Université Aix-Marseille

École Doctorale 184

Faculté des Sciences et Techniques

LSIS UMR CNRS 7296 - Dimag
Cléo UMS CNRS 3287 - OpenEdition

Thèse présentée pour obtenir le grade universitaire de docteur

Discipline : Mathématique et informatique
Spécialité : Informatique

Anaïs OLLAGNIER

Analyse de requêtes en langue naturelle et extraction d'informations bibliographiques pour une recherche de livres orientée contenu efficace.

Soutenue le 29/11/2017 devant le jury :

Pascale SÉBILLOT	Professeure, INSA de Rennes	Rapporteur
Guillaume CABANAC	MCF HDR, Université Toulouse 3	Rapporteur
Brigitte GRAU	Professeure, ENSIEE	Examinateur
Ludovic TANGUY	MCF HDR, Université Toulouse 2	Examinateur
Sébastien FOURNIER	MCF, Université Aix-Marseille	Co-directeur de thèse
Patrice BELLOT	Professeur, Université Aix-Marseille	Directeur de thèse

v2.2 2017-11-29
Ce travail a bénéficié d'une aide de l'État gérée par l'Agence Nationale de la Recherche au titre du programme Investissements d'avenir dans le cadre des projets EquipEx DILOH (ANR-11-EQPX-0013).
Commentaires, corrections, et autres remarques sont les bienvenus à :
anais.ollagnier@univ-amu.fr
Laboratoire des Sciences de L'Information et des Systèmes, LSIS UMR 7296,
Université Aix-Marseille,
Batiment Polytech,
Avenue Escadrille Normandie-Niemen,
13397 MARSEILLE CEDEX 20

Remerciements

Je tiens à remercier tout d'abord l'ensemble des membres de mon jury : Pascale Sébillot et Guillaume Cabanac, qui ont accepté d'être les rapporteurs de mon mémoire, ainsi que Brigitte Grau et Ludovic Tanguy qui ont participé à cette soutenance en tant qu'examinateurs.

J'adresse ensuite mes plus chaleureux remerciements à Patrice Bellot et Sébastien Fournier, mes encadrants de thèse, pour leur confiance et leurs nombreux encouragements. Je remercie Patrice de m'avoir fait confiance afin de mener à bien cette aventure, son soutien, sa ténacité sans faille et ses conseils avisés m'ont été d'une grande aide tout au long de la thèse. Je ne remercierai jamais assez Sébastien de s'être joint à cette aventure, sa présence durant ces trois années a été source d'une grande motivation grâce à sa disponibilité et son soutien. Je sais que je n'ai pas toujours été facile et souvent emprise aux doutes mais vous m'avez permis de tenir et d'avancer.

Bien entendu, cette thèse n'aurait pas été possible sans l'action d'autres personnes et organismes. J'associe ici le Centre pour l'édition électronique ouverte (Cléo), qui m'ont permis de réaliser ces travaux. Je tiens à remercier particulièrement Marin Dacos, fondateur du Cléo, qui m'a accueillie au sein de son unité pendant trois ans, m'offrant ainsi l'opportunité de découvrir l'univers de l'édition électronique et de toutes les technologies associées. Merci à Elodie, Arnaud, Mathieu qui ont été d'une aide précieuse et de bons conseils. J'ai une pensée particulière pour mes amis du 520 qui pour certains d'entre eux ont été présents depuis le début de cette aventure et qui ont su rendre mes passages au Cléo plus légers.

Aussi, je remercie l'ensemble des doctorants, chercheurs, enseignants, personnels que j'ai côtoyé au sein du laboratoire LSIS. Merci à tous pour tous les bons moments passés ensembles et ils ont été nombreux. Chahinez et Shereen qui grâce à leurs conseils avisés m'ont aiguillée au démarrage de cette thèse. Sans oublier ceux qui sont encore là et qui m'ont également beaucoup apportée Amal, Gaël, Adrian, Aznam ... La liste est longue et ils m'excuseront de ne pas tous les citer.

Une pensée particulière pour Moustapha, la thèse n'est pas un long fleuve tranquille, nous le savions, elle nous a permis de mieux nous connaître et de nous soutenir. Je suis convaincue qu'elle nous a rendus plus forts. Nous avons commencé le chemin ensemble et le terminons ensemble. L'avenir est à nous...

Un remerciement tout particulier et évident à ma maman et ma grand-mère. Ma mère sans qui, malgré les difficultés qu'elle a pu rencontrer ces dernières années, a su m'apporter son amour indéfectible et son soutien inconditionnel. Ma grand-mère qui malgré mes années rebelles avec son lot de bétises a toujours cru en moi et m'a poussée à continuer mes études (et ce n'était pas gagné !). Sans oublier mon frère et ma tante pour qui je peux dire sans hésitation que je n'aurais jamais réussi à en arriver là sans votre amour et votre soutien au cours des 28 dernières années.

A toi aussi Julia, ma meilleure amie, qui t'apprête à assister à l'heure la plus longue de ta vie ! Rien ne peut définir ce que tu m'as apportée tout au long de ces années (et ce bien avant la thèse) bonheur, rire mais aussi des pleures. Je ne remercierai jamais assez le destin d'avoir fait que nos chemins se croisent et tu sais que malgré ma mémoire en mousse tu auras toujours une place dans mon coeur.

A toi Olivier, mon BFF, une personne unique que j'ai eu la chance de croiser au détour d'un couloir à la fac de Lettres et avec qui j'ai partagé des moments indescriptibles (comme coller la bibliographie de son mémoire à la colle le jour de sa soutenance...). Tu as toujours su me soutenir et me pousser et pour cela je ne te remercierai jamais assez.

Je ne peux pas terminer ces remerciements sans citer mes amies Elodie (mon boudin), Mary (sans oublier la petite crevette de Milan) et Audrey qui ont également été présentes tout au long de cette aventure (et bien avant !) et qui m'ont apportée plus qu'un soutien une amitié inconditionnelle, plus les années passent et plus je me rends compte que j'ai la chance de côtoyer ces personnes merveilleuses qui me rendent heureuse pour cela je ne vous remercierai jamais assez.

Je sais que ce n'est pas courant mais cela me tient à coeur. J'ai une dernière pensée pour mes animaux (et oui ! Cela ne risque pas de choquer ceux qui me connaissent). Bien que verbalement ils n'aient pu me formuler leur soutien (quoique... je ne sais pas si je dois considérer les heures à dormir à coté de moi pendant que je rédigeais comme un soutien... mental certainement !) leurs présences au quotidien a su me réconforter et m'apaiser. Je ne risque pas de tous les citer car il y en a vraiment beaucoup mais chacun d'entre eux a participé à cette aventure (en marchant sur mon clavier et en arrachant les fils de mon ordinateur entre autre...) et possède une place toute particulière dans mon coeur.

Enfin, bien que le mot merci ne soit pas suffisant ici pour celui qui est tous les jours à mes côtés, qui me soutient sans faiblir, qui m'aime, que j'aime : Merci Baptiste.

En bref, les mots me manquent pour exprimer à quel point je suis heureuse d'être à vos côtés. Je n'y serai jamais arrivée sans vous. De ce fait, il n'y a qu'un mot (enfin plusieurs...) qui puisse me venir à l'esprit pour exprimer tout ce que je ressens : je vous aime de tout mon coeur !

Résumé

Au cours des dernières années, le Web a connu une énorme croissance en matière de contenus et d'utilisateurs. Ce phénomène a entraîné des problèmes liés à la surcharge d'information face à laquelle les utilisateurs ont des difficultés à trouver les bonnes informations. Des systèmes de recommandation ont été développés pour résoudre ce problème afin de guider les utilisateurs dans ce flux d'informations. Les approches de recommandation se sont multipliées et ont été mises en œuvre avec succès, notamment au travers d'approches telles que le filtrage collaboratif. Cependant, il existe encore des défis et des limites qui offrent des opportunités pour de nouvelles recherches. Parmi ces défis, la conception de systèmes de recommandation de lectures est devenue un axe de recherche en pleine expansion suite à l'apparition des bibliothèques numériques.

Traditionnellement, les bibliothèques jouent un rôle passif dans l'interaction avec les lecteurs et ce, faute d'outils efficaces de recherche et de recommandation. Dans ce manuscrit, nous nous sommes donc penchée sur la création d'un système de recommandation de lectures au travers duquel nous tentons d'exploiter les possibilités du numérique en matière d'accès à l'information scientifique. Nos objectifs portent sur :

- améliorer la **compréhension des besoins utilisateurs** exprimés au sein des requêtes en langage naturel de recherches de livres, articles et billets. Ces travaux nécessiteront de mettre en place des procédés capables d'exploiter la structuration des ouvrages et leur dimension ;
- pallier l'absence de liens explicites entre ouvrages et articles de revues par la détection et l'analyse automatique des références bibliographiques afin de **proposer des liens**. Ce travail nécessitera d'établir des procédés permettant de rendre compte de la grande variété de la composition et de la structuration des références bibliographiques présentes souvent de façon incomplète et sans homogénéité de style, dans le corps du texte et dans les notes ;
- parvenir à un **système de recommandation de lectures** s'appuyant sur des données textuelles permettant de fournir une liste de recommandations personnalisées aux utilisateurs actifs, à l'image des systèmes exploitant des profils utilisateurs.

Mots clés : recherche d'information, système de recommandation de lectures, compréhension des besoins utilisateurs, détection des références bibliographiques, mesures bibliométriques

Abstract

In the recent years, the Web has undergone a tremendous growth regarding both content and users. This has led to an information overload problem in which people are finding it increasingly difficult to locate the right information at the right time. Recommender systems have been developed to address this problem, by guiding users through the big ocean of information. The recommendation approaches have multiplied and have been successfully implemented, particularly through approaches such as collaborative filtering. However, there are still challenges and limitations that offer opportunities for new research. Among these challenges, the design of reading recommendation systems has become a new expanding research focus following the emergence of digital libraries.

Traditionally, libraries play a passive role in interaction with users due to the lack of effective search and recommendation tools. In this manuscript, we will study the creation of a reading recommendation system in which we'll try to exploit the possibilities of digital access to scientific information. Our objectives are :

- to improve the **understanding of user needs** expressed in natural language search queries for books, articles and posts. This work will require the establishment of processes capable of exploiting the structures of data and their dimension ;
- to compensate for the absence of explicit links between books and journal articles by automatically detecting and analyzing bibliographic references, and then to **propose links**. This work will require the establishment of methods to heed the wide variety of bibliographic references, often incomplete and without homogeneity of style, in the body of the text and in the notes ;
- to achieve a **reading recommendation system** based on textual data to provide a customized recommendation list to active users, similar to systems already used by users profiles.

Keywords : information retrieval, reading recommender system, understanding user needs, bibliographical references detection, bibliometric measurements

Publications

Articles de conférences internationales

1. **Anaïs Ollagnier**, Sébastien Fournier, et Patrice Bellot. A supervised Approach for detecting allusive bibliographical references in scholarly publications. In : Proceedings of the 6th International Conference on Web Intelligence, Mining and Semantics (WIMS). ACM Digital Library, 2016. p. 36-39.
2. **Anaïs Ollagnier**, Sébastien Fournier, et Patrice Bellot. Linking Task : Identifying Authors and Book Titles in Verbose Queries. In : CLEF (Working Notes). 2016. p. 1064-1067.
3. Chahinez Benkoussas, Patrice Bellot, et **Anaïs Ollagnier**. The Impact of Linked Documents and Graph Analysis on Information Retrieval Methods for Book Recommendation. In : Web Intelligence and Intelligent Agent Technology (WI-IAT), 2015 IEEE/WIC/ACM International Conference on. IEEE, 2015. p. 385-392.
4. Chahinez Benkoussas, **Anaïs Ollagnier**, et Patrice Bellot. Book Recommendation Using Information Retrieval Methods and Graph Analysis. In : CLEF (Working Notes). 2015, p. 64-71.
5. Chahinez Benkoussas, Hussam Hamdan, Shereen Albitar, et **Anaïs Ollagnier**. Collaborative Filtering for Book Recommandation. In : CLEF (Working Notes). 2014. p. 501-509.

Article de revue nationale

1. **Anaïs Ollagnier**, Sébastien Fournier, et Patrice Bellot. Analyse en dépendance et classification de requêtes en langue naturelle, application à la recommandation de livres. Traitement Automatique des Langues. In : Traitement Automatique des Langues. 2015, vol. 56, no 3. p. 23-47.

Articles de conférences nationales

1. **Anaïs Ollagnier**, Sébastien Fournier, et Patrice Bellot. Cascade de CRFs et SVM pour la détection de références bibliographiques diffuses dans les articles scientifiques. In : COnférence en Recherche d'Information et Applications (CORIA 2016). 2016, p 35-47.
2. **Anaïs Ollagnier**, Sébastien Fournier, Patrice Bellot, et Frédéric Béchet. Impact de la nature et de la taille des corpus d'apprentissage sur les performances dans la détection automatique des entités nommées. In : 21e Conférence en Traitement Automatique des Langues Naturelles (TALN 2014). 2014, p. 511-516.

Table des matières

Liste des figures

Liste des algorithmes

Liste des tableaux

Liste des abréviations

- **BNM** Bayésien Multinomial Naïf
- **CCRF** Cascades de champs aléatoires conditionnels
- **CLEF SBS** *Conference and Labs of the Evalution Forum on Social Book Search*
- **Cléo** Centre pour l'édition électronique ouverte
- **CRF** Champs aléatoires conditionnels
- **DCG** *Directed Graph of Citations*
- **DFR** Déviation par rapport à l'aléatoire
- **FI** *Facteur d'Impact*
- **LT** *LibraryThing*
- **MAP** *Mean Average Precision*
- **MMC** Modèles de Markov Cachés
- **MRR** *Mean Reciprocal Rank*
- **NDCG** *Normalized Discounted Cumulative Gain*
- **RFE** Élimination récursive de caractéristiques
- **SHS** Sciences Humaines et Sociales
- **SVM** Séparateurs à Vaste Marge
- **TAL** Traitement Automatique des Langues

Introduction générale

Le domaine de la recherche d'information (RI) est confronté à une diversification massive des contenus et des usages, le *World Wide Web* contenant une énorme quantité d'informations. Aujourd'hui, plus d'un milliard de sites internet sont dénombrés sur le Web [1]. Les statistiques montrent également que le nombre d'internautes s'accroît et se développe rapidement. En effet, environ 40 % de la population mondiale a aujourd'hui une connexion Internet. Le nombre d'utilisateurs a été multiplié par dix de 1999 à 2013. Le premier milliard a été atteint en 2005. Le deuxième milliard en 2010. Le troisième milliard en 2014 [2].

Si le Web est le domaine où cette surabondance de l'information s'observe le plus nettement, d'autres domaines renferment des quantités importantes de données comme les bases de données cinématographiques et musicales. Nous pouvons prendre en exemple les bases de données cinématographiques réalisées par le groupe de recherche GroupLens (RESNICK, IACOVOU et al. 1994) et la compagnie de location de films Netflix (BENNETT, LANNING et al. 2007) qui proposent respectivement environ 15 000 et 130 000 titres.

Ce foisonnement de l'information mais également des interactions des usagers a conduit à des problèmes liés à la surcharge d'information face à laquelle les utilisateurs ont des difficultés à localiser les bonnes données au bon moment (BAEZA-YATES, RIBEIRO-NETO et al. 1999). Face à la diversité des sources d'une part, et des besoins d'autre part, l'objectif principal de tout système de RI est de maximiser le degré de satisfaction de certaines conditions objectives et subjectives, généralement – mais pas uniquement – de la satisfaction des utilisateurs. La recherche et le développement de la RI ont tourné autour de la définition des modèles et d'algorithmes permettant d'atteindre au mieux cet objectif (BRANTS 2003 ; LIOMA et BLANCO 2009 ; MORAL, ANTONIO et al. 2014). De nombreuses recherches ont été réalisées dans le but de fournir aux utilisateurs des services d'information plus proactifs et personnalisés notamment au travers de la création de systèmes de recommandation (RESNICK et VARIAN 1997 ; RICCI, ROKACH et al. 2011 ; GOMEZ-URIBE et HUNT 2016). En effet, par le biais de ces systèmes conçus pour aider les utilisateurs dans leurs tâches d'accès et de récupération d'information, des items leurs sont suggérés en fonction de leurs intérêts.

Dans le milieu universitaire, ces systèmes ont fait l'objet de nombreuses études depuis les années 60. Ce sont des travaux initiés dans les sciences cognitives à la fin des années 1970, par Elaine Rich (RICH 1979), qui sont à l'origine des pre-

1. http://www.internetlivestats.com/total-number-of-websites/
2. http://www.internetlivestats.com/internet-users/

miers questionnements sur la personnalisation des interactions homme-machine. Au cours de ces travaux, des stéréotypes ont été élaborés afin de construire des modèles utilisateurs permettant d'améliorer la pertinence des recommandations effectuées automatiquement. Plus tard, certains travaux dans les domaines de la récupération de l'information et de la théorie des prévisions ont été réalisés et sont considérés aujourd'hui comme le début de la recommandation en tant que domaine de recherche (SALTON 1989 ; ARMSTRONG 2001). De nos jours, des sociétés savantes telles que ACM [3] (*Association for Computing Machinery*) ou encore des campagnes d'évaluation telles que CLEF SBS [4] (*Conference and Labs of the Evaluation Forum on Social Book Search*) organisent des ateliers entièrement dédiés à cette problématique.

Au cours des dernières années, les approches de recommandation se sont multipliées et ont été mises en œuvre avec succès, notamment au travers d'approches telles que le filtrage collaboratif (KONSTAN, MILLER et al. 1997 ; HERLOCKER, KONSTAN et al. 2004 ; SCHOLZ, FORMAN et al. 2016). Cependant, il existe encore des défis et des limites qui offrent des opportunités pour de nouvelles recherches. Parmi ces défis, la conception de systèmes de recommandation de lectures est devenu un axe de recherche en pleine expansion suite à l'apparition des bibliothèques numériques. En effet, au cours des seize dernières années, plus de 200 articles de recherche ont été publiés sur les systèmes de recommandation dédiés, et plus particulièrement, à la suggestion d'articles scientifiques (BEEL, LANGER et al. 2016).

Traditionnellement, les bibliothèques électroniques jouent un rôle passif dans l'interaction avec les utilisateurs et ce, faute d'outils efficaces de recherche et de recommandation. Avec les progrès de la technologie numérique et le développement du Web, les bibliothèques ont évolué afin de répondre aux demandes diversifiées des usagers et dans l'optique d'offrir des services plus personnalisés. Les bibliothèques numériques sont de plus en plus utilisées par différents utilisateurs à des fins diverses. Le partage et la collaboration sont devenus, par ailleurs, des éléments sociaux importants. Au fur et à mesure que les bibliothèques numériques deviennent monnaie courante, les utilisateurs attendent des services plus performants (DI GIACOMO, MAHONEY et al. 2001). Une fonction de recherche traditionnelle fait normalement partie intégrante de toute bibliothèque numérique, or avec un nombre de publications en perpétuelle augmentation et les nouvelles technologies qui ont permis de numériser rapidement de nombreux documents plus anciens, les besoins des utilisateurs sont devenus plus complexes. Ainsi, les bibliothèques numériques doivent passer d'un rôle passif à actif dans l'offre et l'adaptation d'informations aux utilisateurs afin de proposer des services permettant de capturer, de structurer et de partager les connaissances (J.

3. https://recsys.acm.org/
4. http://social-book-search.humanities.uva.nl/

Lu, Dianshuang Wu et al. 2015).

Contexte de recherche et motivation

Cette thèse s'inscrit dans le cadre de l'équipement d'excellence DILOH [5] (*Digital Library of Open Humanities*) dont l'objectif principal est de parvenir à l'élaboration d'une **bibliothèque internationale pour l'édition électronique en libre accès et les humanités numériques**. Ce projet, porté par l'Université d'Aix-Marseille en lien avec le CNRS, l'École des hautes études en sciences sociales (EHESS) et l'Université d'Avignon et des Pays du Vaucluse, est réalisé par le Centre pour l'édition électronique ouverte (Cléo), en partenariat avec le Centre pour la Communication Scientifique Directe (CCSD), le Roy Rosenzweig Center for History and New Media (CHNM), la fondation Open Access Publishing in European Networks (OAPEN) et le Laboratoire des Sciences de l'Information et des Systèmes (LSIS).

Le projet DILOH part du constat que les usages des technologies numériques au sein des sciences humaines et sociales (SHS) sont inégaux. En effet, les pratiques éditoriales dans ce domaine restent marquées par le paradigme de l'imprimé et ne sont pas encore entrées dans un cycle d'innovation leur permettant de tirer parti des nouvelles possibilités qu'offrent les technologies numériques. Les usages des outils du Web 2.0 comme les signets partagés, les blogs et les réseaux sociaux y sont encore très peu répandus. *Via* ce projet, il s'agit de décloisonner un secteur scientifique qui pâtit d'une offre documentaire encore limitée par les contraintes économiques de l'imprimé, les restrictions technologiques et financières d'accès aux contenus, la dispersion des éditeurs et la diversité linguistique des écrits.

Suite à ce constat, le Cléo qui développe et anime le site Openedition.org [6], dont l'une des plateformes Revues.org [7] est le plus ancien portail de revues en SHS en France et qui diffuse près de 300 revues du secteur, sert de base au projet DILOH. L'objectif est de permettre aux plateformes de diffusion du Cléo, dont Revues, mais aussi Calenda [8], Hypotheses [9] et Books [10] de parvenir aux objectifs suivants :

- réunir une masse critique de documents sélectionnés dans les catalogues d'éditeurs de premier plan au niveau international ;

5. http://www.agence-nationale-recherche.fr/?ProjetIA=11-EQPX-0013
6. http://www.openedition.org/
7. http://www.revues.org/
8. http://calenda.org
9. http://hypotheses.org/
10. http://books.openedition.org/

- faire de cette collection une bibliothèque numérique de nouvelle génération, en offrant à ses utilisateurs une palette d'outils innovants exploitant les possibilités du numérique (nouveaux modes de mise en forme et de diffusion de la connaissance, fonctionnalités bibliographiques poussées, système avancé de recommandation, etc.).

Dans ce contexte, nos motivations découlent du fait que les moteurs de recherche du Web et des bibliothèques en ligne sont déclinés en différentes versions ciblant de façon privilégiée la recherche de pages Web, de micro-blogs, d'informations ou d'ouvrages (documents). Mais aucun système de recherche ne répond de façon satisfaisante au besoin d'une navigation personnalisée et contextuelle au sein d'un bibliothèque numérique regroupant un ensemble documentaire aussi varié que celui de DILOH. Le *continuum* itératif de la production scientifique (échanges entre chercheurs sous forme de billets, annonce de colloques et appels à publications, articles de revues, livres) engendre un besoin de ventilation et de croisement spécifique des éléments trouvés en réponse à une requête, des objets informationnels classés par des méthodes automatiques et des liens qu'un système de recommandation peut proposer.

La personnalisation est nécessaire pour rendre accessible un ensemble documentaire aussi varié à une population d'utilisateurs de plus en plus hétérogène. Les bibliothèques numériques se doivent d'adapter leurs services et leurs matériaux à une large gamme d'utilisateurs et ce, afin d'augmenter leur impact et leur utilité. La prochaine génération de bibliothèques numériques doit supporter une large gamme de services personnalisés qui prennent en charge les activités d'un important éventail d'utilisateurs. La mise en place d'un système de recommandation de lectures peut être un premier pas vers ces objectifs en proposant des recommandations adaptées aux goûts, aux besoins ou aux moyens des utilisateurs, afin de les aider à accéder à des ressources utiles ou intéressantes.

Les travaux scientifiques menés au cours de cette thèse s'articuleront donc autour de la création d'un système de recommandation de lectures au travers duquel nous tenterons d'exploiter les possibilités du numérique en matière d'accès à l'information scientifique. Nos objectifs porteront sur :

- améliorer la **compréhension des besoins utilisateurs** exprimés au sein des requêtes en langage naturel de recherches de livres, articles et billets. Ces travaux nécessiteront de mettre en place des procédés capables d'exploiter la structuration des ouvrages et leur dimension ;
- pallier l'absence de liens explicites entre ouvrages et articles de revues par la détection et l'analyse automatique des références bibliographiques afin de **proposer des liens**. Ce travail nécessitera d'établir des procédés permettant de rendre compte de la grande variété de la composition et de la

structuration des références bibliographiques présentes (souvent de façon incomplète et sans homogénéité de style) dans le corps du texte et dans les notes ;

- parvenir à un **système de recommandation de lectures** s'appuyant sur des données textuelles permettant de fournir une liste de recommandations personnalisées aux utilisateurs actifs, à l'image des systèmes exploitant des profils utilisateurs.

Nous proposons d'initier un dispositif innovant non seulement par l'exploitation d'un très riche corpus en SHS associant contenus et méta-données normalisées, mais aussi et surtout par le développement de services incluant une amélioration de la compréhension des besoins utilisateurs, la génération automatique de liens ainsi qu'un système de recommandation s'appuyant sur des données textuelles et des liens entre les documents.

Démarche proposée

Suite à ces constats, nous proposons plusieurs axes de recherche permettant de répondre aux différents enjeux relatifs à : l'amélioration de la compréhension des besoins utilisateurs exprimés au sein des requêtes (i), l'identification des références bibliographiques allusives (ii) et l'élaboration d'un système de recommandation de lectures basé sur des données textuelles (iii).

i La question du but sous-jacent d'une requête est essentielle à la satisfaction du besoin d'information de l'utilisateur. Beaucoup de travaux ont porté sur la compréhension du comportement de recherche des utilisateurs sur le Web en se focalisant plus particulièrement sur leur façon d'effectuer des recherches, et ce, sans finalement connaître leur intention. Or, connaître le « pourquoi » du comportement de recherche de l'utilisateur est essentiel pour satisfaire ses besoins d'information. C'est dans ce contexte que sont apparues les premières typologies des requêtes orientées vers le but des utilisateurs, avec l'hypothèse que les requêtes de même type peuvent être traitées de manière semblable par les moteurs de recherche, menant au développement d'algorithmes et d'interfaces dédiés à l'optimisation d'une recherche selon un type de but (Broder 2002 ; Rose et Levinson 2004). Dans le cadre de l'exploitation de requêtes issues d'une bibliothèque numérique d'articles scientifiques, les taxonomies existantes se sont avérées difficilement utilisables. En effet, nos travaux sont orientés sur des requêtes très spécifiques dont l'intention est purement informationnelle, c'est-à-dire que l'utilisateur exprime un intérêt pour obtenir une information qui, dans notre cas, est relative à la recherche de livres. Inspirée par ces recherches, nous avons donc transposé ces mêmes réflexions dans l'étude des requêtes

de recherche de livres et nous avons élaboré une taxonomie propre au traitement des requêtes de recherche de livres.
Dans ce contexte, nous avons pu observer la présence de verbosité au sein de ce type de requêtes. Au travers de cette verbosité les utilisateurs expriment des informations plus complexes sur leurs besoins. Chaque requête fournit un contexte particulier établi par l'expression des goûts et des centres d'intérêt de l'utilisateur. De plus, ce type de requête définit l'intention qui se cache derrière la demande de l'utilisateur. Le traitement des requêtes exprimées en langage naturel, qualifiées de requêtes verbeuses, a connu un essor considérable. En effet, parvenir à extraire des informations pertinentes de ces requêtes est devenu un enjeu majeur au sein de plusieurs domaines de recherche. Les constations que nous avons pu faire suite à l'étude de la littérature sur le sujet montrent que la majorité de ces travaux met de côté la sémantique de la phrase en exploitant uniquement les mots de façon isolée. Or, dans le cadre de nos recherches nous souhaitons parvenir à interpréter au mieux les besoins d'information exprimés au sein des requêtes par les utilisateurs. Nous proposons donc une approche dans laquelle nous préservons la structure de la phrase via l'utilisation d'un procédé issu du traitement automatique des langues (TAL) : un analyseur en dépendance. Bien que d'autres techniques issues de la linguistique existent comme l'exploitation des rôles sémantiques, nous avons choisi, dans cette thèse, d'explorer uniquement les liens syntaxico-sémantiques. L'exploitation de ces liens nous permet de générer des bigrammes de mots syntaxiquement dépendants dont l'objectif est de préserver au mieux la sémantique de la phrase.

ii Des études portées sur l'écrit scientifique ont permis d'identifier plusieurs types de références (WOUTERS et al. 1999). Parmi elles, certaines sont des références explicites à l'image des références que nous pouvons trouver à la fin des articles ou des livres, tandis que d'autres références, que nous qualifions d'allusives, sont disséminées dans le corps du texte. Ce type de références bibliographiques présentes, souvent de façon incomplète et sans homogénéité de style, dans le corps du texte et dans les notes nécessite des procédés particuliers afin de permettre leur identification et leur extraction. En effet, de nombreux travaux ont porté sur l'identification des références bibliographiques structurées à l'image de celles présentes à la fin des documents (DECONINCK 2010 ; ANZAROOT et MCCALLUM 2013 ; REZAEI et MUNTZ 2013), et ont par ailleurs obtenu des résultats plus que satisfaisants, mais très peu se sont penchés sur l'identification des références disséminées dans le texte. De ce fait, nous nous sommes particulièrement intéressée à ce type de références et nous proposons une approche associant des procédés issus de modèles statistiques dédiés à la classification supervisée et à la reconnaissance de patrons afin de permettre leur détection. En plus de proposer cette méthode, nous savons que les références bibliographiques

sont composées d'éléments permettant l'identification d'un document en tant qu'unité documentaire. Parvenir à exploiter ces éléments peut nous permettre de faire émerger des liens entre des documents thématiquement liés. De ce fait, nous avons étudié les possibles exploitations des références bibliographiques comme outil de liaison entre contenus au travers de la modélisation d'un graphe orienté.

iii Le *continuum* itératif de la production scientifique engendre un besoin de ventilation et de croisement spécifique. Suite à ce constat, nous nous sommes penchée sur la conception d'un système de recommandation de lectures. Notre objectif est de fournir aux utilisateurs des suggestions de lectures directement extraites des thématiques les intéressant. Pour ce faire, nous proposons une approche reprenant les fondements des systèmes de filtrage basés sur le contenu, à savoir, proposer des items relatifs aux thèmes abordés dans les documents par rapport aux thèmes intéressant l'utilisateur. La différence notoire du système que nous proposons, comparativement aux systèmes de filtrage basés sur le contenu, se situe au niveau de l'exploitation du profil de l'utilisateur actif. En effet, au sein de ce système nous n'effectuons pas de comparaison entre des items et un profil utilisateur afin d'établir une recommandation. Nous proposons de déterminer les intérêts de l'utilisateur actif, normalement issus du profil de l'utilisateur, à partir de l'analyse du contenu de l'article sélectionné par ce dernier. Afin d'extraire des thématiques connexes à ce document, et donc relatives aux intérêts de l'utilisateur, nous proposons d'établir un système basé sur des mesures bibliométriques. Nous proposons, *via* la génération d'un indicateur d'impact propre à chaque référence bibliographique, une liste de recommandations de lectures ordonnée selon leurs impacts sur le document ciblé. Par le biais de cette approche, nous suggérons uniquement des documents cités au sein du document ciblé. De ce fait, nous obtiendrons des documents que nous supposons thématiquement liés au document courant. De plus, par le biais de cette approche nous souhaitons également illustrer les interactions bénéfiques pouvant ressortir de l'exploitation de procédés issus de la RI et de la bibliométrie dont les relations sont considérées à certains égards comme symbiotiques (WOLFRAM 2015).

Plan du mémoire

Cette thèse est structurée en trois chapitres principaux : **Classification et enrichissement de la représentation des requêtes verbeuses par analyse en dépendance** (chapitre 1), **Vers une liaison entre contenus *via* l'identification automatique de références bibliographiques** (chapitre 2), **Exploitation des références allusives comme indicateur d'impact au sein d'un système de recommandation** (chapitre 3). Au travers de ces chapitres, nous rendrons compte

plus en détail des recherches que nous avons menées autour de la problématique des systèmes de recommandation de lectures.

Le chapitre 1 s'intéresse à l'enrichissement de la représentation des requêtes et plus particulièrement des requêtes verbeuses. Au sein de ce chapitre, nous proposerons une approche orientée sur l'amélioration de la compréhension des besoins utilisateurs au sein des requêtes de recherche de livres. En particulier, nous nous focaliserons sur des requêtes dans lesquelles l'utilisateur recherche des similitudes avec d'autres livres (auteurs ou encore ensemble d'ouvrages présentant une unité que l'on retrouve sous la forme de collection) que ceux énoncés dans sa requête. La section 1.2 présentera une synthèse des travaux de l'état de l'art relative aux traitements des requêtes verbeuses en RI et sur les avancées scientifiques concernant la compréhension des besoins utilisateurs au sein des requêtes. La section 1.3 proposera une analyse des caractéristiques compositionnelles et structurelles des requêtes verbeuses de recherche de livres que nous avons menée grâce aux données issues des campagnes CLEF SBS [11] (KOOLEN, BOGERS et al. 2016) menées par CLEF Initiative [12]. La section 1.4 détaillera l'approche fondée vers la génération d'un modèle de recherche de livres s'appuyant sur la compréhension des besoins utilisateurs formulés au sein de requêtes analogues. Dans la section 1.5, nous exposerons les résultats obtenus, à la fois, sur la détection des requêtes analogues par une approche de classification automatique supervisée et sur les différentes stratégies d'expansion de ces requêtes. Enfin, ce chapitre se conclura avec une discussion et des conclusions tirées de l'analyse des résultats prélimaires obtenus suite à l'application de l'approche présentée précédemment ainsi que des perspectives de recherche.

Le chapitre 2 s'intéresse à un autre aspect inhérent aux systèmes de recommandation : la représentation des documents. Nous proposerons, dans le cadre de l'utilisation de données issues d'une bibliothèque numérique d'articles scientifiques, d'exploiter les références bibliographiques comme source de liens entre les documents. Notre but ne sera pas d'améliorer la compréhension des documents mais bien d'exploiter les références bibliographiques présentes afin de mettre en perspective de nouveaux liens thématiquement liés au document courant. Afin de parvenir à cet objectif, nous introduirons au cours de la section 2.2 une synthèse des travaux de l'état de l'art se focalisant sur la notion de « référence » en TAL et sur les approches dédiées à la détection et l'analyse de références bibliograhiques. La section 2.3 proposera, d'une part, différentes études menées sur les facteurs influançant la composition ainsi que la structuration des références présentes au sein d'articles scientifiques, tous domaines confondus et d'autre part, nous présenterons les caractéristiques propres aux références allusives que nous avons réussi à dégager suite à leur analyse. Dans la section 2.4,

11. http://social-book-search.humanities.uva.nl/#/overview
12. http://www.clef-initiative.eu/

nous exposerons l'approche que nous avons mis en place dédiée à l'identification automatique des références allusives ainsi que son affiliation au développement du logiciel Bilbo. La section 2.5 introduira les données fournies par OpenEdition ainsi que les résultats obtenus suite à l'application de notre approche. Dans la section 2.6, nous exposerons les résultats obtenus suite à notre participation à la campagne CLEF SBS 2016. La section 2.7 présentera la modélisation, sous la forme d'un graphe orienté, que nous avons établie à partir des références bibliographiques. Enfin, les discussions et conclusions seront présentées dans la section 2.8.

