque numérique, mettant ainsi en cause l'une des règles d'or de la science qui est la propriété de *réfutabilité* Popper (1973).

3. État de l'art

Dans la section précédente, nous avons discuté de l'intérêt épistémologique de pouvoir procéder à des recherches dans un contexte pluridisciplinaire. Nous avons aussi souligné qu'une bibliothèque numérique scientifique pouvait être un outil idéal pour parvenir à de telles fins, mais que le comportement d'une telle mécanique était grippé par le dysfonctionnement d'un élément clé situé à l'interface entre l'humain et la machine : le système de recommandation, biaisé par une tendance à retourner préférentiellement des articles de recherche de la discipline scientifique de l'intéressé.

Les systèmes de recommandation sont des systèmes intelligents cherchant à filtrer l'information présentée à l'utilisateur d'une plateforme numérique, lorsque celle-ci comporte beaucoup d'éléments, en portant à l'attention de cet utilisateur les items qui sont susceptibles de l'intéresser. Dans notre cas particulier, l'utilisateur est le chercheur, la plateforme est l'interface d'accès à la bibliothèque numérique et les items sont des documents scientifiques, le plus souvent des articles de recherche.

Classiquement, les systèmes de recommandation sont répartis entre des systèmes à filtrage collaboratif et des approches basées sur le contenu Bobadilla et al. (2013). La première forme d'approche est dite « collaborative » dans le sens où les utilisateurs collaborent explicitement avec le système en exprimant par des critères d'évaluation leurs avis sur les items recommandés. La seconde approche s'appuie sur le contenu issu de la description des items à recommander et sur les informations concernant les utilisateurs de la plateforme pour créer des représentations et des profils d'utilisateurs. De la sorte, par appariement de profils d'utilisateurs ou d'items présentant des caractéristiques similaires, il sera possible de recommander à un utilisateur donné un item spécifique, c'est-à-dire de proposer un item semblable à un autre item ayant déjà été apprécié dans le passé, ou de recommander un item venant d'être apprécié par quelqu'un dont les préférences sont d'ordinaire semblables à celles d'un utilisateur donné.

Dans le cas plus spécifique des systèmes de recommandation d'articles de recherche, les approches employées peuvent se ranger dans les catégories suivantes Beel et al. (2016) :

1. le stéréotypage, quand un article est recommandé à partir de l'attrait qu'il est possible d'avoir suivant un certain stéréotype donné, mais qui nécessite la création manuelle et laborieuse de tels stéréotypes et qui n'est qu'un pis-aller quand d'autres alternatives sont incapables de proposer des articles de recherche à recommander ;
2. le filtrage basé sur le contenu, qui est l'approche la plus utilisée dans le domaine de la recommandation des articles scientifiques, et qui utilise des éléments (que ceux-ci soient des mots simples, des n-grammes, des concepts...) dérivés du texte de l'article (titre, résumé, mots-clés, texte brut...) pour servir de variables, notamment via une allocation latente de Dirichlet Blei et al. (2003), suivant un procédé d'apprentissage automatique ;

3. le filtrage collaboratif, qui pour procéder à ses recommandations nécessite des évaluations explicites des articles par des chercheurs qui utilisent ces plateformes – ce qui est assez rarement le cas dans les bibliothèques numériques scientifiques – ou qui déduit l'évaluation d'un article à travers le nombre de fois où un article donné est cité (ce qui nécessite la création d'un graphe de citations) ;
4. les recommandations par co-occurrence, qui se basent sur l'apparition simultanée de plusieurs unités linguistiques semblables entre deux articles, que ce soit au sein du texte brut de ces articles ou limité aux seules citations ;
5. les approches basées sur des graphes, quand des structures comportant des nœuds et des arêtes peuvent être générées entre des co-auteurs, à partir de co-citations, voire de termes communs dans les titres des articles, et qui, à partir du moment où le graphe a été créé, emploient des métriques propres aux graphes et aux réseaux pour trouver des articles candidats à recommander ;
6. la pertinence globale, quand les articles les plus populaires sont recommandés sans distinction à tout le monde ;
7. les approches de recommandation hybride, quand les algorithmes de recommandation combinent plus d'une approche parmi les six citées précédemment.

Notons que, parmi ces approches, la part faite aux recommandations censées favoriser la diversité et la pluridisciplinarité n'est nullement présente. Au contraire, les approches s'appuyant sur la stéréotypie, les graphes de co-auteurs ou la popularité des articles ne feront que renforcer l'enfermement dans la bulle de filtres des chercheurs par leurs tendances à recommander des articles de recherche qui sont caractéristiques de leurs disciplines.

De manière générale, parmi les méthodes développées dans le domaine des systèmes de recommandation, les travaux mettant en avant la diversité ou la nouveauté existent Castells et al. (2015), mais ils demeurent anecdotiques comparés à la littérature écrasante des travaux qui cherchent à optimiser la pertinence ou la précision. Citons néanmoins, dans le domaine de la recommandation d'articles scientifiques, des travaux qui se basent sur des approches de graphes (p. ex. des graphes de citations) et qui introduisent du hasard (p. ex. par des procédés de marche aléatoire) ou de la diversité (p. ex. en favorisant des critères de différence) pour favoriser l'apparition de la sérendipité par recommandation d'articles d'auteurs aux profils très dissimilaires depuis un graphe de co-auteurs Sugiyama et Kan (2011), ou pour proposer une recherche à facettes permettant aux utilisateurs de filtrer l'ensemble des articles scientifiques en choisissant un ou plusieurs critères (les facettes) comme *FeRoSA* Chakraborty et al. (2016), ou encore pour rechercher un équilibre entre la pertinence des recommandations d'articles et leur diversité à travers une marche aléatoire renforcée sur des nœuds du graphe de citations comme *FairScholar* Anand et al. (2017).

Ajoutons que si, pendant des années, les approches de recommandation à filtrage collaboratif ont eu le vent en poupe, notamment boostées par le prix Netflix proposé en 2007, les approches de recommandations basées sur le contenu ont connu un regain d'intérêt ces dernières années.

Une des raisons de ce revirement produit au cours des années 2010 est liée au développement du domaine du traitement automatique des langues naturelles, aux technologies du *web* sémantique et à la disponibilité de nombreuses sources de connaissances ouvertes et utilisables, telles que *Wikipedia*, *DBpedia* ou *BabelNet*, et permettant de travailler sur des concepts plutôt que sur des mots-clés de Gemmis et al. (2015). Comme déjà mentionné à la fin de l'introduction de cet article, nous désignerons par la suite les approches issues de ce domaine de « sémantiques ». Ces approches donnent lieu à une repré-