Le chapitre 3 s'intéresse à l'exploitation des références bibliographiques allusives dans le cadre de la réalisation d'un système de recommandation basé sur des mesures bibliométriques. Pour ce faire nous proposerons, en premier lieu, une méthode destinée à la construction d'un indicateur d'impact à partir des références bibliographiques allusives. En second lieu, nous proposerons d'intégrer cet indicateur d'impact dans le cadre de la réalisation d'un système de recommandation de lectures. Afin de parvenir à ces objectifs, nous présenterons aux cours de la section 3.2 un état de l'art synthétisant les travaux réalisés sur les domaines de la bibliométrie et des systèmes de recommandation. Nous nous pencherons également sur les contributions mutuelles relevées entre la RI et la bibliométrie. Dans la section 3.3, nous exposerons en détail les facteurs d'impact permettant d'établir nos indicateurs d'impact. La section 3.4 introduira les processus d'intégration des indicateurs d'impact dans le cadre de la réalisation d'un système de recommandation de lectures. Dans la section 3.5, nous exposerons la méthode mise en place afin d'évaluer ce système de recommandation de lectures. La section 3.6 présentera les résultats obtenus suite à nos expérimentations ainsi qu'une analyse des documents suggérés par le système de recommandation de lectures. Ce chapitre se conclura sur une discussion et des perspectives de recherche initiées par ces travaux.

En conclusion, nous présenterons un résumé des recherches réalisées au cours de cette thèse. Enfin, nous présenterons nos futurs travaux au travers de perspectives envisageables à court, moyen et long terme.

1. Classification et enrichissement de la représentation des requêtes verbeuses par analyse en dépendance

1.1. Introduction

Au cours des dernières années, les interactions des usagers avec les services Web ont évolué passant de simples mots-clés à des requêtes de plus en plus « sophistiquées et spécifiques » (M. GUPTA et BENDERSKY 2015). En effet, au travers d'applications comme les moteurs de recherche, l'utilisation uniquement de mots-clés ne permet pas toujours de véhiculer les besoins informationnels des utilisateurs de ce fait, ils ont recours à des requêtes plus longues et détaillées. Comme le démontre cette requête extraite des données fournies lors des campagnes d'évaluation TREC (*Text REtrieval Conference* [1]) :

« Provide information on all kinds of material international support provided to either side in the Spanish Civil War »

De nombreuses études ont révélé que ces requêtes, qualifiées de verbeuses, représentent une partie significative du flux de requêtes au sein des recherches effectuées sur le Web mais aussi dans d'autres applications comme les systèmes de questions-réponses, les systèmes de dialogue ou encore les systèmes de reconnaissance vocale (HUSTON et CROFT 2010 ; DI BUCCIO, MELUCCI et al. 2014). De ce fait, afin de s'orienter vers des systèmes de plus en plus performants, le traitement des requêtes verbeuses a connu un essor considérable. Plus récemment, des travaux portant sur l'exploitation des informations issues des médias sociaux comme les forums en ligne, dont la croissance exponentielle depuis plusieurs années suscite un vif intérêt, s'intéresse également aux traitements de ces requêtes (CHAA, BELLOT et al. 2017).

La complexité du sujet abordé ou encore l'expression de l'intention de l'utilisateur peuvent être des causes de verbosité. En effet, au travers de ce type de requêtes les utilisateurs expriment des besoins d'information complexes ou hautement spécifiques. Des études montrent qu'il est souvent difficile pour les utilisateurs d'exprimer un besoin spécifique d'information au sein d'un moteur

1. http://trec.nist.gov/

de recherche (BISKRI et ROMPRE 2012) ce qui explique entre autres l'émergence des médias sociaux qui permettent aux utilisateurs d'intérargir et de dépasser les limites d'une interaction homme-machine. Les recherches portant sur le traitement des requêtes verbeuses ont été largement explorées *via* la mise en place de procédés tels que la réduction, la pondération ou encore l'expansion afin d'obtenir une représentation plus efficace de ce type de requêtes. Cependant, la majorité de ces travaux mettent de côté la sémantique de la phrase en exploitant uniquement les mots de façon isolée (sous la forme d'approches « *sacs de mots* »).

Au cours de ce chapitre, nous présentons une approche orientée sur l'amélioration de la compréhension des besoins utilisateurs au sein des requêtes verbeuses de recherche de livres. En particulier, nous nous focalisons sur des requêtes dans lesquelles l'utilisateur recherche des similitudes avec d'autres livres, auteurs ou collections, que ceux énoncés dans sa requête. Par commodité, nous nommerons ce type de requêtes des requêtes *analogues* tout au long de ce chapitre en raison de leur signifié connotatif qui renvoie à la recherche de caractéristiques semblables entre des choses. Nous proposons un cadre applicatif orienté sur la recommendation de lectures *via* l'exploitation des données fournies lors des campagnes CLEF SBS [2] (KOOLEN, BOGERS et al. 2016) menées par CLEF Initiative [3]. Ces données extraites du forum *LibraryThing* [4] (LT) se présentent sous la forme de requêtes longues et détaillées exprimant finement le besoin informationnel de l'utilisateur. Notre objectif est de parvenir à une meilleure compréhension des besoins informationnels exprimés par les utilisateurs *via* l'introduction de procédés issus du TAL.

Notre contribution porte sur une meilleure préservation de la sémantique de la phrase *via* l'utilisation d'un analyseur en dépendance appliqué au traitement des requêtes analogues. Bien que d'autres procédés issus de la linguistique existent comme les rôles sémantiques, nous avons choisi, dans cette thèse, d'explorer uniquement les liens syntaxico-sémantiques. L'exploitation de ces liens permettant de générer des bigrammes de mots syntaxiquement dépendants, nous supposons ainsi préserver au mieux la sémantique de la phrase.

Ce chapitre est structuré comme suit : dans la section 1.2 nous effectuons un état de l'art se focalisant sur les approches relatives aux traitements des requêtes verbeuses en RI et sur les avancées scientifiques concernant la compréhension des besoins utilisateurs au sein des requêtes. La section 1.3 présente une analyse des requêtes de recherche de livres. La section 1.4 détaille l'approche fondée vers la génération d'un modèle de recherche de livres s'appuyant sur la compréhension des besoins utilisateurs formulés au sein de requêtes analogues. Dans

2. http://social-book-search.humanities.uva.nl/#/overview
3. http://www.clef-initiative.eu/
4. https://www.librarything.com/

la section 1.5 nous exposons les résultats obtenus, à la fois, sur la détection des requêtes analogues par une approche de classification automatique supervisée et sur les différentes stratégies d'expansion de ces requêtes. Enfin, les conclusions sont présentées dans la section 1.6.

1.2. État de l'art

Dans cette section, nous présentons un état des lieux sur deux domaines connexes aux travaux de recherche présentés dans ce chapitre, à savoir, les approches relatives aux traitements des requêtes verbeuses en RI et les avancées scientifiques concernant la compréhension des besoins utilisateurs au sein des requêtes.

1.2.1. Le traitement des requêtes verbeuses en recherche d'information

Depuis plusieurs années, l'objet de nouvelles applications de recherche est passé de requêtes constituées de mots-clés à des requêtes verbeuses exprimées en langage naturel. Les exemples incluent les systèmes de questions-réponses, les systèmes de dialogue, la reconnaissance vocale et les moteurs de recherche d'entités comme le graphe de recherche de Facebook ou le graphe de connaissance de Google [5]. Depuis plusieurs années, nous constatons un intérêt grandissant afin de trouver des traitements efficaces pour ce type de requêtes. Dans un premier temps, nous définissons ce qu'est une requête verbeuse ainsi que ses propriétés. Dans un second temps, nous présentons les différentes approches dédiées aux traitements de ce type de requête. Dans un troisième temps, nous étudions leur exploitation au sein des modèles de recherche de livres.

1.2.1.1. Les propriétés

Une première difficulté se pose quant à la définition d'une requête verbeuse, en effet, aucun concensus n'est clairement établi permettant de distinguer une requête longue d'une requête verbeuse. Dans la littérature, ces deux appellations ont souvent la même signification, la première étant essentiellement caractérisée par le nombre de mots-clés et la seconde par son aspect détaillé. Cependant, l'emploi du qualificatif « verbeuse » ne signifie pas en réalité qu'une requête doit nécessairement être longue pour être verbeuse, elle peut être succincte et une requête courte peut être verbeuse (Di Buccio, Melucci et al. 2014). En effet, nous pouvons avoir des requêtes détaillées composées de mots-clés sans structure apparente et inversement. Afin de lever toute ambiguïtée, nous qualifions les requêtes présentées dans cette thèse comme des requêtes verbeuses car nous

5. https://www.google.com/intl/fr_fr/insidesearch/features/search/knowledge.html

exploitons des requêtes exprimées en langage naturel qui sont longues, au niveau de leur densité textuelle et détaillées, au niveau de l'expression des besoins de l'utilisateur.

Nous avons pu constater que les requêtes verbeuses sont produites dans de multiples domaines, la complexité du sujet abordé ou encore l'expression de l'intention de l'utilisateur peuvent être des causes de verbosité. Nous avons pu noter que certaines de ses caractéristiques, comme sa longueur, fluctuent selon le domaine de recherche.

Dans le domaine des moteurs de recherche plusieurs estimations ont été faites. Par exemple, Yahoo ! qui en 2006 établit que 17% des requêtes contiennent cinq mots et plus [6]. (KUMARAN et CARVALHO 2009) observe également, sur la collection TREC123 [7], que plus de 15% du volume total des requêtes varie de trois à six mots. Dans le domaine des assistants vocaux, une comparaison intéressante a été faite sur les logs du moteur de recherche de Yahoo !. Cette étude montre que les requêtes écrites ont une longueur moyenne de 9,54 mots et de 7,48 mots sans les mots outils, tandis que les requêtes formulées à l'oral ont une moyenne de 23,07 mots et de 14,33 mots sans les mots outils (YI et MAGHOUL 2011). L'étude des échanges sur les médias sociaux, comme les forums en ligne, révèle une forte utilisation des requêtes verbeuses. L'étude que nous avons menée sur les données fournies lors de la campagne CLEF SBS 2016 [8] nous a permis d'estimer la longueur moyenne des requêtes extraites du forum LibraryThing [9] (LT) à 64,53 mots. Bien évidemment dans ce cas particulier, nous ne sommes pas face à des interactions homme-machine ce qui explique la taille importante des requêtes. En effet, dans ce contexte précis, les utilisateurs échangent entre eux ce qui tend vers des interactions plus détaillées.

Concernant les caractéristiques compositionnelles (BENDERSKY et CROFT 2009), suite à l'analyse des logs de recherche de MSN (environ 15 millions), les auteurs ont observé que 7,5% des requêtes qualifiées comme étant verbeuses (définies comme verbeuses de par leurs longueurs dans l'article) sont composées d'adverbes et pronoms interrogatifs, 5,5% utilisent des opérateurs booléens, 64% sont composées de fragments de requêtes courtes, 14,7% utilisent des noms et 8,3% emploient des phrases avec des verbes. L'étude que nous avons menée sur les requêtes extraites de LT nous a permis d'établir les classes grammaticales les plus fréquentes sur l'intégralité du corpus : les noms avec taux de représentation de 15%, les verbes avec 14,5% et les prépositions et/ou conjonction de subordination avec 9,8%. Dans le domaine des assistants vocaux, nous ne possédons pas

6. http://www.zdnet.com/article/yahoo-searches-more-sophisticated-and-specific/
7. http://trec.nist.gov/data/test_coll.html
8. http://social-book-search.humanities.uva.nl/#/overview
9. https://www.librarything.com/

de telles études cependant nous pouvons supposer, au vu de la longueur plus importante des requêtes vocales, que les utilisateurs ont tendance à se rapprocher d'une élocution que nous pouvons retrouver lors d'interactions humaines avec beaucoup de mots de fonction.

Dans les sous-sections suivantes, nous présentons, premièrement, différentes approches dédiées à la transformation des requêtes dont les applications ont été constatées sur les requêtes verbeuses. Secondement, nous effectuons un tour d'horizon des différents modèles de RI ainsi que des campagnes d'évaluation dédiées à l'expertise de ces modèles.

1.2.1.2. Les approches

Plusieurs approches ont été proposées pour le traitement des requêtes verbeuses essentiellement basées sur la transformation de la requête comme la réduction, la pondération, l'expansion, la reformulation et la segmentation. Nous allons brièvement présenter chacune de ces transformations.

La réduction de requête consiste à réduire la longueur de la requête initiale afin d'obtenir une ou plusieurs sous-requêtes. Dans les approches dédiées à la génération d'une seule sous-requête plusieurs types de traits sont utilisés comme les traits statistiques (KUMARAN et CARVALHO 2009), les traits linguistiques (BALASUBRAMANIAN, KUMARAN et al. 2010) et plus récemment des traits basés sur les logs de requêtes (B. YANG, PARIKH et al. 2014). Quant à la sélection de la meilleure sous-requête, les approches se basent généralement sur des modèles simples de classification ou de régression. Cependant, des travaux ont également démontré leur efficacité par le biais d'approches basées sur les marches aléatoires (« *random walks* ») (MAXWELL et CROFT 2013) ou encore *via* les structures d'arrêt (« *stop structures* » [10]) (HUSTON et CROFT 2010). Des approches se sont également intéressées à la génération de plusieurs sous-requêtes permettant, à la différence des approches précédentes, de saisir les dépendances entre les mots et les phrases et de considérer non seulement la meilleure requête reformulée, mais aussi d'autres options. Le principe de ce type d'approche est de reduire la requête initiale en plusieurs sous-requêtes et d'affecter un poids à chacune. Trois principales méthodes sont utilisées : CRF-perf, ListNet et les arbres de reformulation (« *Reformulation Trees* »). Les arbres de reformulation et ListNet ont montré une plus grande efficacité que CRF-perf (XUE et CROFT 2011 ; XUE et CROFT 2012).

La pondération des mots ou des concepts d'une requête consiste à pondérer chaque mot ou concept de la requête. Ce type d'approche fournit des mécanismes

10. une structure d'arrêt se définit comme une phrase qui ne fournit aucune information sur les besoins d'information de l'utilisateur. (CALLAN et CROFT 1993)

puissants et flexibles pour traiter les requêtes verbeuses. Traditionnellement, ce sont des modèles « sacs de mots » qui sont utilisés comme BM25 ou la déviation par rapport à l'aléatoire (« *Divergence From Randomness* »). Avec le temps, les recherches se sont de plus en plus orientées vers des méthodes basées sur une pondération supervisée des termes comme la régression de rang (« *Regression rank* ») ou la nécessité de termes (« *Term necessity* »). Les dernières avancées dans ce domaine proposent des méthodes comme l'utilisation d'hypergraphes de requêtes afin de modéliser les dépendances entre les concepts d'une même requête (BENDERSKY et CROFT 2012) ou encore l'utilisation de la notion de centralité que l'on retrouve généralement lors des tâches de résumé de textes (PAIK et OARD 2014).

L'expansion de requête, généralement utilisée sur des requêtes courtes, a démontré son efficacité sur les requêtes verbeuses (SYMONDS, BRUZA et al. 2014). Ce type de transformation peut être manuel, automatique ou interactif. Pour chacune de ces expansions, plusieurs sources sont nécessaires. Il en existe au moins deux types : les résultats de recherche et les bases de connaissances lexicales. Dans le cas de bases de connaissances lexicales dépendantes des données, nous avons affaire à des algorithmes de modification des mots (suppression de suffixes, recherche de similarité entre les mots, des regroupements, etc.). Dans le cas de bases de connaissances lexicales indépendantes, ce sont des thésaurus ou dictionnaires/lexiques spécifiques au domaine qui sont employés. Parmi les techniques d'expansion observées sur les requêtes verbeuses, nous pouvons citer : l'ajout de labels ODP (*Open Directory Project*) (BAILEY, R. WHITE et al. 2010), l'ajout de concepts latents pondérés (BENDERSKY, METZLER et al. 2011) et l'expansion interactive de type pseudo-retour de pertinence (KUMARAN et ALLAN 2008).

La reformulation de requête est un moyen utile de traiter les requêtes verbeuses surtout quand il existe un décalage entre la requête originale et le vocabulaire du corpus. En effet, si nous prenons l'exemple présenté dans l'article de (X. WANG et ZHAI 2008), un utilisateur peut émettre une requête du type « *auto wash* » ou « *vehicle wash* », qui ne fournit généralement pas de bons résultats auprès d'un moteur de recherche. Or, par le biais de méthodes permettant la reformulation de requête de nouvelles requêtes sont proposées, telle que « *car wash* » plus couramment utilisée au sein des pages web. Cinq approches sont principalement utilisées : les probabilités de traduction entre la requête de l'utilisateur et les paires questions-réponses disponibles dans les logs (XUE, JEON et al. 2008), l'utilisation des marches aléatoires sur les termes de la requête (SHELDON, SHOKOUHI et al. 2011), les graphes d'URL de requête (« *query-URL graphs* »), les logs de requête (XUE, TAO et al. 2012) et les ancres de lien (DANG et CROFT 2010). Ces différentes approches ont permis de démontrer leur efficacité à réduire l'impact négatif de la différence de vocabulaire entre requêtes et documents.

La segmentation de requête : une requête verbeuse contient souvent plusieurs concepts ou éléments d'information. Plutôt que de réduire, agrandir ou reformuler la requête, il peut être utile de diviser la requête en plusieurs segments et de traiter chaque segment séparément, ce traitement s'appelle la segmentation de requête. Quatre approches sont plus largement plébiscitées pour la segmentation de requêtes verbeuses : les approches statistiques basées généralement sur la fréquence des termes ou l'information mutuelle entre des paires de termes (JONES, REY et al. 2006), les approches supervisées dans lesquelles les termes sont modélisés sous la forme d'une classification binaire (BERGSMA et Q. I. WANG 2007), les approches génératives basées sur l'extraction de concepts sous-jacents (TAN et PENG 2008) et les approches fondées sur du TAL qui exploitent les dépendances entre les annotations (BENDERSKY, CROFT et SMITH 2011). Des travaux récents ont porté sur une méthode statistique orientée sur la segmentation spécifique de requêtes de e-commerce en utilisant des informations sur la fréquence via les logs de recherche extraits des données des acheteurs et des données des vendeurs sur les produits (PARIKH, SRIRAM et al. 2013).

De nombreux efforts sont ainsi faits pour améliorer la représentation des requêtes verbeuses. Cependant, aucune méthode ne met réellement en perspective l'emploi du TAL comme moyen de préserver la structure de la phrase et donc sa sémantique. Nous pensons pourtant que le TAL peut apporter une résolution nouvelle pour ce type de requête en mettant en avant des caractéristiques sémantiques qui vont au-delà du lexique seul. Concernant la méthode d'intégration des informations extraites via des procédés issus du TAL, nous avons choisi d'utiliser l'expansion de requête. Notre choix s'est porté sur cette méthode car cette dernière nous permet, d'une part, d'élargir la requête originale afin d'augmenter les correspondances potentielles avec des documents supplémentaires, et d'autre part, de permettre l'ajout d'informations permettant une meilleure représentation de la requête originale.

1.2.1.3. Modèle de recherche de livres basé sur l'analyse de requêtes verbeuses

Il existe un grand nombre de modèles de RI pouvant s'appliquer à la recherche de livres. La principale distinction entre chacun de ces modèles s'opère sur la façon dont les informations sont représentées et par la façon d'interroger la base documentaire. Les principaux modèles sont : les modèles booléens, les modèles algébriques ou appelés aussi vectoriels, les modèles probabilistes, les modèles de langue et les modèles bayésiens.

Le premier modèle, nommé modèle booléen (SALTON 1973 ; MANNING, RAGHAVAN et al. 2008), est une représentation mathématique du contenu d'un do-

cument, selon une approche ensembliste. Les documents sont représentés par des ensembles de termes et les requêtes traitées comme des expressions logiques. Considérant un vocabulaire $T = t_1, \ldots, t_m$, un document est caractérisé par la présence ou l'absence de chaque t_i dans son contenu. La requête s'exprime alors avec des opérateurs logiques selon le formalisme de l'algèbre de Boole. Un document du corpus est ainsi considéré comme pertinent uniquement quand son contenu est vrai pour l'expression de la requête.

Le second modèle, nommé modèle vectoriel (SALTON 1979 ; CASTELLS, FERNANDEZ et al. 2007), est une représentation mathématique du contenu d'un document, selon une approche algébrique. L'ensemble de représentation des documents est un vocabulaire comprenant des termes d'indexation. Ceux-ci sont typiquement les mots les plus significatifs du corpus considéré. Ils peuvent éventuellement être des constructions plus élaborées comme des expressions ou des entités sémantiques. Chaque contenu est ainsi représenté par un vecteur $\vec{v}$, dont la dimension correspond à la taille du vocabulaire. Chaque élément v_i du vecteur $\vec{v}$ consiste en un poids associé au terme d'indice i et à l'échantillon de texte. Un exemple simple est d'identifier v_i au nombre d'occurrences du terme i dans l'échantillon de texte. La composante du vecteur représente donc le poids du mot i dans le document. L'un des schémas de pondération le plus utilisé est le TF-IDF.

Le troisième modèle, nommé modèle probabiliste (MARON et KUHNS 1960 ; ROBERTSON, VAN RIJSBERGEN et al. 1980), est fondé sur le calcul de la probabilité de pertinence d'un document pour une requête. La similarité entre un document et une requête est mesurée selon le rapport entre la probabilité qu'un document d donné soit pertinent pour une requête Q, notée $p(d/Q)$. Ces probabilités sont estimées par les probabilités conditionnelles de la présence ou non d'un terme de la requête dans le document. Les documents et les requêtes sont représentés par des vecteurs booléens dans un espace à n dimensions.

Le quatrième modèle, nommé modèle de langue (BERGER et LAFFERTY 1999 ; PANG, LEE et al. 2008), consiste à ordonner chaque document d de la collection C selon leur capacité à générer la requête de l'utilisateur q. Ainsi, il s'agit d'estimer la probabilité de génération $P(q/d)$. Le dernier modèle, nommé modèle bayésien (TURTLE et CROFT 1991 ; K.-M. KIM, HONG et al. 2007), est basé sur un graphe de dépendances, direct et acyclique. Dans ce graphe, les nœuds représentent des variables aléatoires propositionnelles et les arcs des relations causales entre les nœuds. Ainsi, si le nœud p représente une proposition qui cause ou implique la proposition représentée par le nœud q, on trace un arc allant de p vers q. Les nœuds sont pondérés par des valeurs de probabilités conditionnelles.

Des collections de tests sont souvent utilisées pour juger de l'efficacité des systèmes de RI et ainsi faire évoluer leur performance. Parmi elles, nous pou-

vons citer : la collection TREC [11], CLEF SBS [12] et NTCIR (*NII Test Collection for IR Systems* [13]). Dans le cadre de nos travaux, nous nous sommes particulièrement intéressée à la collection proposée lors des campagnes CLEF SBS introduite en 2010. Cette collection s'oriente plus particulièrement sur l'évaluation des modèles de recherche de livres basés sur l'analyse de requêtes verbeuses.

CLEF SBS se découpe en deux sous-tâches : la première dédiée à la suggestion et la seconde relative à l'interactivité. Les travaux présentés lors de ces campagnes exploitent trois types d'informations : les informations sociales, les informations descriptives et les informations sur le contenu textuel. Les informations sociales englobent tous les avis émis par les utilisateurs à propos d'un livre (commentaires, évaluations, etc.). Les informations descriptives renvoient aux métadonnées contenant des informations sur un livre telles que l'auteur, le prix, le nombre de pages, etc. Les informations sur le contenu textuel correspondent au contenu même du livre (pouvant renvoyer à version intégrale ou non). Toutes ces caractéristiques donnent lieu à de nombreuses combinaisons propres aux systèmes de RI dédiés à la recherche de livres.

La majorité des pistes abordées se fonde sur le principe de correspondance de mots en faisant varier le nombre de champs à considérer tout en jouant sur les poids des paramètres des mots ou groupes de mots. Des techniques de reclassement, qui consistent à classer de nouveau les résultats initiaux fournis par les modèles de recherche, ainsi que des techniques d'enrichissement des résultats fournis par le modèle de recherche sont également utilisées.

La grande majorité des techniques de reclassement présentées consiste à établir un score social (fondé sur les métadonnées générées par les utilisateurs) qui est ensuite combiné aux résultats du modèle de recherche sur le contenu. Parmi les scores présentés, (Bonnefoy, Deveaud et al. 2012) émettent l'hypothèse que plus un livre a de commentaires, et si ses notes sont généralement bonnes, plus il doit être un bon livre. (Benkoussas et Bellot 2013) effectuent une pondération entre l'évaluation de chaque commentaire et les votes attribués par les utilisateurs sur l'utilité de ce même commentaire (*helpful vote*).

D'autres travaux tentent également d'effectuer un reclassement mais cette fois *via* des méthodes d'apprentissage d'ordonnancements [14], ce sont par ailleurs ces

11. http://trec.nist.gov/
12. http://social\discretionary{-}{}{}book\discretionary{-}{}{}search.humanities.uva.nl/#/overview
13. http://research.nii.ac.jp/ntcir/index-en.html
14. *Learning to rank* : apprentissage d'ordonnancements. L'apprentissage d'ordonnancements est l'application de techniques d'apprentissage dans la construction de modèles de classement pour les systèmes de recherche d'information (Hang 2011).

travaux qui ont obtenu les meilleures performances en 2014 (B.-W. ZHANG, YIN et al. 2014). En 2015, c'est également une technique de reclassement fondée sur des méthodes d'apprentissage d'ordonnancements qui a obtenu les meilleurs résultats (GÄDE, HALL et al. 2015). L'originalité de cette approche est de tenir compte de métadonnées disponibles pour chacun des livres de la collection telles que le prix et la longueur du livre généralement peu exploitées dans ce contexte. (IMHOF 2016) s'est également interressée à ce type de métadonnée *via* l'intégration d'indicateurs basés sur le prix moyen des livres déjà lus par l'utilisateur. En 2016, les meilleures performances ont été obtenues grâce à l'analyse des profils utilisateurs *via* la génération d'un index basé sur les livres les plus fréquemment répertoriés au sein des profils (S.-H. FENG, B.-W. ZHANG et al. 2016). À partir de ces profils, ces mêmes chercheurs ont également élaboré une technique de reclassement incluant des informations comme le nombre de lecteurs ayant déjà lu des livres de la collection. Concernant l'enrichissement des résultats, nous avons présenté, lors de notre participation, un modèle de RI classique dont les résultats sont enrichis par une analyse effectuée par des graphes (BENKOUSSAS, OLLAGNIER et al. 2015).

Certains chercheurs se sont intéressés plus spécifiquement aux besoins exprimés par les utilisateurs en implémentant une base de connaissances. (S.-H. WU, LIAO et al. 2014) et (S.-H. WU, HSIEH et al. 2016) utilisent un jeu d'amorces extraites de requêtes dans lesquelles l'utilisateur cherche des livres similaires à ceux énoncés dans sa requête. Un filtre a été mis en place afin d'éliminer les livres non nécessaires lors de la recommandation pour ce type de requêtes.

Dans la sous-section suivante, nous proposons d'étudier les intentions des utilisateurs au travers de taxonomies réalisées dans la littérature ainsi que les méthodes de classification introduites à cet effet.

1.2.2. Déterminer l'intention des utilisateurs au sein des requêtes verbeuses

Beaucoup de travaux ont porté sur la compréhension du comportement de recherche des utilisateurs sur le Web en se focalisant plus particulièrement sur la façon dont les utilisateurs font des recherches, et ce, sans finalement connaître l'intention des utilisateurs. Connaitre le « pourquoi » du comportement de recherche de l'utilisateur est essentiel pour satisfaire ses besoins d'information. Dans les sections suivantes, nous présentons les taxonomies réalisées à cet effet ainsi que les méthodes de classification mises en place.

1.2.2.1. Les taxonomies de requêtes de recherche

De nombreuses taxonomies ont été créées en vue de classifier chaque requête selon le besoin exprimé par l'utilisateur. La première taxonomie réalisée suite à l'étude de requêtes issues du Web est celle de (BRODER 2002), elle comprend trois catégories (voir tableau 1.1).

Intention utilisateur	**Objectif**	**Exemple**
Navigationnelle	Atteindre un site particulier que l'utilisateur a à l'esprit.	*aéroport de Chicago*
Informationnelle	Trouver des informations présumées disponibles sur le Web.	*Comment faire une demande de passeport*
Transactionnelle	Effectuer une interaction dans un site.	*Imprimer les cartes des comtés de NC*

Tableau 1.1. – Taxonomie des intentions utilisateurs de Broder

Les requêtes navigationnelles : ce type de requêtes contient des informations sur l'organisation, le nom d'entreprises ou d'universités, des suffixes de domaine comme « .com », « .org » ou des préfixes comme « www » ou « http ». Certaines de ces requêtes contiennent des URL ou des parties d'URL. Le but de ces requêtes est d'atteindre un site particulier que l'utilisateur a à l'esprit, soit parce qu'il l'a visité dans le passé, soit parce qu'il suppose qu'un tel site existe.

Les requêtes informationnelles : la principale caractéristique de ce type de requêtes est d'être exprimée en langage naturel. Les requêtes pour une telle recherche peuvent consister en des termes informatifs tels que « liste » ou « liste de lectures », contenir des mots interrogatifs comme « qui » ou « quand ». Les termes de recherches peuvent être associés à des conseils, une aide ou des directives comme « FAQs » ou « comment faire [...] », ou encore des idées ou des suggestions. Des recherches de données multimédia, d'informations historiques, sur des célébrités, sur les sciences ou encore la médecine sont considérées comme des requêtes informationnelles.

Les requêtes transactionnelles : le but de ces requêtes est d'atteindre un site où une interaction supplémentaire se produira. Les principales catégories de ces requêtes sont le shopping, la recherche de divers services Web, le téléchargement de divers types de fichiers (images, chansons, etc.), l'accès à certaines bases de données (par exemple, les données du type Pages jaunes), la recherche de serveurs (par exemple pour le jeu), etc. Ce type de requêtes peut contenir des termes de divertissement tels que « photos », « jeux » ou encore des termes qui suggèrent une interaction tels que « acheter », « chat », « commander ». Il est également possible de trouver des termes relatifs à des téléchargements comme « logiciel » ou encore des extensions de fichier comme « jpeg » ou « zip ».

Beaucoup d'autres recherches (KATHURIA, JANSEN et al. 2010 ; HERNÁNDEZ, Parth GUPTA et al. 2012 ; MOHASSEB, EL-SAYED et al. 2014) ont basé leurs travaux sur la taxonomie de Broder. (ROSE et LEVINSON 2004 ; JANSEN, BOOTH et SPINK 2008) proposent des versions plus étendues de la classification de Broder *via* l'ajout de sous-catégories aux requêtes informationnelles, navigationnelles et transactionnelles. (LEWANDOWSKI, DRECHSLER et al. 2012) proposent deux nouvelles catégories, *Commerciale* et *Locale*, la première catégorie permet d'identifier les requêtes avec un potentiel commercial comme la requête : « offre commerciale » et la seconde catégorie correspond à des requêtes dans lesquelles l'utilisateur recherche des informations proches de sa position géographique actuelle. (BHATIA, BRUNK et al. 2012) proposent une toute autre classification se divisant en quatre catégories : Ambigüe, des requêtes non ambiguës mais sous-spécifiées, Recherche d'information et Divers. (CALDERON-BENAVIDES, GONZALES-CARO et al. 2010 ; ASHKAN, CLARKE et al. 2009) proposent une classification des requêtes basée sur des facettes. Ces facettes sont extraites des requêtes de l'utilisateur afin d'aider à identifier son intention lors de la recherche comme le genre, l'objectif, la spécificité, la portée, le sujet, la tâche, la sensibilité d'autorité ou encore la sensibilité spatiale et temporelle.

Toutes ces taxonomies répondent à des applications bien précises que sont les moteurs de recherche. Après une étude appronfondie de la littérature, nous n'avons pu trouver de taxonomie propre au traitement des requêtes de recherche de livres. Cependant, tous ces travaux posent des bases solides que nous comptons exploiter afin d'interpréter les intentions des utilisateurs au sein de nos requêtes.

1.2.2.2. Les méthodes de classification

La classification de requêtes basée sur l'intention de l'utilisateur consiste à classer les requêtes dans des catégories relatives aux besoins exprimés par les utilisateurs. (JANSEN et BOOTH 2010) définissent « *l'intention de l'utilisateur comme l'expression d'un objectif affectif, cognitif et situationnel avec un moteur de recherche* »[15]. Ce type de classification est différent de ceux employés dans la classification de documents. En effet, les requêtes Web sont très courtes et beaucoup de ces requêtes sont ambigües et il est courant qu'une requête appartienne à plusieurs catégories. Suite aux différentes taxonomies présentées dans la section précédente, de nombreux travaux ont tenté de réaliser des modèles de classification dédiés à la classification automatique de requêtes selon l'intention des utilisateurs. Ces approches se divisent en trois catégories (CAO, HAO HU et al. 2009). La première catégorie consiste à étendre la requête par le biais de

15. « *user intent as the expression of an affective, cognitive, or situational goal in an interaction with a Web Search Engine.* »

données externes comme des résultats de recherche retournés pour des requêtes similaires, des informations provenant d'un corpus existant ou encore *via* une taxonomie intermédiaire. La seconde catégorie utilise des données non annotées afin d'améliorer l'exactitude de l'apprentissage supervisé. La troisième catégorie augmente le taux de données d'apprentissage *via* l'annotation automatique de requêtes. Des ressources telles que les logs, les ancres textuelles ou encore les résultats retournés par les moteurs de recherche sont généralement utilisées afin d'extraire des traits permettant l'amélioration de la représentation de la requête. Parmi les dernières avancées nous pouvons citer, (LEWANDOWSKI, DRECHSLER et al. 2012) qui proposent d'analyser les données de clics afin de déterminer si les requêtes sont *commerciales* ou *navigationnelles*, une méthode de crowdsourcing est ensuite utilisée pour classer un grand nombre de requêtes. (HERNÁNDEZ, Parth GUPTA et al. 2012) introduisent une solution permettant de classer automatiquement les requêtes en utilisant uniquement le texte inclus dans la requête, en fonction des traits et des caractéristiques décrits par (BRODER 2002 ; JANSEN, BOOTH et SPINK 2008 ; Dayong WU, Y. ZHANG et al. 2010). Plus récemment, (MOHASSEB, EL-SAYED et al. 2014) exploitent des modèles de type de recherche construits à partir d'un ou plusieurs termes et (FIGUEROA 2015) utilisent des attributs issus de la linguistique.

Dans la section suivante, nous présentons une analyse des requêtes de recherche de livres *via* l'étude des propriétés ainsi que des caractéristiques, à la fois, compositionnelles et structurelles de ces dernières. Nous exposons, ensuite, une taxonomie dédiées aux types de besoins exprimés au sein des requêtes de recherche de livres que nous avons élaborée.

1.3. Analyse de requêtes de recherche de livres

Ces dernières années, la croissance des médias sociaux a encouragé l'utilisation de l'intelligence collective dans un esprit de collaboration en ligne. Grâce à ces médias sociaux, les utilisateurs peuvent obtenir des opinions, des suggestions ou des recommandations d'autres membres. Ces médias sociaux permettent aux utilisateurs d'exprimer des besoins d'information complexes ou hautement spécifiques grâce à des requêtes détaillées. L'intérêt pour le traitement de ce type de requêtes s'accroît. En effet, les études démontrent qu'il est souvent difficile pour les utilisateurs d'exprimer un besoin spécifique d'information au sein d'un moteur de recherche (BISKRI et ROMPRE 2012). Comme le présente l'exemple qui suit, dans le contexte de requêtes de recherche de livres, les requêtes fournissent des informations qui peuvent s'apparenter à des profils utilisateurs comme les goûts, les centres d'intérêt mais aussi leurs attentes.
« *I am looking for a good biography on Lenin and also a biography on Franklin D Roosevelt. Must be quite recently published, informative and readable. Can anyone*

suggest any good titles ? »

Les sections suivantes vont, dans un premier temps, déterminer les propriétés des requêtes de recherche de livres. Dans un second temps, établir leurs caractéristiques, à la fois, compositionnelles et structurelles. Dans un troisième temps, présenter une taxonomie dédiée aux types de besoins exprimés au sein des requêtes de recherche de livres.