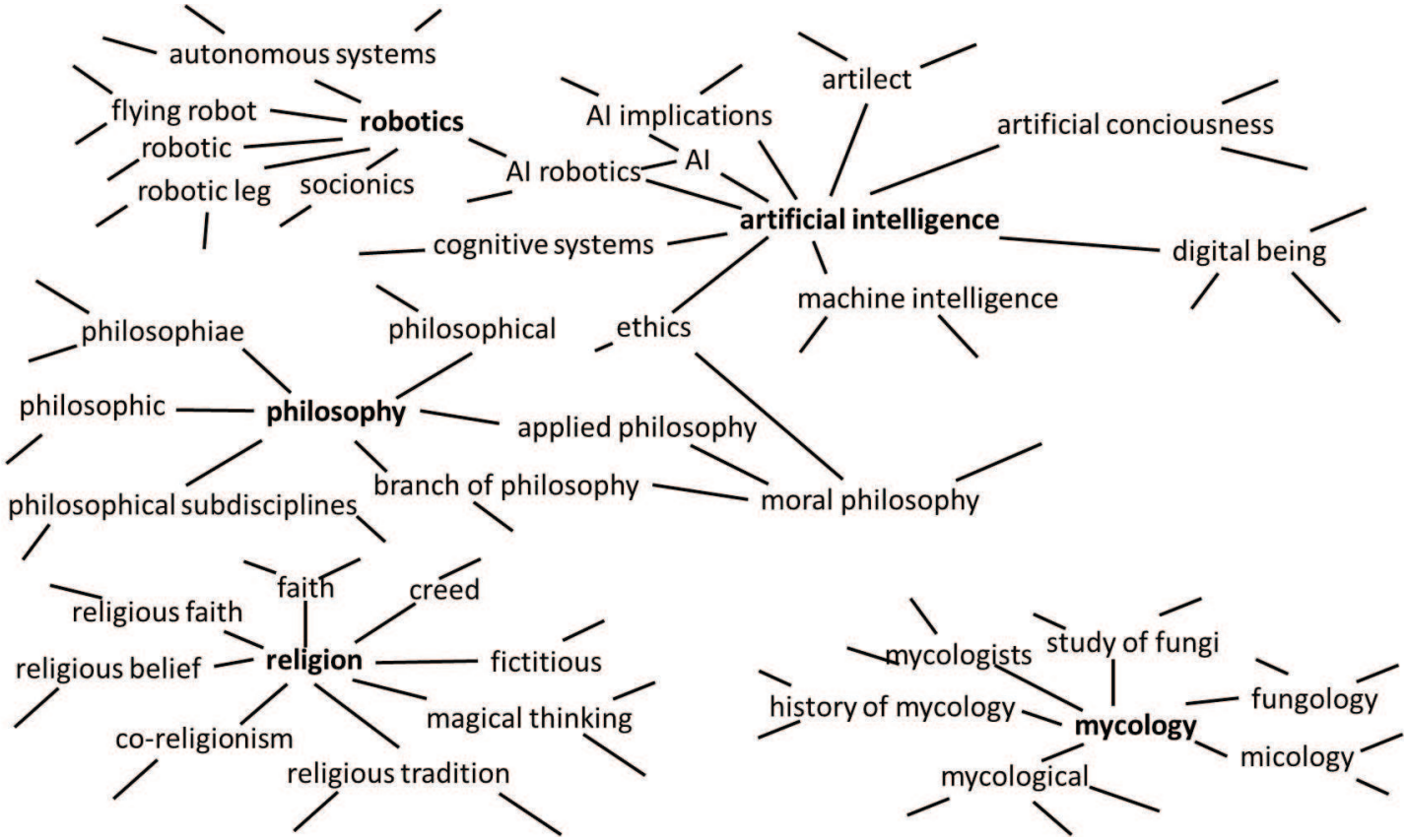

Figure 1. Représentation des termes **artificial intelligence**, **robotics**, **philosophy**, **religion** et **mycology** dans le réseau sémantique *BabelNet*. Cette représentation est caractéristique des approches dites « sémantiques ».

sentation des mots ou concepts liés par des relations de liens sémantiques dans un graphe, comme cela est présenté sur la Figure 1. Les mots ou concepts sont liés en réseaux sémantiques, ce qui permet de retrouver l'ensemble des synonymes (appelé aussi « synset » pour *synonym set*) de ces derniers, et ceci afin de faciliter la récupération des équivalences de termes à des fins de traitements linguistiques. Sur la Figure 1, nous présentons un extrait du réseau sémantique *BabelNet* avec quelques termes en langue anglaise (**intelligence artificielle**, **robotique**, **philosophie**, **religion** et **mycologie**). Les autres concepts directement liés à ces termes par une arête dans le réseau constituent leurs *synsets*.

Une autre raison de l'intérêt pour l'exploitation du contenu dans l'établissement des recommandations est à mettre sur le compte de l'arrivée de l'apprentissage profond et des chamboulements que cette approche a créés dans le domaine de l'intelligence artificielle Sejnowski (2018).

Là encore, l'histoire donne raison aux transferts de connaissances entre disciplines, puisque c'est en s'inspirant du mode de fonctionnement des cerveaux et de la façon dont les neurones communiquent entre eux et sont modifiés par l'expérience qu'ont été établis les modèles de réseaux de neurones McCulloch et Pitts (1943) ; Rosenblatt (1962) ; Rumelhart (1989). Puis ces modèles ont été modifiés, de nouvelles structures ont été développées pour reproduire certains particularités du traitement visuel Hubel et Wiesel (1962), des architectures nouvelles de connexions entre les diverses couches des réseaux ont été mises au point avec l'emploi d'un grand nombre de transformations – d'où l'aspect « profond » de ce type d'apprentissage – opérées sur les données entre la couche d'entrée et la couche de sortie.

Adapté au domaine qui nous intéresse plus spécifiquement dans cet article, cette approche a permis l'émergence de systèmes de recommandation à base d'apprentissage profond, qui tirent bénéfice des

grandes quantités de données extraites pour décrire les items à recommander sans connaître *a priori* l'importance relative de telle ou telle variable descriptive. Diverses directions de recherche ont été tracées dans cette optique, telles que la réalisation d'un apprentissage de représentations combinant les informations de contenu massivement extraites à la fois des profils des utilisateurs et des items. Une autre piste consiste à travailler dans la direction de l'explicabilité des modèles d'apprentissage profond, car les recommandations sont d'autant plus acceptées par l'utilisateur que ces choix lui sont rendus de manière explicite. Pour conclure ces quelques exemples de pistes de recherche menées en vue d'améliorer les performances des modèles de recommandation par apprentissage profond, indiquons que de meilleurs résultats peuvent être attendus en cherchant à augmenter la profondeur et la dimension des architectures de ces modèles S. Zhang et al. (2019). Citons enfin que des récents travaux sur les systèmes de recommandation basés sur le contenu combinent les techniques des réseaux de neurones profonds avec les données ouvertes liées issues des travaux dans le domaine du *web* sémantique Musto et al. (2018)

Complémentaires à ces approches neuronales, les méthodes de plongement lexical (*word embedding*) tirent aussi parti de la grande disponibilité des données sous forme numérique pour produire un apprentissage automatique, ici l'apprentissage d'une représentation de mots à partir de corpus de textes Lebret et Collobert (2013) ; Mikolov, Chen, et al. (2013) ; Mikolov, Sutskever, et al. (2013). Les mots utilisés dans les textes et composant un dictionnaire sont représentés par des vecteurs de nombres réels en fonction de leurs occurrences au sein des textes, autrement dit, à partir de la fréquence d'apparition d'un mot au voisinage d'un autre mot, une représentation condensée (par réduction de la dimensionnalité) va être réalisée et permettra de retrouver une valeur de proximité entre des mots qui apparaissent dans des contextes similaires.

En raison de l'emploi de propriétés statistiques (les fréquences d'apparition des mots) dans la constitution de cette représentation vectorielle, nous parlerons d'approches « statistiques » pour désigner les techniques issues du plongement lexical. L'intérêt du plongement lexical est de faciliter l'analyse sémantique des mots car il existe un lien fort entre des mots qui apparaissent statistiquement souvent dans des contextes similaires et les propriétés sémantiques qui lient ces mots.

Sur la Figure 2, nous donnons une idée graphique de la représentation vectorielle obtenue par avec la technique de plongement lexical *Word2Vec* Mikolov, Chen, et al. (2013) des mêmes termes en langue anglaise que ceux employés dans la Figure 1. La proximité sémantique entre les différents termes est déduite de leurs représentations vectorielles au moyen d'une mesure de similarité (par exemple, la similarité cosinus).

Quelques utilisations de l'approche du plongement lexical ont déjà été réalisées avec succès dans le domaine des systèmes de recommandation à partir du contenu Manotumruksa et al. (2016) ; Musto et al. (2016, 2015) mais aussi dans le domaine des systèmes de recommandation par filtrage collaboratif Krishnamurthy et Puri (2016) ; F. Zhang et al. (2016), voire en combinant les systèmes de recommandation à partir de contenu et ceux basés sur le filtrage collaboratif Lian et al. (2017).