1.3.1. Qu'est-ce qu'une requête verbeuse de recherche de livres ?

Dans le cadre de nos travaux, nous nous sommes intéressée aux requêtes dans lesquelles les utilisateurs recherchent des livres *via* un forum un ligne. Comme nous avons pu le voir dans la section 1.2.1.1, les requêtes verbeuses exprimées en langage naturel sont caractérisées par leur aspect long, au niveau de leur densité textuelle, et détaillé, au niveau de l'expression des besoins de l'utilisateur. Les requêtes sur lesquelles nous nous sommes penchée sont extraites des données fournies lors des campagnes CLEF SBS 2014, 2015 et 2016 [16]. Les requêtes fournies lors de ces campagnes sont extraites du forum LT dominé par la langue anglaise et sont exprimées en langage naturel. La figure 1.1 illustre un exemple de requête rencontrée au sein des corpus des campagnes CLEF SBS.

```
<message>
  <date>Aug 15, 2006, 9:38pm </date>
  <text>I'm interested in recommendations for good legal/courtroom mystery series.

Two I like are John Lescroart's Dismas Hardy series that starts with Dead Irish, and Barbara Parker's Gail Connor series that starts
with Suspicion of Innocence.

There are at least a dozen books in the Dismas Hardy series, which is set in San Francisco. I like them because they are meaty
stories (usually around 500 pages) with lots of recurring characters. There are even a couple where Dismas Hardy steps aside and
other characters take the protagonist role.

The Gail Connor series appeals to me because, although she is a lawyer, the books are not courtroom mysteries. She is more of an
amateur detective who happens to be a lawyer, than a lawyer who solves mysteries. Also, her relationship with her hot Cuban
boyfriend, Anthony Quintanna, makes for an interesting side story throughout the series.

But I need a new series, since I am almost completely through with these two. Any suggestions?
</text>
    <postid>1</postid>
    <threadid>1030</threadid>
    <username>RoseCityReader</username>
</message>
```

Figure 1.1. – Exemple de requête extraite du corpus CLEF SBS 2016

Ces requêtes sont destinées à d'autres êtres humains et non à un moteur de recherche, ce qui explique leur longueur ainsi que le niveau de détail exprimant le besoin informationnel de l'utilisateur. Dans ce cas précis, nous pouvons observer que l'utilisateur fournit un certain nombre d'informations dont des exemples de livres bien précis qui sont utilisés à titre d'exemple et permettent ainsi d'aiguiller précisement les autres utilisateurs sur le type de littérature recherchée. L'utilisateur poursuit ensuite sur des explications relatives aux aspects intrinsèques des

16. http://social-book-search.humanities.uva.nl//#/data/suggestion

livres cités qui font que ce dernier recherche une similarité avec lesdits ouvrages. L'utilisateur conclut sur l'expression de son intention suite à la rédaction de cette requête *via* la locution : « *Any suggestions?* ». Comme le reflète cette requête, les requêtes émises dans le cadre d'une recherche de livres expriment des informations plus complexes sur le besoin de l'utilisateur. Chaque requête fournit un contexte particulier établi par l'expression des goûts et des centres d'intérêt de l'utilisateur. De plus, ce type de requête définit clairement l'intention qui se cache derrière la demande l'utilisateur.

1.3.2. Quelles sont les caractéristiques compositionnelles et structurelles des requêtes verbeuses de recherche de livres

Dans le cas de forum en ligne, nous sommes face à des interactions bien différentes de celles que nous pouvons rencontrer entre homme-machine. En effet, les requêtes de recherche de livres sont longues, au niveau de leur densité textuelle, et détaillées, au niveau de l'expression des besoins de l'utilisateur, avec des structures bien souvent complexes. Suite à l'étude des corpus fournis lors des campagnes CLEF SBS, nous avons pu relever que les requêtes sont composées en moyenne de 64,53 mots et de 3 phrases. Comme nous avons pu le relever dans la partie 1.2.1.1, nous sommes face à des requêtes de taille bien supérieure aux requêtes rencontrées traditionnellement en RI. Afin d'approfondir l'étude que nous avons menée sur les caractéristiques compositionnelles et structurelles de ces requêtes, nous nous sommes penchée plus particulièrement sur l'analyse du corpus CLEF SBS 2016 [17]. Cette étude a été réalisée *via* l'utilisation de la librairie Python *Natural Language ToolKit* [18] (NLTK) . Le tableau 1.2 décrit l'ensemble des classes grammaticales utilisées lors de l'annotation.

Balise	**Signification**	**Exemple en anglais**
CC	conjonction de coordination	*and, both, but, etc.*
CD	numéral et nombres cardinaux	*mid-1890, nine-thirty, forty-two, etc.*
DT	déterminant	*all, an, another, etc.*
EX	"existential there"	*there*
IN	prépositions ou conjonctions de subordination	*astride, among, uppon, etc.*
JJ	adjectif comparatif, numéral et superlatif	*third, bleaker,calmest, etc.*

17. http://social-book-search.humanities.uva.nl/data/topics/sbs16suggestion.topics.xml.gz
18. http://www.nltk.org/

MD	auxiliaire modal	*can, cannot, could, etc.*
NN	nom commun singulier, masse et pluriel	*common-carrier, cabbage, undergraduates, etc.*
NNP	nom propre singulier	*Motown, Venneboerger, Czestochwa, etc.*
PDT	pré-déterminant	*all, both, half, etc.*
POS	marque du génitif	*', 's*
PRP	pronom personnel et possesif	*I, herself, him, etc.*
RB	adverbe, adverbe comparatif et superlatif	*occasionally, further, biggest, etc.*
RP	particle	*further, gloomier, grander, etc.*
TO	« to » comme préposition ou marque de l'infinitif	*to*
UH	interjection	*Goodbye, Goody, Gosh, etc.*
VB	verbe	*ask, assembled, speaks, etc.*

Tableau 1.2. – Liste des balises utilisées lors de l'annotation des classes grammaticales

La première étude, illustrée par la figure 1.2, présente le taux d'utilisation des différentes classes grammaticales au sein de l'intégralité du corpus. La seconde étude présentée par la figure 1.3, illustre la fréquence moyenne d'apparition des différentes classes grammaticales au sein d'une requête.

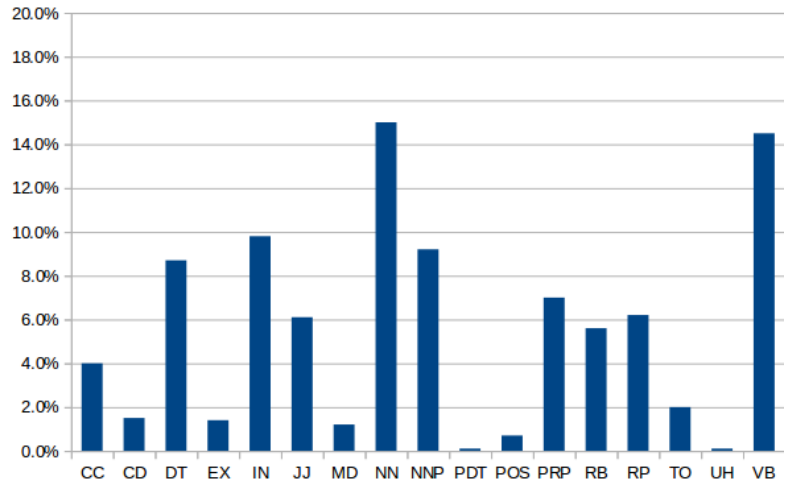


Figure 1.2. – Taux de répartition des classes grammaticales sur l'ensemble du corpus

Ces deux figures, nous permettent de constater une corrélation entre la répartition des classes grammaticales sur l'ensemble du corpus et leur fréquence

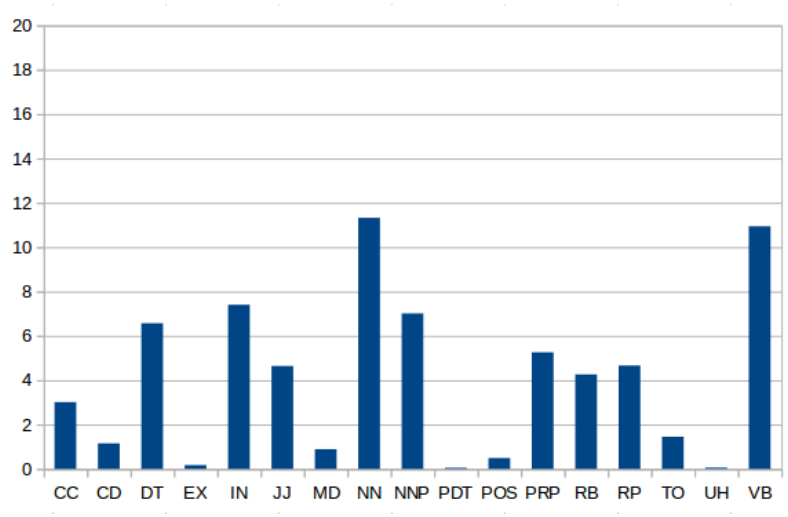


Figure 1.3. – Fréquence moyenne d'apparition des classes grammaticales par requête

d'apparition au sein d'une requête. En effet, la figure 1.2 nous permet d'observer très clairement une forte utilisation des noms communs (NN) avec un taux de 15%, des verbes (VB) avec un taux de 14,5% ainsi que des prépositions et conjonctions de subordination (IN) avec un taux de 9,8%. Concernant la figure 1.3, le constat est le même avec comme classes grammaticales les plus utilisées les NN avec 11,6 noms communs employés en moyenne par requête suivi de 10,7 VB et 6,8 IN. En linguistique, les noms ou plutôt ce que nous définissons comme *noms substantifs* (LITTRÉ 1874) réfèrent à des entités. Il peut s'agir d'êtres (fille, maçon, chat, arbre...), de choses (maison, pierre, feu, livre...), de sensations (lumière, peur, musique, goût...), de notions (force, idée, âge, style...), etc. Leur rôle de descripteur participe à la construction du sens discursif, tandis que le verbe a le pouvoir d'exprimer une action. D'un point de vue morphosyntaxique, le verbe joue un rôle majeur dans l'organisation de la plupart des phrases. En mettant en relation les autres éléments constitutifs d'une proposition, selon son sens propre et des règles morphosyntaxiques propres à chaque langue, le verbe fait de la proposition un ensemble signifiant dont il constitue le noyau. Ces premières constations sont intéressantes car les verbes et les noms sont les classes grammaticales qui composent les prémisses d'une phrase et sont donc des éléments capables de porter l'énoncé d'une proposition. Ceci nous permet d'attester que les requêtes de recherche de livres se rapprochent d'expressions émises en langage naturel. La présence importante de IN est également très intéressante car elle met en évidence la complexité structurelle des requêtes de recherche de livres. En effet, en linguistique, la conjonction de subordination sert à relier deux éléments syntaxiques de nature différente et la préposition permet une incidence c'est-à-dire qu'elle établit un rapport logique entre les mots. Ces classes

employées dans le cadre d'une phrase permettent d'obtenir différents réseaux de relations plus ou moins complexes. Nous pouvons donc attester, *via* l'utilisation importante de ces trois classes grammaticales, que nous sommes face à des structures complexes composées de prémisses de phrase pouvant posséder des réseaux de relations.

Au total, 17 classes grammaticales ont été identifiées au sein du corpus et ce de façon plus ou moins marquée. Pour la figure 1.2, nous dénotons une oscillation du taux de répartion des classes grammaticales sur l'ensemble du corpus entre 15% et 0,1% dont le taux le plus faible est obtenu à la fois par les interjections (UH) et les pré-déterminants (PDT). Et pour la figure 1.3, nous observons une oscillation de la moyenne des apparitions comprise entre 11,6 et et 0,3 avec le taux le plus faible également obtenu par les UH et les PDT. Ces études ont permis de caractériser la diversité des classes grammaticales au sein des requêtes verbeuses de recherche de livres. Cette diversité engendre des variations à la fois structurelles et compositionnelles importantes ce qui traduit la complexité qui réside au sein de ce type de requêtes.

Dans la sous-section qui suit, nous présentons une taxonomie dédiée à la représentation des besoins informationnels des utilisateurs au sein des requêtes de livres.

1.3.3. Quels sont les types de requêtes de recherche de livres et comment les identifier ?

Comme nous avons pu le relever dans la partie 1.2.2.2, des recherches ont porté sur la mise en place de taxonomies permettant d'identifier les besoins utilisateurs (BRODER 2002 ; ROSE et LEVINSON 2004). Or, ces taxonomies très orientées sur l'analyse des requêtes Web s'appliquent difficilement à l'analyse de requêtes de recherche de livres. En effet, nos travaux sont orientés sur des requêtes très spécifiques dont l'intention est purement informationnelle, c'est-à-dire que l'utilisateur exprime un intérêt pour obtenir une information qui, dans notre cas, est exclusivement relative à la recherche de livres. Inspirés par ces recherches, nous avons transposé ces mêmes réflexions dans l'étude des requêtes de recherche de livres *via* l'étude du corpus CLEF SBS 2014. Afin d'établir un panel des plus exhaustifs, nous avons choisi d'étudier ce corpus car ce dernier contient un nombre plus important de requêtes comparativement aux autres années, soit 680 requêtes. À partir de ce corpus, nous avons pu établir 5 taxons principaux composés pour certain de sous-taxons permettant de retranscrire les différents besoins informationnels des utilisateurs. La taxonomie est la suivante :

- **Requêtes orientées** : requête qui exprime une thématique de recherche.

  - Thématiques générales : expression d'un concept plus ou moins précis.
    « *I've lived in LA for about 6 months now, and I always like to get to know a city through the fiction that has been written about it or from within it. Any suggestions on quintessential LA novels?* »
  - Thématiques spécifiques : expression d'un concept associé à un ou plusieurs attributs.
    « *anyone got any suggestions on books dealing with early pre-wwii gay movements, especially outside germany and britain ?* »
  - Personnages : soit un livre d'un auteur particulier ou un livre parlant d'un personnage (fictif ou non).
    « *I just watched a documentary, "Bukowski: Born Into This," and was wondering if anyone could recommend a book of poems. I'm looking to start reading him and curious as to where to start. This author was recommended to me by dylanwolf. I do have one of her novels on my "must buy" list. What do others think of her? Which of her novels do you like best?* »
- **Requêtes analogues** : requêtes visant à chercher des similitudes avec d'autres auteurs, livres, collections.
  Are we going to pick a new book to group read soon? I'd like to propose a Brandon Sanderson book since he just joined Green Dragon. My preference is Elantris . Does anyone else have any thoughts or suggestions? »
- **Requêtes dissemblables** : dans ce type de requêtes l'utilisateur émet des contre-indications.
  « *I'm looking for a good gritty wilderness based fantasy/fantasy series. No George R. R. Martin, Robert Jordan, Terry Goodkind, or Terry Brooks. Thanks Rich* »
- **Requêtes basées sur le contenu** : l'utilisateur fournit une description du contenu du livre.
  « *My girlfriend and I are trying to figure out the name of a book she read in high school several years ago. It was a novel about an English schoolboy having some sort of mental breakdown in the late 19th or early 20th Century. He starts to believe that his friend's mother is Eve, and her yard is literally the garden of Eden. It was very much a coming of age story about trying to find one's self in a world that tries to define you. Though she read it in high school, she does not think it was published for young adults. She does not remember when it was published, but is sure it was not new when she read it. What she does recall very clearly that the title was the name of the main character, and probably started with an 'h'. The cover was an image of the boy in his school uniform. Thanks in advance, folks!* »
- **Autres requêtes** : requêtes trop courtes sans apport d'informations permettant d'identifier le besoin de l'utilisateur ou requêtes dont l'intention de l'utilisateur n'est pas clair.
  « *I've looked her series over several times. I'm a fan of both fantasy and historical fiction... but the "romance" aspect of her Outlander novels has always*

turned me away. Once again I'm considering picking these up. Any recommendations or advice? »

Cette taxonomie permet d'illustrer les différents besoins informationnels exprimés par les utilisateurs au sein des requêtes de recherche de livres. Le premier taxon *requêtes orientées* regroupe toutes les requêtes exprimant une thématique de recherche. Suite à l'étude du corpus CLEF SBS 2014, cette dénomination générique a nécessité des spécialisations *via* la création de sous-taxons. En effet, nous avons pu constater l'expression de thématiques plus ou moins spécifiques comportant des termes permettant de particulariser la recherche. Nous avons donc décidé de distinguer ces particularités *via* trois sous-taxons qui tiennent compte de la modélisation des données exprimées au sein de ces requêtes. Pour ce type de requêtes, nous considérons une modélisation des données basée sur des concepts dans laquelle des attributs peuvent venir caractériser le concept. Comme l'illustre l'exemple présenté dans la taxonomie que nous proposons : « *pre-wwii gay movements* » renvoit au concept et « *outside germany and britain* » à l'attribut qui lui est associé. Le premier sous-taxon *thématiques générales* regroupe des requêtes possédant uniquement des concepts qui ne sont pas des personnages, il représente 8,4 % des requêtes au total. Le second sous-taxon *thématiques spécifiques* représentant 23,5 % des requêtes, regroupe des requêtes exprimant des concepts associés à un ou plusieurs attributs qui permettent de préciser le concept initial (ex : date, lieu, point de vue, classe sociale, langues, etc.). Le troisième sous-taxon *personnages* répertorie des requêtes relatives à des thématiques de recherches portant sur des personnages fictifs ou non, il représente 4,7 % des requêtes au total. Nous avons choisi de dissocier les personnages des concepts dits génériques car les méthodes d'extraction d'information qui seront potentiellement mises en place s'orienteraient, dans ce cas-ci, vers l'utilisation d'un détecteur d'entités nommées. Le second taxon *requêtes analogues* englobe toutes les requêtes dans lesquelles l'utilisateur exprime un besoin visant à rechercher des similitudes entre des livres, des auteurs ou encore des collections, il représente 44,2 % des requêtes au total. Dans ce type de requêtes, nous retrouvons essentiellement l'expression des goûts de l'utilisateur qui peuvent s'exprimer sous diverses formes sans nécessairement employer des mots-clés tels que « *similaire à* ». Le troisième taxon *requêtes dissemblables*, avec un pourcentage de représentation de 1,8 %, regroupe les requêtes dans lesquelles les utilisateurs expriment des contre-indications sur les livres et les auteurs qu'ils ne veulent pas lire. Le quatrième taxon *requêtes basées sur le contenu* repertorie les requêtes dans lesquelles les utilisateurs effectuent une description du contenu du livre car ils ne se souviennent plus du titre et/ou de l'auteur. Ce taxon représente 13,3 % des requêtes au total. Le dernier taxon *autres requêtes* est composé de requêtes ne comportant qu'un apport très faible d'information généralement sous la forme de propos laconiques. Pour ce type de requêtes, il est courant que des informations sous-jacentes soient induites par le contexte d'apparition de la requête comme

l'appartenance de l'utilisateur à un groupe de lecture précis. Ce type de requêtes peut également référer à des cas complexes dans lesquels l'intention de l'utilisateur n'est pas clairement exprimée. Ce taxon représente 3,8 % des requêtes au total.

Suite à l'étude de ce corpus et à la réalisation de cette taxonomie, nous avons pu rencontrer de nombreuses difficultés car comme toute taxonomie, cette dernière a ses limites et ses contraintes. En effet, les limites et les contraintes d'une taxonomie peuvent intervenir à différents niveaux et notamment lors de la création des différents taxons mais aussi lors de la classification des requêtes au sein de ces derniers. L'une des problématiques fondamentales est de parvenir à déterminer quels sont les aspects discriminants qui vont permettre d'établir chaque taxon tout en préservant une unification de la taxonomie. Ensuite, il est impératif de bien définir pour chaque taxon quelles sont ses limites d'application. Sur ce point dans le cadre de nos travaux, nous avons pu constater que la délimitation des frontières entre chaque taxon est parfois complexe. En effet, nous avons pu rencontrer des requêtes que nous qualifions d'ambigües soit de par l'ambivalance dans leur formulation ou soit de par une expression imprécise des intentions de l'utilisateur. La requête suivante illustre le type d'ambigüité qu'il est possible de rencontrer :

« *Here are some things I've been thinking about: -Anyone know of any good South Carolina authors out there writing for independent presses? -What book have you acquired recently (either for yourself or for your public library) that has been truly outstanding? -Anyone out there read Un Lun Dun ? (I recommend it if you haven't). Our library has it classified as an adult book. I personally think it's more of a YA, but I know some libraries have it as juvenile. Where has your library put it, or where do you think it belongs? -Anyone out there who works with YAs or is attuned to the YA readership- help! It's almost summer, and we want to get them in here for something besides MySpace! Suggestions? -How can you do something special to promote adult reading during the summer without being completely cheesy? Right now all we have is a Kurt Vonnegut tribute display, and it seems a little morbid to wait for authors to die to create displays. What have you done in the past, or what would you like to see?* »

Dans cet exemple l'intention principale de l'utilisateur est difficile à cerner, ce dernier exprime un certain nombre d'informations et pose un certain nombre de questions. Dans ce cas-ci, il est délicat de mettre en exergue un des aspects qui permettrait de caractériser cette requête. Nous avons pu également constater que certaines requêtes peuvent, étant donné les intentions dégagées par l'utilisateur, s'intégrer dans plusieurs classes. Dans ce premier exemple, l'utilisateur exprime clairement plusieurs besoins :

« *Would anyone like to suggest a good book on appeasement, or on Neville Chamberlain, the British Conservative government of the 1930s, or attitudes to Hitler, Fascism and Japan prior to the war?* »

Cette requête peut à la fois convenir aux taxons *thématiques générales*, *thématiques spécifiques* et *personnages*. En effet, l'utilisateur exprime son désir de trouver des livres à propos de l'apaisement (thématique générale), de Neville Chamberlain (personnage) ou encore de sujets plus spécifiques comme le gouvernement conservateur britannique des années 1930 (thématique spécifique). Dans ce second exemple, la requête peut s'intégrer dans plusieurs classes de par l'interprétation de l'annotateur lors de la classification :

« *I'd like suggestions for good Merchant/Trader SF? I've read Cherryh's Merchanter books, and I'm currently reading the Andre Norton Solar Queen books (both of which are good in their own ways) Any other SF with Traders/Merchants as the primary focus of the story? Thanks* »

Peut-on considérer cette requête comme une recherche sur une thématique spécifique ou plutôt une recherche de livres similaires à ceux énoncés. Dans le cadre d'une classification mono-label que nous souhaitons mettre en place, cette requête marque bien l'importance de délimiter les frontières de chaque taxon mais aussi l'importance d'un accord interannotateur afin de confronter les différents points de vue.

Concernant les limites de la taxonomie proposée, bien que nous ayons identifié les principales intentions des utilisateurs, les besoins exprimés au sein des requêtes peuvent évoluer et nécessiter des révisions et des mises à jour. L'évolution des techniques est également un facteur prépondérant dans l'élaboration d'une taxonomie. En effet, l'évolution des moteurs de recherche, par exemple, a permis d'envisager des taxonomies orientées sur des aspects tels que le « *besoin derrière la requête* ».

La section suivante portera sur l'application de cette taxonomie dans le contexte de la réalisation d'un système de recommandation de lectures. Cependant, nous avons réduit cette taxonomie à seulement deux taxons afin de fournir une solution de traitement plus générique. Comme nous le soulignions au début de ce chapitre, à des fins expérimentales, nous nous focalisons sur certains types de requêtes mais il est évident que l'approche proposée est généralisable.

1.4. Analyse en dépendance et classification de requêtes en langue naturelle : application à la recommandation de livres

Au cours de cette section, nous présentons la méthode d'intégration de l'approche que nous proposons, suite à la mise en place de la taxonomie dédiée à la représentation des besoins informationnels des utilisateurs au sein des requêtes de recherche de livres, ainsi qu'à l'utilisation d'un analyseur en dépendance dans la représentation de ces requêtes.

1.4.1. Cadre applicatif

Afin d'évaluer les modèles de recherche réalisés, nous utilisons les données fournies dans le cadre des campagnes d'évaluation CLEF SBS 2014. Pour notre travail, nous nous orientons vers la tâche de suggestions, qui consiste à établir une liste des livres les plus pertinents en fonction d'une requête émise par un utilisateur. Les données utilisées se décomposent en un jeu de 680 requêtes longues et détaillées [19] exprimées en langue naturelle posées par les utilisateurs de LT, dont un exemple est présenté par la figure 2.12. À la base, LT est une application web de catalogage social destinée à enregistrer et partager des bibliothèques personnelles et des listes de livres. Or, LT dispose de services supplémentaires dont les *Groupes* qui permettent, par le biais de forums de discussions, aux membres de se réunir. C'est à partir de ces *Groupes* que les requêtes ont été extraites afin de constituer le corpus. Toutes les requêtes fournies sont en anglais. Pour chaque requête nous disposons de cinq champs : <title>, <mediated_query>, <group>, <narrative>, <catalog>. Les champs <title> et <narrative> sont générés par l'utilisateur tandis que le champ <mediated_query> créé par un annotateur permet une représentation plus succincte reflétant les aspects présents dans la requête de l'utilisateur. Les champs <group> et <catalog> correspondent respectivement à la communauté dans laquelle la requête est adressée et au catalogue personnel de l'utilisateur qui a écrit la requête. Il est important de préciser que les requêtes formulées sont destinées à d'autres êtres humains et non à un moteur de recherche, ce qui engendre des requêtes longues et détaillées aux structures complexes.

La collection de livres est constituée de 2,8 millions de descriptions de livres extraites d'Amazon [20], elle est composée de 64 champs XML (un exemple est présenté figure 2.14). Parmi ces champs, nous distinguons :

— les métadonnées : <book>, <isbn>, <title>, <authorid>, etc.

19. http://inex.mmci.uni-saarland.de/protected/books/inex2014sbs.topics.xml.gz
20. http://www.amazon.fr/

```
<topic id="1584">
  <title>Great alternative history books?</title>
  <mediated_query>alternate history and alternative histories</mediated_query>
  <group>Time Travel, Alternate Histories and Parallel Worlds</group>
  <narrative>    I love alternative histories - two great ones I've enjoyed are
Robert Harris's  Fatherland  and Kim Stanley Robinson's  Years of Rice and Salt .
Any other recommendations?  John     </narrative>
```

Figure 1.4. – Exemple de requête de CLEF SBS 2014

— les informations sociales : <reviews>, <summary>, <tags>, <rating>, etc.

Un ensemble de 93 976 profils anonymes d'utilisateurs est également fourni *via* LT. Ces profils contiennent chacun le catalogue personnel de l'utilisateur. Ce catalogue se compose de l'ensemble des commentaires effectués sur un livre avec son évaluation et éventuellement un jeu de balises relatif à ses thématiques.

```
<book><isbn>0001360000</isbn><title>Mog's Kittens</title><ean>9780001360006</ean>
<binding>Board book</binding><label>HarperCollins UK</label><listprice>$6.99
</listprice><manufacturer>HarperCollins UK</manufacturer><publisher>HarperCollins UK
</publisher><readinglevel>Ages 4-8</readinglevel><releasedate/><publicationdate>
1994-09-01</publicationdate><studio>HarperCollins UK</studio><edition/><dewey/>
<numberofpages>16</numberofpages><dimensions><height>55</height><width>461</width>
<length>469</length><weight>22</weight></dimensions><reviews><date>2007-11-27</date>
<summary>Cute Book</summary><content>cute board book for the cat lover or animal lover
 author of "Hobo Finds A Home"</content><rating>5</rating><totalvotes>0</totalvotes>
<helpfulvotes>0</helpfulvotes></review></reviews>
```

Figure 1.5. – Exemple d'un livre de la collection de CLEF SBS 2014

L'étude des travaux effectués au cours des campagnes CLEF SBS présentée dans la section 1.2.1.3, nous permet de constater que très peu de travaux se penchent sur une interprétation sémantique du contenu de la requête. Bien que (S.-H. Wu, Liao et al. 2014) et (S.-H. Wu, Hsieh et al. 2016), se soient intéressés à la compréhension des besoins des utilisateurs, leurs travaux ne s'orientent toutefois pas vers une interprétation sémantique de la requête avec une réelle interprétation des besoins exprimés par les utilisateurs.

Ces dernières années, les travaux vont plus vers une compréhension des profils utilisateurs, ainsi que des métadonnées générées par ces derniers, sans cependant tenir compte de manière approfondie des besoins d'informations exprimés au travers des requêtes. Nous pensons donc que l'exploitation du TAL peut s'avérer bénéfique et permettre de mieux comprendre les besoins exprimés au sein des requêtes et ainsi améliorer la recommandation. Préserver la structure de la phrase peut être une étape vers l'amélioration de la compréhension des besoins utilisateurs. Au vu des données disponibles *via* CLEF SBS, nous proposons un modèle qui se veut une hybridation des systèmes de RI et des systèmes de recommandation classiques. En effet, le modèle proposé exploite à la fois des informations, issues des commentaires et des évaluations émis par les utilisateurs, ainsi que l'analyse de requêtes, que nous qualifions de verbeuses, exprimant le

besoin informationnel de l'utilisateur. Ces requêtes longues et détaillées se retrouvent bien loin de la suite de mots-clés que l'on utilise en RI et sont porteuses d'éléments relatifs aux informations exploitées par les systèmes de recommandation comme les goûts ou encore les centres d'intérêt des utilisateurs qui peuvent, par ailleurs, s'apparenter à des profils utilisateurs.

Dans la section suivante, nous présentons l'architecture générale du système de recherche de livres que nous proposons.

1.4.2. Architecture du système de recommandation de livres

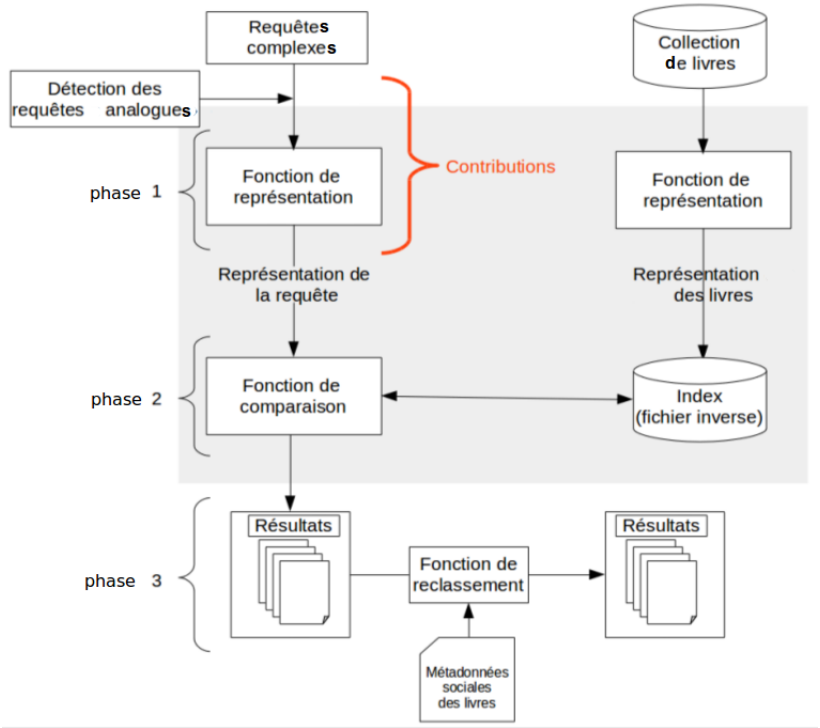


Figure 1.6. – Architecture du système de recommandation de livres

La figure 1.6 représente l'architecture générale du système de recherche des livres dans lequel nos travaux s'inscrivent. Nous distinguons deux processus, d'un côté le traitement des requêtes verbeuses et de l'autre, le traitement de la collection de livres. Au niveau de la phase 1, les fonctions de représentation sont employées à la fois sur les requêtes et sur la collection. Nous implémentons l'approche proposée *via* un prétraitement des requêtes qui consiste en une approche de classification automatique supervisée, suivant la taxonomie présentée dans la section 1.3.3. Afin de générer les modèles de classification, nous avons eu recours

à deux outils Weka [21] et SVMLight [22]. Nous utilisons deux classes qui se nomment respectivement analogue et non analogue. Une fois cette classification effectuée, nous établissons plusieurs stratégies de représentation des requêtes analogues qui sont intégrées dans les fonctions de représentation des requêtes. Pour effectuer ces différentes représentations nous utilisons un outil nommé Stanford Dependencies [23] (DE MARNEFFE, MACCARTNEY et al. 2006). Cet outil nous permet d'obtenir pour chaque requête une analyse en dépendance sous forme de bigrammes de mots associés au nom de la dépendance correspondante. Les informations supplémentaires apportées par ces bigrammes prennent la forme d'une nouvelle balise qui est utilisée lors de la fonction de représentation. La figure 1.7 présente un exemple d'une requête comprenant l'ajout de l'analyse en dépendance.

```
<topic>
<nb>1584
<query>alternate history and alternative histories
<title>Great alternative history books?
<group>Time Travel, Alternate Histories and Parallel Worlds
<narrative>    I love alternative histories - two great ones I've enjoyed are Robert Harris's
Fatherland  and Kim Stanley Robinson's  Years of Rice and Salt .  Any other recommendations? John
<narrative_analyze> I love alternative histories two ones great ones ones enjoyed ones Fatherland
I enjoyed 've enjoyed are Fatherland Robert Harris Harris Fatherland Kim Robinson Stanley Robinson
Robinson Years Any recommendations other recommendations
</topic>
<topic>
```

Figure 1.7. – Exemple de représentation d'une requête *analogue* utilisée lors de la fonction de représentation

Du côté de la collection, la figure 1.8 présente l'exemple d'un livre. Nous avons choisi d'indexer tous les champs de chaque livre et de les représenter sous la forme de sacs de mots.

```
<book>
<isbn>1871034000</isbn>
<text>1871034000 Medicine in Early Mediaeval England 9781871034004 Paperback Manchester Centre
for Anglo-Saxon Studies Manchester Centre for Anglo-Saxon Studies Manchester Centre for
Anglo-Saxon Studies 1989-05-01 Manchester Centre for Anglo-Saxon Studies 40 31 551 795 18 Marilyn
Deegan Editor D.G. Scragg Editor History Medical Subjects Books Refinements Binding (binding)
Paperback Format (feature_browse-bin) Printed Books</text>
```

Figure 1.8. – Exemple de représentation d'un livre utilisée par la fonction de représentation

Lors de la phase 2, le modèle de RI InL2 est utilisé *via* le logiciel Terrier [24]. Nous choisissons ce modèle car lors de la campagne CLEF SBS 2014 nous avons obtenu la deuxième position au regard de la mesure NDCG (*Normalized discounted cumulative gain*) à 10 (BENKOUSSAS, HAMDAN et al. 2014). InL2 est un modèle DFR qui nous permet de calculer un score de pertinence estimé pour chaque

21. http://www.cs.waikato.ac.nz/ml/weka/
22. http://svmlight.joachims.org/
23. http://nlp.stanford.edu/software/stanford-dependencies.shtml
24. http://terrier.org/

livre de la collection selon la requête posée. Les modèles DFR considèrent que, dans un document donné, plus la fréquence d'un mot s'écarte de sa fréquence d'apparition moyenne dans les documents de la collection, plus ce mot est représentatif du document considéré [25] (ROBERTSON, VAN RIJSBERGEN et al. 1980). Dans InL2, L renvoie à la succession de Laplace (E. WILSON 1927) pour la première normalisation et le 2 à la normalisation de fréquence des termes. Le poids de chaque mot est calculé de la manière suivante :

$$w_d(t,d) = \frac{1}{tf+1}\left(tf.log_2\frac{N+1}{n_t+0,5}\right)$$

où tf est la fréquence du terme t dans le document d. N le nombre de documents dans une collection D et n_t le nombre de documents contenant t.

Lors de la phase 3, une fois les résultats obtenus *via* InL2, nous effectuons un nouveau classement des documents *via* la mise en place d'un score social (score d'ordonnancement). Ce score d'ordonnancement est lié à une information précise qui est, dans notre cas, les votes (évaluations) émis par les utilisateurs. En d'autres termes, pour chaque document extrait pour une requête donnée nous calculons un score à partir des informations sociales générées par les utilisateurs telles que les votes et les commentaires. Ce score se fonde sur l'idée que plus un livre a de critiques et de bonnes évaluations plus il est intéressant.

$$Score\_d'ordonnancement(D) = \frac{\sum_{r\in R_D} r}{|commentaire_D|}$$

où, R_D est l'ensemble de tous les votes donnés par les utilisateurs pour le document D et $|commentaire_D|$ est le nombre de votes. Après cela, nous effectuons une pondération entre les scores fournis par InL2 et le score social pour chacun des documents. Comme nous le soulignions précédemment, nos contributions sont présentes au niveau de la phase 1 et plus particulièrement du côté de la requête.

La section suivante présente plus en détail les travaux réalisés afin de classer les requêtes de recherche de livres selon l'intention de l'utilisateur.

25. « *The more the divergence of the within-document term-frequency from its frequency within the collection, the more the information carried by the word t in the document d.* »

1.4.3. Classification supervisée et analyse des requêtes verbeuses

L'utilisation d'une classification des requêtes peut permettre de mieux caractériser les besoins et ainsi d'adapter le traitement employé selon le type de requêtes. Dans cette section nous présentons, dans un premier temps, les classes que nous avons conservées afin de générer un modèle de classification orienté sur la détection de l'intention de l'utilisateur, et dans un second temps, nous étudions différents types de représentations potentielles de requêtes de recherche de livres.

1.4.3.1. Classification de requêtes verbeuses par approche supervisée pour la recommandation de lectures

Suite à la taxonomie présentée dans la section 1.2.2.2, nous avons choisi à des fins expérimentales de ne conserver que deux classes l'une correspondant au taxon *requêtes analogues* et l'autre regroupant le reste des taxons. Les deux classes que nous avons définies se nomment analogues et non analogues. Les requêtes analogues englobent toutes les requêtes dans lesquelles l'utilisateur exprime un besoin visant à rechercher des similitudes entre des livres, des auteurs ou encore des collections. À l'opposé, les requêtes non analogues expriment des besoins variés référant à tous les types de besoins identifiés dans les taxons *requêtes orientées*, *requêtes dissemblables*, *requêtes basées sur le contenu* et *autres requêtes*.

L'intégralité du jeu de requêtes des utilisateurs de LT est classé manuellement par trois annotateurs, chaque classe est ensuite choisie suite à un accord inter-annotateur majoritaire. Pour la classe *analogue*, nous recensons 300 requêtes de ce type soit 44,2 % du corpus fourni par CLEF SBS 2014 et pour la classe non analogue 380 requêtes, soit 55,8 % du corpus. Cette statistique suggère que le corpus est relativement équilibré.

1.4.3.2. Représentation des requêtes analogues

Suite à la classification établie dans la section 1.4.3.1, nous nous penchons plus particulièrement sur la représentation des requêtes analogues. Dans le cadre de la réalisation d'un système de recommandation de lectures, nous présentons plusieurs stratégies de représentation des requêtes analogues *via* l'utilisation d'un analyseur en dépendance. Ces différentes représentations sont incluses sous la forme d'expansions jointes à la requête originale.

Nous supposons que parmi les dépendances trouvées certains types peuvent permettre de caractériser les besoins exprimés. Les dépendances que nous utili-

sons pour les expansions sont réduites, c'est-à-dire que les dépendances impliquant les prépositions, les propositions conjointes, ainsi que des informations sur les référents de clauses relatives sont réduites pour obtenir des dépendances directes entre les mots. Par exemple, pour les dépendances impliquant la préposition « *in* », nous avons :

$$prep((based, 7), (in, 8)), pobj((in, 8), (LA, 9)) \Longrightarrow prep\_in((based, 7), (LA, 9))$$

Ces réductions s'effectuent grâce à des listes préalablement définies de prépositions, de propositions conjointes, ainsi que sur les référents de clauses relatives. Ces réductions se présentent sous la forme de bigrammes associés à la dépendance correspondante. Les nombres présents à côté des mots correspondent aux indices (ou index) de chaque terme au sein de la requête.

Nous présentons différentes études menées afin de raffiner la sélection des dépendances à des éléments permettant de caractériser le besoin d'informations exprimé par un utilisateur au sein des requêtes *analogues*. L'objectif de ces études est de nous fournir des bigrammes de mots permettant d'améliorer la représentation des requêtes et ainsi d'élargir les correspondances potentielles avec l'index fondé sur la représentation des livres.

1.4.3.2.a. Étude fréquentielle

Nous effectuons une analyse fréquentielle qui nous permet d'avoir une vision globale des types de dépendances les plus redondants au sein des requêtes analogues. L'hypothèse émise est que, d'une part, certains types de dépendances sont moins porteurs d'informations, et que d'autre part, une redondance au niveau de certains types de dépendances peut être un vecteur d'informations sur les caractéristiques structurelles de ce type de requêtes et nous permettre ainsi de mettre en exergue les besoins informationnels. Nous retenons les dépendances nominales (nn) qui composent 7,29 % des dépendances présentes dans les requêtes analogues sur un total de 23 728 dépendances (3,38 % dans les requêtes non analogues). Les dépendances composées de prépositions sont également retenues car les prépositions sont des mots qui permettent une incidence, c'est-à-dire qu'ils établissent un rapport logique. Parmi les prépositions jugées pertinentes pour la compréhension des besoins utilisateurs, nous pouvons citer : les prépositions composées avec *of* (prep_of) (2,63 % des dépendances, requêtes non analogues : 2,15 %), *to* (prep_to) (0,67 % des dépendances, requêtes non analogues : 0,53 %) et *about* (prep_about) (0,45 % des dépendances, requêtes non analogues : 0,36 %).

La figure 1.9 présente l'exemple d'une requête après analyse, avec en gras, les

types de dépendances relevés précédemment comme étant pertinents. Nous pouvons observer que l'analyse des requêtes effectuée par Stanford Dependencies nous fournit des bigrammes de mots associés à la dépendance correspondante. Concernant les nombres présents à côté des mots, nous n'utilisons pas cette information, nous la supprimons lors d'un prétraitement effectué sur la requête afin de ne conserver que les bigrammes de mots ainsi que le nom de la dépendance. Cette analyse nous permet d'extraire des bigrammes de mots, pouvant être non consécutifs, correspondant à des liens syntaxico-sémantiques potentiellement représentatifs des besoins d'informations exprimés dans les requêtes analogues que nous intègrerons par une expansion pour chacune des requêtes.

nsubj(enchanted−4, I−1), cop(enchanted−4, was−2), advmod(enchanted−4, completely−3), root(ROOT −0, enchanted−4), det(story−7, the−6), prep_by(enchanted−4, story−7), **prep_of(story-7, Katherine-9)**, **prep_of(story-7, John-11)**, conj_and(Katherine−9, John−11), prep(story−7, of−12), dep(of−12, Gaunt−13), dobj(wondering−17, Gaunt−13), nsubj(wondering−17, I−15), aux(wondering−17, m−16), rcmod(Gaunt−13, wondering−17), mark(recommend−21, if−18), nsubj(recommend−21, anybody−19), aux(recommend−21, could−20), advcl(wondering−17, recommend−21), amod(books−23, historical−22), dobj(recommend−21, books−23), amod(quality−26, similar −25), prep_with(recommend−21, quality−26), **nn(characters-31, writing-28)**, conj_and(writing−28, charismatic−30), **nn(characters-31, charismatic-30)**, **prep_of(quality-26, characters-31)**, ccomp(enchanted−4, T hank−33), dobj(Thank−33, you−34)

Figure 1.9. – Résultat de l'analyse en dépendance pour la requête : "I was completely enchanted by the story of Katherine and John of Gaunt. I'm wondering if anybody could recommend historical books with similar quality of writing and charismatic characters. Thank you!"

1.4.3.2.b. Étude des modèles de classification

Avec cette étude, nous souhaitons montrer que l'analyse des modèles de classification permet d'extraire des éléments caractéristiques des requêtes analogues. Nous pensons que les éléments caractéristiques des requêtes analogues peuvent s'assimiler à des amorces référant à l'expression des besoins de l'utilisateur (par exemple, l'expression « *similar to* »). De ce fait, extraire ces informations ainsi que les éléments voisins peut permettre de préciser les informations que l'utilisateur cherche à obtenir.

1.4.3.2.1. Étude du modèle généré par un classifieur bayésien multinomial naïf (BMN) Un classifieur bayésien naïf est un modèle à caractéristiques statistiquement indépendantes. En termes simples, ce classifieur suppose que l'existence d'une caractéristique pour une classe, est indépendante de l'existence d'autres caractéristiques. Dans le cas de BMN, ce classifieur modélise explicitement le nombre de mots par le biais d'une distribution multinomiale (McCallum et Nigam 1998).

Afin d'extraire les bigrammes de mots les plus représentatifs des requêtes analogues, nous découpons ces requêtes ainsi que les requêtes non analogues *via* des

fenêtres glissantes de deux mots afin d'obtenir des requêtes uniquement constituées de bigrammes. À partir de ces requêtes, nous appliquons le modèle BMN qui nous fournit un score de pertinence pour chaque bigramme de chacune des classes. Le tableau 1.3 présente une liste non exhaustive des bigrammes les plus représentatifs des requêtes analogues extraits par le modèle BMN.

other, suggestions	other, books
I, enjoy	I, loved
just, finished	been, reading
I, finished	ve, read

Tableau 1.3. – Liste non exhaustive des bigrammes extraits par le modèle BMN

Une fois ces bigrammes extraits, nous les comparons aux dépendances produites par Stanford Dependencies afin d'extraire aussi tous les types de dépendances contenant ces bigrammes. Le voisinage direct de ces bigrammes est également extrait. À partir de ces groupes de bigrammes, nous établissons plusieurs représentations des requêtes *via* l'ajout d'une expansion. Pour ce modèle de classification, l'extraction du bigramme de mots précédant le bigramme courant donne les meilleures performances lors de la recommandation par rapport aux autres combinaisons. Le tableau 1.4 présente un exemple d'expansion fondée sur des bigrammes de mots.

Requête	Expansion
I love alternative histories - two great ones I've enjoyed are Robert Harris's Fatherland and Kim Stanley Robinson's Years of Rice and Salt . Any other recommendations?	Salt Fatherland Any recommendations, Salt Fatherland I enjoyed, ones Fatherland Any recommendations

Tableau 1.4. – Exemple de requête avec une expansion fondée sur des bigrammes de mots basée sur les résultats du modèle BMN

Nous avons également, à partir des dépendances extraites *via* ce modèle de classification, établi des bigrammes fondés sur des catégories syntaxiques. Nous avons, à partir du voisinage du bigramme de catégories syntaxiques courant, établi une liste des bigrammes de catégories syntaxiques les plus récurrents. Une fois ces bigrammes de catégories syntaxiques repérés nous extrayons leur contenu afin de ne garder que les mots qui les composent. Dans ce cas-ci, les meilleures

performances obtenues lors de la recommandation sont constatées suite à l'extraction des deux bigrammes de catégories syntaxiques précédant et suivant le bigramme de catégories syntaxiques courant. Ce qui donne, par exemple, des bigrammes de catégories syntaxiques du type :

- *parataxis*[26]*, conjonction and, déterminant, modificateur adjectival, dépendant / objet direct, sujet nominal, auxiliaire, modificateur de rapport de clauses / nom, préposition by, marqueur, modificateur adverbial, auxiliaire*.

Le tableau 1.5 présente un exemple d'expansion fondée sur des bigrammes de mots réalisés à partir de bigrammes de catégories syntaxiques.

Requête	Expansion
I love alternative histories - two great ones I've enjoyed are Robert Harris's Fatherland and Kim Stanley Robinson's Years of Rice and Salt. Any other recommendations?	've enjoyed, ones Fatherland, I enjoyed, ones enjoyed, enjoyed ones

Tableau 1.5. – Exemple de requête avec une expansion fondée sur des bigrammes de catégories syntaxiques basée sur les résultats du modèle BMN

1.4.3.2.2. Étude du modèle généré par C4.5 (J48) J48 est un algorithme de classification supervisée, de type arbre de décision, fondé sur l'algorithme ID3 (BALTIÉ 2002) auquel il apporte plusieurs améliorations. Cet algorithme se fonde sur une mesure de l'entropie dans l'échantillon d'apprentissage pour construire son modèle.

Tout comme BMN, nous découpons les requêtes analogues et non analogues *via* des fenêtres glissantes de deux mots afin d'obtenir des requêtes uniquement constituées de bigrammes. À partir de ces requêtes, nous appliquons J48 qui construit un arbre de décision. Nous sélectionnons ensuite les branches composées de bigrammes dont le poids est le plus important pour la classe analogue. Le tableau 1.6 présente un extrait des bigrammes établis par ce modèle de classification.

Tout comme pour BMN, une fois ces bigrammes extraits nous les comparons aux dépendances produites par Stanford Dependencies afin d'extraire tous les

26. Coordination des phrases et des clauses sans conjonction de coordination.

to, start	other, suggestion
I, reading	other, by
read, have	anyone, recommended
read, liked	finished, having

Tableau 1.6. – Liste non exhaustive des bigrammes extraits par le modèle J48

types de dépendances contenant ces bigrammes. Nous extrayons ensuite les bigrammes de mots avoisinant le bigramme courant.

Pour ce modèle de classification, l'extraction des bigrammes de mots précédant et suivant le bigramme courant donne les meilleures performances lors de la recommandation. Le tableau 1.7 présente un exemple d'expansion fondée sur des bigrammes de mots.

Requête	Expansion
British/Irish authors I've read include Susan Cooper, C.S. Lewis and J.R.R. Tolkein. Can anyone recommend something strongly that might fit more or less into this vein?	British/Irish authors, authors read, authors include

Tableau 1.7. – Exemple de requête avec une expansion fondée sur des bigrammes de mots basée sur les résultats du modèle J48

Nous avons également, à partir des dépendances extraites *via* ce modèle de classification, établi une liste de bigrammes de catégories syntaxiques les plus récurrents. Pour ce modèle de classification, l'extraction des deux bigrammes de catégories syntaxiques précédant et suivant le bigramme courant donne les meilleures performances lors de la recommandation. Ce qui donne, par exemple, des bigrammes de catégories syntaxiques du type :

- *nom, nom, préposition by / sujet nominal, auxiliaire, modificateur adverbial / conjonction and, auxiliaire, modificateur adverbial.*

Le tableau 1.8 présente un exemple d'expansion fondée sur des bigrammes de mots réalisé à partir de bigrammes de catégories syntaxiques.

Les résultats obtenus par ces expansions sont présentés dans la section suivante.

Requête	Expansion
British/Irish authors I've read include Susan Cooper, C.S. Lewis and J.R.R. Tolkein. Can anyone recommend something strongly that might fit more or less into this vein?	Tolkein recommend, Can recommend, anyone recommend

Tableau 1.8. – Exemple de requête avec une expansion fondée sur des bigrammes de catégories syntaxiques basée sur les résultats du modèle J48

1.5. Expérimentations

Dans cette section, nous présentons les mesures d'évaluation. Ensuite, nous détaillons les résultats obtenus par les différentes approches de classification supervisée des requêtes analogues et non analogues. Enfin, nous présentons les différentes indexations effectuées sur les requêtes analogues et l'impact sur la tâche de recommandation.

1.5.1. Mesures d'évaluation

Le tableau 2.1 présente les mesures d'évaluation utilisées. Concernant la classification, nous trouvons la précision, le rappel ainsi que la F-mesure qui sont les mesures usuelles de la tâche CLEF SBS. Pour la classification, supposons une classe i dans laquelle nous devons classer nos requêtes et supposons que le système donne pour cette classe **vp** requêtes vraies positives, **vn** requêtes vraies négatives, **fp** requêtes fausses positives, **fn** requêtes fausses négatives. Concernant l'évaluation des différents modèles de recherche de livres, nous utilisons les mesures suivantes : *Mean Reciprocal Rank* (MRR) et *Mean Average Precision* (MAP).

Afin d'attester de la significativité des résultats entre nos différents modèles de recherche de livres, nous utilisons le test des rangs signés de Wilcoxon [27] (Hull 1993).

27. Le test de Wilcoxon est un test statistique non paramétrique qui permet de tester l'hypothèse selon laquelle la distribution des données est la même dans deux groupes. Comme tous les tests statistiques, il consiste à partir de ce qui est observé à mettre en évidence un évènement dont on connait la loi de probabilité (au moins sa forme asymptotique). La valeur obtenue, si elle est peu probable selon cette loi, suggèrera de rejeter l'hypothèse nulle. Le résultat de ce test est exprimé par une valeur p. Le procédé est généralement utilisé pour comparer la valeur de p à un seuil préalablement défini (typiquement 5 %). Si la p-valeur est inférieure au seuil, nous rejetons l'hypothèse nulle en faveur de l'hypothèse alternative, et le résultat du test est « statistiquement significatif ». Sinon, si la p-valeur est supérieure au seuil, nous ne rejetons pas l'hypothèse nulle, et nous ne pouvons rien conclure sur les hypothèses formulées.

Mesures pour la classification

Nom	Formule	Description
Précision	$P = \frac{vp}{vp+fp}$	Proportion de solutions trouvées qui sont pertinentes. Mesure la capacité du système à refuser les solutions non pertinentes.
Rappel	$R = \frac{vp}{vp+fn}$	Proportion des solutions pertinentes qui sont trouvées. Mesure la capacité du système à donner toutes les solutions pertinentes.
F1-mesure	$F = \frac{2PR}{P+R}$	Moyenne harmonique de la précision et du rappel. Mesure la capacité du système à donner toutes les solutions pertinentes et à refuser les autres.

Tableau 1.9. – Présentation des mesures d'évaluation pour la classification

Mesures pour les modèles de recherche de livres

Nom	Formule	Description
Mean Reciprocal Rank (MRR)	$MRR = \frac{1}{\|Q\|}\sum_{i=1}^{\|Q\|}\frac{1}{Rank_i}$	Le MRR est une mesure pour évaluer une liste de réponses possibles à un échantillon de requêtes, ordonné par la probabilité d'exactitude. Le rang inverse d'une réponse est l'inverse du rang de la première bonne réponse. Le rang inverse moyen est la moyenne des rangs réciproques pour un échantillon de requêtes Q.
Mean Average Precision (MAP)	$MAP = \frac{\sum_{q=1}^{AveP(q)}}{Q}$	La MAP correspond à la précision moyenne pour un ensemble de requêtes Q. En d'autres termes, la MAP est la moyenne des scores moyens de précision pour chaque requête q.

Tableau 1.10. – Présentation des mesures d'évaluation pour les modèles de recherche de livres

1.5.2. Expérimentations sur la classification automatique des requêtes

Dans le cadre de ces expérimentations, nous comparons trois techniques de classification : « Machine à vecteurs de support » (SVM), BMN et J48. Pour l'implémentation du SVM, nous utilisons l'outil SVMLight et pour l'implémentation de BMN et J48 l'outil Weka. L'objectif de ces expérimentations est de parvenir à

détecter les requêtes analogues afin d'arriver à repérer les caractéristiques structurelles qui les composent.

Concernant les paramétrages effectués, nous établissons pour SVM une liste de mots les plus caractéristiques de chaque classe que nous utilisons comme attributs. Cette liste est réalisée grâce à deux algorithmes : *GainRatioAttribute* (GRA) et *InfoGainAttribute* (IGA). GRA consiste en une modification du gain de l'information qui permet de réduire sa polarisation. IGA est utilisé pour réduire le biais vers les attributs à valeurs multiples en prenant en compte le nombre et la taille des branches lors du choix d'un attribut. Après plusieurs tests, nous choisissons pour GRA d'utiliser une fréquence minimale d'apparition des termes de 3 (Fq3) combinée à une liste comprenant tous les mots *All Words* (AW). Pour IGA, nous avons choisi d'utiliser une fréquence minimale d'apparition des termes de 1 (Fq1) combinée à une liste pour laquelle nous avons supprimé les mots dont le score « Élimination récursive de caractéristiques »[28] (RFE) est égal à 0 (AW-0). Concernant les paramètres utilisés lors de la classification *via* BMN, nous optons pour des descripteurs binaires qui sont associés aux mots du vocabulaire *via* l'utilisation, sans paramètre, de la fonction StringToWordVector de Weka. La fréquence minimale d'un terme est réglée sur 1. Pour J48, nous optons également pour la fonction StringToWordVector. Nous choisissons de convertir tous les mots en minuscules. La fréquence minimale d'un terme est réglée sur 1. Concernant les paramètres internes de J48, nous avons, suite à des expériences menées au préalable, modulé le facteur de confiance qui est utilisé pour configurer la taille de l'arbre de décision à 0,10.

Paramètres	**Précision**	**Rappel**	**F-mesure**
SVM IGA-AW-0Fq1	**96,8 %**	71,4 %	82,2 %
SVM GRA-AWFq3	90,5 %	**92,7 %**	**91,6 %**
BMN	79,7 %	79,3 %	79,3 %
J48	71,2 %	71,8 %	71,5 %

Tableau 1.11. – Évaluations de la classification des requêtes analogues et non analogues

Dans le tableau 1.11, nous pouvons observer que les meilleures performances sont obtenues par SVM, BMN et J48 présentent des résultats beaucoup plus faibles pour chacune des mesures. La meilleure F-mesure ainsi que le meilleur rappel sont obtenus suite à l'utilisation de la combinaison GRA-AWFq3 avec SVM. La meilleure précision, quant à elle, est observée sur la combinaison IGA-AW-0Fq1 avec SVM. Le net écart de performances entre SVM et les autres classi-

28. *Recursive Feature Elimination* : élimination récursive de caractéristiques.

fieurs peut s'expliquer par la très grande robustesse du SVM face à des données hétérogènes (AURIA et MORO 2008).

Les résultats obtenus lors de ces évaluations nous permettent d'établir que des caractéristiques structurelles permettent de qualifier les requêtes analogues. Dans la suite de nos expérimentations, nous exploitons ces caractéristiques structurelles extraites de l'étude des modèles générés par BMN et J48 afin d'établir plusieurs représentations des requêtes analogues.

1.5.3. Expérimentations sur la représentations des requêtes analogues

Dans cette section, nous présentons les résultats obtenus suite aux différentes modulations de représentation des requêtes analogues effectuées au sein du modèle de recherche de livres. Tous les modèles présentés utilisent le modèle de RI InL2. Le tableau 1.12 donne une description des caractéristiques des différents modèles utilisés.

Nom	**Description**
Baseline	- Indexation de tous les champs des livres - Indexation de tous les champs des requêtes
FullDep	- Indexation de tous les champs des livres - Indexation de tous les champs des requêtes dont une balise contenant l'ensemble des dépendances générées lors de l'analyse en dépendance
SelectDep	- Indexation de tous les champs des livres - Indexation de tous les champs des requêtes dont une balise contenant les résultats de l'étude fréquentielle
BMN	- Indexation de tous les champs des livres - Indexation de tous les champs des requêtes dont une balise contenant les bigrammes de mots résultant de l'étude du modèle de classification BMN
BMN Pattern	- Indexation de tous les champs des livres - Indexation de tous les champs des requêtes dont une balise contenant les patrons extraits de l'étude du modèle de classification BMN
J48	- Indexation de tous les champs des livres - Indexation de tous les champs des requêtes dont une balise contenant les bigrammes de mots résultant de l'étude du modèle de classification J48

J48 Pattern	- Indexation de tous les champs des livres - Indexation de tous les champs des requêtes dont une balise contenant les patrons extraits de l'étude du modèle de classification J48

Tableau 1.12. – Description des modèles de recherche de livres

Le tableau 1.13 présente les résultats obtenus suite à nos différentes expérimentations. Nous constatons que les meilleurs résultats sont obtenus suite à l'utilisation du modèle J48. Le test de Wilcoxon nous permet de constater que la différence entre les moyennes des deux échantillons de requêtes est statistiquement significative, suggérant une plus grande pertinence du modèle J48 par rapport au modèle de référence Baseline avec une p-valeur pour la MAP de 0,04919 et une p-valeur pour le Recip_rank de 0,01968. Ces résultats nous confortent dans le fait que l'apport d'informations supplémentaires comme les liens syntaxico-sémantiques permet d'améliorer la pertinence de la recommandation. Le deuxième modèle obtenant des résultats sensiblement supérieurs au modèle de référence Baseline est J48 Pattern, le test de Wilcoxon donne pour la MAP une p-valeur de 0,04764 et pour le Recip_rank une p-valeur de 0,04788. Nous pouvons constater que l'exploitation des bigrammes générés par les différentes modulations effectuées sur le modèle de classification J48 est une piste prometteuse. Ce sont ces bigrammes qui permettent de mieux révéler les caractéristiques structurelles des requêtes analogues et ainsi améliorer les performances lors de la recommandation. Concernant les autres modèles, nous pouvons noter des performances sensiblement similaires sauf dans le cas du modèle SelectDep. Les performances peu concluantes de SelectDep peuvent s'expliquer par le choix des types de dépendance comme les noms qui ont pour effet de restreindre la représentativité des requêtes. En effet, les requêtes analogues sont composées de beaucoup de noms de livres et d'auteurs.

Dans l'ensemble, ces résultats nous permettent d'établir que l'apport d'informations *via* l'utilisation de liens syntaxico-sémantiques permet d'améliorer sensiblement les performances lors de la recommandation. Bien que, sur l'ensemble des modèles, les résultats présentent des performances sensiblement similaires au modèle de référence Baseline ces résultats restent encourageants. Nous savons que par rapport à des tâches de RI *ad hoc* ou de RI Web qui obtiennent des MAP aux alentours de 0,3, les résultats présentés ne démontrent pas un gain très important. Cependant, nous sommes face à une tâche qui n'est étudiée que depuis quelques années. Ces premières expérimentations nous encouragent, cependant, à exploiter le TAL comme moyen d'interprétation des besoins utilisateurs au sein des requêtes verbeuses.

Modèle	Recip_Rank	MAP
Baseline	0,1587	0,0339
FullDep	0,1481	0,0305
SelectDep	0,1366	0,0292
BMN	0,1549	0,0324
BMN Pattern	0,1510	0,0320
J48	**0,1616**	**0,0374**
J48 Pattern	0,1590	0,0348

Tableau 1.13. – Évaluations sur les différentes stratégies d'intégration de l'analyse en dépendance dans la représentation des requêtes analogues

Dans la sous-section suivante, nous présentons une analyse qualitative des livres suggérés pour une requête analogue par chacun des modèles.

1.5.3.1. Analyse qualitative des livres suggérés pour une requête analogue

Dans cette sous-section, nous proposons l'exemple d'une analyse qualitative des livres suggérés pour une requête analogue par chacun des modèles présenté précédemment. Cette démarche a pour objectif, à plus grande échelle, de voir si certains modèles fournissent des résultats plus performants selon la structure de certaines requêtes analogues. Cette constatation nous permettrait d'induire l'existence de sous-classes ou de facettes au sein même de la classe des requêtes analogues et ainsi de raffiner nos traitements lors de la phase de représentation des requêtes. Dans le cadre de la campagne CLEF SBS 2014, un fichier contenant des livres associés à des valeurs de pertinence pour chaque requête est fourni (Qrels). Le processus de sélection des livres, lors de CLEF SBS 2014, se fonde sur les suggestions proposées par les utilisateurs de LT. Les valeurs de pertinence attribuées pour chaque livre sont calculées en s'appuyant sur un arbre de décision qui regroupe tous les cas de figure possibles. Par exemple, un livre jugé positivement par un utilisateur et dont les caractéristiques répondent aux besoins exprimés au sein de la requête obtient une valeur de pertinence de 6. À l'inverse, un livre ne correspondant pas aux besoins exprimés au sein de la requête obtient une valeur de pertinence égale à 0. Les Qrels 2014 se composent de 8 918 livres associés à des valeurs de pertinence avec une moyenne de 8 livres par requête. La liste qui suit présente un extrait des livres obtenus pour la requête :

- « *I love alternative histories - two great ones I've enjoyed are Robert Harris's Fatherland and Kim Stanley Robinson's Years of Rice and Salt. Any other recommendations ? John* ».

Concernant les deux livres cités dans cette requête, tous les deux font référence

à des histoires alternatives, c'est-à-dire, à des récits dans lesquels l'histoire est réécrite à partir de la modification d'un événement du passé. *Fatherland* conte une histoire dans laquelle les nazis ont gagné la guerre et *Years of Rice and Salt* remonte au XIV^{e} siècle et imagine que la peste a tué 99 % de la population. Nous présentons pour chaque modèle les trois premiers livres retournés par le modèle de recherche accompagnés de commentaires permettant d'identifier la nature du livre.

- *Qrels 2014*
  1 - **Titre :** *Pavane* **Auteur :** Keith Roberts. **Commentaires :** histoire alternative, fondée sur une histoire ramifiée autour de la mort de la reine Elizabeth et de l'Armada espagnole qui a réussi à conquérir l'Angleterre.
  2 - **Titre :** *The Yiddish Policemen's Union : A Novel* **Auteur :** Michael Chabon. **Commentaires :** À la fois polar, histoire d'amour, hommage aux Noirs des années 1940, et une exploration des mystères de l'exil et la rédemption.
  3 - **Titre :** *Fatherland* **Auteur :** Robert Harris. **Commentaires :** clairement cité dans la requête.
- *Baseline*
  1 - **Titre :** *The Years of Rice and Salt* **Auteur :** Kim Stanley Robinson. **Commentaires :** clairement celui cité dans la requête.
  2 - le même livre dans une autre édition.
  3 - le même livre dans une autre édition.
- *FullDep*
  1 - **Titre :** *The First Men In : U.S. Paratroopers and the Fight to Save D-Day* **Auteur :** Ed Ruggero. **Commentaires :** histoire d'une dangereuse mission confiée à un parachutiste durant la Seconde Guerre mondiale.
  2 - **Titre :** *Combat Jump : The Young Men Who Led the Assault into Fortress Europe, July 1943* **Auteur :** Ed Ruggero. **Commentaires :** fiction autour de la Seconde Guerre mondiale fondée sur des entrevues avec des anciens combattants de la 82e division aéroportée.
  3 - le même livre dans une autre édition.
- *SelectDep*
  1 - **Titre :** *The Years of Rice and Salt* **Auteur :** Kim Stanley Robinson. **Commentaires :** clairement celui cité dans la requête.
  2 - le même livre dans une autre édition.
  3 - le même livre dans une autre édition.

- *BMN*
  1 - **Titre :** *The Iron Lance* **Auteur :** Stephen Lawhead. **Commentaires :** trilogie épique du combat d'une noble famille écossaise pour son existence et sa foi au cours de l'âge des grandes croisades.
  2 - **Titre :***La Herejia* **Auteur :** Romain Sardou. **Commentaires :** thriller qui se déroule au Moyen Âge.
3 - **Titre :** *Russia and the Soviet Union : An Historical Introduction from the Kievan State to the Present* **Auteur :** John M Thompson. **Commentaires :** Une introduction historique de l'État de Kiev à aujourd'hui.
- *BMN Pattern*
  1 - **Titre :** *The Iron Lance* **Auteur :** Stephen Lawhead. **Commentaires :** trilogie épique du combat d'une noble famille écossaise pour son existence et sa foi au cours de l'âge des grandes croisades.
  2 - **Titre :** *La Herejia* **Auteur :** Romain Sardou. **Commentaires :** thriller qui se déroule au Moyen Âge.
  3 - **Titre :** *Summer Lightning* **Auteur :** Judith Richards. **Commentaires :** fiction fondée autour des camps de concentration.
- *J48*
  1 - **Titre :** *The master of the High Castle* **Auteur :** Philip K. Dick. **Commentaires :** fiction autour de la Seconde Guerre mondiale après la capitulation des alliés.
  2 - **Titre :** *Letters Back To Ancient China* **Auteur :** Herbert Rosendorfer. **Commentaires :** mandarin chinois du X^{e} siècle qui se déplace vers le XX^{e} siècle dans sa machine à voyager dans le temps.
  3 - **Titre :** *Summer Lightning* **Auteur :** Judith Richards. **Commentaires :** fiction fondée sur des camps de concentration.
- *J48 Pattern*
  1 - **Titre :** *Inside GHQ : The Allied Occupation of Japan and Its Legacy* **Auteur :** Eiji Takemae. **Commentaires :** compte rendu après l'occupation du Japon donnant un aperçu de l'état japonnais contemporain.
  2 - **Titre :** *Letters Back To Ancient China* **Auteur :** Herbert Rosendorfer. **Commentaires :** mandarin chinois du X^{e} siècle se déplace vers le XX^{e} siècle dans sa machine à voyager dans le temps.
  3 - le même livre dans une autre édition

Suite à l'étude de ces résultats, nous pouvons établir la synthèse suivante :

- *Qrels 2014* ne répond que partiellement aux besoins exprimés dans la requête. Seule, le premier livre renvoie à une histoire alternative. Le deuxième est du type roman noir et le dernier fait clairement référence à *Fatherland*

qui est déjà cité dans la requête.
- *Baseline* retourne trois fois le même livre, *The Years of Rice and Salt*, mais avec trois ISBN différents. Le titre de ce livre fait partie des livres présents dans la requête.
- *FullDep* renvoie à des livres de fiction dont les thématiques se rapprochent de *Fatherland* cité dans la requête.
- *SelectDep* présente les mêmes résultats que *Baseline*.
- *BMN* retourne deux livres correspondant à la thématique de *The Years of Rice and Salt*. Le troisième est, quand à lui, loin de répondre aux besoins exprimés dans la requête car la thématique et le genre littéraire ne sont pas corrects.
- *BMN Pattern* renvoie à deux livres de fiction dont les thématiques se rapprochent de *The Years of Rice and Salt* cité dans la requête. Le troisième, quand à lui, correspond aux besoins exprimés dans la requête étant donné sa thématique et son genre littéraire.
- *J48* fournit deux livres correspondant aux besoins exprimés dans la requête. Seul, le deuxième s'éloigne au niveau thématique et genre littéraire de ce qui est stipulé dans la requête.
- *J48 Pattern* fournit un premier livre correspondant à la thématique de *Fatherland*. Les autres livres ne correspondent pas aux besoins exprimés dans la requête car la thématique et le genre littéraire ne sont pas corrects.

Ce que nous pouvons retenir de cette analyse est que les résultats varient selon le type d'expansion choisi. L'utilisation d'une expansion comme apport d'informations supplémentaires provoque bel et bien un impact non négligeable sur la recommandation. Concernant plus particulièrement les résultats obtenus pour cette requête, il est intéressant de constater que la grande majorité des livres sélectionnés tiennent compte d'au moins un des aspects exprimés dans la requête : le genre littéraire et la thématique. Ces résultats sont également intéressants car ils reflètent deux problématiques que l'on peut rencontrer en RI : la difficulté de se détacher des différentes unités lexicales de la requête et le fait que chaque document est jugé indépendamment des autres ce qui provoque la présence de doublons.