L'amélioration des recommandations d'articles scientifiques peut être également obtenue à travers une stratégie qui ne va pas porter sur l'algorithme de recommandation en tant que tel, mais sur l'enrichissement thématique des articles présents dans la bibliothèque numérique. Avec des mots-clés couvrant davantage de thématiques pour caractériser un article scientifique donné, les systèmes de recherche d'information qui utilisent ces mots-clés parviennent à de meilleurs résultats et permettent de dépasser les

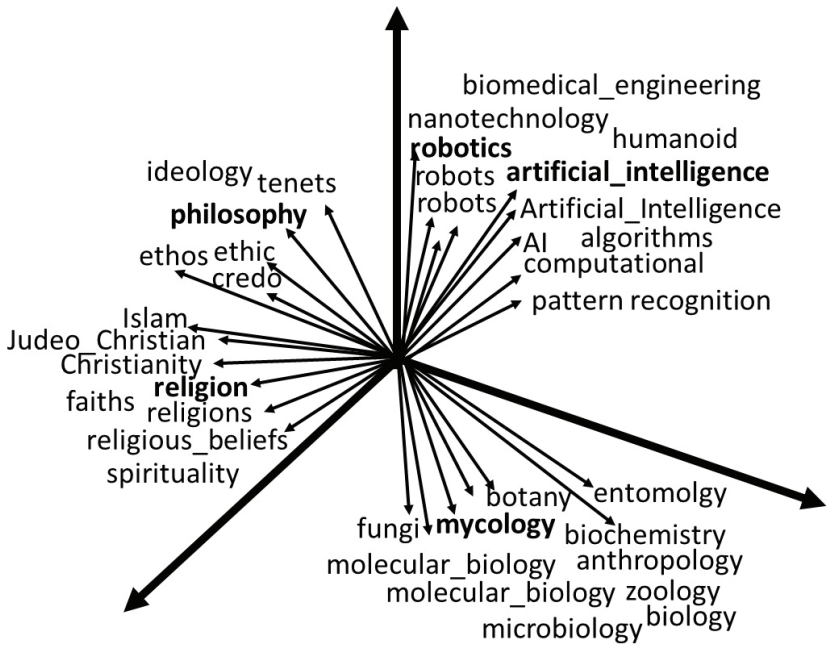

Figure 2. Représentation des termes **artificial intelligence**, **robotics**, **philosophy**, **religion** et **mycology**, ainsi que leurs voisins les plus proches, dans la représentation vectorielle issue de plongement lexical par *Word2Vec*. Cette représentation est caractéristique des approches dites « statistiques ».

cloisonnements disciplinaires. Néanmoins, ces enrichissements (*tags*, mots-clés ou catégories de sujets) manquent d'une taxonomie standardisée et sont pénalisés par la subjectivité du jugement des personnes impliquées dans le processus d'annotation manuel Abrizah et al. (2013). Étendre le champ sémantique des articles des bibliothèques numériques pour améliorer leur lien sémantique avec des documents ou des mots-clés fournis en entrée d'un moteur de recherche ou en tant que requête d'un système de recherche d'information peut se faire de différentes manières. Là encore, une première solution consiste à aborder le problème suivant une approche « sémantique » et ainsi utiliser des bases de données lexicales comme *WordNet* Miller (1995) ou des bases de connaissances telles que *BabelNet* Navigli et Ponzetto (2012), *DBpedia* Lehmann et al. (2015) ou *YAGO* Mahdisoltani et al. (2015). L'autre solution est d'employer l'angle « statistique » pour trouver une solution en utilisant des techniques de plongement lexical Bojanowski et al. (2017) afin de retrouver des terminologies sémantiquement similaires. Malgré l'avantage de ces techniques, celles-ci ne fournissent pas de critère permettant de définir précisément une proximité et ainsi de concevoir qu'un terme proche dans la projection puisse être encore considéré comme étant sémantiquement proche du terme initial. Les modèles thématiques, tels que l'*allocation de Dirichlet latente* ou LDA Blei et al. (2003), sont parmi ceux qui semblent être les plus appropriés pour résoudre le problème qui nous intéresse, bien que difficiles à mettre en œuvre sur des applications réelles comportant des millions de documents.

4. Bibliothèque numérique scientifique pluridisciplinaire et science ouverte

L'objectif de cet article est de s'intéresser à la question épistémologique des conditions qui favorisent la production des connaissances scientifiques et de fournir des pistes de recherche sur des procédés permettant d'améliorer la recherche de documents scientifiques en cherchant à promouvoir la recherche pluridisciplinaire.

Il va sans dire qu'une condition nécessaire à la réponse à apporter à cette problématique est l'existence préalable de sources de documents scientifiques pluridisciplinaires accessibles, comme cela est le cas au sein des institutions françaises depuis la plateforme numérique ISTEX. L'initiative d'**ex**cellence en **i**nformation **s**cientifique et **t**echnique, ou ISTEX, est un projet qui a pour objectif d'offrir un accès en ligne aux collections rétrospectives de la littérature scientifique dans toutes les disciplines à l'ensemble de la communauté de l'enseignement supérieur et de la recherche. Le site ISTEX[1] propose une plateforme permettant un accès à la plupart des collections scientifiques utilisées dans le domaine de la recherche, ainsi que des services à valeur ajoutée permettant d'en optimiser l'exploitation.

Mis en place dans le cadre d'une initiative nationale, le projet ISTEX a été pensé dans une optique de science ouverte Scientific and Technical Information Department – CNRS (2016). Afin d'offrir aux chercheurs de toute institution de recherche publique (française) un accès équivalent aux ressources scientifiques dans un but éthique de droit au savoir, la bibliothèque numérique scientifique ISTEX contient des millions de ressources obtenues à la suite d'une politique massive d'achats centralisée d'archives scientifiques suivant un principe de licences nationales. Ces licences et les corpus collectés ont été obtenus par des accords auprès des éditeurs internationaux et francophones majeurs, en veillant à ce que les accords passés correspondent aux besoins de l'ensemble des communautés scientifiques, avec un équilibre assuré entre les différentes disciplines.

Nuançons néanmoins la portée de ces accords car les éditeurs privés n'ont le plus souvent pas laissé d'accès aux publications des années les plus récentes, les réservant à leurs plateformes (payantes) propres, or, dans le domaine de la recherche, et plus particulièrement dans certaines disciplines où les évolutions vont très vite, il est essentiel d'accéder rapidement aux toutes dernières publications.

À l'heure où nous écrivons ces lignes, la bibliothèque numérique ISTEX comporte plus de 20 millions de documents, concerne plus de 9 000 revues scientifiques et héberge près de 350 000 *ebooks*.

Dans les expérimentations que nous avons menées et que nous présenterons dans la suite, nous nous sommes restreints aux seuls articles scientifiques (publiés en revues ou dans des actes de conférences), rédigés en langue anglaise, publiés après 1990 et dont le résumé comportait une taille suffisante (de 35 à 500 mots). De la sorte, nous avons pu extraire de la bibliothèque numérique pluridisciplinaire ISTEX un ensemble de méta-données concernant plus de 4,17 millions d'articles scientifiques.

5. Illustrations : études expérimentales de la combinaison des approches statistiques et sémantiques pour favoriser la recherche pluridisciplinaire

Le problème qui nous intéresse concerne les recherches documentaires qui échouent à dépasser la seule discipline scientifique de l'utilisateur en raison de la terminologie spécifique employée par celui-ci lors de l'interrogation des bibliothèques numériques scientifiques.

Nous illustrons dans la suite, à travers trois études expérimentales, différents angles d'attaque permettant d'aborder ce problème. Notre objectif n'est pas d'apporter dans cet article une description exhaustive d'une méthode particulière mais plutôt de fournir, à partir de quelques exemples représentatifs, des idées sur certaines pistes envisageables pour répondre à ce problème, ainsi que de trouver des programmes met-

1. https://www.istex.fr/

tant en œuvre concrètement les pistes explorées. Le lecteur intéressé par une description plus complète de ces études (protocoles expérimentaux, résultats obtenus, analyses et discussions) trouvera matière à répondre à ses interrogations à la lecture des publications des auteurs, ainsi qu'en la possibilité de réutiliser le code (en Python) développé pour mener à bien ces expérimentations et accessible en ligne :

– sur l'évaluation de la similarité entre phrases : Al-Natsheh, Martinet, Muhlenbach, et Zighed (2017) ; Al-Natsheh (2019)[2]
– sur le système de recommandation des articles scientifiques fonctionnant par recherche sémantique à partir d'exemples : Al-Natsheh, Martinet, Muhlenbach, Rico, et Zighed (2017)[3]
– sur l'enrichissement de la bibliothèque numérique par étiquetage thématique des articles : Al-Natsheh et al. (2018).[4]

5.1. *Évaluation de la similarité sémantique entre des phrases*

5.1.1. Méthode d'évaluation de la similarité par approche sémantique

Avant de pouvoir procéder concrètement à des améliorations possibles des systèmes de recommandation des articles scientifiques en vue de promouvoir la diversité disciplinaire, il nous paraissait nécessaire de nous intéresser au préalable à l'évaluation de la similarité sémantique entre des documents (puisque nous souhaitons travailler sur des articles de recherche), et même, pour commencer, entre des phrases.