Il est important de relever que dans l'évaluation des systèmes de recherche de livres fondés sur des requêtes, il est difficile de fournir une réponse optimale. En effet, le fait de partir d'une requête longue et détaillée exprimée par un utilisateur peut engendrer plusieurs interprétations de cette dernière. Il est donc difficile du côté des annotateurs qui établissent les jugements de pertinence et du côté des systèmes de fournir une réponse optimale face à toutes les contraintes énoncées dans la requête. De plus, dans le cadre plus spécifique des requêtes analogues, il est parfois difficile de juger quels sont les aspects intrinsèques du livre qui font que l'utilisateur recherche une similarité avec ledit

ouvrage. Cette analyse nous conforte dans le fait que des sous-classes ou des facettes sont envisageables selon les caractéristiques de certaines requêtes analogues afin d'employer le modèle le plus adapté aux types de requêtes analogues rencontrés.

1.6. Conclusion

Ce premier chapitre a permis de détailler les travaux menés sur l'un des axes de recherches présentés au cours de cette thèse, à savoir, l'analyse des requêtes verbeuses de recherche de livres ainsi que l'exploitation des liens syntaxico-sémantiques en vue d'enrichir leurs représentations au sein d'un système de recommandation de lectures.

Nous avons pu observer que ce type de requêtes possède un certain nombre de particularités dont une longueur importante ce qui la différencie entre autres des requêtes à base de mots-clés traditionnellement rencontrées en RI. Nous avons pu constater que ce type de requêtes est l'expression de besoins détaillés au travers desquels il est possible de dégager des informations qui sont d'ordinaire rencontrées au sein de profils utilisateurs. L'analyse, à la fois compositionnelle et structurelle de ces requêtes, a permis de mettre en évidence l'une des caractéristiques principales de ces requêtes, à savoir, son rapprochement vers des expressions émises en langage naturel nativement réalisées par des êtres humains. Les classes grammaticales révélées lors de nos études ont servi à attester cette constation. En effet, l'emploi important de noms, de verbes ainsi que de prépositions et conjonctions de subordination a permis de mettre en évidence l'utilisation de phrases comportant des réseaux de relations plus ou moins complexes. Nous avons pu également rendre compte de la diversité des besoins informationnels exprimés par les utilisateurs via la création d'une taxonomie propre au traitement des requêtes de recherche de livres. Cette étude, nous a permis d'établir 5 taxons principaux composés pour certains de sous-taxons permettant de retranscrire les différents besoins informationnels des utilisateurs.

La suite de nos travaux a ensuite porté sur des solutions de traitements de ces requêtes selon un besoin particulier. À des fins expérimentales, nous nous sommes concentrée sur un type de requêtes bien particulier à savoir les requêtes analogues. Au cours de ce chapitre, nous avons présenté une approche de recommandation fondée sur l'analyse de ces requêtes basée sur des procédés issus de la classification supervisée et du TAL. Comme nous avons pu le constater dans plusieurs domaines de recherche, la formulation des requêtes et le processus de récupération sont souvent considérés comme une tâche simple, et les systèmes de recommandation ne dérogent pas à la règle. Toutefois, les intentions implicites cachées derrière les requêtes sont plus complexes et le processus de recherche

devrait être orienté vers une caractérisation de ces tâches notamment par la prise en compte des besoins formulés par les utilisateurs.

Dans ce chapitre nous avons tenté de nous concentrer sur la compréhension ainsi que sur la représentation des besoins utilisateurs au sein des requêtes analogues. Nous avons pu observer que la méthode de classification des requêtes selon la taxonomie que nous avons établie offre des résultats plus que satisfaisants avec une précision moyenne relevée pour nos deux algorithmes à plus de 90 % sur les classes : analogue et non analogue (cf : tableau 1.11). Concernant l'application d'un analyseur en dépendance, nous avons pu constater que son utilisation dans l'analyse des requêtes analogues nous permet de faire un pas vers une meilleure interprétation des besoins exprimés par les utilisateurs. En effet, l'analyse plus fine des livres restitués pour une requête nous a permis de constater que l'apport d'informations supplémentaires par expansion permet d'améliorer la pertinence des résultats. Les tests de Wilcoxon nous ont permis d'observer des degrés de significativité différents sur certaines requêtes. Ce phénomème nous amène à penser que des sous-classes ou des facettes sont envisageables selon les caractéristiques de certaines requêtes analogues afin d'employer le modèle le plus adapté. De plus, l'analyse détaillée des résultats obtenus pour une requête analogue sur chacun de nos modèles nous permet de corroborer cette hypothèse. Dans nos futurs travaux, nous tenterons d'établir les caractéristiques des requêtes analogues en fonction des meilleures performances obtenues selon les modèles employés. À plus long terme, nous envisageons de mettre en place une stratégie de représentation propre au type de requêtes en fonction de la classe prédite lors de la classification automatique. Nous pensons, par exemple, pour les requêtes orientées extraire les termes qui viennent particulariser la requête.

Dans ce chapitre nous avons vu l'utilisation de liens syntaxico-sémantiques comme moyen d'enrichir la représentation des requêtes de recherche de livres et plus particulièrement des requêtes analogues. Cependant, nous rappellons que nous avons appliqué ces travaux sur des requêtes mais il est évident que nous pouvons utiliser cette approche sur d'autres types de données textuelles exprimés en langage naturel.

Au cours du chapitre suivant, nous passons à un autre aspect inhérent des systèmes de recommandation : la représentation des documents. Lors de ces études nous n'appliquerons pas de procédés comme nous avons pu le présenter dans ce chapitre basés sur des aspects linguistiques. Nous proposons, au vu de notre cadre applicatif cette fois orienté sur une bibliothèque d'articles scientifiques, d'exploiter les références bibliographiques comme source de liens entre les documents. Notre but ne sera pas d'améliorer la compréhension des documents mais bien d'exploiter les références bibliographiques présentes afin de mettre en perspective de nouveaux liens thématiquement liés au document courant.

2. Vers une liaison entre contenus *via* l'identification automatique de références bibliographiques

2.1. Introduction

Au sein de publications scientifiques, les références bibliographiques permettent la construction de la connaissance *via* leur valeur pouvant être, à la fois, rhétorique et argumentative. Des études portées sur l'écrit scientifique ont permis d'identifier plusieurs types de références (CRONIN 1984 ; WOUTERS et al. 1999 ; TUTIN et GROSSMANN 2013). Parmi elles, certaines sont des références explicites à l'image des références que nous pouvons trouver à la fin des articles ou des livres, tandis que d'autres références, que nous qualifions d'allusives, sont disséminées dans le corps du texte. Au cours de ce chapitre, nous nous sommes particulièrement intéressée à l'identification automatique des références allusives dans le cadre de l'exploitation d'articles scientifiques issus d'une bibliothèque numérique dédiée au Sciences Humaines et Sociales. Ce type de références, essentiellement utilisé dans le corps du texte et les notes de bas de page, a une structure très particulière (cf : section 2.3) et peut se trouver dispersé en plusieurs segments dans une phrase ou un paragraphe. Leur caractéristique est d'être disséminée dans le texte selon un degré plus ou moins fort d'implicite. Par la notion d'implicite, nous sous-entendons des références pouvant s'intégrer ou non à la syntaxe du discours. Le premier objectif de ce chapitre est de parvenir à identifier automatiquement ce type de référence.

Pour rappel, l'extraction d'informations bibliographiques consiste à identifier automatiquement chaque mot dans une chaîne de références bibliographiques comme l'un des champs bibliographiques prédéfinis tels que l'auteur, le titre, la date, etc. L'extraction des informations bibliographiques est souvent limitée aux zones plus formelles, à savoir, les références présentes à la fin des articles ou des livres. Les résultats dans ce domaine montrent, par ailleurs, des performances très satisfaisantes (Y.-M. KIM, BELLOT et al. 2011 ; ANZAROOT et MCCALLUM 2013). Nous introduisons au cours de ce chapitre une méthode dédiée à l'indentification automatique des références allusives, qui consiste d'une part, à identifier les paragraphes qui contiennent des références *via* un processus de classification supervisée et d'autre part, dans l'application de CCRF afin de détecter plus précisément les zones bibliographiques et d'annoter leurs contenus.

L'étude des travaux dédiés à l'analyse des écrits scientifiques nous a égale-

ment permis d'établir que l'utilisation de références bibliographiques est un élément crucial dans la création et la diffusion de l'information (CRONIN 1984). Ces dernières années, l'exploitation des références bibliographiques fait l'objet d'un sursaut d'intérêt parmi la communauté scientifique ainsi que dans les milieux industriels. En effet, l'exploitation des références permet, au travers de leur utilisation, d'étudier les tendances scientifiques, d'établir des mesures bibliométriques mais aussi d'identifier les relations ainsi que l'influence des œuvres et des auteurs (BELTER 2015). Au cours de ce chapitre, nous présentons un autre moyen d'exploiter les références bibliographiques et ce, dans le cadre de la réalisation d'un système de recommandation de lectures.

Lors du chapitre précédent, nous nous sommes intéressée à l'enrichissement de la représentation des requêtes verbeuses, cette fois nous passons à un autre aspect inhérent aux systèmes de recommandation : la représentation des documents. Nous proposons, au vu de notre cadre applicatif, d'exploiter les références bibliographiques comme source de liens entre les documents. Notre but ne sera pas d'améliorer la compréhension des documents mais bien d'exploiter les références bibliographiques présentes afin de mettre en perspective de nouveaux liens thématiquement liés au document courant. L'exploitation des liens entre contenus est cruciale dans les approches de recommandation (YAZDANFAR et THOMO 2013). Dans le cadre d'une bibliothèque d'articles scientifiques, les références bibliographiques peuvent s'avérer être une source de liens majeure. En effet, à partir des références bibliographiques, il est possible d'obtenir des informations permettant l'identification d'un document en tant qu'unité documentaire. Parvenir à exploiter ces informations peut nous permettre de faire émerger des liens entre des documents dont les thématiques sont connexes. Cette hypothèse donne lieu au second objectif de ce chapitre qui est, à partir de l'identification automatique des références bibliographiques, d'établir des liens entre contenus. Cette première étude, menée *via* l'exploitation des références présentes dans les zones bibliographiques, est effectuée en vue d'intégrer des informations quantitatives basées sur des observations telles que la fréquence d'apparition et la granularité de distribution des références allusives que nous présentons dans le chapitre suivant.

L'originalité de notre contribution est, dans un premier temps, de proposer une approche dédiée à la détection automatique des références que nous avons qualifié d'allusives. Dans un second temps, nous proposons, dans le cadre de la réalisation d'un système de recommandation, d'exploiter les références bibliographiques comme outil de liaison entre contenus.

Ce chapitre est structuré comme suit : dans la section 2.2 nous effectuons un état de l'art se focalisant sur la notion de « référence » en TAL et sur les approches dédiées à la détection et l'analyse de références bibliograhiques. La section 2.3

propose différentes études sur l'identification des caractéristiques des références bibliographiques présentes au sein d'articles scientifiques. Dans la section 2.4, nous exposons une approche permettant l'identification automatique des références allusives ainsi que son affiliation au développement du logiciel Bilbo. La section 2.5 introduit le cadre applicatif d'OpenEdition ainsi que les résultats obtenus suite à l'application de l'approche proposée. Dans la section 2.6, nous exposons les résultats obtenus suite à notre participation à la campagne CLEF SBS 2016 au cours de laquelle nous avons évalué notre approche face à d'autres approches dédiées à la recherche d'information au sein de collections de livres et de forums en ligne. La section 2.7 est dédiée à la présentation de la modélisation, sous la forme d'un graphe orienté, des références bibliographiques comme outil de liaison entre contenus. Enfin, les conclusions sont présentées dans la section 2.8.

2.2. État de l'art

Au cours de ces sections nous introduisons, d'abord, la notion de « référence » en TAL afin d'établir un parallèle entre les travaux menés sur la résolution des chaînes de référence et l'identification des références bibliographiques allusives. Puis, nous proposons un aperçu des approches et outils dédiés à la détection et l'analyse de références bibliographiques.

2.2.1. La notion de « référence » en Traitement Automatique des Langues

La définition de la notion même de **référence** est un sujet largement étudié. En effet, la question de la référence est une des préoccupations majeures des réflexions sur le langage : philosophes, logiciens, linguistes et psychologues s'y sont depuis longtemps intéressés. Car, finalement, à quoi réfère un mot ? Et comment se fait-il que nous puissions renvoyer au monde par l'intermédiaire du langage ? Afin d'établir un lien plus étroit avec nos sujets de recherches, nous nous sommes concentrée plus spécifiquement sur la notion de référence proposée en linguistique dont de nombreux travaux ont porté sur leurs expressions et leurs structures textuelles.

Définie en linguistique moderne comme « un mot ou un syntagme qui, dans un énoncé, assure une reprise sémantique d'un précédent segment appelé antécédent. » (APOTHÉLOZ 1995), la référence peut désigner plusieurs types d'entités comme les anaphores qui permettent de reprendre sans le répéter un élément du contexte précédent (ex : *J'ai vu **mon professeur**, **il** avait l'air distrait.*) ou les cataphores qui est un procédé consistant à annoncer par un substitut une partie du contexte à venir (ex : *Si tu **la** vois, tu diras à **Julia** que j'ai retrouvé son livre.*). Sous

la forme de composant syntaxique de l'énoncé, les références rendent compte du fait que la plupart des textes évoquent des individus de façon constante mais au moyen d'expressions référentielles aussi diversifiées que les noms propres, pronoms et groupes nominaux plus ou moins variés (LEROY 1999). Différents courants, comme les travaux sur la théorie des chaînes de référence (SCHNEDECKER 1997) ou les travaux sur les théories énonciatives (CHARAUDEAU, MAINGUENEAU et al. 1985), ont eu pour objet le classement des expressions référentielles selon le type de repérage ou de détermination opéré en vue de dresser une typologie textuelle.

Le TAL s'est notamment intéressé à l'identification de ces différents types de références *via* l'intégration de procédés issus de la linguistique afin d'améliorer la qualité des résultats. Au cours des deux dernières décennies, l'ingénierie des langues a connu des avancées spectaculaires qui ont permis l'émergence de nombreuses applications opérationnelles et ce notamment dans les domaines de la RI et de l'indexation de documents. La qualité des outils d'indexation ou d'interrogation développés pour ces tâches dépend, dans une large mesure, de leur robustesse en matière de détection des entités nommées (MUZERELLE, SCHANG et al. 2013). Une entité nommée est une unité linguistique qui désigne un élément précis de l'univers du discours. La détection des entités nommées, désignant le plus souvent les éléments sur lesquels porte le discours, est donc essentielle dans les applications d'extraction ou de recherche d'information textuelle. Afin d'améliorer les capacités de ces applications dans la résolution des références, cette problématique est devenue un axe de recherche important (MUZERELLE, SCHANG et al. 2012 ; LONGO 2013 ; DÉSOYER, LANDRAGIN et al. 2016). Par exemple, la problématique de résolution des chaînes de référence a conduit à l'émergence de nombreux travaux qui ont fait l'objet de campagnes d'évaluation internationales telles que *Message Understanding Conference* (MUC) [1] et *Sementic Evaluation* (SemEval) [2]), ou certaines francophones comme *DÉfi Fouille de Textes* (DEFT) [3]. Ces recherches ont toutefois porté majoritairement sur des documents ou des messages électroniques (langage écrit). La communauté « parole », quant à elle, s'est plutôt intéressée à la problématique de l'anaphore pronominale, très présente en dialogue oral homme-machine. Les avancées continues du traitement de la parole amènent désormais les chercheurs à s'intéresser à une recherche d'information dans des flux oraux ou vidéos équivalente à celle réalisée sur les documents textuels.

Dans le cadre des travaux réalisés au cours de ce chapitre, nous travaillons sur les références bibliographiques allusives. Ce type de références peut également tenir le rôle de composant syntaxique de l'énoncé au sein d'un discours scienti-

1. www.itl.nist.gov/iaui/894.02/related_projects/muc/
2. semeval2.fbk.eu/semeval2.php
3. deft.limsi.fr/index.php?id=1&lang=fr

fique comme nous le constaterons lors de la section 2.3.2. En tenant compte de la définition au sens large du terme référence, à savoir, « son action de référer » et donc, dans le contexte de références bibliographiques de « renvoyer à un document », nous établissons un parallèle entre les travaux menés sur la résolution des chaînes de référence et l'identification des références bibliographiques allusives. De plus, des caractéristiques similaires se dégagent dans la résolution de ces deux tâches, à savoir, les difficultés dans l'appréhension de la diversité des structures des chaînes référentielles ainsi que dans leur identification et la mise en relation avec leur référant.

2.2.2. Détection et analyse de références bibliographiques

Depuis les années 90, l'analyse automatique de références suscite beaucoup d'attention pas seulement pour leur utilisation dans les processus d'indexation mais aussi pour améliorer les performances lors de la récupération et l'extraction d'information. CiteSeer[4] qui est un moteur de recherche ainsi qu'une bibliothèque numérique pour les articles scientifiques appartenant au domaine informatique, est l'un des premiers systèmes d'indexation automatique de citations. Depuis, plusieurs outils, tels que *Google Scholar*[5], permettent l'analyse des réseaux de citation de la littérature scientifique avec une vocation disciplinaire ou pluridisciplinaire (De Bellis 2009). Au niveau disciplinaire, des bases de données fournissent aux spécialistes du domaine des réseaux de citation *via* des systèmes autonomes d'analyse de citations, notamment ceux fournis par le *Chemical Abstract Service*[6] (CAS) pour la chimie et d'autres sciences connexes ; le *SAO/NASA Astrophysics Data System*[7] (ADS) pour l'astronomie ; *IEEE Xplore*[8] pour le domaine informatique. Au niveau multidisciplinaire, nous pouvons citer *Google Scholar*, *Web of Science*[9], *Scopus*[10], *ResearchGate*[11] et depuis 2017, en partenariat avec la fondation Wikimedia, OpenCitations[12].

Le problème de l'analyse des références bibliographiques est considéré comme un problème de labellisation de séquences dans lequel la référence est considérée comme une chaîne de caractères. La récupération et l'extraction d'information basée sur les citations a permis de contribuer à la résolution d'autres tâches telles que le résumé de documents scientifiques (Qazvinian et Radev 2008), la récupération de texte (Ritchie, Robertson et al. 2008) et le regrou-

4. http://citeseerx.ist.psu.edu/index
5. https://scholar.google.fr/
6. https://www.cas.org/
7. http://ads.harvard.edu/
8. http://ieeexplore.ieee.org/Xplore/home.jsp?reload=true
9. https://www.webofknowledge.com/
10. https://www.scopus.com/
11. https://www.researchgate.net/
12. https://i4oc.org/

pement de documents (LAVERGNE, CAPPÉ et al. 2010). Beaucoup de travaux se sont penchés sur l'annotation de séquences selon des classes prédéfinies, nous pouvons classer ces travaux en trois grandes approches : **expressions régulières basées sur des heuristiques**, **algorithmes d'apprentissage** et des **systèmes à base de connaissances**. La première approche fondée sur des expressions régulières basées sur des heuristiques, comme le définissent (HUANG, HO et al. 2004), consiste à extraire des séquences redondantes *via* l'utilisation d'algorithmes heuristiques. Ces séquences permettent ensuite d'établir des modèles pour l'analyse des séquences (CABANAC 2014). L'utilisation de ces approches basées sur des heuristiques sont un bon compromis entre la qualité des solutions trouvées et la rapidité de la mise en oeuvre du procédé, toutefois, ils ne garantissent pas nécessairement des solutions optimales selon la complexité de la tâche à réaliser (DECONINCK 2010). La seconde approche fondée sur des algorithmes d'apprentissage emploie des algorithmes qui opèrent en construisant un modèle basé sur un corpus donné en entrée. Le modèle construit procède ensuite à des prédictions ou des décisions afin d'analyser les séquences. Bien que ces approches aient une bonne capacité d'adaptation et de bons résultats (ANZAROOT et MCCALLUM 2013), elles ont des limites notamment au niveau du corpus d'apprentissage souvent hautement dépendant des données d'application. La troisième approche fondée sur des systèmes à base de connaissances, construit une ontologie dédiée à la description des données d'intérêt. Cette connaissance va alors inclure des informations comme les relations, les caractéristiques lexicales et des mots clés contextuels (REZAEI et MUNTZ 2013). Par le biais d'une analyse de l'ontologie, plusieurs règles et extracteurs peuvent être ainsi générés. Ces règles et extracteurs sont ensuite utilisés pour effectuer l'extraction de l'information. Cependant, l'expertise d'un expert du domaine est nécessaire afin de maintenir la base de connaissances à jour.

Dans le cadre de nos travaux, nous avons choisi d'utiliser une approche basée sur des algorithmes d'apprentissage car ces derniers représentent une alternative très attrayante à la construction manuelle de règles d'extraction. De plus, la littérature nous a permis d'attester de leurs bonnes adaptabilités ainsi que leur bonne performance pour l'étiquetage des séquences (LAFFERTY, MCCALLUM et al. 2001 ; COUNCILL, GILES et al. 2008). Dans les sections suivantes, nous examinons trois techniques majeures d'apprentissage automatique : les modèles de Markov cachés (MMC), les champs aléatoires conditionnels (CRF) et les machines à vecteurs de support (SVM).

2.2.2.1. Modèle de Markov caché

Le modèle de Markov caché (ou Hidden Markov Model -HMM-) (RABINER 1989) est l'une des techniques traditionnelles utilisée lors de l'étiquetage de séquences. Il s'agit d'un modèle probabiliste génératif qui suppose un processus de

Markov [13] dans la modélisation des données séquentielles. La fonction objective des MMC est la distribution conjointe des entrées et des étiquettes. Cela signifie trouver un ensemble optimisé de paramètres et de distributions qui maximisent la distribution des mots co-occurrents et des étiquettes indiquées dans les données d'entrée. Une séquence d'entrée, appelée séquence d'observations, s'écrit : $\mathbf{x} = \{x_1, x_2, ..., x_T\}$. Une séquence d'étiquettes appelée suite d'états se note comme suit : $\mathbf{y} = \{y_1, y_2, ..., y_T\}$. Un ensemble de N échantillons est écrit : $D = \{\mathbf{x}^{(i)}, \mathbf{y}^{(i)}\}_{i=1}^{N}$. Dans les MMC, la relation entre des variables aléatoires est simplifiée telle que chaque état y_{t-1} et chaque observation x_t dépendent uniquement de l'état correspondant y_t. On peut aussi dire que les observations dans un MMC sont générées par des états cachés. Le mot caché ne signifie pas que la valeur des étiquettes n'est pas donnée même dans l'ensemble d'entraînement. Cela signifie que la nature de la séquence d'état modélisée est seulement accessible *via* la séquence d'observations. La distribution conjointe d'une séquence d'observations et d'une séquence d'états peut s'écrire : $p(\mathbf{y}, \mathbf{x}) = \prod_{t=1}^{T} p(y_t|y_{t-1})p(x_t|y_t)$. Une fois le modèle appris *via* l'application d'un algorithme approprié, l'attribution d'étiquette la plus probable est calculée pour une séquence d'observations donnée : $\mathbf{y}^* = \arg\max_{y \in \mathbf{y}} p(y|\mathbf{x})$.

2.2.2.2. Champs aléatoires conditionnels

Dans sa définition originale, un champ aléatoire conditionnel (Conditional Random Fields –CRF–) est un processus stochastique qui modélise les dépendances entre un ensemble d'observations discrètes réalisées sur une séquence discrète (à l'origine, une séquence de mots) et un ensemble d'étiquettes (analyse morphosyntaxique). Un CRF est un modèle probabiliste discriminant qui ne fait pas l'hypothèse qu'une observation conditionnée par son étiquette est indépendante des observations voisines (LAFFERTY, MCCALLUM et al. 2001). En nous appuyant sur (SUTTON et MCCALLUM 2011), nous rappelons les propriétés principales d'un CRF, telles que :
$\mathbf{X} = x_1, x_2, ..., x_T$ est une séquence de T observations discrètes.
$\mathbf{Y} = y_1, y_2, ..., y_T$ est la séquence des T étiquettes associées aux observations de la séquence X.
L est l'ensemble des étiquettes possibles (les valeurs possibles pour les y_t).
O est l'ensemble des observations (les valeurs possibles pour les x_t, par exemple un lexique discret si les observations sont des mots).
Un CRF est défini par :

$$p(\mathbf{Y}|\mathbf{X}) = \frac{1}{Z(\mathbf{x})} \exp\left(\sum_{t=1}^{T}\sum_{k=1}^{K} \lambda_k f_k\left(y_{t-1}, y_t, x, t\right)\right) \tag{2.1}$$

13. En probabilité, un processus stochastique vérifie la propriété de Markov si et seulement si la distribution conditionnelle de probabilité des états futurs, étant donnés les états passés et l'état présent, ne dépend en fait que de l'état présent et non pas des états passés (NORRIS 1998).

Comme nous pouvons le voir, la probabilité qu'une séquence d'étiquettes particulière **Y** soit associée à la séquence d'observations **X** est obtenue par une combinaison linéaire de poids λ_k associée à des fonctions f_k sur la séquence d'observations. Dans le cas d'un CRF discret, ces fonctions sont généralement binaires. Les poids (ou potentiels) λ_k sont les paramètres du modèle et peuvent être interprétés comme l'importance ou la fiabilité de l'information apportée par la fonction binaire f_k. $f_k(y_{t-1}, y_t, x, t)$ est la notation générale des fonctions f_k appelées fonctions de caractéristique, qui rendent compte chacune de l'occurrence d'une combinaison d'observation(s) et de label(s) particulière. Par exemple :

$$f_k\left(y = l_i, x = o_j\right) = \begin{cases} 1 \text{ si } y_t = l_i \text{ et } x_t = o_j \\ 0 \text{ sinon} \end{cases} \tag{2.2}$$

Les fonctions de caractéristique sont définies par l'utilisateur. Elles reflètent la connaissance de l'utilisateur dans le domaine d'application. En ne faisant pas d'hypothèse d'indépendance des observations entre elles conditionnellement à leurs étiquettes, le modèle CRF permet de déterminer tout un ensemble de caractéristiques descriptives. Ces fonctions de caractéristique descriptives sont établies à l'aide de modèles de combinaison contextuels (ou pattern dans la littérature anglophone) entre observation(s) et étiquettes(s). Un modèle de combinaison caractérise une combinaison contextuelle d'une ou plusieurs observations et d'une ou plusieurs étiquettes. Par exemple, le modèle de combinaison contextuel $f(y_t, x_t)$ va prendre en compte tous les couples (étiquette, observation) à chaque position t dans la séquence. Ce modèle peut générer au maximum $|O| \times |L|$ fonctions de caractéristique binaires.

2.2.2.3. Machines à vecteurs de support

Les Support Vector Machines, souvent traduits par l'appellation de Séparateur à Vaste Marge (SVM), sont une classe d'algorithmes d'apprentissage initialement définis pour la discrimination c'est-à-dire la prévision d'une variable qualitative binaire (Vladimir VAPNIK 1999). Ils ont été ensuite généralisés à la prévision d'une variable quantitative. Dans le cas de la discrimination d'une variable dichotomique, ils sont basés sur la recherche de l'hyperplan de marge optimale qui, lorsque c'est possible, classe ou sépare correctement les données tout en étant le plus éloigné possible de toutes les observations. Le principe est donc de trouver un classifieur, ou une fonction de discrimination, dont la capacité de généralisation (qualité de prévision) est la plus grande possible. La première idée clé est la notion de marge maximale, i.e. la distance entre la frontière de séparation et les échantillons les plus proches est maximisée (BURGES 1998 ; GUYON, WESTON et al. 2002). L'autre idée maîtresse des SVM est de transformer, grâce à une fonction noyau, l'espace de représentation des données d'entrée en un espace de plus grande dimension, dans lequel il est probable qu'il existe un sé-

parateur linéaire. Formellement, considérons l'ensemble d'apprentissage suivant composé de n exemples et p attributs :

$$D = (x_i, c_i) \, | x_i \in \mathbb{R}^p, c_i \in [-1, +1], i \in [1, n] \tag{2.3}$$

Où, c_i indique la classe à laquelle le vecteur réel $p-$ dimensionnel x_i appartient. N'importe quel hyperplan qui peut diviser les points ayant $c_i = +1$, à partir de ceux ayant $c_i = -1$, peut être écrit comme un ensemble de points qui vérifient la condition suivante :

$$w \cdot x - b = 0 \tag{2.4}$$

Où, w est normal à l'hyperplan, $|b|/||w \cdot x||$ est la distance perpendiculaire de l'origine à l'hyperplan. $||w||$ est la norme euclidienne de w et b est une constante. La recherche de la marge optimale permettant de déterminer les paramètres w et b de l'hyperplan conduit à un problème d'optimisation quadratique qui consiste (dans le cadre général) à minimiser :

$$||w||^2 + C \sum_i \epsilon_i | y_i \left(w.\Phi \left(x_i \right) + b \right) \geq 1 - \epsilon_i, \epsilon_i \geq 0 \tag{2.5}$$

Où C est un paramètre de compromis entre la marge et les erreurs [14], ϵ_i est une variable ressort associée à l'observation x_i , et Φ est une transformation. Le problème peut être résolu par la méthode Lagrangienne d'optimisation quadratique avec contraintes (formulation duale) pour maximiser la marge (V. N. VAPNIK et Vlamimir VAPNIK 1995).

$$\sum_i \alpha_i \left(\frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j K \left(x_i x_j \right) | 0 \leq \alpha_i \leq C, \sum_i \alpha_i y_i = 0 \right) \tag{2.6}$$

Où, α_i est le multiplicateur Lagrangien associé aux vecteurs x_i. Si la valeur de α_i est non-nulle alors x_i est un vecteur de support et $K(x_i, x_j) = \Phi(x_i).\Phi(x_j)$ est le noyau de transformation. Le noyau d'un SVM est une fonction symétrique définie-positive qui permet de projeter les données dans un espace transformé de grande dimension dans lequel s'opère plus facilement la séparation des classes.

La décision est obtenue selon (le signe de) la fonction :

$$f \left(\mathbf{x} \right) = sign \left[\alpha_i y_i k \left(x_i, x \right) + b \right] \tag{2.7}$$

Dans le cadre de nos travaux, nous avons choisi d'utiliser un modèle discriminant (CRF) plutôt qu'un modèle génératif (MMC) bien que les deux algorithmes soient efficaces, plusieurs différences sont observables. Premièrement, les CRF permettent de modéliser un grand nombre de caractéristiques. De plus, les caractéristiques modélisées peuvent représenter des attributs pour une même ob-

14. Le choix de C est critique si les données sont bruitées.

servation à différents niveaux de granularité. À l'inverse, les MMC sont limités sur les types de fonctionnalités pouvant être modélisés. Deuxièmement, les MMC effectuent la normalisation par état, tandis que les CRF effectuent la normalisation sur l'ensemble de la séquence. Troisièmement, la différence la plus notable, est au niveau de la distribution qui est conjointe dans le cas des MMC et conditionnelle pour les CRF. Pour une distribution conjointe $p(x, y)$, un modèle $p(x)$ est généré tandis que pour une distribution conditionnelle $p(y|x)$ la modélisation de $p(x)$ n'est pas nécessaire. Cette différence est significative car la modélisation de $p(x)$ contient souvent de nombreuses caractéristiques hautement dépendantes qui sont difficiles à modéliser (SUTTON et MCCALLUM 2006). De plus, les modèles génératifs efficaces nécessitent souvent des hypothèses strictes d'indépendance conditionnelle (SUTTON 2008). Les caractéristiques non indépendantes des entrées, telles que la capitalisation, les suffixes et les mots environnants, sont importantes pour traiter les mots invisibles dans la formation, mais ils sont difficiles à représenter dans les modèles génératifs (SHA et F. PEREIRA 2003).

Concernant les SVM, nous n'avons pas opté pour les SVM pour étiqueter les séquences, bien qu'ils exhibent de meilleures performances (dans leur version structurée) que les CRF pour détecter les entités nommées (TANG, Y. FENG et al. 2015). Nous avons souhaité modéliser, au vu de nos données, des caractéristiques descriptives riches or, dans la littérature les performances obtenues par les SVM lors de l'inclusion d'un trop grand nombre de caractéristiques descriptives sont moins bonnes que celles obtenues *via* les CRF (D. LI, KIPPER-SCHULER et al. 2008). De plus, les CRF peuvent considérer toutes les dépendances liées aux transitions d'état ainsi que celles liées à la définition de caractéristiques d'entrée (caractéristiques contextuelles) tandis que les SVM ne considèrent pas de telles dépendances.

Pour résumer, les CRF offrent plusieurs avantages par rapport aux MMC pour de telles tâches, y compris la capacité d'assouplir les hypothèses d'indépendance fortes faites dans ces modèles. Alors que les MMC doivent faire des hypothèses d'indépendance très strictes sur les observations (LAFFERTY, MCCALLUM et al. 2001). Concernant les SVM, nous avons pu noter que par rapport aux CRF l'inclusion d'un nombre trop important de caractéristiques descriptives présentait de moins bonnes performances. De ce fait, suite à ces observations et au vu des données utilisées dans nos travaux, notre choix s'est porté sur l'emploi des CRF qui est mieux adapté à l'inclusion de riches fonctionnalités pouvant se chevaucher. Nous avons choisi d'utiliser plus particulièrement des cascades de CRF (CCRF) qui reprennent les principes généraux présentés dans la section 2.2.2.2 tout en permettant une hiérarchisation des processus (OLLAGNIER, FOURNIER et al. 2016). En ce qui concerne les SVM bien qu'utilisés pour le marquage des séquences, nous avons décidé de l'utiliser dans le cadre de l'apprentissage automatique d'un classifieur que nous souhaitons combiner avec les CCRF afin d'établir

un modèle hybride pour le traitement des références allusives présentes dans le corps du texte.

Suite à l'élaboration de ces différents algorithmes de nombreux outils, accessible *via* le Web, ont été élaborés afin de répondre spécifiquement à la tâche d'analyse des références bibliographiques. Ces outils sont présentés dans la section suivante.

2.2.2.4. Outils dédiés à la labéllisation de références bibliographiques

Nous introduisons, dans cette section, plusieurs outils d'analyse de références connus et accessibles en ligne. La plupart des outils recensés, dédiés à la tâche d'analyse des références bibliographiques, sont basés sur des algorithmes d'apprentissage, en particulier sur des CRF. Parmi ces outils, nous pouvons citer ParsCit [15] (COUNCILL, GILES et al. 2008) qui, sous la forme d'une boite à outil, permet l'annotation des références présentes dans les zones bibliographiques d'articles scientifiques. Cet outil, *via* des CRF entrainés sur le jeu de données Cora [16], fournit l'analyse de références à partir de différents formats de documents tels que pdf, xml, texte brut, etc. Biblio Citation Parser [17], anciennement appelé Paratool, a été developpé au cours de la même période que ParsCit dans le cadre du projet OpCit [18]. Malheureusement, les informations disponibles à son sujet n'étant plus accessibles nous n'avons pu trouver d'informations sur sa composition, nous savons juste que ce dernier permet l'annotation des références présentes dans les zones bibliographiques d'articles scientifiques. Freecite [19], également inspiré par ParsCit, utilise des CRF combinés à des bibliothèques CRF++. Tout comme ParsCit, l'ensemble de ses données d'apprentissage est basé sur le jeu de données Cora. L'une des différences notables avec ParsCit dans sa conception est l'inclusion de caractéristiques provenant d'informations lexicales établies à partir du *Directory of Research and Researchers at Brown* [20] (DRR-B). Grobid [21] (LOPEZ 2009), quant à lui, permet d'extraire, analyser et restructurer des informations bibliographiques issues de publications techniques et scientifiques à partir de documents bruts tels que les PDF. *Via* cet outil, il est possible d'obtenir des documents struturés au format TEI. Un point intéressant est que cet outil peut enrichir l'annotation *via* l'utilisation de données externes comme *Crossref* [22]. Bibpro [23] (C.-C. CHEN, K.-H. YANG et al. 2008) capture les propriétés structurelles

15. http://aye.comp.nus.edu.sg/parsCit/
16. https://hpi.de/naumann/projects/repeatability/datasets/cora-dataset.html
17. http://paracite.eprints.org/developers/
18. http://opcit.eprints.org/
19. http://freecite.library.brown.edu/welcome
20. https://vivo.brown.edu/
21. https://github.com/grobid/grobid
22. https://www.crossref.org/
23. https://github.com/ice91/BibPro

et transforme des propriétés en un modèle de séquences. Son modèle est basé sur l'alignement de séquences, des algorithmes d'apprentissage et des systèmes à base de connaissances. Bilbo [24] (Y.-M. KIM, BELLOT et al. 2012), dont les travaux menés au cours de cette thèse s'inscrivent dans le cadre de son développement, permet l'annotation des références structurées présentes dans les zones bibliographiques et dans les notes de bas de page. Ses modèles sont basés sur l'utilisation de CRF mais également de SVM dédiés à l'identification des notes de bas de page contenant des références bibliographiques structurées. D'autres outils existent comme, Pdf-extract [25] qui permet d'identifier et extraire des régions sémantiquement significatives d'un article en PDF. Cet outil utilise une technique « visuelle » basée sur un ensemble d'heuristiques permettant d'identifier les zones sémantiquement importantes d'un fichier PDF. De nombreux autres outils ont été réalisés afin de permettre également l'analyse de références. Parmi ces outils nous pouvons citer : le service web *AnyStyle* [26] qui permet d'analyser les références académiques dans le but de les rendre compatibles avec des logiciels de gestion de références (Endnote, Zotero, etc.) ou encore *Gargantext* [27] qui permet, entre autres, à partir d'un corpus de données textuelles une exploration avancée ainsi qu'une visualisation des réseaux.

Dans la section suivante, nous introduisons deux études effectuées sur l'analyse des caractéristiques bibliographiques au sein de la littérature scientifique.

2.3. Identification des caractéristiques des références bibliographiques

Dans cette section, nous présentons, tout d'abord, une première étude générale, dont l'objectif est de mettre en perspective les différents facteurs influançant la composition ainsi que la structure des références présentes au sein d'articles scientifiques, tous domaines confondus. Une seconde étude, plus spécifique, effectuée dans le cadre d'un cas pratique qui est l'exploitation des données d'OpenEdition, est ensuite présentée. À partir de ces données nous tentons de dégager des caractéristiques propres aux références allusives.

24. https://github.com/OpenEdition/bilbo
25. http://labs.crossref.org/pdfextract/
26. https://anystyle.io/
27. https ://gargantext.org/

2.3.1. Étude sur les références bibliographiques au sein de la littérature scientifique

Au cours de cette section nous présentons une étude sur les facteurs influençant la structure ainsi que la composition des références dans la littérature scientifique. Afin d'éliminer toute ambiguïté, nous définissons tout d'abord ce que nous considérons comme une **référence bibliographique**. Une référence bibliographique correspond à l'ensemble des éléments de données nécessaire permettant l'identification d'un document ou d'une partie de tout document sur tout support (livre, article, site Web, etc.) (MALCLÈS 1977). Moins complète qu'une notice bibliographique ou un catalogue, couramment utilisés au sein des bibliothèques traditionnelles ou numériques, les références sont utilisées dans les bibliographies, les notes de bas de page ou le corps du texte. Au cours de ce chapitre, nous utiliserons le terme « référence bibliographique » en tant que notion correspondant à la définition donnée précédemment.

Nous avons pu constater que le format des références suit généralement des conventions stylistiques définies par un certain nombre d'organisations (*Crosscite* [28] recense 1178 formats de référence). Certaines conventions sont plus largement utilisées selon les domaines de recherches comme celle de l'institut d'ingénieurs en électricité et électronique (*Institute of Electrical and Electronics Engineers*, IEEE [29]) orientée sur l'application de styles propres aux génies informatiques, électriques et logiciels ou encore celle de l'association américaine de psychologie (*American Psychological Association*, APA [30]) focalisée sur l'application de styles propres à d'autres domaines d'ingénierie. Cependant force est de constater une marge de personnalisation, un certain nombre d'organisations ont établi des conventions stylistiques spécifiques afin de répondre à leurs besoins. Parmi les conventions existantes, deux catégories se distinguent : les styles orientés vers les SHS et ceux pour les sciences dures [31] (bien que l'on observe un chevauchement considérable entre ces deux catégories). Le tableau 2.1 présente quelques styles utilisés par ces deux catégories.

Dans les sections qui suivent, l'utilisation d'une convention stylistique est pré-

28. https://citation.crosscite.org/
29. http://www.ieee.org/documents/ieeecitationref.pdf
30. https://www.library.cornell.edu/research/citation/apa
31. Nous avons choisi d'employer l'expression populaire « sciences dures », couramment utilisée dans la littérature, afin de désigner au sein d'un même ensemble les sciences de la nature et les sciences formelles (SOLER 2009).
32. http://www.chicagomanualofstyle.org/home.html
33. https://www.library.cornell.edu/research/citation/mla
34. http://www.mhra.org.uk/Publications/Books/StyleGuide/download.shtml
35. http://library.williams.edu/citing/styles/acs.php
36. http://www2.ametsoc.org/ams/index.cfm/publications/authors/journal-and-bams-authors/journal-and-bams-authors-guide/references/
37. http://www.nlm.nih.gov/pubs/formats/recommendedformats.html

Sciences humaines et sociales

Dénomination de la convention stylistique	Description
Association américaine de psychologie (APA)	couramment utilisée dans les domaines des sciences humaines et des sciences du comportement.
Chicago Manuel de Style (CMOS [32])	utilisée en histoire, économies et sciences sociales.
Association des langues modernes (MLA [33])	employée dans les arts, sciences humaines et surtout dans les études littéraires anglaises.
Association de recherche sur les humanités modernes (MHRA [34])	plus largement utilisée dans les arts et les humanités au Royaume-Uni.

Sciences mathématiques, ingénieries, physiologiques et médicales

Dénomination de la convention stylistique	Description
Société américaine de chimie (ACS [35])	utilisée en chimie et physique.
Institut d'ingénieurs en électricité et électronique (IEEE)	employée dans les domaines des sciences informatiques, électriques et logiciels.
Société américaine de mathématique (AMS [36])	couramment utilisée dans les domaines des sciences mathématiques.
Bibliothèque nationale de médecine (NLM [37])	utilisée dans le domaine médical et en particulier au sein de la plate-forme PubMed.

Tableau 2.1. – Quelques conventions stylistiques en SHS et sciences dures

sentée comme un facteur affectant la structure ainsi que la composition des références bibliographiques. Nous étudions l'impact que génère la multiplication des conventions stylistiques, à la fois, sur la structure générale des références bibliographiques et sur le formatage des différents champs.

2.3.1.1. Les conventions stylistiques : facteur d'influence dans la composition et la structure des références bibliographiques

Comme souligné précédemment, chaque domaine scientifique utilise plusieurs conventions stylistiques spécifiques. Le choix d'un style implique de suivre des normes de formatage, notamment au niveau des champs bibliographiques, propres à chacun d'entre eux. Plusieurs champs, présentés dans les sections suivantes, sont régis par l'utilisation d'un style spécifique au domaine dont l'influence va être constatée, à la fois, dans leur composition et leur structure. Pour rappel, un champ bibliographique se rapporte aux différentes zones d'information composant une référence bibliographique (auteur, date, titre, etc.).

Formatage général

Le choix du style va définir le format général d'une référence bibliographique *via* la spécification des normes suivantes :

- **Les champs obligatoires et facultatifs**
  Les attributs standards pour chaque type de document ont leur propre motif de citation composé de champs obligatoires et facultatifs. La mention des champs obligatoires n'est requise que si elle est applicable au document indiqué et si l'information est facilement disponible à partir du document lui-même ou de la documentation d'accompagnement. Évidemment,

dans la pratique, les systèmes sont souvent confrontés à un ou plusieurs champs non renseignés même dans les publications imprimées, pourtant profitant d'une longue tradition éditoriale qui en a fixé des règles plus ou moins rigoureuses. De ce fait, la notion de champ obligatoire n'est pas une obligation, mais une recommandation. Quant aux champs facultatifs, ces informations doivent être incluses si les informations requises sont disponibles. Leur présence peut également signifier que les informations supplémentaires fournies, *via* l'ajout de ces champs facultatifs, sont pertinents afin de localiser la source. (KYHENG 2003) publie un tableau répertoriant les schémas de compositions des références par types de documents référencés dont un extrait est présenté dans le tableau 2.2.

Champs	**Ouvrage**	**Partie d'ouvrage**	**Contribution à un ouvrage**
Titre	obligatoire	obligatoire	obligatoire
Responsabilité principale du document hôte	-[38]	-	obligatoire
Titre du document hôte	-	obligatoire	obligatoire
Responsabilité secondaire[39]	facultatif	facultatif	-
Collection	facultatif	-	-
Type de support[40]	obligatoire	obligatoire	obligatoire

Tableau 2.2. – Exemple de schémas de compositions des références par type de documents

La première colonne identifie les champs d'information concernés, la seconde fait référence aux normes de citation pour un ouvrage, la troisième aux normes de citation établies pour la citation d'une partie d'un ouvrage et la dernière colonne fait référence aux normes de citation employées dans le cas d'une contribution à un ouvrage.

- **Périodicité des publications**
  La périodicité est un élément important qui entraîne la présence ou l'absence de certains champs bibliographiques. Trois types de périodicité, présentés dans le tableau 2.3, peuvent être distingués.

Ces catégories ont une incidence sur le nombre de champs présents dans

38. Le tiret signifie que ces champs ne doivent pas être renseignés
39. L'italique signifie que ces champs sont facultatifs pour tous types de documents.
40. Propre aux documents électroniques

Type	Description
Périodique	tout support d'information figurant en fascicules ou en volumes successifs (journaux, magazines, etc.).
Non-périodique	publication complète en un seul volume ou destinée à être complétée dans un nombre limité de volumes (textes, actes d'événements scientifiques, etc.).
Publications de systèmes de communications électroniques	périodicité difficilement identifiable : une liste de diffusion, par exemple, publiée périodiquement, mais dont la « périodicité » est assez irrégulière.

Tableau 2.3. – Les différents types de périodicité

une référence bibliographique telle que l'inclusion d'un champ dédié à la désignation du volume dans le cas d'une publication non périodique.

- **La cible d'une référence** Une référence peut renvoyer à une publication complète ou à l'une de ses composantes : partie (préface, introduction, chapitre, section, etc.) ou contribution (préface, article, message, etc.). Le tableau 2.4 illustre les configurations possibles selon la cible de la référence :

Cible	Exemples
Ouvrage	SCHLEIERMACHER, Friedrich Daniel Ernst. Herméneutique. Traduit de l'allemand par Christian Berner. Paris : Editions du Cerf ; Québec : Presses de l'Université Laval, 1987, XVIII-202 p. Collection Passages.
Partie d'ouvrage	SCHLEIERMACHER, Friedrich Daniel Ernst. Herméneutique. Traduit de l'allemand par Christian Berner. Paris : Editions du Cerf ; Québec : Presses de l'Université Laval, 1987. Les discours prononcés à l'Académie, p. 153-188
Contribution à un ouvrage	BERNER, Christian. Notes sur la présente édition. In SCHLEIERMACHER, F. D. E. Herméneutique. Paris : Editions du Cerf ; Québec : Presses de l'Université Laval, 1987, p. I-XVIII

Tableau 2.4. – Les différentes cibles d'une reference (convention stylistique utilisée : revue théologique de Louvain)

Telle que présentée dans ce tableau, la cible d'une référence dénote une réelle influence sur le nombre de champs présents dans une référence bibliographique.

- **Zone d'apparition**
Comme indiqué au début de cette section, des références peuvent être employées dans le corps du texte, dans les notes de bas de page ainsi que dans les zones bibliographiques. Les références se trouvant dans le corps du texte

dont la caractéristique est d'être disséminées dans ce dernier permettent la construction de l'argumentaire de l'auteur. Elle fournissent généralement les informations nécessaires à la localisation de la source présente dans la zone bibliographique du document. Ces références peuvent être exclues de la syntaxe du discours et prendre la forme d'une appel, i.e. (Berner, 1987), ou bien faire partie intégrante de la syntaxe de ce dernier, i.e. *« Ce n'est qu'en **1931**, avec la parution de son ouvrage **Le Symbolisme de la Croix**, que **René Guénon** dévoile sa filiation à l'ésotérisme soufi »*. Dans ce cas précis, leur utilisation ne répond à aucune norme relative à leur formatage. En ce qui concerne les références présentes dans les notes de bas de page, leur degré de précision peut varier selon la fonction de cette dernière. En effet, sa fonction est soit de citer seulement une référence, soit d'introduire des informations complémentaires, soit d'ajouter un commentaire. Chacune de ces fonctions peut contenir des références plus ou moins structurées. En guise d'illustration, prenons l'exemple du traitement des notes dans deux guides différents :

- Le style MHRA utilise les notes afin de citer toute référence sans avoir besoin de se référer à la zone bibliographique.
- Le style MLA utilise des références abrégées entre parenthèses, sous la forme d'un appel, indiquant l'auteur et la page (Smith 395), de sorte qu'il n'est pas nécessaire de consulter la bibliographie en lisant le reste des détails de la publication.

Certains styles tels que le style AMS utilisent même des abréviations alphanumériques, par exemple, [AB90] qui nécessitent un renvoi à la bibliographie.

Comme nous avons pu le constater suite à l'étude des facteurs influant sur le formatage général d'une référence bibliographique plusieurs aspects génèrent des particularités propres à la convention bibliographique choisie. Les principaux facteurs sont : les champs stipulés comme étant obligatoires et facultatifs, la périodicité de la publication, la cible de la référence ainsi que sa zone d'apparition. Chacun de ces aspects va influer de façon parfois combinatoire sur la structure de la référence, qui peut s'avérer plus ou moins formatée (cf : la zone d'apparition), et sur le nombre de champs composant la référence bibliographique (cf : la cible d'une référence).

Formatage des champs bibliographiques

Après avoir identifié un certain nombre de facteurs affectant le formatage général, nous allons déterminer quels sont les facteurs impactant la formation des champs ainsi que les éléments qui les composent.

- **Ordre des champs bibliographiques**

L'ordre d'apparition de l'intégralité des champs (auteur, date, titre, etc.) est propre à chaque convention bibliographique. Par exemple, dans le style IEEE, l'auteur est présenté avec le prénom suivi du nom. Tandis que dans le style MLA le nom de l'auteur est présenté en premier. D'autres exemples sont à signaler tels que la position de l'année de publication qui pour le style ACS est situé à la fin de la référence tandis que pour le style APA, il se situe après le nom de l'auteur. Cet ordre peut varier selon l'inclusion de spécificités évoquées lors du formatage général comme l'ajout de champs facultatifs.

- **Typographie et ponctuation**
  Comme énoncé précédemment, les références contiennent des champs spécifiques placés dans un ordre défini préalablement en fonction du style choisi. Une subtilité vient également s'ajouter à ce formalisme, une typographie particulière associée à la ponctuation standard qui va régir la structure, à la fois, entre chaque champs et à l'intérieur de ces derniers. Cependant, bien que toutes les typographies et les ponctuations puissent être utilisées, leurs utilisations répondent à une logique globale :
  - utilisation des mêmes caractères pour toute la liste ;
  - respect des mêmes styles d'écriture ;
  - suivre une ponctuation cohérente, à la fois, dans la séparation des champs ainsi qu'à l'intérieur de ces derniers. Dans l'approche standard, chaque champ est séparé du suivant par une seule marque de ponctuation (suivi d'un espace, d'une virgule, suivi d'un espace, d'un tiret et suivi et précédé d'un espace, etc.). Dans un champ, les occurrences différentes sont séparées par le même signe de ponctuation généralement une virgule suivie d'un espace.

Suite à cette étude, nous pouvons constater de nombreuses variations possibles au sein même des champs bibliographiques. Les principaux facteurs de variation sont : l'ordre des champs, l'emploi d'une typographie, la ponctuation. Chacun de ces aspects combiné aux facteurs influant sur le formatage général, va engendrer de nombreuses variations à la fois, sur la structure et la composition de la référence.

Cette étude nous permet de constater que les références peuvent apparaître comme des éléments structurés, cependant, de nombreux facteurs influençant leur structure rendent la généralisation des moyens de traitrement plus complexe. En effet, cette analyse a permis de révéler qu'en plus du nombre important de conventions existantes s'ajoutent des personnalisations dont la portabilité impacte tous les constituants d'une référence. Ces constatations permettent d'illustrer les difficultés qui résident dans la création d'un processus permettant un traitement de ce type de données.

Dans la section suivante, nous allons étudier plus particulièrement les références bibliographiques allusives afin d'identifier la présence de facteurs influençant leur formatage.

2.3.2. Cas pratique : analyse des références bibliographiques allusives dans OpenEdition

Comme nous venons de le voir, bien que des conventions stylistiques soient utilisées afin de normaliser la présentation d'œuvres écrites ainsi que leur citation, différents facteurs influent sur leur composition et leur structure. Parmi ces facteurs, nous avons pu constater que la zone d'apparition de la référence engendre de nombreuses représentations pouvant être plus ou moins struturées. Les études menées sur les écrits scientifiques ont permis d'établir l'existence de références pouvant être intégrées ou non dans la syntaxe du discours. Par le biais de ces références, l'auteur peut véhiculer différents objectifs comme l'adhésion à une communauté ou à un courant spécifique (SWALES 1990) mais aussi afin de mettre en perspective la nouveauté d'une approche (TUTIN 2010). Tous ces éléments peuvent être véhiculés au travers de ces références ce qui les investit d'une valeur rhétorique et argumentative.

Dans le cadre de nos travaux, nous nous sommes penchée sur l'étude de ce type de références que nous qualifions de **références bibliographiques allusives**. Comme nous venons de le souligner, leur caractéristique est d'être disséminées dans le texte selon un degré plus ou moins fort d'implicite. Par la notion d'implicite nous sous-entendons des références qui ne sont pas énoncées de façon formelle. En effet, nous avons pu observer l'utilisation de références qui ne s'effectue pas dans le cadre d'un renvoi, par le biais d'informations exhaustives, à une unité documentaire. Ces références sont plutôt utilisées afin de ponctuer l'argumentaire de l'auteur, elles font donc partie intégrante du discours (sans être pour autant incluses dans la syntaxe de ce dernier) et ne sont donc pas toujours formulées aussi explicitement que les références que nous pouvons trouver dans les zones bibliographiques. Au cours des sections qui suivent, nous allons tenter de les caractériser *via* l'étude des données fournies par OpenEdition établie par le biais de l'analyse du corpus que nous présentons dans la section 2.5.1.

2.3.2.1. Existe t-il différents types de références bibliographiques allusives ?

Suite à l'étude du corpus dédié aux références allusives fournis par OpenEdition, composé de 42 articles scientifiques issus de Revues.org, que nous présenterons plus en détail dans la section 2.5.1, nous avons constaté que lors de la rédaction d'un article scientifique, l'auteur cite des références afin de ponctuer son argumentaire tout au long du document et ces références peuvent prendre de nombreuses formes. Parmi elles, certaines sont des références explicites comme

celles présentes à la fin des articles ou des livres, tandis que d'autres références, que nous avons qualifiées de références bibliographiques allusives, sont disséminées dans le texte ou dans les notes de bas de page selon un degré d'implicite plus ou moins fort. Compte tenu de ces observations, nous avons pu établir une typologie basée sur la variation des schémas normatifs de composition des références bibliographiques allusives en fonction du degré d'implicite constaté.

- **Références bibliographiques allusives suivant un schéma normatif** : plus couramment utilisées dans les notes de bas de page lors de l'apparition de la première occurrence. Ce type de références fournit un certain nombre d'éléments informationnels permettant d'identifier et localiser l'unité documentaire correspondante. La structure de ces références s'approche des normes bibliographiques qui les rendent plutôt explicites. Elles consistent en une succession de champs bibliographiques fournissant des informations d'identification très précises et détaillées et ne s'intègrent généralement pas à la syntaxe du discours. La figure 2.1 présente un exemple de référence suivant un schéma normatif extrait du corpus annoté présenté dans la section 2.5.1.

```
<bibl><author>DRIESMAN Dominique</author>, <title>Les principes de Cesare Brandi appliqués
à la conservation-restauration de la céramique et du verre : progression, évolution,
révolution</title>, dans <bibl><author>GESCHE-KONING Nicole</author> et <author>
PERIER-D'IETEREN Cathelin</author> (éd.), <title>Cesare Brandi (1906-1988) : Sa pensée et
l'évolution des pratiques de restauration</title>, <title> Série spéciale des Annales
d'Histoire de l'Art et Archéologie de l'ULB</title>, <title>Cahier d'études</title>
<biblScope>X</biblScope></bibl>.</bibl>
```

Figure 2.1. – Exemple d'une référence bibliographique allusive suivant un schéma normatif

- **Références bibliographiques allusives suivant partiellement un schéma normatif** : ces références se retrouvent, à la fois, dans les notes et dans le texte. Les références affiliées à ce taxon ont la particularité d'avoir une partie de leurs éléments dont la distribution est aléatoire tandis que l'autre partie suit une norme bibliographique. Comme présenté dans la figure 2.2, également extraite du corpus annoté présenté dans la section 2.5.1, cette référence présente des éléments bibliographiques disséminés dans le corps du texte intégrés à la syntaxe du discours auxquels sont rattachés des informations d'identification plus précises mais aussi plus normées qui ne sont cette fois pas incluses dans la syntaxe du discours. Avec cette configuration, un degré plus élevé d'implicite se pose.

```
Dans <bibl><title>Les formes de la vie religieuse</title> (<author><surname>Durkheim</surname>
</author>, <date>1998</date> [<date>1912</date>])</bibl>
```

Figure 2.2. – Exemple d'une référence bibliographique allusive suivant partiellement un schéma normatif

- **Références bibliographiques allusives ne suivant pas de schéma normatif** : plus couramment utilisées dans le corps du texte, ce type de références présente une distribution non régulière de ses éléments bibliographiques. Elles peuvent être composées d'un à plusieurs éléments. Ces références présentent un degré important d'implicite car en plus d'utiliser divers champs bibliographiques, leur distribution peut s'étendre sur plusieurs lignes ou paragraphes. Ce type de références fait partie intégrante de la syntaxe du discours et les informations relatives à l'identification de l'unité documentaire sont décrites *via* le langage naturel par l'auteur. La figure 2.3 présente un exemple de référence ne suivant pas de schéma normatif extrait du corpus annoté présenté dans la section 2.5.1.

```
Ce n'est qu'en <bibl><date>1931</date>, avec la parution de son ouvrage <title>Le Symbolisme de
la Croix</title>, que <author>René Guénon</author></bibl> dévoile sa filiation à l'ésotérisme
soufi.
```

Figure 2.3. – Exemple d'une référence bibliographique allusive ne suivant pas de schéma normatif

Cette typologie permet d'identifier trois types de structure différents. Nous avons pu observer que ces caractéristiques structurelles sont fortement liées au degré d'implicite de la référence. En effet, nous avons pu noter une corrélation entre le degré d'intégration à la syntaxe du discours et la structure des informations bibliographiques. Comme nous avons pu le relever les informations bibliographiques relatives aux références allusives ne suivant pas de schéma normatif sont énoncées au sein même du discours et font parties intégrante de la syntaxe de ce dernier. De ce fait, la distribution des éléments bibliographiques n'est pas régulière ce qui engendre un degré d'implicite plus important. À l'inverse dans le cas des références bibliographiques allusives suivant un schéma normatif, les éléments bibliographiques s'approchent des normes bibliographiques et ne sont pas intégrées à la syntaxe du discours ce qui les rend explicites.

La section suivante se penche sur l'analyse des références bibliographiques allusives, à la fois, au niveau de leurs répartitions et de leurs compositions.

2.3.2.2. Études fréquentielles sur la répartition et la composition des références bibliographiques allusives

Dans cette section, nous présentons deux études fréquentielles basées sur l'analyse des références allusives dans le corpus OpenEdition dont nous présenterons les caractéristiques au cours de la section 2.5.1. La première étude consiste à quantifier l'utilisation des références allusives au sein d'articles scientifiques, tandis que la seconde étude se focalise sur une analyse détaillée de la composition de ces dernières. Chacune de ces analyses est menée au préalable sur les para-

graphes et ensuite sur les notes de bas de page. Cette distinction, nous permettra de voir si des caractéristiques spécifiques se dégagent.

2.3.2.2.a. Étude fréquentielle de la répartition des références bibliographiques allusives

Pour établir la répartition des références bibliographiques allusives présentes dans le corps du texte, les documents ont été coupés en 10 portions égales en fonction du nombre moyen de mots dans les documents. La figure 2.4 présente la répartition de références bibliographiques allusives présentes dans le corps du texte en fonction de leur nombre.

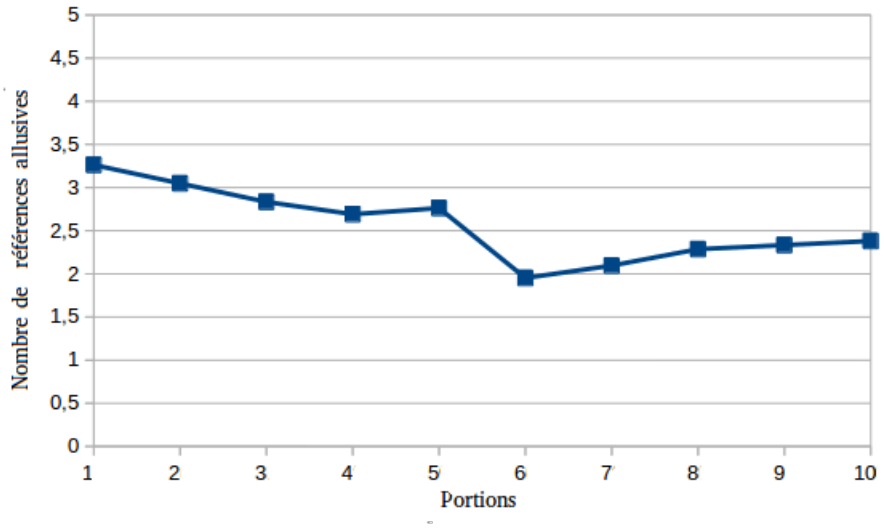


Figure 2.4. – Nombre moyen de références bibliographiques allusives dans le corps du texte

Cette figure nous permet d'observer que les références allusives sont présentes tout au long du document. Leur nombre varie entre 1,9 et 3,3 références allusives par portion. Une certaine constance émerge des portions 1 à 5 ainsi que des portions 7 à 10. Ces portions se réfèrent respectivement aux parties consacrées à l'introduction et à la conclusion. Ces parties sont généralement dédiées à la contextualisation du travail de recherche ainsi qu'à la valorisation des contributions scientifiques ce qui explique, entre autre, le taux plus important de références allusives dans ces zones. Le taux le plus faible se présente sur les portions médianes, ce phénomène peut s'expliquer par la méthodologie de rédaction propre à un article scientifique dans laquelle la partie centrale est souvent destinée à la présentation du sujet de recherche introduit par le biais d'une méthodologie, d'une approche ou encore d'expérimentations. Via ces constatations nous pouvons faire écho aux travaux présentés dans l'article (BERTIN et

ATANASSOVA 2014) au sujet de la structure standard IMRaD[41] dans lequel il est question d'une structure standardisée des articles scientifiques. Cependant, nous ne pouvons affirmer l'existence d'une telle rigueur dans l'organisation des éléments d'un article dans les SHS pour lesquels l'IMRAD n'est pas employé.

En ce qui concerne la répartition des références allusives dans les notes de bas de page, cette étude se concentre sur le nombre de notes contenant une ou plusieurs références bibliographiques allusives. La méthode de mesure correspond à celle utilisée pour mesurer le nombre de références dans le corps du texte. La figure 2.5 présente la répartion des références allusives dans les notes en fonction de leur nombre.

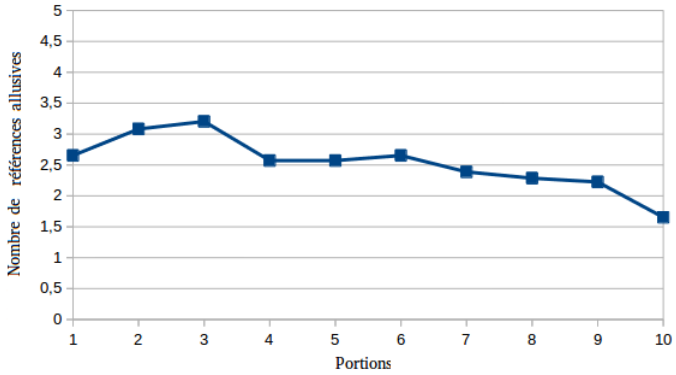


Figure 2.5. – Nombre moyen de références bibliographiques allusives dans les notes de bas de page

La figure 2.5 révèle une utilisation significative des références dans les notes de bas de page sur l'ensemble du document. Le taux de répartition varie entre 1,6 et 3,2 avec un pic plus important entre les portions 1 à 3. L'étude réalisée sur le nombre moyen de références allusives dans le corps du texte a illustré une forte utilisation des références dans l'introduction, cette même observation est faite pour les notes. Ce phénomène peut également s'expliquer par les mêmes arguments que l'étude précédente *via* l'utilisation de notes comme zone de citation d'une référence, d'inclure des arguments ailleurs que dans le corps du texte ou encore, afin d'ajouter un commentaire.

41. IMRaD (Introduction, Method, Result and Discussion) réfère à la struture Introduction, Méthode, Résultat et Discussion. Cette structure donne un aperçu rhétorique de l'écriture scientifique qui a commencé à prédominer en 1965. L'article de (SOLLACI et M. G. PEREIRA 2004) montre notamment que ce standard a été introduit comme norme en 1979 notamment dans le domaine biomédical.

Les figures 2.4 et 2.5, nous ont permis d'observer un nombre moyen de références bibliographiques plus important dans les premières parties. Bien que des règles tacites de composition des publications scientifiques prédisposent certaines parties, telles que l'introduction, à inclure un plus grand nombre de références allusives, l'utilisation de références allusives est faite tout au long du document. Cette présence récurrente en fait un motif d'exploitation en vue de s'en servir comme vecteur d'informations. De plus *via* cette configuration, un premier axe de travail se dessine concernant l'identification d'indicateurs structurels des références bibliographiques plutôt que des méthodes de positionnement permettant d'identifier des paragraphes spécifiques comme ceux dédiés à l'introduction.

2.3.2.2.b. Étude fréquentielle sur la distribution des champs bibliographiques dans les références bibliographiques allusives

Dans cette partie, l'objectif est d'observer la distribution des principaux champs bibliographiques au sein des références allusives aussi bien dans les paragraphes que dans les notes de bas de page. Nous avons choisi d'étudier la distribution des titres, des auteurs et des dates qui sont les champs les plus courants (cf : tableau 2.8). Bien que la ponctuation soit fortement présente, elle n'est pas un indicateur suffisant dans la définition de motifs se rapportant aux références allusives. En effet, au vu de la distribution irrégulière des informations bibliographiques selon le degré d'implicite des références nous nous sommes penchée sur des éléments bibliographiques communs à leurs schémas de compositions et ce, en tenant compte des différentes structures que nous avons pu relever au cours de la section 2.3.2.1. Afin d'établir le taux de distribution des champs bibliographiques, à la fois, dans les paragraphes et les notes, les mêmes portions que celles utilisées pour mesurer le taux de répartition des références allusives sont employées.

La figure 2.6 présente le taux de distribution des champs bibliographiques dans les paragraphes. Cette figure nous permet d'observer une fréquence d'utilisation sensiblement similaire entre les différents champs tout au long du document. Pour les trois champs, des fréquences supérieures sont constatées entre les portions 1 à 3. Les auteurs sont, cependant, plus largement utilisés. Leur distribution varie entre 1,6 et 2,7. Les auteurs sont largement cités à des fins d'argumention ou de comparaison. Il est d'usage de voir, une fois le travail d'un auteur ciblé, des références à la même œuvre ne citant que le nom de l'auteur. Bien que le champ auteur ait une fréquence substanciellement plus importante, nous notont un certain chevauchement. Ce phénomène s'explique par la fonction des références allusives présentes dans le texte. En effet, ce type de citation donne au lecteur un chemin à suivre pour trouver où chercher l'information empruntée.

La figure 2.7 présente les taux de distribution des champs bibliographiques dans les notes de bas de page. Contrairement à l'étude sur la fréquence des

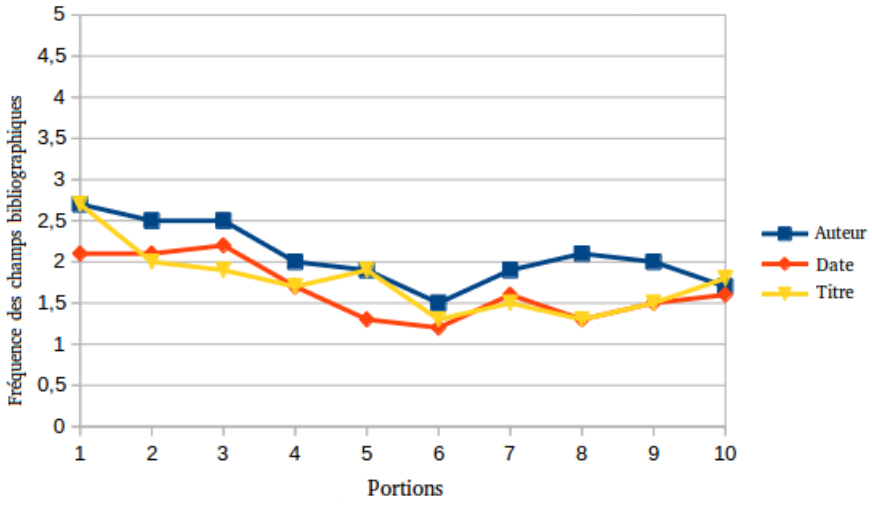


Figure 2.6. – Fréquence de distribution des champs bibliographiques dans les paragraphes

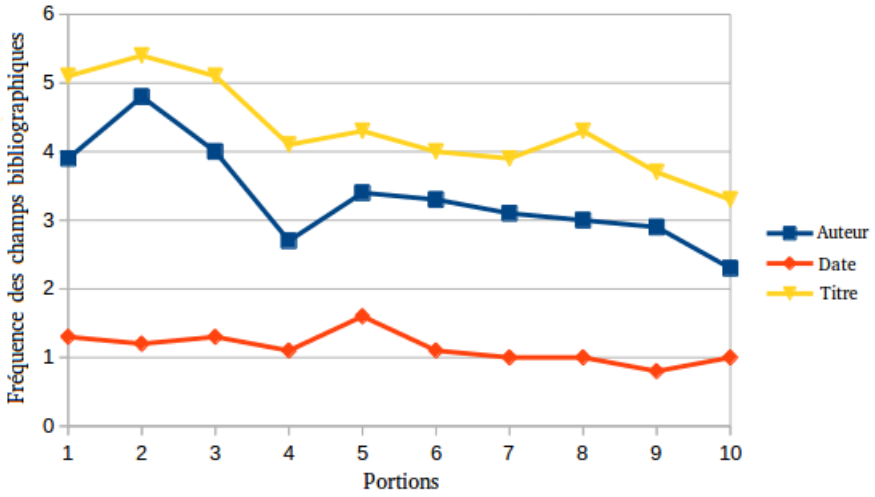


Figure 2.7. – Fréquence de distribution des champs bibliographiques dans les notes de bas de page

champs bibliographiques dans les paragraphes, les courbes montrent très clairement un taux de distribution différent entre chaque champ. Les titres sont plus largement utilisés dans les notes, leur répartition varie entre 3,3 et 5,4, avec un pic plus important entre les portions 1 à 3. Les auteurs sont également largement utilisés avec une oscillation comprise entre 2,3 et 4,7. Ce phénomène peut s'expliquer par le fait que la fonction d'une référence allusive au sein d'une note en SHS s'utilise principalement dans le contexte de la citation de livres dans le but de fournir des informations supplémentaires. Souvent, les notes permettent

de fournir un état de la recherche sur le sujet. En tant que tel, la présentation des références bibliographiques dans les notes est différente des références présentes dans une section bibliographique. En effet, les références dans les notes fournissent certains éléments d'identification alors que les références en fin de document sont complètes.