Nous avons démarré notre étude par une preuve de concept consistant à vérifier que nous étions en mesure de distinguer des textes similaires d'autres textes qui ne l'étaient pas. Pour cela, nous avons utilisé un analyseur syntaxique de la nature d'un mot (*Part-of-Speech parser*, ou *PoS*) afin de trouver des quadruplets de type « sujet-verbe-objet-adverbe » (approche « SVOA ») extraits de dépendances syntaxiques. La base de données lexicale *WordNet* est également utilisée pour agréger les similitudes de phrases afin d'extraire des synonymes et des antonymes (dans le cas où des termes de négation tels que « non » sont retrouvés avant le terme étiqueté par l'analyseur syntaxique). Ce modèle d'analyse syntaxique Amir et al. (2017) utilise l'analyseur de dépendance de Stanford Chen et Manning (2014).

Afin de tester les performances en identification de la similarité sémantique de notre approche, nous avons appliqué celle-ci sur un petit texte décrit de manières diverses. Pour cela, nous avons utilisé les *Exercices de styles* de Raymond Queneau Queneau (1947). Dans cet ouvrage, une histoire simple est racontée de 99 façons différentes. Il y a ainsi le même contenu sémantique, sauf que l'auteur ne met pas en avant les mêmes parties du texte, n'emploie pas les mêmes termes, ou ne procède pas au même agencement chronologique des événements.

Lors de nos expérimentations, afin de pouvoir ensuite étendre nos travaux aux articles scientifiques des bibliothèques numériques, majoritairement écrits en anglais, nous avons choisi d'utiliser la traduction anglaise par Barbara Wright des textes de Raymond Queneau Wright et Queneau (1986). Pour effectuer nos tests, nous avons préparé des petits textes de taille équivalente à ceux de Queneau et Wright mais sans relation avec le contenu sémantique initial, à savoir des extraits de *Google News* en anglais. Pour

2. *Sentence Similarity Estimator* : `https://github.com/natsheh/sensim`
3. *Semantic Search-by-Examples* : `https://github.com/ERICUdL/ISTEX_MentalRotation`
4. *Scientific Topic Semantics Tagging* : `https://github.com/ERICUdL/stst`

notre protocole expérimental, nous sélectionnions en tant que requête l'un des 99 textes d'*Exercices de style*, nous calculions une mesure de similarité sémantique issue d'une similarité cosinus entre des représentations par les quadruplets SVOA, puis nous cherchions à voir combien, parmi les 98 autres textes aux styles différents, étaient ceux qui étaient classés devant les autres documents aléatoires de taille similaire. Afin de classer les résultats de la requête, nous avons utilisé la similarité en cosinus entre la requête et les documents. Ce modèle est comparé aux résultats obtenus avec les mêmes données pour un modèle de type TF–IDF tenant compte de n-grammes avec des valeurs de n de 2 ou de 3.

Les résultats obtenus (en F-mesure) pour notre méthode n'étaient pas meilleurs que ceux obtenus par TF–IDF, mais ils nous ont permis de découvrir que :

– il était nécessaire de disposer d'un grand nombre de documents et d'un vocabulaire suffisamment important pour avoir de bons résultats en recherche d'information ;
– dans notre modèle, les résultats pouvaient grandement varier suivant le mode de pondération des différentes variables des quadruplets SVOA (par exemple, donner un poids plus fort aux sujets ou aux verbes qu'aux adverbes) ;
– il était extrêmement difficile d'obtenir des bonnes estimations de la similarité dans des cas particuliers, comme l'écriture de textes sur un mode métaphorique, or les liens sémantiques présents dans une métaphore sont parfois minces et échappent même à la compréhension de certains esprits humains.

5.1.2. Méthode d'évaluation de la similarité par approche hybride

Afin de réaliser des évaluations de similarité sémantique entre phrases, nous avons cherché à améliorer notre modèle initial en le combinant à des informations issues de l'approche statistique. Cette hybridation s'est faite au départ très simplement en ajoutant, aux variables issues de l'analyse de la nature des mots (PoS) et aux entités nommées permettant d'obtenir une mesure de similarité moyenne sur les vecteurs de mots de chaque paire de phrases, d'autres variables de nature plus « statistiques » telles qu'une variable issue de la transformation de chaque phrase en une représentation par sac-de-mots, ou de n-grammes de 2 ou 3 caractères, pondérée par TF–IDF, ou une autre variable d'appariement se basant sur la longueur des phrases. À partir de ces variables, nous avons produit deux modèles d'estimateurs de type régression : les forêts aléatoires, d'une part, et la méthode Lasso d'autre part.

Les modèles que nous avons développés ont d'abord été entraînés avec le petit jeu de données issu de la version anglaise d'*Exercices de style* Wright et Queneau (1986). Encouragés par les bons résultats obtenus par rapport à la méthode de base, et afin d'avoir un caractère plus générique pour nos modèles, nous avons utilisé pour l'entraînement différentes bases de test, notamment celles issues des anciennes éditions des compétitions *SemEval* (sur l'évaluation de la similarité sémantique) organisées par l'ACL (*Association for Computational Linguistics*). En procédant à une validation croisée, nous avons ainsi pu sélectionner les paramètres fournissant les meilleurs résultats dans l'évaluation de la similarité sémantique pour chaque méthode d'apprentissage par régression, à savoir le nombre d'estimateurs et la profondeur pour les forêts aléatoires, et le nombre d'itérations pour Lasso.

Ces deux modèles ont ensuite été testés lors de la compétition SemEval 2017 sur une tâche de similarité sémantique de textes (STS) entre des paires de phrases en langue anglaise Al-Natsheh, Martinet, Muhlenbach, et Zighed (2017).

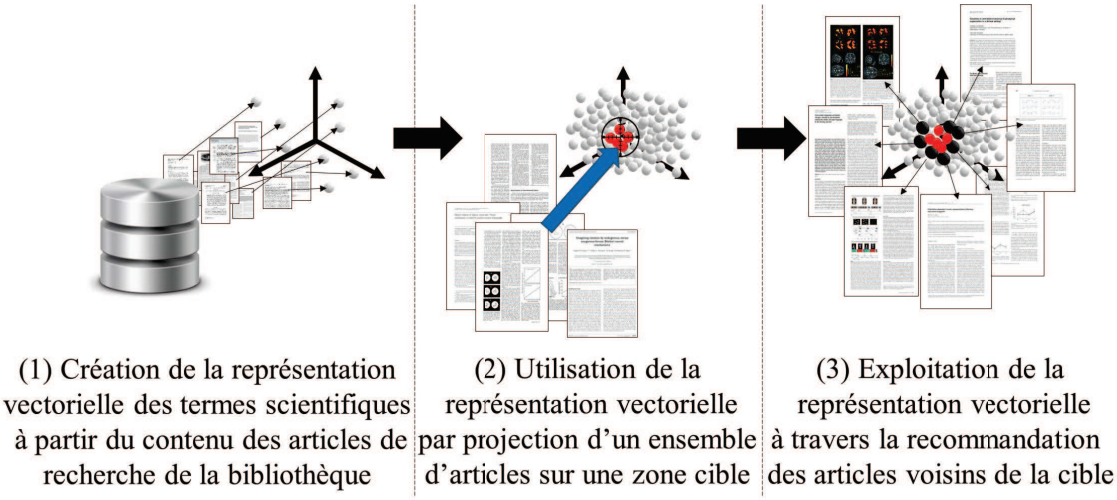

Figure 3. Fonctionnement du modèle *SSbE*. En étape (1) : création de la représentation vectorielle des termes scientifiques à partir du contenu des articles de recherche de la bibliothèque numérique. Cette représentation est réalisée au moyen de techniques de plongement lexical. En étape (2) : utilisation de la représentation vectorielle des termes de la bibliothèque numérique scientifique dans le cadre d'une recommandation à partir d'un ensemble de textes traitant d'un sujet donné. Les similarités sémantiques entre les documents proposés vont amener à une projection dans l'espace de représentation vectorielle dans une zone d'intérêt donnée (ciblée sur les points en rouge). En étape (3) : exploitation de la représentation vectorielle par extension de la zone d'intérêt issue des documents proposés en tant que requête à la bibliothèque numérique scientifique. À partir de la zone d'intérêt obtenue par projection dans l'espace de représentation vectorielle des termes, le système va recommander des documents scientifiques considérés comme proches, au sens d'une similarité vectorielle (les points noirs voisins des rouges).

Bien que nos propositions n'aient pas remporté la compétition, les résultats obtenus notamment par le modèle avec forêts aléatoires avaient un coefficient de corrélation de Pearson de 80% avec des annotations humaines, ce qui était bien meilleur que la moyenne des résultats proposés (70,82%) ou la médiane des résultats (77,75%), mais qui était moins bon que le meilleur résultat obtenu par une autre équipe (avec un coefficient de corrélation de 85,47%).

5.2. *Système de recommandation d'articles de recherche par extension de corpus*

5.2.1. Modèle de recommandation proposé

Le modèle d'interrogation et de recommandation que nous avons développé est appelé *SSbE* pour "Semantic Search-by-Examples" Al-Natsheh, Martinet, Muhlenbach, Rico, et Zighed (2017). Son mode de fonctionnement est atypique car il n'exploite pas une requête avec des mots-clés saisis par l'utilisateur mais un ensemble d'exemples d'articles scientifiques fournis par celui-ci. De là, le modèle procède à une recherche d'autres exemples dans une représentation sémantique issue d'un plongement lexical.

Dans un premier temps, les articles de la bibliothèque numérique scientifique (ici, les 4 millions d'articles retenus d'ISTEX) sont utilisés pour construire une représentation vectorielle sémantique (à partir de leurs titres et résumés), comme représenté en Figure 3–(1). La méthode de vectorisation que nous avons employée est la transformation en sac-de-mots pondérée par TF–IDF. Pour faire émerger les propriétés sémantiques de cette représentation, il est nécessaire de condenser celle-ci, ce que nous avons fait par analyse sémantique latente (LSA). Lorsque la bibliothèque numérique scientifique est utilisée par un chercheur en phase exploratoire, il est demandé à l'utilisateur de fournir au système d'interrogation un ensemble de documents pertinents traitant du sujet qui l'intéresse et qu'il souhaiterait approfondir

par d'autres sources. Titres, résumés, mots-clés et autres méta-données sont extraits des articles fournis et ces caractéristiques sont transformées en représentation vectorielle suivant le même procédé que dans l'étape vue précédemment. Afin de mieux délimiter la zone d'intérêt présente dans ces articles considérés comme étant positifs, un nombre équivalent d'articles négatifs sont sélectionnés aléatoirement dans la bibliothèque numérique, en veillant à ne pas avoir d'article correspondant aux articles positifs. De cette façon, un apprentissage supervisé peut être opéré. Après divers tests, notre choix s'est porté sur les forêts aléatoires comme méthode d'apprentissage. Cet apprentissage supervisé permet de tracer une limite entre les articles scientifiques positifs qui concernent la zone d'intérêt de l'utilisateur du système et les articles négatifs qui se situent en dehors de la zone cible, comme cela est indiqué sur la Figure 3–(2). Cette zone d'intérêt cible peut alors être activée, et le système peut ainsi recommander des articles considérés comme susceptibles de traiter de sujets semblables en retrouvant des articles scientifiques dont la représentation vectorielle se situe dans la zone d'intérêt ou en est très proche (Figure 3–(3)). Ajoutons que nous avons aussi fait des tests en réutilisant les évaluations des utilisateurs sur la pertinence des recommandations opérées suivant un procédé d'apprentissage actif, cherchant à voir si les retours utilisateurs permettaient d'améliorer les résultats.

5.2.2. Protocole expérimental et cas d'étude

Pour nos expérimentations, nous avons décidé de nous intéresser à un champ de recherche volontairement éloigné du nôtre (l'informatique) afin de ne pas connaître les résultats attendus et de ne pas pouvoir interférer sur ces derniers. Notre choix s'est porté pour les sciences du sport (dénommées aussi « STAPS », les sciences et techniques des activités physiques et sportives). Ce domaine étant par essence interconnecté avec d'autres domaines (tels que la physiologie, la psychologie, l'anatomie, la biomécanique, la biochimie, etc.), il nous semblait propice à d'éventuelles découvertes transdisciplinaires.

Nous avons soumis notre idée auprès de collègues enseignants-chercheurs du département STAPS de l'Université Lyon 1 qui nous ont proposé, comme cas d'étude, des articles traitant de la « rotation mentale ». Le sujet de la rotation mentale concerne une tâche psychologique permettant de rendre compte de certaines aptitudes mentales à manipuler les images Shepard et Metzler (1971). Une telle tâche de rotation mentale mesure le temps que prennent des sujets humains pour déterminer une identité formelle entre des représentations d'objets pris suivant différentes orientations. L'intérêt des chercheurs en sciences du sport pour la rotation mentale est multiple : les transformations d'images mentales impliquent parfois des processus moteurs et parfois non Kosslyn et al. (2001), les performances de rotation mentale sont liées aux capacités motrices, et des études expérimentales suggèrent qu'une meilleure performance en rotation mentale favoriserait la capacité à effectuer rapidement des rotations motrices complexes, comme on peut le voir dans les mouvements corporels des athlètes professionnels Francuz (2010). Le sujet de recherche de la « rotation mentale » est étudié par des disciplines et communautés de chercheurs très variées (p. ex. les sciences cognitives, l'aérodynamique ou les sciences du sport), et chacune de ces disciplines examine ce sujet selon son propre point de vue, en employant le plus souvent un vocabulaire différent.

Notre approche (modèle *SSbE*) a été comparée avec la méthode *More-Like-This* (*MLT*), la fonction présente dans le serveur de recherche d'*Elasticsearch* Dixit (2017) et qui permet de retourner les documents similaires à un document d'entrée donné suivant une stratégie basée sur TF–IDF. Nous avons considéré les 100 premiers résultats obtenus par chacune des deux approches *SSbE* et *MLT* à partir des documents concernant la rotation mentale donnés en entrée par l'utilisateur. Précisons qu'aucun de ces

200 articles ne devait contenir l'expression "mental rotation" dans son titre ou son résumé pour être candidat aux articles de recherche recommandés, car nous voulions montrer qu'il était possible d'aller au-delà des capacités des systèmes de recherche d'information classiques qui procèdent par mots-clés. Ces 200 articles ont été mélangés et proposés à l'aveugle à deux experts afin qu'ils décident si ces articles étaient pertinents pour le sujet « rotation mentale » ou pas, ou s'ils n'étaient pas en mesure de décider de la pertinence de tels articles. Cette évaluation humaine n'était pas une étape aisée car, comme les articles candidats ne présentaient pas explicitement la mention "mental rotation" en titre ou résumé, il arrivait fréquemment que les experts du domaine soient obligés de lire les articles dans leur intégralité avant de pouvoir se prononcer sur une évaluation de leur pertinence avec le sujet. Les résultats ont montré que, pour les premiers articles retournés par les deux méthodes, *MLT* fournissait une meilleure précision pour les premiers résultats, mais que les valeurs de précision dégringolaient vite dès que la méthode avait proposé plus de 10 articles candidats, alors que les résultats en précision de la méthode *SSbE* avaient une précision moyenne mais parvenaient à la conserver longtemps tout au long du déroulement de la liste des articles recommandés Al-Natsheh, Martinet, Muhlenbach, Rico, et Zighed (2017).

Les articles scientifiques trouvés par le modèle *SSbE* et recommandés aux chercheurs ne sont pas toujours considérés comme pertinents. Cependant, étant donné que les articles proposés contiennent des similitudes sémantiques avec ceux utilisés en entrée (c'est-à-dire avec le corpus initial), les articles recommandés partagent certains liens thématiques avec les articles fournis en entrée et ouvrent sur de nouvelles pistes de recherche. Dans notre étude, il est arrivé à plusieurs reprises que des articles recommandés aient surpris les experts chargés d'évaluer la pertinence de ces documents et leur ont donné des idées pour des recherches futures dans de nouvelles directions. Une première piste qui a suscité l'intérêt des experts en sciences du sport concernait un article qui, sans mentionner la tâche de rotation mentale, évoquait un thème proche concernant les études sur l'aptitude à lire une carte dans différentes orientations Tlauka (2006). Cette découverte a conduit les chercheurs en sciences du sport à voir des extensions de leur travail sur la rotation mentale dans le domaine de la course d'orientation, un sport qui requiert des compétences de navigation utilisant la carte et la boussole pour naviguer sur un terrain inconnu. Un autre exemple de découverte transdisciplinaire faite par les experts en rotation mentale est le suivant : ces derniers ont constaté qu'il existe des liens entre la rotation mentale et la langue des signes, notamment une similitude au niveau des activations de certaines zones cérébrales Sadato et al. (2005). En effet, la langue des signes et la lecture labiale utilisées par les sourds lorsqu'ils communiquent sont des actions qui nécessitent des capacités de rotation mentale pour lire la communication manuelle. D'après ces experts, aucun pont scientifique n'avait déjà été établi entre le sujet de la rotation mentale et ces différents domaines d'études alors qu'il y avait vraisemblablement des choses très intéressantes à découvrir en croisant les connaissances de ces différentes approches.