Une autre étude menée a permis d'estimer quelles sont les structures les plus redondantes au sein des références allusives selon leur zone d'apparition. Le tableau 2.5 présente le pourcentage d'apparition de ces structures (la somme des poucentages n'est pas de 100% car nous ne présentons pas l'intégralité des structures observées au sein des références bibliographiques allusives).

Position	Structure	Pourcentage
Texte	titre	35%
	titre + auteur	23%
Note	auteur + titre + date	7%
	auteur + titre	2%

Tableau 2.5. – Principales structures observées au sein des références bibliographiques allusives

En moyenne, deux éléments sont utilisés pour composer une référence allusive dans le texte ce qui restreint les variations possibles et donc explique les pourcentages plus importants. Tandis que pour les notes de bas de page, cinq éléments sont utilisés en moyenne ce qui augmente les combinaisons possibles entre les champs. De ce fait, un nombre d'éléments plus important engendre nécessairement des variations structurelles importantes qui expliquent des pourcentages plus faibles. Tandis que, les références dans le texte présentent des structures plus récurrentes. Cependant, ces éléments sont souvent éparpillés sous la forme de références non structurées ce qui les rend difficile à identifier. À l'inverse, les notes composées de plus d'éléments prennent souvent des formes plus structurées engendrées par une densité textuelle plus faible, à la différence de celle des paragraphes. Ces constatations nous permettent d'identifier un second axe de travail, tenant compte de l'utilisation redondante de certaines étiquettes, basé sur l'identification de traits discriminants permettant de caractériser ces étiquettes. Un troisième axe s'oriente sur l'exploitation de motifs permettant de capturer les amorces d'apparition de ces structures. Une perspective également intéressante se dégage concernant l'exploitation plus particulière des titres que nous retrouvons au sein de toutes les structures.

Ces études nous ont permis de mettre en exergue les différences de composition existantes au sein des références allusives présentes dans les paragraphes et dans les notes de bas de page. En effet, nous avons vu que les étiquettes présentes au sein des paragraphes étaient souvent utilisées de manière combinée

mais sur un espace textuel plus grand. À l'inverse, les notes sont plus largement ponctuées des étiquettes titre et auteur sur une densité textuelle plus faible. Cependant, des champs communs ainsi que des structures communes ont été recensés ce qui nous encourage à envisager des traitements plus génériques.

En tenant compte de ces considérations, nous présentons l'approche dédiée à l'identification des références bibliographiques allusives que nous avons réalisée et son intégration au logiciel Bilbo.

2.4. Approche supervisée dédiée à la détection des références bibliographiques allusives au sein de publications scientifiques

Comme nous avons pu le constater au cours de la section 2.2, l'extraction de l'information bibliographique est souvent limitée aux zones les plus formelles, à savoir les références présentes à la fin des articles ou des livres. Les résultats dans ce domaine montrent, par ailleurs, des performances très satisfaisantes (autour des 90 % de F-mesure pour les labels référant aux auteurs) (Y.-M. KIM, BELLOT et al. 2011 ; ANZAROOT et MCCALLUM 2013). Cependant comme nous avons pu le relever lors de la section 2.3, les références bibliographiques peuvent également se trouver à divers niveaux du document comme dans les notes de bas de page ou dans le corps du texte. Ce type de références que nous avons qualifié d'allusives se caractérise par des particularités à la fois compositionnelles et structurelles. L'étude de la littérature ne nous a pas permis de trouver d'outil permettant l'annotation des références allusives disséminées dans le corps du texte. Au cours des sections qui suivent, nous introduisons, premièrement, l'approche que nous avons mise en place afin de détecter les références allusives au sein d'articles scientifiques et en second lieu, son intégration au sein du logiciel Bilbo.

2.4.1. SVM et cascade de CRF dédiés à la détection de références bibliographiques allusives dans des articles scientifiques

Inspiré par les travaux menés dans le cadre du développement du logiciel Bilbo que nous présentons dans la section 2.5 et au vu des résultats probants obtenus dans l'identification de références structurées au sein d'article scientifique en SHS (Y.-M. KIM, BELLOT, TAVERNIER et al. 2012), nous avons opté pour l'utilisation d'un modèle SVM combiné à des CCRF. L'approche proposée consiste, d'une part, à identifier les paragraphes qui contiennent des références *via* un processus de classification supervisée et, d'autre part, dans l'application de CCRF afin de

détecter plus précisément les zones bibliographiques et d'annoter leurs contenus.

Compte tenu du grand nombre de paragraphes au sein des articles scientifiques, dont nous avons évalué la moyenne à 39 paragraphes par article sur la collection de revues d'OpenEdition, et dont une importante partie ne comporte pas de références bibliographiques (62,3 % en moyenne), nous avons décidé d'effectuer un pré-filtrage *via* l'utilisation d'une approche de classification supervisée. Ensuite, notre choix s'est porté sur l'utilisation de CCRF qui nous permet de tenir compte distinctement de deux facteurs influençant les performances d'analyse des références, à savoir, le positionnement et la composition des références. De plus, les modèles CRF, de par leur portabilité et leur capacité d'inclure de riches caractéristiques, sont les plus adaptés au traitement des références bibliographiques issues d'un domaine aux thématiques variées.

En moyenne un paragraphe est composé de 219 mots et 6 phrases ce qui nous donne une densité textuelle relativement importante et qui associée à un degré plus ou moins fort d'implicite engendre des variations de répartitions importantes. De plus, nous avons vu au cours de la section 2.3.2.2.b que le nombre d'éléments peut osciller en moyenne entre deux et cinq éléments ce qui engendre également de nombreuses variations structurelles. Le fait d'utiliser un premier CRF, qualifié de CRF de bas niveau, destiné à l'identification des zones d'apparition des références bibliographiques nous permet d'affiner la zone d'application du second CRF, qualifié de CRF de haut niveau. Nous supposons que cette méthode peut permettre de diminuer la propagration des erreurs dans la chaîne de traitement.

La figure 2.8 présente une modélisation de la méthode de traitement des références allusives. Comme nous pouvons l'observer, lors de la phase d'apprentissage, trois processus sont exécutés consécutivement, un premier destiné à l'entrainement du modèle SVM et deux autres dédiés aux CRF. Le modèle SVM est ici dédié à l'identification des paragraphes contenant une référence bibliographique. Pour chaque paragraphe un identifiant de classe est généré : 0 pour les non références et 1 pour les références. Pour chacune des références présente dans les paragraphes et à partir des annotations fournies au préalable, un premier CRF est construit permettant d'identifier la zone d'apparition de la référence au sein d'un paragraphe. Le second CRF est, quant à lui, destiné à identifer chacun des champs composant la référence. Lors de la phase d'annotation, chaque paragraphe est extrait et classé en fonction du modèle SVM précédemment appris. Pour chacun des paragraphes contenant une référence le CRF de bas niveau est appliqué afin d'identifier la zone d'apparition de la référence au sein du paragraphe. Ensuite, le CRF de haut niveau procède à l'identification des champs bibliographiques.

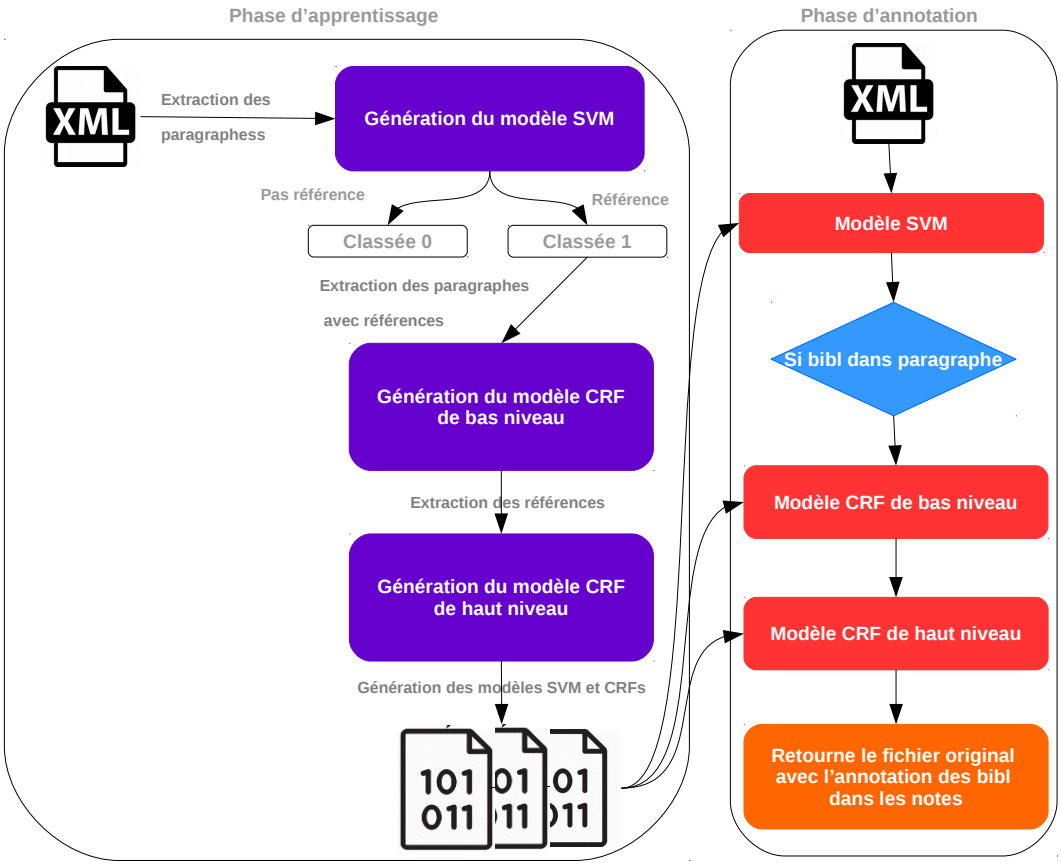


Figure 2.8. – Modélisation du traitement des références allusives. La couleur violette renvoie au processus marquant la génération des modèles, la couleur bleue à l'application de conditions, la couleur rouge aux processus relatifs à l'application des modèles et la couleur orange au(x) produit(s) résultant de l'application des modèles.

Dans la section suivante, nous présentons en détail les vecteurs de caractéristiques que nous avons choisis afin de construire les modèles des CRF.

2.4.1.1. Vecteurs de caractéristiques

Au cours de la conception de l'approche que nous proposons, de profondes réflexions ont été menées afin d'établir une modélisation efficace des CRF notamment au niveau des caractéristiques permettant la construction des fonctions descriptives. L'objectif au travers de ces réflexions est de parvenir à définir des traits permettant de modéliser au mieux les données d'entrée. Le terme « carac-

téristique » est parfois ambigu. Essentiellement utilisé pour définir des instances d'entrée, il peut également renvoyer à la définition de traits spécifiques, propres aux instances d'entrée telles que les caractéristiques morphologiques des mots. Afin de tenir compte de ces deux désignations, deux types de caractéristiques ont été définis : les *caractéristiques contextuelles* et les *caractéristiques locales*.

2.4.1.1.a. Caractéristiques contextuelles

Une fois la séquence d'entrée segmentée en unité lexicale (aussi appelée *token*), chaque unité est traitée séparément en tant que fonction descriptive. Grâce à la capacité des CRF à encoder toutes les informations relatives à l'observation en cours, plusieurs fonctionnalités ont été ajoutées *via* des observations sur le contexte d'apparition de l'unité lexicale courante. Des expériences menées au préalable sur les données d'OpenEdition (Y.-M. KIM, BELLOT et al. 2011) ont permis de démontrer que l'inclusion de trois marqueurs précédents et trois suivants comme caractéristiques supplémentaires donnait de meilleures performances lors de la modélisation des fonctions descriptives. Le tableau 2.6 recense les caractéristiques contextuelles conservées afin de modéliser les fonctions descriptives. Bien que de nombreuses autres caractéristiques puissent être utilisées telles que les N-grammes de mots, des expériences réalisées au cours de la conception de notre approche ont montré une diminution de la précision.

Catégorie	Nom	Description
Mot d'entrée brut	-	mot courant en lettre minuscule
Mots suivants et précédents	-	trois mots précédents et trois suivants le mot courant

Tableau 2.6. – Description des caractéristiques contextuelles

2.4.1.1.b. Caractéristiques locales

L'étude de la littérature (SAUL, WEISS et al. 2005 ; FOX et SILVA TORRES 2014) a permis d'attester de l'efficacité, dans l'analyse de référence, d'inclure des caractéristiques locales au sein des fonctions descriptives. Cependant l'inclusion de caractéristiques trop détaillées peut engendrer une trop forte dépendance aux données d'application ainsi qu'un surapprentissage (SHALEV-SHWARTZ et BEN-DAVID 2014). En tenant compte de ces aspects, le choix des caractéristiques locales a été fait en vue d'offrir une portabilité permettant une exploitation de l'approche proposée sur un large panel de données. Le tableau 2.7 fournit une description des caractéristiques locales utilisées dans la modélisation des fonctions descriptives.

Caractéristiques morphologiques

Catégorie	**Nom**	**Description**
Nombre	ALLNUMBERS	tous les caractères sont des nombres
	NUMBERS	un ou plusieurs caractères sont des nombres
	DASH	un ou plusieurs traits d'union sont inclus dans les nombres
Capitalisation	ALLCAPS	tous les caractères sont en lettre capitale
	FIRSTCAP	le premier caractère est en lettre capitale
	ALLSAMLL	tous les caractères sont en minuscules
	NONIMPCAP	les types de capitalisations sont mélangés
Expression régulière	INITIAL	expression pour l'identification d'initiale
	WEBLINK	expression régulière pour les pages web
Emphase	ITALIC	caractères en italique
Stemme	-	radical du mot courant
Lemme	-	forme canonique du mot courant

Caractéristiques de localisation

Location	BIBL_START	position dans le premier tiers de la référence
	BIBL_IN	position entre le premier tiers et le deuxième tiers de la référence
	BIBL_END	position entre les deux tiers et la fin de la référence

Caractéristiques lexicales

Lexique	POSSEDITOR	abréviations d'éditeur
	POSSPAGE	abréviations de pages
	POSSMONTH	abréviations pour les mois
	POSSBIBLSCOP	abréviations pour les extensions bibliographiques
	POSSROLE	abréviations pour les rôles
Liste externe	SURNAMELIST	liste de noms
	FORENAMELIST	liste de prénoms
	PLACELIST	liste de lieux
	JOURNALLIST	liste de revues
POS Simple	Jeu de labels	*part of speech* harmonisé
POS Détaillé	Jeu de labels	*part of speech* détaillé

Ponctuation		
Ponctuation	COMMA	description des marques de ponctuation
	POINT	
	LINK	
	PUNC	
	LEADINGQUOTES	
	ENDINGQUOTES	
	PAIREDBRACES	

Tableau 2.7. – Description des caractéristiques locales

Les caractéristiques locales se divisent en quatre sous-catégories : les caractéristiques morphologiques, de localisation, lexicales et syntaxiques.

- Les caractéristiques morphologiques permettent de définir la forme des mots (p. ex. : majuscule/minuscule).
- Les caractéristiques de localisation définissent la position des champs dans la séquence.
- Les caractéristiques lexicales correspondent à l'exploitation de listes de mots prédéfinis.
- Les caractéristiques syntaxiques réfèrent à la ponctuation.

Certaines caractéristiques ne peuvent décrire simultanément un même mot telles que les caractéristiques présentes dans la catégorie Capitalisation, mais de nombreuses autres caractéristiques peuvent être utilisées conjointement pour décrire un mot. Bien que ces caractéristiques aient été sélectionnées suite à l'étude de la littérature, il existe certaines particularités à leurs exploitations dans ce contexte de travail. Parmi les principales particularités, nous pouvons citer la présence de plusieurs caractéristiques explicites comme les traits d'union présents dans la catégorie Nombre. Cette information propre au traitement des références bibliographiques peut s'avérer utile pour mieux caractériser les ensembles de nombres tels que les années et les numéros de page. L'identification des suites d'initiales est également propre aux données bibliographiques et permet la détection d'expressions simples mais aussi d'expressions plus complexes incluant des minuscules et des tirets. La position de chaque élément est définie par trois positions différentes (BIBL_START, BIBL_IN et BIBL_END). Bien que (COUNCILL, GILES et al. 2008) utilise 12 positions différentes leurs applications sur des références issues des SHS diminuent la qualité d'analyse. Ce phénomène peut être dû à la diversité des structures des références présentes qui ne permet pas de définir des emplacements proportionnels réguliers pour chaque champs. Les diacritiques induites par la nature multilingue des publications scientifiques sont également prises en compte dans les fonctions descriptives. Concernant le traitement des

signes de ponctuation, nous nous sommes bornée aux signes de ponctuation les plus récurrents présents au sein du corpus d'apprentissage. Au total, 6 caractéristiques ont été choisis. Les caractéristiques COMMA et POINT représentent le signe de ponctuation lui-même et LINK correspond aux marques reliant deux phrases telles que le point-virgule. LEADINGQUOTES, ENDINGQUOTES et PAIREDBRACES représentent les signes de ponctuation eux-mêmes soit, les guillemets ouvrants et fermants ainsi que les accolades. La caractéristique PUNC est, quant à elle, utilisée pour les autres signes de ponctuation ou caractères spéciaux.

Dans le cadre de l'approche que nous proposons, nous n'avons, pour le moment, pas employé de vecteurs de caractéristiques propres à chacun des niveaux du modèle de CCRF. Bien que chacun des CRF utilisés s'oriente sur des applications différentes, nous nous sommes basée, afin de sélectionner les caractéristiques dédiées à la constructions des fonctions descriptives, sur des constatations effectuées suite aux études menées sur la composition des références bibliographiques allusives que nous avons menées au cours de la section 2.3.2.2.b. De plus, notre souhait s'orientait, dans un premier temps, sur des caractéristiques offrant une portabilité permettant une exploitation de cette approche sur un large panel de données. Dans nos travaux futurs, nous nous pencherons sur l'étude des caractéristiques les plus influentes en fonction du niveau d'application de chacun des CRF.

Dans les sections qui suivent nous présentons les caractéristiques principales de Bilbo.

2.4.2. Bilbo : Annotation automatique de références bibliographiques

Développé dans le cadre d'un projet de R&D du laboratoire OpenEdition Lab depuis 2011, le logiciel Bilbo [42], disponible en libre accès, permet l'annotation des références structurées présentes dans les zones bibliographiques et les notes de bas de page. Notre objectif au cours de cette thèse était notamment d'en améliorer la portabilité afin que Bilbo soit en mesure d'annoter les références que nous avons qualifiées d'allusives.

Dans les sections suivantes nous présentons succinctement les deux modèles réalisés afin de permettrer le traitement des références structurées présentes dans les zones bibliographiques et celles présentes dans les notes de bas de page. Bien que certains procédés soient communs aux deux fonctionnalités, nous avons choisi de les présenter distinctement par souci de compréhension.

42. https://github.com/OpenEdition/bilbo

2.4.2.1. Traitement des références dans les zones bibliographiques

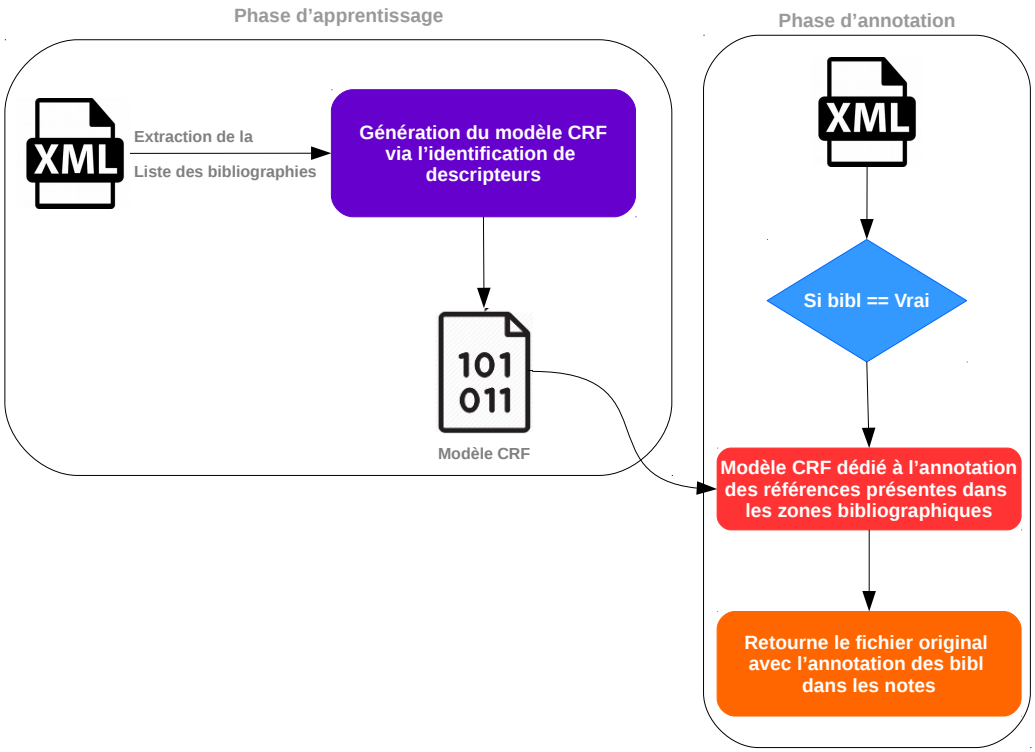


Figure 2.9. – Modélisation du traitement des références présentes dans les zones bibliographiques. La couleur violette renvoie au processus marquant la génération des modèles, la couleur bleue à l'application de conditions, la couleur rouge aux processus relatifs à l'application des modèles et la couleur orange au(x) produit(s) résultant de l'application des modèles.

La figure 2.9 présente une modélisation du traitement des références présentes dans les zones bibliographiques par Bilbo. Lors de la phase d'apprentissage, Bilbo prend comme entrée un document XML au format TEI dans lequel les références bibliographiques sont préalablement annotées. À partir des informations fournies par la TEI, Bilbo est en mesure d'identifier la zone dédiée aux références bibliographiques. Pour chacune des références présente dans cette zone et à partir des annotations fournies, un modèle est construit *via* l'identification des caractéristiques. Lors de la phase d'annotation des références, Bilbo prend également

comme entrée un document XML au format TEI. Une fois la zone bibliographique identifiée, le modèle CRF précédemment appris est exécuté sur chaque référence afin d'identifier chacun des champs bibliographiques. Ces deux processus peuvent s'opérer consécutivement mais aussi séparement selon les options définies lors du lancement de Bilbo. Un modèle par défaut est disponible pour le lancement de la phase d'annotation sans apprentissage préalable.

2.4.2.2. Traitement des références dans les notes de bas de page

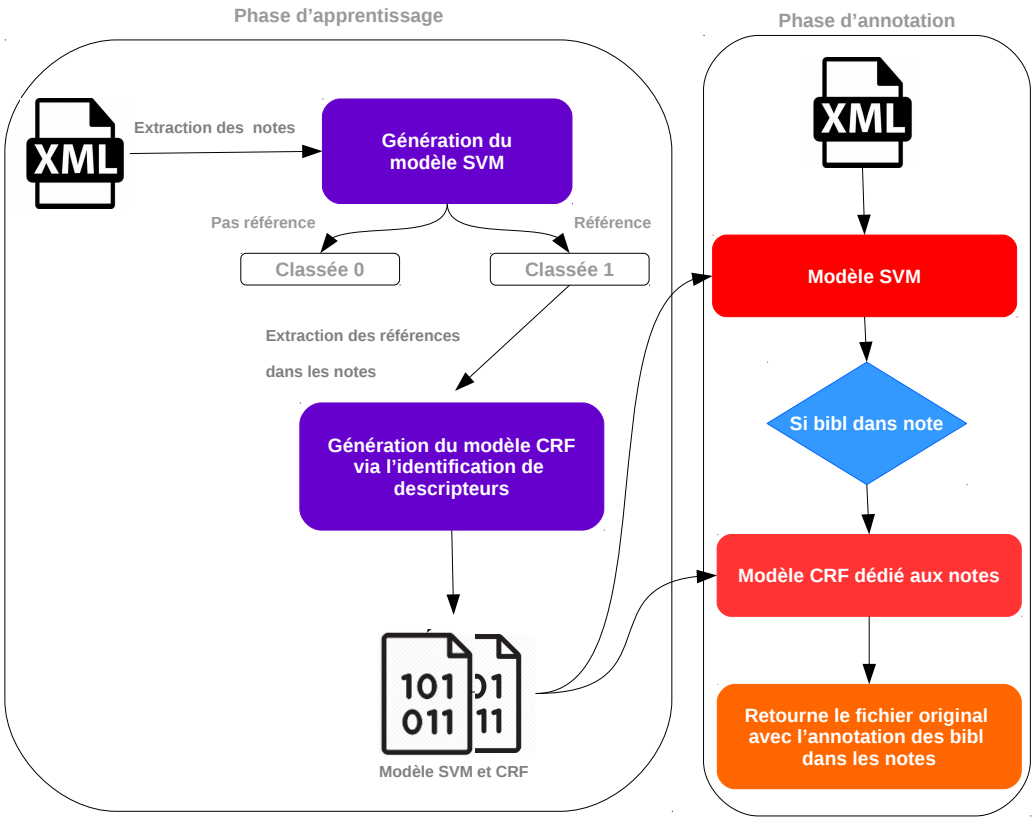


Figure 2.10. – Modélisation du traitement des références présentes dans les notes de bas de page. La couleur violette renvoie au processus marquant la génération des modèles, la couleur bleue à l'application de conditions, la couleur rouge aux processus relatifs à l'application des modèles et la couleur orange au(x) produit(s) résultant de l'application des modèles.

La figure 2.10 présente une modélisation du traitement des références pré-

sentes dans les notes de bas de page par Bilbo. Comme présenté dans la section précédente, deux processus distincts s'opèrent au cours du traitement de ces références, un processus dédié à la phase d'apprentissage et un autre destiné à l'annotation des références. Lors de la phase d'apprentissage, deux processus d'apprentissage sont exécutés consécutivement afin d'entrainer le modèle SVM et CRF. Le modèle SVM est ici dédié à l'identification des notes contenant une référence bibliographique. Pour chaque note un identifiant de classe est généré : 0 pour les non références et 1 pour les références. Pour chacune des références présentes dans les notes et à partir des annotations fournies au préalable, un modèle CRF est construit suivant les procédés présentés dans la section précédente. Lors de la phase d'annotation, chaque note est extraite et classée en fonction du modèle SVM précédemment appris. Pour chacune des notes contenant une référence le modèle CRF est appliqué afin d'identifier chacun des champs bibliographiques. L'apprentissage du modèle SVM et CRF peut s'opérer consécutivement mais aussi séparement selon les options définies lors du lancement de Bilbo, ce qui est également le cas pour la phase d'apprentissage et d'annotation. Un modèle SVM et CRF par défaut est disponible pour le lancement de la phase d'annotation sans apprentissage préalable.

Dans la section qui suit, nous proposons d'appliquer l'approche présentée dans la section 2.4 dans le cadre du traitement des données OpenEdition.

2.5. Identification des références allusives : application aux données d'OpenEdition

Au cours de cette section, nous présentons, en premier lieu, le cadre applicatif dans lequel s'intègre nos travaux, à savoir les données d'OpenEdition, et en second lieu l'évaluation des performances de l'approche proposée appliquée sur ces données.

2.5.1. Cadre Applicatif

Dans le cadre de nos travaux, le *Centre pour L'édition Électronique Ouverte* (Cléo) nous a fourni des données extraites de son portail OpenEdition.org [43]. Ce portail regroupe quatre plateformes complémentaires consacrées aux livres, revues, blogs de recherche et annonces scientifiques. La majorité du contenu proposé par ces plateformes est disponible en libre accès. Fondée en 1999, la plateforme Revues.org [44] contient 400 revues en ligne dont 95% sont disponibles

43. http://www.openedition.org/
44. http://www.revues.org/

en texte intégral. La plateforme Calenda [45], créée en 2000, est un calendrier en libre accès qui informe les étudiants, les enseignants et les chercheurs de l'actualité de la recherche. Fondée en 2009, la plateforme Hypothèses [46] regroupe près de 1 000 carnets de recherche animés par une communauté de 10 000 blogueurs provenant de tous les pays. Inaugurée en 2013, la plateforme OpenEdition Books [47] rassemble près de 1 700 livres provenant de 45 éditeurs et dont plus de la moitié sont en libre accès. Dans le cadre de nos travaux, nous nous sommes plus particulièrement intéressée à la plateforme Revues.org. Cette plateforme dédiée aux publications scientifiques fournit une large illustration des divers emplois que peuvent revétir les références bibliographiques allusives.

Dans le cadre d'un programme de R&D dirigé par le laboratoire OpenEdition Lab [48], un corpus destiné à mettre en évidence les références allusives, à la fois, dans le texte et dans les notes de bas de page a été recueilli. La figure 2.11 présente l'extrait d'un article provenant de la plateforme Revues.org contenant des références bibliographiques allusives soulignées en gras.

En **1967**, quatre étudiants publient un manifeste **« Pourquoi des sociologues ? »**. Ils s'interrogent sur le fait qu'aussi bien aux États-Unis, en France, en Allemagne, qu'en Pologne et en Tchécoslovaquie, l'agitation universitaire se développe surtout dans les sections de sciences sociales. Pour eux, « Le passage d'une sociologie académique, vassale de la philosophie, à une sociologie indépendante, à prétentions scientifiques, correspond au passage du capitalisme concurrentiel au capitalisme organisé ». Critiquant à la fois les prises de position d'Alain Touraine et la passivité de la grande majorité des étudiants, les auteurs appellent à dissiper « l'illusion des mots d'ordre stalino-tourainiens sur un mouvement étudiant de masse aux intérêts convergents. Par leur origine sociale comme par leur acceptation de devenir des salariés des différents appareils autoritaires (État, entreprise, firme publicitaire, etc.), la majorité des étudiants sont déjà conservateurs ». Et pourtant les auteurs de ce pamphlet seront parmi les « leaders » du 22 mars et du mouvement de Mai… Dans un très beau texte, **L'irruption, de Nanterre au sommet**, **Henri Lefebvre** analyse le départ des événements de Mai à Nanterre.

Figure 2.11. – Extrait de la revue CEDREF

Ce corpus [49] a été manuellement annoté suivant le formalisme de la TEI [50]. La TEI est un consortium qui développe, définit et maintient un langage de balisage afin de décrire les caractéristiques structurelles et conceptuelles des textes. Actuellement, la plupart des articles de publication électronique en SHS suivent les directives de la TEI. Afin d'annoter ce corpus, des lignes directrices ont été établies de manière très détaillée afin d'encoder les informations bibliographiques.

45. http://calenda.org
46. http://hypotheses.org/
47. http://books.openedition.org/
48. http://lab.hypotheses.org/
49. https://lab.hypotheses.org/
50. http://www.tei-c.org/index.xml

Le tableau 2.8 établit une liste des balises utilisées lors de l'annotation des références bibliographiques allusives. Au total, 10 sous-types ont été définis en fonction des champs bibliographiques rencontrés.

Type	Sous-type (nombre d'entités)
Auteur ou Editeur	Nom (918) Prénom (339)
Titre	Titre (712) or Conférence
Ponctuation	Ponctuation (1881)
Publication Marqueur	Date (230) Éditeur (37) Édition (17) Édition détail (110)
Etc.	Abbréviation (203)

Tableau 2.8. – Balises définies pour le corpus OpenEdition

Les revues sélectionnées ont été choisies aléatoirement dans la base de données OpenEdition indépendamment de l'affiliation disciplinaire des revues ou des conventions stylistiques de ces dernières. Les articles extraits sont essentiellement en français. 42 articles de Revues.org ont été sélectionnés, totalisant 607 références bibliographiques allusives, à la fois, dans les textes et dans les notes de bas de page. Le tableau 2.9 donne un aperçu des statistiques effectuées sur les publications du corpus réalisé à partir des données de Revues.org. Chaque article est composé en moyenne de 39 paragraphes et 65 notes. En moyenne, les paragraphes comptent 219 mots et 6 phrases et les notes de bas de page comptent 60 mots et 2 phrases.

Nombre minimum de mots/ article	3378
Nombre maximum de mots/ article	23909
Nombre moyen de mots/ article	8272
Nombre moyen de paragraphes	39
Nombre moyen de notes de bas de page	65

Tableau 2.9. – Statistiques sur un échantillon ($N = 42$) des publications scientifiques du corpus OpenEdition

Dans la section qui suit, nous présentons les résultats obtenus suite à l'appli-

cation du modèle présenté dans la section 2.4 sur les données d'OpenEdition.

2.5.2. Évaluation sur les données d'OpenEdition

Dans cette section, nous exposons les résultats obtenus suite aux évaluations menées sur le corpus présenté au cours de la section précédente. Nous présentons premièrement, trois expérimentations effectuées sur chacun des processus de notre modèle, à savoir, la classification supervisée de paragraphes *via* un modèle SVM et l'identification des zones bibliographiques ainsi que l'annotation de leurs contenus par le biais d'un modèle CCRF. Nous avons fait le choix d'évaluer les performances de chacun de nos procédés afin d'attester de leur robustesse. Deuxièmement, nous proposons l'évaluation globale de ce même modèle *via* l'éxecution consécutive des trois processus cités précédemment.

2.5.2.1. Classification supervisée des paragraphes contenant des références bibliographiques

Afin de procéder à une classification supervisée, un corpus a été préparé au préalable contenant des exemples pour chacune des classes : « zone bibliographique » et « zone non bibliographique ». Au total, le corpus comporte 27,8 % de paragraphes comportant des références bibliographiques sur un ensemble de 2067 paragraphes. Pour l'implémentation du SVM, nous utilisons l'outil *SVM-Light* [51] développé par (THORSTEN 1999). Concernant les paramétrages effectués, l'utilisation de l'outil *Weka* [52] nous a permis d'établir une liste des mots les plus caractéristiques de chaque classe que nous utilisons comme attributs. Cette liste est réalisée grâce à l'algorithme *InfoGainAttribute* (IGA). IGA est utilisé pour réduire le biais vers les attributs à valeurs multiples en prenant en compte le nombre et la taille des branches lors du choix d'un attribut. Nous avons choisi d'utiliser une fréquence minimale d'apparition des termes de 3 combinée à une liste pour laquelle nous avons conservé les mots dont le score « Élimination récursive de caractéristiques » [53] (RFE) est égal à 0. Nos choix se sont porté sur ces paramètres suite à des expérimentations dont les résultats sont disponibles en annexe A. Nous effectuons 10 validations croisées afin d'évaluer la façon dont les résultats se généralisent à un ensemble de données. Le tableau 2.10 présente les résultats obtenus suite à la classification supervisée.

51. http://svmlight.joachims.org/
52. http://www.cs.waikato.ac.nz/ml/weka/
53. *Recursive Feature Elimination* : La sélection de caractéristiques par élimination récursive de caractéristiques est un processus récursif au cours duquel chaque caractéristique est classée selon son degré d'importance. À chaque itération l'importance des caractéristiques est mésurée et les moins pertinentes sont supprimées (GRANITTO, FURLANELLO et al. 2006).