5.3. *Étiquetage thématique automatisé de corpus*

5.3.1. Modèle d'étiquetage sémantique de thèmes scientifiques

Afin de donner une réponse au problème de l'accès aux documents de disciplines variées lors de l'interrogation d'une bibliothèque numérique avec une approche plus classique, c'est-à-dire au moyen de mots-clés, nous avons décidé d'apporter une proposition de solution consistant à enrichir les articles hébergés dans cette bibliothèque avec de nouvelles catégories. En procédant ainsi, nous souhaitions fa-

voriser les transferts de connaissances entre disciplines, les nouveaux mots-clés ajoutés étant censés faciliter l'accès à des documents dépassant les barrières disciplinaires.

Notre modèle procède par la combinaison de deux méthodes. Dans notre étude, les données employées sont les 4,17 millions d'articles scientifiques issus de la bibliothèque numérique *ISTEX*. Les catégories à attribuer aux articles sont issues de la collection "Web of Science"[5] contenant plus de 250 sujets faisant consensus dans le monde de la recherche. Dans notre étude expérimentale, nous n'avons retenus que 33 catégories de sujets (p. ex. *Artificial Intelligence*, *Biomaterials*, *Biophysics*, *Ceramics*, *Condensed Matter*, *Emergency Medicine*, *Immunology*, etc.)

La première méthode, que nous qualifions de « sémantique », repose sur le simple emploi d'un réseau sémantique (dans notre étude, nous avons utilisé *WordNet* Miller (1995)) afin d'obtenir pour chaque catégorie une liste de synonymes conséquente (un ensemble de synonymes, ou « synset »). Pour chaque catégorie de sujet, nous construisons un ensemble d'entraînement au moyen d'une requête sélectionnant les articles de la bibliothèque scientifique numérique comportant les différentes catégories associées (issues de *Web of Science*) dans leur titre ou leur résumé. Cette requête s'effectue par le moteur de recherche *Elasticsearch* Dixit (2017) implanté dans la plateforme d'*ISTEX*. L'ensemble de test est construit par des articles contenant les mots recherchés dans leur liste de mots-clés ou de sujets, mais absents de leur titre et résumé.

La seconde méthode, que nous qualifions de « statistique » procède par une projection vectorielle afin d'obtenir une représentation sémantique. Au préalable, tous les articles de la bibliothèque sémantique numérique sont transformés dans leur représentation dans un espace sémantique vectoriel (avec utilisation de LSA Halko et al. (2011), sur la matrice de sac de bi-grammes et uni-grammes de mots). Seuls les mots ayant une fréquence d'apparition d'au moins 20 occurrences sont conservés pour l'entraînement du modèle de classement. Nous avons réalisé un modèle de classement pour chacune des 33 catégories et, après avoir testé plusieurs méthodes d'apprentissage supervisé, notre choix s'est porté sur les forêts aléatoires avec un entraînement sur des ensembles d'exemples positifs et négatifs de même taille. Les exemples positifs sont extraits du corpus *ISTEX* avec *Elasticsearch* et les exemples négatifs sont retournés de façon aléatoire. Tous les articles du corpus sont ainsi ordonnés suivant leur probabilité d'appartenir au sujet recherché en une liste dont nous ne conservons que les 100 000 premières réponses.

Notre modèle d'étiquetage final combine les deux méthodes que nous venons de décrire et fusionne, de manière ordonnée, les résultats des listes obtenues par l'approche sémantique et par l'approche statistique.

5.3.2. Protocole expérimental et esquisse des résultats obtenus

La qualité des résultats de chaque méthode est évaluée au moyen du rappel. Notons que pour savoir si une réponse est correcte ou non pour un article donné, il faudrait avoir une évaluation humaine experte dans tous les domaines, ce qui n'est pas envisageable. Nous avons ainsi utilisé pour nos expérimentations un petit jeu de test déjà étiqueté (au minimum 100 articles par sujet). En testant différentes combinaisons de fusion entre les listes fournies par les approches statistique et sémantique (c.-à-d. en

5. `https://images.webofknowledge.com/images/help/WOS/hp_subject_category_terms_tasca.html`

essayant l'approche sémantique seule, ou autant d'articles fournis par les deux approches, ou deux fois plus d'articles de l'approche statistique que de l'approche sémantique, etc.), nos résultats expérimentaux Al-Natsheh et al. (2018) ont indiqué que la combinaison des deux approches produit les meilleurs taux de rappel (15,82%) pour le plus grand nombre de catégories testées (24/33) plutôt qu'une méthode seule. Quelques cas indiquent de meilleurs résultats pour la seule méthode d'utilisation du réseau sémantique, mais ne se rencontrent que si les catégories sont décrites par un seul terme (p. ex. « psychologie », à la différence de « intelligence artificielle »).

À titre d'exemple, un article médical faisant partie de la catégorie 'Transplantation' (c.-à-d. « greffe ») s'est retrouvé étiqueté par notre système avec 'Transplantation' mais aussi 'Mycology' (discipline étudiant les champignons). À la lecture de l'article, on comprend qu'il s'agit d'une étude de greffe de reins et que la catégorie 'Mycology' y est tout à fait appropriée puisqu'il est question de 'fungi' (des « champignons ») dans cet article. Ainsi, dans de nombreux cas, l'attribution de nouvelles catégories permettra une meilleure identification des articles pertinents pour un appariement par mots-clés et possiblement favoriser les recherches interdisciplinaires.

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 combinant les deux approches retourne les meilleurs résultats.

6. Conclusion et perspectives

Dans cet article, nous avons montré l'intérêt complémentaire des approches sémantiques et des approches statistiques pour évaluer la similarité sémantique entre des documents scientifiques. À partir de cette proximité sémantique, nous avons donné des illustrations de travaux personnels permettant d'ajouter de la diversité dans les retours de l'interrogation d'une bibliothèque numérique scientifique afin de promouvoir la recherche scientifique pluridisciplinaire, que ce soit dans le processus de recommandation des articles ou dans l'étiquetage des articles scientifiques hébergés dans cette bibliothèque.

De façon synthétique, on peut considérer que les approches sémantiques vont surtout fonctionner par des liens de synonymie (établis initialement par une expertise manuelle), donnant lieu à des relations sémantiques exactes. Dans le cas des approches statistiques, au contraire, les relations sémantiques sont dérivées de propriétés statistiques des termes dans les documents, issues de représentations vectorielles condensées calculées par ordinateur, et ne sont donc que le produit d'une propriété que l'on pourrait qualifier d'« accidentelle ». Il en résulte que les méthodes et techniques issues du plongement lexical ne possèdent pas intrinsèquement l'exactitude des approches sémantiques. Combiner les approches statistiques et sémantiques, c'est parvenir à ajouter une sorte de flou sémantique autour des termes, phrases ou documents. Nous considérons que parvenir à générer une telle zone de transition sémantique est particulièrement utile car elle peut éventuellement capturer dans son voisinage des concepts semblables mais décrits par des termes différents et apparaissant dans des contextes différents.

En perspective de recherche, mentionnons le fait qu'il existe d'autres angles d'attaque pour chercher à favoriser les recherches entre différentes communautés scientifiques à partir d'une bibliothèque numérique pluridisciplinaire. En particulier, des travaux se basent sur des exploitations de graphes issus des

articles (p. ex. réseaux de citation d'articles, réseau de collaborations entre chercheurs, etc.) pour déduire des relations de proximité thématique entre articles, chercheurs ou disciplines scientifiques. Ces graphes permettent aussi la construction de cartographies conceptuelles et dynamiques de la science, des outils qui permettent de retrouver des connexions dans le système en constante évolution des connaissances scientifiques Small (1997). Mentionnons l'emploi d'une telle cartographie de la science déduite de modèles de mélanges d'analyseurs de facteurs (*mixtures of factor analyzers*) pour gestion de la polysémie et de la synonymie Kwakkel et Cunningham (2009), ou l'utilisation d'une carte conceptuelle déduite d'un graphe construit par l'extraction des thématiques d'un document scientifique par des modèles thématiques comme LDA Blei et al. (2003) afin de combler l'écart existant entre les connaissances de base d'un utilisateur et les connaissances cibles visées par celui-ci Zhao et al. (2016).

Enfin, signalons que les expérimentations présentées ici ont porté sur des documents scientifiques en anglais "scientifique" (qui peut être parfois éloigné de l'anglais britannique, américain ou d'une autre zone linguistique du monde anglo-saxon ou plus globalement anglophone) mais peuvent se généraliser à l'ensemble des langues, pourvu qu'il existe des corpus suffisamment importants dans une langue donnée et des outils d'exploitation des relations sémantiques disponibles dans cette langue (c.-à-d. des réseaux sémantiques ou ontologies lexicalisées). À l'heure actuelle, les articles de recherche sont souvent rédigés en langue anglaise pour favoriser une lecture possible par des locuteurs du monde entier, tout comme le furent à leurs époques et sur de vastes étendues géographiques le grec, le latin, l'arabe, la *lingua franca*, le français ou l'allemand. Néanmoins, la spécialisation de la recherche professionnelle en disciplines et sous-disciplines a donné lieu à l'apparition d'une multitude de jargons spécifiques, des sortes de dialectes disciplinaires dont les concepts demeurent incompréhensibles pour ceux qui ne font pas partie de la même communauté scientifique. Ce phénomène est analogue au mythe de la *Tour de Babel*, lorsque les hommes auraient décidé d'entreprendre la construction d'une tour dont le sommet toucherait le ciel. Une intervention divine aurait ainsi brouillé leur langue afin qu'ils ne se comprennent plus et que cesse cette construction.

Les travaux présentés dans cet article constituent une avancée dans le domaine de l'accès au savoir, cherchant par différentes stratégies sémantiques et statistiques à favoriser les échanges pluridisciplinaires, les analogies entre disciplines scientifiques et les transferts de connaissance. Gageons que l'application des outils proposés ici puisse servir au sein des bibliothèques numériques scientifiques afin de faciliter la construction de cet ouvrage dirigé, le plus haut possible, vers l'établissement de la connaissance scientifique universelle, et ceci à travers une véritable compréhension mutuelle entre les hommes et femmes de toute science !

7. Remerciements

Les auteurs tiennent à exprimer leur gratitude pour leurs collègues du laboratoire ERIC de Lyon, en particulier Djamel A. Zighed pour avoir proposé l'idée de tester les similarités sémantiques sur les textes de Raymond Queneau, ainsi que Lucie Martinet et Fabien Rico pour leurs contributions aux parties expérimentales. De vifs remerciements sont aussi adressés à Patrick Fargier et à Raphaël Massarelli du Centre de Recherche et d'Innovation sur le Sport de Lyon qui ont contribué aux parties expérimentales en tant qu'experts du domaine des sciences du sport.

Bibliographie

Abrizah A., Zainab A. N., Kaur K., & Raj R. G. (2013). LIS journals scientific impact and subject categorization : a comparison between Web of Science and Scopus. *Scientometrics*, *94*(2), 721–740.

Al-Natsheh H. T., Martinet L., Muhlenbach F., Rico F., & Zighed D. A. (2017). Semantic search-by-examples for scientific topic corpus expansion in digital libraries. In R. Gottumukkala et al. (Eds.), *2017 IEEE International Conference on Data Mining Workshops, ICDM Workshops 2017, New Orleans, LA, USA, November 18-21, 2017*, pp. 747–756. IEEE Computer Society.

Al-Natsheh H. T., Martinet L., Muhlenbach F., Rico F., & Zighed D. A. (2018). Metadata enrichment of multi-disciplinary digital library : A semantic-based approach. In E. Méndez, F. Crestani, C. Ribeiro, G. David, & J. C. Lopes (Eds.), *Digital Libraries for Open Knowledge, 22nd International Conference on Theory and Practice of Digital Libraries, TPDL 2018, Porto, Portugal, September 10-13, 2018, Proceedings.*, Vol. 11057, pp. 32–43. Springer.

Al-Natsheh H. T., Martinet L., Muhlenbach F., & Zighed D. A. (2017). UdL at SemEval-2017 Task 1 : Semantic textual similarity estimation of english sentence pairs using regression model over pairwise features. In S. Bethard, M. Carpuat, M. Apidianaki, S. M. Mohammad, D. M. Cer, & D. Jurgens (Eds.), *Proceedings of the 11th International Workshop on Semantic Evaluation, SemEval@ACL 2017, Vancouver, Canada, August 3-4, 2017*, pp. 115–119. Association for Computational Linguistics.

Al-Natsheh H. T. (2019). *Text mining approaches for semantic similarity exploration and metadata enrichment of scientific digital libraries*. Thèse de doctorat, Université de Lyon, Université Lumière Lyon 2, France.

Amir S., Tanasescu A., & Zighed D. A. (2017). Sentence similarity based on semantic kernels for intelligent text retrieval. *Journal of Intelligent Information Systems (JIIS)*, *48*(3), 675–689.

Anand A., Chakraborty T., & Das A. (2017). Fairscholar : Balancing relevance and diversity for scientific paper recommendation. In J. M. Jose et al. (Eds.), *Advances in Information Retrieval - 39th European Conference on IR Research, ECIR 2017, Aberdeen, UK, April 8-13, 2017, Proceedings*, Vol. 10193, pp. 753–757.

Beel J., Gipp B., Langer S., & Breitinger C. (2016). Research-paper recommender systems : a literature survey. *International Journal on Digital Libraries*, *17*(4), 305–338.

Berger-Tal O., Nathan J., Meron E., & Saltz D. (2014). The exploration-exploitation dilemma : A multidisciplinary framework. *PLoS ONE*, *9*(4 : e95693).

Blei D. M., Ng A. Y., & Jordan M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Bobadilla J., Ortega F., Hernando A., & Gutiérrez A. (2013). Recommender systems survey. *Knowledge Based Systems*, *46*, 109–132.

Bojanowski P., Grave E., Joulin A., & Mikolov T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics (TACL)*, *5*, 135–146.

Brenner A. (2003). *Les origines françaises de la philosophie des sciences*. Paris : Presses Universitaires de France.

Brenner A., & Gayon J. (Eds.). (2009). *French studies in the philosophy of science : Contemporary research in France*. Dordrecht : Springer Netherlands.

Castells P., Hurley N. J., & Vargas S. (2015). Novelty and diversity in recommender systems. In F. Ricci, L. Rokach, & B. Shapira (Eds.), *Recommender Systems Handbook*, 2e éd., pp. 881–918. Springer.

Chakraborty T., Krishna A., Singh M., Ganguly N., Goyal P., & Mukherjee A. (2016). Ferosa : A faceted recommendation system for scientific articles. In J. Bailey, L. Khan, T. Washio, G. Dobbie, J. Z. Huang, & R. Wang (Eds.), *Advances in Knowledge Discovery and Data Mining - 20th Pacific-Asia Conference, PAKDD 2016, Auckland, New Zealand, April 19-22, 2016, Proceedings, Part II*, Vol. 9652, pp. 528–541. Springer.

Chen D., & Manning C. D. (2014). A fast and accurate dependency parser using neural networks. In A. Moschitti, B. Pang, & W. Daelemans (Eds.), *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pp. 740–750. ACL.

de Gemmis M., Lops P., Musto C., Narducci F., & Semeraro G. (2015). Semantics-aware content-based recommender systems. In F. Ricci, L. Rokach, & B. Shapira (Eds.), *Recommender Systems Handbook*, pp. 119–159.

Dixit B. (2017). Chapter 2. The Improved Query DSL. In *Mastering Elasticsearch 5.x*, pp. 74–141. Birmingham, UK : Packt Publishing, Limited.

Francuz P. (2010). The influence of body position and spatial orientation of an object on mental rotation task's performance. *Procedia Social and Behavioral Sciences, 5*, 896–900.

Goetz J., Lapoix S., & Poulain H. (2016). Privés de savoir ? (ep.63). In *#DATAGUEULE*. Consulté sur `https://wiki.datagueule.tv/Priv%C3%A9s_de_savoir_%3F_(EP.63)` (Durée : 10 :18, disponible aussi sur `https://www.youtube.com/watch?v=WnxqoP-c0ZE`)

Halko N., Martinsson P.-G., & Tropp J. A. (2011). Finding structure with randomness : Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review, 53*(2), 217–288.

Hey T., Tansley S., & Tolle K. (2009). *The fourth paradigm : Data-intensive scientific discovery*. Microsoft Research.

Hubel D. H., & Wiesel T. N. (1962, Jan.). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of Physiology, 160*(1), 106–154.2.

Kosslyn S. M., Thompson W. L., Wraga M., & Alpert N. M. (2001, Aug.). Imagining rotation by endogenous versus exogenous forces : distinct neural mechanisms. *Neuroreport, 12*(11), 2519–2525.

Kresh D. (2007). *The whole digital library handbook* (D. Kresh, Ed.). American Library Association.

Krishnamurthy B., & Puri N. (2016). Learning vector-space representations of items for recommendations using word embedding models. In M. Connolly (Ed.), *International Conference on Computational Science 2016, ICCS 2016, 6-8 June 2016, San Diego, California, USA*, Vol. 80, pp. 2205–2210. Elsevier.

Kwakkel J. H., & Cunningham S. W. (2009). Managing polysemy and synonymy in science mapping using the mixtures of factor analyzers model. *Journal of the American Society for Information Science and Technology (JASIST), 60*(10), 2064–2078.

Lahire B. (2012). Des effets délétères de la division scientifique du travail sur l'évolution de la sociologie. *SociologieS, Débats, La situation actuelle de la sociologie*.

Langley P. W., Simon H. A., Bradshaw G., & Zytkow J. M. (1987). *Scientific discovery : Computational explorations of the creative process*. Cambridge, MA : The MIT Press.

Lebret R., & Collobert R. (2013). Word emdeddings through hellinger PCA. *CoRR, abs/1312.5542*.

Lehmann J., Isele R., Jakob M., Jentzsch A., Kontokostas D., Mendes P. N., et al. (2015). DBpedia – A large-scale, multilingual knowledge base extracted from Wikipedia. *Semantic Web, 6*(2), 167–195.

Lian J., Zhang F., Xie X., & Sun G. (2017). CCCFNet : A content-boosted collaborative filtering neural network for cross domain recommender systems. In R. Barrett, R. Cummings, E. Agichtein, & E. Gabrilovich (Eds.), *Proceedings of the 26th International Conference on World Wide Web Companion, Perth, Australia, April 3-7, 2017*, pp. 817–818. ACM.

Mahdisoltani F., Biega J., & Suchanek F. M. (2015). YAGO3 : A knowledge base from multilingual wikipedias. In *CIDR 2015, Seventh Biennial Conference on Innovative Data Systems Research, Asilomar, CA, USA, January 4-7, 2015, Online Proceedings*. www.cidrdb.org.

Manotumruksa J., MacDonald C., & Ounis I. (2016). Modelling user preferences using word embeddings for context-aware venue recommendation. *CoRR, abs/1606.07828*.

Marchionini G. (2006). Exploratory search : From finding to understanding. *Communications of the ACM (CACM), 49*, 41–46.

McCulloch W. S., & Pitts W. (1943). A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics, 5*, 115–133.

Mikolov T., Chen K., Corrado G., & Dean J. (2013). Efficient estimation of word representations in vector space. *CoRR, abs/1301.3781*.

Mikolov T., Sutskever I., Chen K., Corrado G. S., & Dean J. (2013). Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, Z. Ghahramani, & K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems 26 : 27th Annual Conference on Neural Information Processing Systems (NIPS) 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pp. 3111–3119.

Miller G. A. (1995). WordNet : A lexical database for English. *Communications of the ACM (CACM), 38*(11), 39–41.

Musto C., Franza T., Semeraro G., de Gemmis M., & Lops P. (2018). Deep content-based recommender systems exploiting recurrent neural networks and linked open data. In T. Mitrovic, J. Zhang, L. Chen, & D. Chin (Eds.), *Adjunct Publication of the 26th Conference on User Modeling, Adaptation and Personalization, UMAP 2018, Singapore, July 08-11, 2018*, pp. 239–244. ACM.

Musto C., Semeraro G., de Gemmis M., & Lops P. (2016). Learning word embeddings from wikipedia for content-based recommender systems. In N. Ferro et al. (Eds.), *Advances in Information Retrieval - 38th European Conference on IR Research, ECIR 2016, Padua, Italy, March 20-23, 2016. Proceedings*, Vol. 9626, pp. 729–734. Springer.

Musto C., Semeraro G., Gemmis M. de, & Lops P. (2015). Word embedding techniques for content-based recommender systems : An empirical evaluation. In P. Castells (Ed.), *Poster Proceedings of the 9th ACM Conference on Recommender Systems, RecSys 2015, Vienna, Austria, September 16, 2015*, Vol. 1441. CEUR-WS.org.

Navigli R., & Ponzetto S. P. (2012). BabelNet : The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence, 193*, 217–250.

Pariser E. (2011). *The filter bubble : What the internet is hiding from you*. New York, NY : Penguin Press.

Popper K. R. (1973). *Logique de la découverte scientifique*. Payot. (Traduction française par Nicole Thyssen-Rutten et Philippe Devaux de „Logik der Forschung. Zur Erkenntnistheorie der modernen Naturwissenschaft" paru en 1934. Rééd. Payot, Coll. « Bibliothèque scientifique », 1995.)

Queneau R. (1947). *Exercices de style*. Paris : Gallimard, Collection Folio. (Nouvelle édition : 16 mars 1982.)

Rosenblatt F. (1962). *Principles of neurodynamics : Perceptrons and the theory of brain mechanisms*. Spartan Books.

Rumelhart D. E. (1989). *Parallel distributed processing, 9th edition*. Cambridge, MA : The MIT Press.

Sadato N., Okada T., Honda M., Matsuki K.-I., Yoshida M., Kashikura K.-I., et al. (2005, August). Cross-modal integration and plastic changes revealed by lip movement, random-dot motion and sign languages in the hearing and deaf. *Cerebral Cortex, 15*(8), 1113–1122.

Scientific and Technical Information Department – CNRS. (2016). *White paper – open science in a digital republic*. Marseille : OpenEdition Press.

Sejnowski T. J. (2018). *The deep learning revolution*. Cambridge, MA : The MIT Press.

Shepard R. N., & Metzler J. (1971, Feb.). Mental rotation of three-dimensional objects. *Science, New Series, 171*(3972), 701–703.

Simon H. A. (1986). Préface. In C. Bonnet, J.-M. Hoc, & G. Tiberghien (Eds.), *Psychologie, intelligence artificielle et automatique*, pp. 5–7. Liège : Pierre Mardaga.

Simon H. A. (1996). *Models of my life*. Cambridge, MA : The MIT Press.

Small H. G. (1997). Update on science mapping : Creating large document spaces. *Scientometrics, 38*(2), 275–293.

Sugiyama K., & Kan M. (2011). Serendipitous recommendation for scholarly papers considering relations among researchers. In G. Newton, M. J. Wright, & L. N. Cassel (Eds.), *Proceedings of the 2011 Joint International Conference on Digital Libraries, JCDL 2011, Ottawa, ON, Canada, June 13-17, 2011*, pp. 307–310. ACM.

Tlauka M. (2006). Orientation dependent mental representations following real-world navigation. *Scandinavian Journal of Psychology, 47*, 171–176.

Van House N. A. (2003). Digital libraries and collaborative knowledge construction. In A. Peterson Bishop, N. A. Van House, & B. P. Buttenfield (Eds.), *Digital library use : Social practice in design and evaluation*, pp. 271–295. Cambridge, MA : The MIT Press.

Wright B., & Queneau R. (1986). *Exercises in style*. Gaberbocchus Press. (Traduction anglaise d'« Exercices de style » de Raymond Queneau, paru en 1947.)

Zhang F., Yuan N. J., Lian D., Xie X., & Ma W. (2016). Collaborative knowledge base embedding for recommender systems. In B. Krishnapuram, M. Shah, A. J. Smola, C. C. Aggarwal, D. Shen, & R. Rastogi (Eds.), *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pp. 353–362. ACM.

Zhang S., Yao L., Sun A., & Tay Y. (2019). Deep learning based recommender system : A survey and new perspectives. *ACM Computing Surveys, 52*(1), 5 :1–5 :38.

Zhao W., Wu R., & Liu H. (2016). Paper recommendation based on the knowledge gap between a researcher's background knowledge and research target. *Information Processing and Management, 52*(5), 976–988.