Exactitude	Précision	Rappel	F-mesure
76,9 %	79,8 %	77,2 %	78,5 %

Tableau 2.10. – Résultats de la classification supervisée des paragraphes

Les performances obtenues sont satisfaisantes avec une exactitude de 76.9 % soit 481 paragraphes incorrectement classés. Une part des erreurs commises peut s'expliquer par la grande hétérogénéïté présente à partir desquelles nous avons sélectionné des attributs. Cette hétérogénéïté, induite par les types de documents, entraîne une grande diversité, à la fois structurelle et compositionnelle au sein des paragraphes. Nous savons que la représentation des données *via* la sélection d'attributs est une tâche essentielle impactant fortement les performances d'un classifieur. Nous supposons qu'une analyse approfondie des attributs sélectionnés afin d'identifier, à la fois leur pertinence et leur portabilité, peut permettre d'optimiser les performances. D'autres pistes sont également envisageables au vu de la densité textuelle des données comme la mise en place d'extracteur de séquences ainsi que l'élaboration de motifs syntaxiques utilisés dans le traitement des requêtes verbeuses (ETTALEB, LATIRI et al. 2016).

2.5.2.2. Détection des zones bibliographiques *via* un modèle CRF de bas niveau

Cette section expose les résultats obtenus suite à l'identification des séquences contenant des champs bibliographiques au sein des paragraphes. Pour effectuer les expérimentations, nous établissons un corpus composé uniquement de paragraphes comprenant des références bibliographiques, c'est-à-dire, que ce corpus est établi indépendamment de la classification présentée précédemment. Nous présentons des expérimentations basées sur 10 validations croisées composées de 70 % de corpus d'apprentissage et de 30 % de corpus de test.

Précision	Rappel	F-mesure
78,8 %	79,8 %	79,3 %

Tableau 2.11. – Résultats pour la détection des zones bibliographiques

Le tableau 2.11 introduit les résultats obtenus suite à la détection des zones bibliographiques. Les résultats présentés sur cette expérimentation sont satisfaisants. Nous pouvons observer une certaine constance entre les résultats obtenus, ce qui nous permet d'attester de la portabilité des caractéristiques choisies afin de construire nos vecteurs d'informations et ce même en présence de séquence inconnue. Les taux de rappel et précision obtenus, proches des 80 %, nous confortent dans la capacité des caractéristiques descriptives à identifier des zones d'apparition pertinentes. Bien que des degrés plus ou moins forts d'implicite ont été relevés, induisant des références bibliographiques à la structure

complexe, la combinaison de caractéristiques contextuelles et locales s'avère concluante. De plus, ces résultats démontrent une bonne adaptabilité de ce type de caractéristiques sur des données hétérogènes.

2.5.2.3. Détection des champs bibliographiques *via* un modèle CRF de haut niveau

Dans cette section, nous exposons les résultats obtenus suite à l'identification des différents champs bibliographiques. Pour cette expérimentation, nous choisissons de présenter les résultats obtenus suite à l'utilisation de deux modèles. Le premier modèle (Corpus de référence) a été construit *via* l'extraction de références allusives et le second à partir des références bibliographiques présentes dans les zones bibliographiques (Corpus structuré). Nous choisissons d'établir cette comparaison afin de constater l'impact des performances d'un modèle appris sur des données fortement structurées et de son application sur des données de nature différente. Cette analyse nous permet également de comparer l'approche que nous proposons aux approches à base de CRF présentées précédemment dans la section 2.2.2. Nous précisons également que cette expérimentation est faite indépendamment des résultats obtenus lors de la détection des zones bibliographiques. Les corpus utilisés ne découlent donc pas directement des résultats présentés lors de l'expérimentation précédente. Nous proposons des expérimentations basées sur 10 validations croisées composées de 70% de corpus d'apprentissage et de 30% de corpus de test.

Corpus	Précision	Rappel	F-mesure
Corpus de référence	85,0%	78,2%	82,4%
Corpus structuré	57,4%	54,6%	56,1%

Tableau 2.12. – Résultats pour la détection des champs bibliographiques

Le tableau 2.12 nous permet d'observer une nette dégradation des performances suite à l'apprentissage effectué *via* le modèle établi sur le Corpus structuré. Ce phénomène s'explique par la structure très formelle que l'on retrouve dans les références bibliographiques présentes dans les zones bibliographiques. En effet, ce type de références répond à des conventions très strictes à la différence des références allusives pour lesquelles il n'existe aucun consensus. Cette expérience nous permet de mettre en évidence une perte importante de performance lorsque l'on change le cadre applicatif sur lequel les modèles ont été appris. Concernant les performances obtenues suite à l'apprentissage effectué *via* l'extraction de références allusives, nous avons pu observer des résultats satisfaisants. Nous ponvons noter une précision plus importante ce qui nous permet d'attester que l'utilisation de caractéristiques locales et contextuelles permet d'obtenir des annotations pertinentes malgré des références aux compositions très variées. Le taux de rappel plus faible peut s'expliquer quant à lui part la

diversité des étiquettes (10 au total) ce qui peut potentiellement engendrer des étiquettes aux caractéristiques redondantes. De plus, la variété importante des combinaisons d'étiquette associée à des compositions pouvant osciller entre 2 et 5 champs suscite des difficultés supplémentaires. Bien que les résultats obtenus soient encourageants des pistes permettant l'optimisation des performances sont envisageables comme l'étude approfondie de l'influence de chacune des caractéristiques lors de l'apposition d'une étiquette.

2.5.2.4. Évaluation de la globalité du système de détection des références allusives

Dans cette section, nous présentons les résultats obtenus suite à l'exécution consécutive des trois processus présentés précédemment, à savoir, l'identification des paragraphes contenant des références *via* un modèle SVM suivi de l'application du CRF de bas de niveau et de haut niveau. Afin d'effectuer cette évaluation, nous avons utilisé exactement les mêmes données que celles présentées aux cours des expérimentations précédentes. Nous présentons des expérimentations basées sur 10 validations croisées composées de 70% de corpus d'apprentissage et de 30% de corpus de test.

Précision	Rappel	F-mesure
58,8 %	52,1 %	54,9 %

Tableau 2.13. – Résultats de l'évaluation globale du système de détection des références allusives

Les résultats obtenus suite à cette évaluation sont assez mitigés. Bien que les expérimentations menées précédemment aient permis d'attester d'une certaine robustesse lors de l'évaluation de chacun de nos processus, leur exécution consécutives engendre des pertes. Nous supposons que les erreurs induites consécutivement par chacun des procédés impactent les performances globales. En effet, lors de la classification supervisée des paragraphes contenant des références sont occultés ce qui fait qu'ils ne sont pas annotés par les CRF. Ces occultations se répercutent sur les performances globales. De plus, le bruit et le silence émis par le CRF de bas niveau impactent directement le CRF de haut niveau ce qui engendre également des pertes.

Pour conclure, nous avons pu observer, suite à ces expérimentations, des résultats satisfaisants concernant l'exécution individuelle de chacun des procédés. Ces résultats nous ont permis d'attester la robustesse de l'approche proposée. En effet, les caractéristiques choisies afin d'établir nos modèles ont démontré une capacité d'adaption intéressante au vu des données hétérognènes que nous traitons. Cependant, l'éxécution consécutive de tous les procédés a révélé des

résultats plus mitigés induits par des erreurs générées au cours de chaque processus. Nous avons pu observer que les erreurs de traitement faites au cours de chaque processus impactent nécessairement le suivant. Comme nous avons pu le souligner au cours des sections précédentes, des solutions d'amélioration sont envisageables afin d'éviter la propagation des erreurs.

Dans la section suivante, nous présentons l'application de l'approche que nous proposons sur les données de la campagne d'évaluation CLEF SBS 2016.

2.6. Identification des références allusives : application à la recherche de titres de livres

Afin de comparer les performances du système que nous proposons mais aussi de tester la portabilité de notre approche *via* l'utilisation d'un jeu de données aux conventions stylistiques et à la langue différentes des données d'OpenEdition, nous avons participé à la campagne d'évaluation CLEF SBS 2016[54] (KOOLEN, BOGERS et al. 2016). En effet, le corpus fourni lors de la campagne CLEF SBS 2016 orienté sur des fils de discussion provenant de forums en ligne se distingue des données d'OpenEdition, à la fois, au niveau de la structure des références bibliographiques et au niveau de leurs situations d'énonciation. Les sections suivantes présentent plus en détail ce jeux de données ainsi que les résultats que nous avons obtenus.

2.6.1. Les données de CLEF SBS 2016

Dans le cadre des campagnes CLEF SBS différents jeux de données sont proposés. Le but de ces campagnes est d'évaluer des approches permettant de faciliter tous types de recherches d'information au sein de collections de livres et de forums en ligne. Plusieurs pistes de recherche sont évaluées permettant l'exploitation de données contenant des métadonnées issues de livres ou encore des contenus associés générés par les utilisateurs telles que les pistes de suggestion (*Suggestion Track*) et d'extraction (*Mining Track*). Dans le cadre des travaux réalisés, nous avons exploité les données fournies lors de la piste de recherche orientée sur l'extraction de 2016. Cette piste comprend deux tâches : la tâche de classification (*classification task*) et la tâche de liaison (*linking task*). Afin d'apparenter nos travaux réalisés dans le cadre de l'identification automatique des références bibliographiques allusives, nous nous sommes concentrée sur la tâche de liaison, qui consiste à reconnaître des titres de livres au sein de fils de discussion et à les associer à l'identifiant unique du livre correspondant. Pour solutionner cette tâche, il n'est pas nécessaire d'identifier l'expression exacte qui

54. http://social-book-search.humanities.uva.nl/#/overview

se réfère au livre mais d'obtenir, *via* les éléments (généralement titre et auteur) extraits des requêtes, l'identifiant unique du livre correspondant dans la collection.

Le corpus fourni lors de l'année 2016 se compose d'un ensemble de 200 fils de discussion, orientés sur des demandes de recommandation de livres, présentés sous la forme d'une suite de requêtes consécutives sur un même thème, classés de manière arborescente. Au total, ce corpus dénombre 3619 requêtes[55] exprimées en langue naturelle posées par les utilisateurs de LT, dont un exemple est présenté par la figure 2.12. Dans cet extrait, nous avons un exemple de fil de discussion présent dans le corpus, avec le post 1 qui marque le début de la discussion *via* l'expression du besoin d'un utilisateur et le post 2 qui correspond à une réponse d'un autre utilisateur.

```
  <message>
    <text>I'd love to get some books for my 4 year old to teach him Libertarian beliefs.
I remember checking out Yurtle the Turtle and one about a free-spirited family that was going
 to get run out of the neighborhood by the snoopy neighbors. Can't remember the name though.
"The -somethings-" Anyone have others that they thought were good?
</text>
    <postid>1</postid>
    <username>kkirkhoff</username>
    <threadid>3408</threadid>
    <date>Oct 30, 2006, 11:39pm </date>
  </message>
  <message>
    <text>I've heard good things about Richard Maybury's Whatever Happened to Penny Candy?
- though probably not for a 4yr old :)</text>
    <postid>2</postid>
    <username>Misesean</username>
    <threadid>3408</threadid>
    <date>Oct 31, 2006, 12:40am </date>
  </message>
```

Figure 2.12. – Exemple d'un fil de discussion extrait de CLEF SBS 2016

Toutes les requêtes fournies sont en anglais à la différence des données d'OpenEdition (cf : section 2.5.1) essentiellement en français ce qui, par ailleurs, n'empêche pas l'approche que nous proposons de s'adapter. Pour chaque requête nous disposons de cinq champs : <*text*>, <*postid*>, <*username*>, <*threadid*> et <*date*>. Le premier champ contient la requête de l'utilisateur. Le second correspond à l'identifiant de la requête par rapport à son émission dans le flux de discussion. Le troisième correspond au nom de l'utilisateur. Le quatrième renvoie à l'identifiant du fil de discussion et le cinquième correspond à la date de mise en ligne de la requête. Ce corpus ne possède aucune annotation. Il est important de préciser que les requêtes formulées sont destinées à d'autres êtres humains et non à un moteur de recherche, ce qui engendre des requêtes longues et détaillées aux structures complexes. Ce type de requête nous permet, par ailleurs, de travailler sur une densité textuelle semblable aux paragraphes présents au

55. http://social-book-search.humanities.uva.nl/#/data/mining

sein d'articles scientifiques. Le tableau 2.14 donne un aperçu des statistiques effectuées sur les fils de discussion extrait de CLEF SBS 2016. Chaque fil de discussion est composé en moyenne de 18 requêtes. En moyenne, les requêtes comptent 153 mots et 3 phrases.

Nombre minimum de mots/ requête	1
Nombre maximum de mots/ requête	2480
Nombre moyen de mots/ requête	153
Nombre moyen de requête/ fil de discussion	18

Tableau 2.14. – Statistiques sur les fils de discussion

Afin de pouvoir établir une correspondance entre les titres de livres trouvés et la collection de livres, nous disposons d'un fichier contenant les identifiants uniques de chaque livre, des titres de base et des métadonnées. Un exemple est présenté dans la figure 2.13.


```
{"workID": "19425", "versions":
[{"author": "Leo Walmsley", "ISBN": "0001831208", "booktitle": "Foreigners"}]}
{"workID": "2474294", "versions":
[{"author": "Judith Kerr", "ISBN": "0001360000", "booktitle": "Mog's Kittens"}]}
```


Figure 2.13. – Extrait du fichier permettant la liaison avec la collection de livres

Pour chaque livre nous disposons de son identifiant unique, de son ISBN [56], de son titre et parfois du nom de l'auteur. Nous disposons également de métadonnées supplémentaires *via* la mise à disposition d'une collection constituée de 2,8 millions de descriptions de livres extraites d'Amazon [57]. Chaque livre se compose de 64 champs XML (un exemple est présenté figure 2.14). Parmi ces champs, nous distinguons :

— les métadonnées : <*book*>, <*isbn*>, <*title*>, <*authorid*>, etc.
— les informations sociales : <*reviews*>, <*summary*>, <*tags*>, <*rating*>, etc.

La section suivante présente la méthode que nous avons mise place afin de solutionner la tâche de liaison.

2.6.2. Méthode

Afin de résoudre la tâche de liaison nous proposons l'intégration de notre approche de détection des références bibliographiques allusives présentée dans la section 2.4. Une fois l'identification des références allusives effectuée, une liste

56. *International Standard Book Number* ou Numéro international normalisé du livre est un numéro international qui permet d'identifier de manière unique chaque édition de chaque livre publié, que son support soit numérique ou sur papier.
57. http://www.amazon.fr/

<book><isbn>0001360000</isbn><title>Mog's Kittens</title><ean>9780001360006</ean>
<binding>Board book</binding><label>HarperCollins UK</label><listprice>$6.99
</listprice><manufacturer>HarperCollins UK</manufacturer><publisher>HarperCollins UK
</publisher><readinglevel>Ages 4-8</readinglevel><releasedate/><publicationdate>
1994-09-01</publicationdate><studio>HarperCollins UK</studio><edition/><dewey/>
<numberofpages>16</numberofpages><dimensions><height>55</height><width>461</width>
<length>469</length><weight>22</weight></dimensions><reviews><date>2007-11-27</date>
<summary>Cute Book</summary><content>cute board book for the cat lover or animal lover
author of "Hobo Finds A Home"</content><rating>5</rating><totalvotes>0</totalvotes>
<helpfulvotes>0</helpfulvotes></review></reviews>

Figure 2.14. – *Exemple d'une fiche descriptive d'un livre extrait de la collection d'Amazon*

des livres recommandés par les utilisateurs est obtenue pour chaque requête. La distance de Levenshtein est ensuite utilisée afin de relier chaque élément de cette liste à son ID de livre unique présent dans les métadonnées des livres d'Amazon. Deux variations de la distance de Levenshtein sont proposées ici, une première tenant compte en tant que facteur de la distance d'alignement la plus courte entre les séquences et une seconde tenant compte de la distance d'alignement la plus longue. Ensuite, chaque titre de livre est comparé à l'ensemble des titres de livre extrait des métadonnées d'Amazon. Pour chaque titre de livre, une liste de livres est obtenue et triée selon une distance de Levenshtein normalisée. Pour chaque titre de livre, le résultat le plus proche de 1 est conservé. Ensuite, l'ID unique du livre estimé comme étant le plus probable est récupéré. L'algorithme 1 présente la chaîne de traitement proposée afin de résoudre cette tâche et dans laquelle nous avons intégré l'approche dédiée à la détection des références allusives. Supprimer les éléments en double est une contrainte de la tâche de liaison. En effet, il est possible de trouver à plusieurs reprises la même référence dans la même requête et dans ce cas de figure il n'est pas nécessaire de conserver le duplicata.

Considérons T comme l'ensemble des fils de discussion, tel que :

$$T = \{t_1, t_2, ..., t_i\} \tag{2.8}$$

Considérons Rq comme l'ensemble des requêtes de recommandation, étant lui-même un sous ensemble de T, tel que :

$$Rq = \{rq_1, rq_2, ..., rq_i\} \tag{2.9}$$

Chaque t_i et rq_i ont un identifiant unique défini respectivement par les variables T_{id} et Rq_{id}. rq_i est utilisé comme paramètre, à la fois dans la fonction de classification ($SVM(rq_i)$) et dans la fonction d'analyse ($CCRF(rq_i)$).

$SVM(rq_i)$ donne la valeur 1 si rq_i est identifié comme référence. Le résultat de $SVM(rq_i)$ est stocké dans la variable $class$. ($CCRF(rq_i)$) annote rq_i *via* les modèles préalablement appris.

Definition 1. (**Fonction d'appariement**) Cette fonction recherche si le nom d'un auteur (a) est proche d'un titre de livre (b) dans la même rq_i qui est préa-

lablement tokenizée (w_i). α correspond à une fenêtre glissante fixée empiriquement à 4.

$$appariement_{\alpha}(a,b) = \begin{cases} 1 \text{ si } a \in \{w_1\ w_2 ... w_i\} \text{ tel que } w_1 ... w_\alpha\ b\ w_{\alpha+1} ... w_{2\alpha} \\ 0 \end{cases} \tag{2.10}$$

Definition 2. (**Fonction de similarité**) $Sim_{\theta,\omega}(i,j)$ correspond à l'application de la distance de Levenshtein. Cette fonction concerne les éléments θ qui réfèrent à b ou ab présents dans les requêtes ($R = r_i$) extraites de Rq et ω qui correspond à b' (title de livre) ou ab' (nom d'auteur) extraits des métadonnées des livres d'Amazon (BM). i et j se rapportent à la longueur des séquences.

Definition 3. (**Fonction de correspondance**) Cette fonction traite les résultats de $Sim_{\theta,\omega}()$. Une valeur normalisée de la distance de Levenshtein est attribuée pour chaque comparaison entre θ et ω.

$$B_{\theta} = b_{id}(\omega) \text{tel que } \exists\omega \text{ avec } \max_{\omega \in BM}(Sim_{\theta,\omega}(i,j)) \tag{2.11}$$

Algorithme 1 Identification des noms d'auteur et des titres de livres au sein de requêtes verbeuses

Prérequis: T, Rq, $SVM(rq_i)$ et $CCRF(rq_i)$.
Garantie: Liste des titres de livre liés avec leur ID de livre unique.

Pour chaque rq_i **faire**
    class = $SVM(rq_i)$
    Si class == 1 **alors**
        $CCRF(rq_i)$
        Pour chaque $rq_i \in t_i$ **faire**
            extraire T_{ID} et b
            Si appariement(a,b) **alors**
                $ab \leftarrow a \cup b$
                Pour chaque a'_i et $b'_i \in BM$ **faire**
                    $ab' \leftarrow a' \cup b'$
            supression des doublons de a et b dans rq_i
            Si ab **alors**
                calculer $Sim_{ab,ab'}(i,j)$
            Sinon calculer $Sim_{b,b'}(i,j)$
            $B_{ID} \leftarrow B_{id}(Sim_{b,b'|ab,ab'})$
            Pour chaque $rq_i \in t_i$ **faire**
                récupération de T_{ID}, Rq_{ID} et B_{ID}
            supression des doublons de r_i pour Rq_i

La section qui suit expose les résultats obtenus dans le cadre de notre participation à la campagne CLEF SBS 2016.

2.6.3. Évaluation sur les données de CLEF SBS 2016

Dans cette section, nous présentons les résultats obtenus suite à notre participation à la tâche de liaison organisée au cours de la piste d'extraction de CLEF SBS 2016. Au préalable, des données ont été extraites du corpus d'entraînement fourni pour la tâche afin d'établir un modèle de classification propre à ces données. Nous avons utilisé les mêmes classes que celles présentées dans la section 2.5.2.1, à savoir, « zone bibliographique » et « zone non bibliographique ». La classe « zone bibliographique » contient ici des exemples de requêtes possédant des références bibliographiques allusives et la classe « zone non bibliographique » des exemples de requêtes ne contenant pas de références allusives. La classe « zone bibliographique » contient 184 requêtes et la classe « zone non bibliographique » 153 requêtes. Une liste des mots les plus caractéristiques de chaque classe a également été constituée afin d'être utilisée comme attributs. En ce qui concerne la construction du modèle CCRF, nous nous sommes basée sur les mêmes données que celles utilisées pour la construction du modèle SVM. 133 requêtes ont été annotées manuellement afin de distinguer les zones bibliographiques ainsi que les champs bibliographiques (titres de livre et noms d'auteur uniquement). Au total, 264 titres de livres et 203 noms d'auteurs sont annotés.

Chaque exécution présente dans le tableau 2.15 résulte, à la fois, du processus de classification et du processus d'annotation, les variations effectuées se concentrent sur la combinaison des champs bibliographiques et sur la configuration de la distance de Levenshtein. Les lignes *Title_LV1* et *Title_LV2* diffèrent par la configuration de la distance de Levenshtein. *LV1* correspond à la configuration du facteur sur la longueur de l'alignement le plus court et *LV2* à la configuration du facteur sur la longueur de l'alignement le plus long. Pour chaque exécution, seuls les titres de livre sont utilisés. L'exécution *TitleU_LV1* présente le même paramètre que *Title_LV1* avec l'ajout d'une nouvelle fonctionnalité aux CRF de haut et bas niveau permettant la description détaillée des signes de ponctuation. Pour les deux dernières exécutions *TitleAu_LV1* et *TitleAu_LV2*, si un nom d'auteur est situé à une distance maximale de quatre mots d'un titre de livres, nous les combinons. Pour l'expérience, 217 fils de discussion ont été utilisés, soit 5097 titres de livres identifiés dans 2117 requêtes.

Notre meilleure exécution est *TitleAu_LV2*, classée seconde au regard de la mesure F-score et première au regard de la précision lors de notre participation à la campagne d'évaluation (BOGERS, HENDRICKX et al. 2016). L'agrégation des noms d'auteur augmente la performance avec un gain de 2,18 points concernant le F-score et un gain de 3,43 points pour la précision entre *TitleAu_LV2* et *Title_LV1*. Ce phénomène nous amène à penser que l'ajout d'informations supplémentaires sur le sujet de la recherche améliore la pertinence des résultats et

Exécution	Rappel	Précision	Fscore
Meilleure_exécution_2016 (Know)	41.14	28.26	33.50
TitleAu_LV2	**26,99**	**38,23**	**31,64**
TitleAu_LV1	26,54	37,58	31,11
Title_LV2	26,01	35,39	29,98
TitleU_LV1	26,34	34,50	29,87
Title_LV1	25,54	34,80	29,46

Tableau 2.15. – Résultats officiels de CLEF SBS 2016. Les exécutions sont classées selon le F-score

en particulier la précision. Nous supposons que l'agrégation de telles caractéristiques singulières comme le nom d'auteur augmente l'efficacité au cours de l'extraction d'information. Nous observons également que la configuration de la distance de Levenshtein est meilleure lorsqu'elle est définie sur la longueur de l'alignement plus long entre les séquences, à la fois uniquement sur les titres de livres et la combinaison des titres de livres et des noms d'auteurs. Par rapport à la meilleure exécution de 2016, l'ensemble de nos exécutions obtient une meilleure précision. Plusieurs hypothèses peuvent expliquer le manque de rappel. Tout d'abord, le processus de classification peut occulter des requêtes contenant des références. Deuxièmement, la quantité de données d'apprentissage peut ne pas être suffisante pour être représentative de tous les cas de figure.

Pour conclure, nous rappelons que notre objectif *via* la participation à cette campagne d'évaluation est d'estimer la portabilité de l'approche dédiée à la détection des références allusives sur un jeu de données aux conventions stylistiques et à la langue différentes des données d'OpenEdition. Bien que les résultats présentés ne se concentrent pas sur l'évaluation des performances de détection des références allusives la résolution de cette tâche est liée à la juste identification de ces dernières. Nous ne pouvons proposer une évaluation propre à la détection des références allusives car ni le jeu de test ni le jeu d'entraînement ne possèdent d'annotation. Cependant, les résultats obtenus nous confortent dans la capacité d'adaption de notre approche. Par ailleurs, les solutions que nous apportons *via* l'identification des références allusives permet d'obtenir une précision supérieure à la meilleure exécution de 2016. Bien que les résultats puissent s'avérer faibles dans leur ensemble, nous rappelons que cette tâche n'a été introduite pour la première fois qu'en 2016 de ce fait, de nombreuses pistes de recherche sont envisageables.

Dans la section suivante, nous présentons une modélisation sous la forme d'un graphe orienté à partir de laquelle nous avons voulu illustrer le fait que les références bibliographiques peuvent permettre d'établir des liens entre les docu-

ments.

2.7. Les références bibliographiques comme outil de liaison entre contenus

Comme nous l'avons souligné au début de ce chapitre, l'exploitation des liens entre contenus est cruciale dans les approches de recommandation. Dans le cadre d'une bibliothèque d'articles scientifiques, les références bibliographiques peuvent s'avérer être des sources de liens. En effet, à partir des références bibliographiques il est possible d'obtenir des informations permettant l'identification d'un document en tant qu'unité documentaire. Parvenir à exploiter ces informations peut nous permettre de faire émerger des liens entre des documents thématiquement liés.

Nous présentons une modélisation sous la forme d'un graphe orienté dont les arcs sont établis *via* les références bibliographiques présentes dans les zones bibliographiques. Nous nous sommes tout d'abord orientée sur ce type de références à cause de leurs conventions plus formelles qui les rendent plus facilement exploitables. De plus, les références présentes dans ces zones sont aussi censé reprendre exhaustivement l'ensemble des références citées au cours du document.

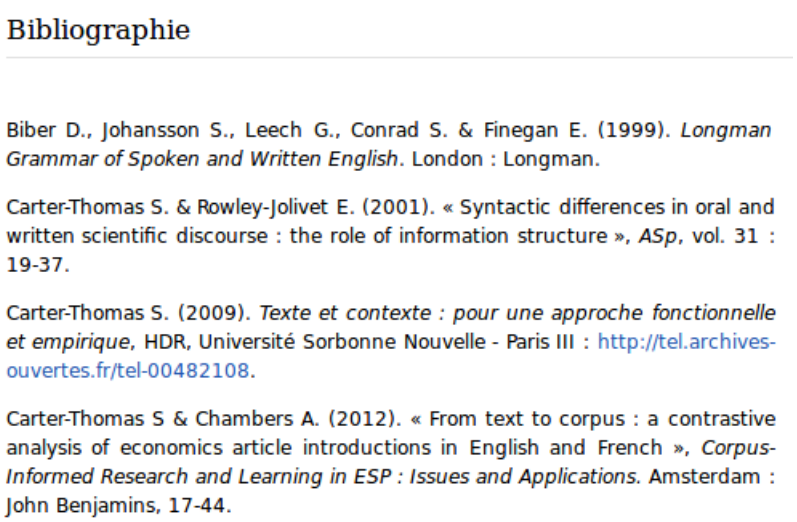
Dans les sections suivantes, nous présentons, dans un premier temps, la revue *Corpus* extraite de la plateforme OpenEdition sur laquelle nous avons réalisé nos modélisations. Dans un second temps, nous proposons, *via* l'exploitation des références bibliographiques de cette revue, une modélisation sous la forme d'un graphe orienté des liens générés entre contenus.

2.7.1. Présentation de la revue *Corpus*

Afin d'établir la représentation de nos données sous forme de graphe, notre choix s'est porté sur la revue *Corpus* [58] consacrée à la linguistique de corpus envisagée sous tous ses aspects : théoriques, épistémologiques et méthodologiques. Le choix de cette revue ne s'est pas fait en tenant compte d'une affiliation disciplinaire particulière mais par rapport à des expérimentations menées sur le nombre de retours positifs de l'index d'OpenEdition sur les références bibliographiques présentes dans les articles. En effet, nous avons souhaité savoir, suite à la construction du corpus présenté dans la section 2.5.1, quelles étaient les revues possédant le plus de références bibliographiques présentes au sein de l'index OpenEdition. L'objectif *via* cette expérimentation est, dans le cadre de la génération d'un système de recommandation de lectures, de permettre en

58. https://corpus.revues.org/

plus de l'exploitation des références bibliographiques d'envisager d'utiliser des algorithmes dédiés au parcours de graphes. Ce type d'application peut nous permettre de faire émerger des documents pertinents dont la portée irait au-delà du document ciblé. Les résultats de cette expérimentation sont présentés en annexe B. Un extrait d'une bibliographie que l'on peut retrouver au sein de la revue *Corpus* est présentée par la figure 2.15.


Bibliographie

Biber D., Johansson S., Leech G., Conrad S. & Finegan E. (1999). *Longman Grammar of Spoken and Written English*. London : Longman.

Carter-Thomas S. & Rowley-Jolivet E. (2001). « Syntactic differences in oral and written scientific discourse : the role of information structure », *ASp*, vol. 31 : 19-37.

Carter-Thomas S. (2009). *Texte et contexte : pour une approche fonctionnelle et empirique*, HDR, Université Sorbonne Nouvelle - Paris III : http://tel.archives-ouvertes.fr/tel-00482108.

Carter-Thomas S & Chambers A. (2012). « From text to corpus : a contrastive analysis of economics article introductions in English and French », *Corpus-Informed Research and Learning in ESP : Issues and Applications*. Amsterdam : John Benjamins, 17-44.

Figure 2.15. – Extrait de la bibliographie de l'article « *Valeurs et fonctions des éléments initiaux commentaires. Analyse contrastive d'un corpus d'articles de recherche en économie* »

Afin de réaliser nos modélisations, nous avons extrait au total 172 articles de cette revue de la plateforme OpenEdition. En moyenne, nous avons pu recenser 13 références bibliographiques par article.

2.7.2. Modélisation des liens entre contenus à l'aide d'un graphe

Afin d'établir les liens entre contenus à l'aide d'un graphe, nous avons, pour chacun des 172 articles provenant de la revue *Corpus*, extrait les références bibliographiques présentes dans la zone bibliographique. Au sein de ce graphe, que nous avons appelé « *Directed Graph of Citations* » (DGC), chaque nœud correspond à un article donné. Chaque nœud possède l'ensemble des propriétés suivantes :

1. ID : l'url de l'article duquel est extrait les références bibliographiques ;
2. Attribut : Une référence bibliographique composée du nom du premier auteur ainsi que du titre de l'article.

Les relations dans le DGC sont orientées et correspondent à la citation au sein d'un article d'une référence bibliographique. Étant donnés les deux nœuds A, B, si A pointe vers B, B est un article citant la même référence que l'article A. Via ce graphe, nous procédons donc aux couplages bibliographiques au sens de (KESSLER 1963) c'est-à-dire que les articles sont couplés bibliographiquement lorsque différents auteurs citent un ou plusieurs articles en commun. Dans la figure 2.16, nous montrons un exemple de la structure du DGC. Ce dernier comporte au total 2963 nœud et 3131 relations.

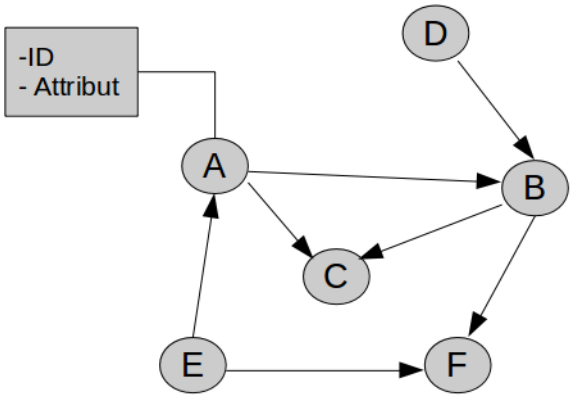


Figure 2.16. – Exemple du « *Directed Graph of Citations* » (DGC)

La figure 2.17 présente un extrait du graphe DGC où chaque nœud est représenté par l'url de l'article de revue (*corpus.revues.org/identifiant*). Comme nous pouvons l'observer les liaisons s'effectuent entre chaque nœud par le biais d'un attribut commun représenté par une référence bibliographique. Dans cet exemple, nous pouvons noter que la référence « *Quirk A comprehensive grammar of the english language* » est commune aux articles possédant les identifiants : 2076, 2110 et 2446. Nous pouvons également observer que l'article possédant l'identifiant 2446 est également lié aux articles 2110 et 2502 par le biais des références « *Combette La construction détachée en français* » et « *Huddleston The cambridge grammar of the english language* ».

Afin de procéder à la construction du graphe DGC, nous avons utilisé un fichier au format graphML à partir duquel il possible de décrire les propriétés structurales d'un graphe *via* le langage XML. Une fois le fichier graphML construit, nous avons utilisé l'outil *Gephi* [59] afin d'obtenir une visualisation du graphe DGC.

59. https://gephi.org/

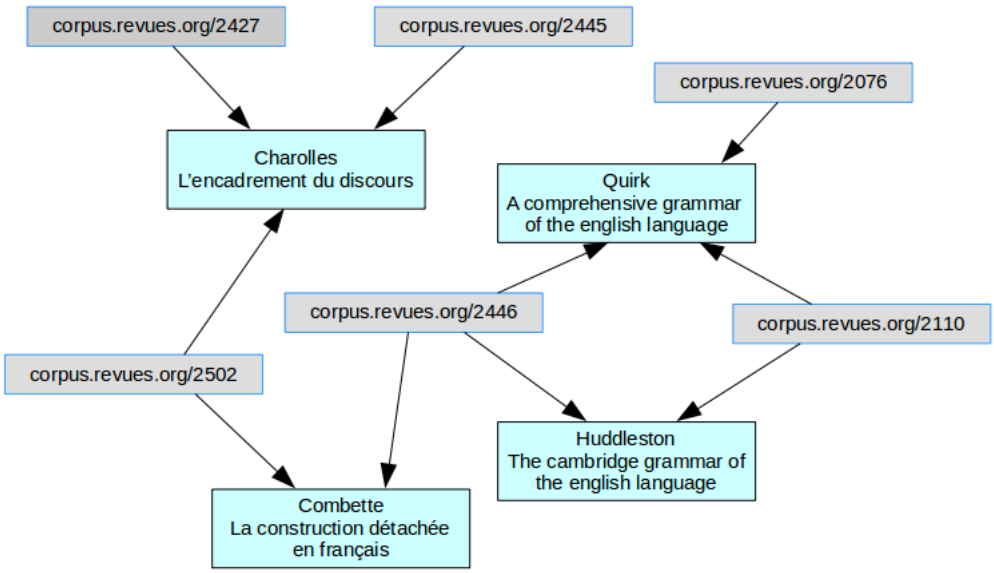


Figure 2.17. – Extrait du « *Directed Graph of Citations* » (DGC) sous la forme d'un graphe biparti

Notre choix s'est porté sur cet outil car ce dernier fournit une compréhension très intuitive du processus de mise en page et de ses paramètres. De plus, à la différence d'outils tels que TouchGraph [60] ou Pajek [61], Gephi à la capacité d'adapter la mise en page à l'échelle et de permettre une exploration dynamique.

Nous avons opté pour l'algorithme de spatialisation nommé *force Atlas 2* (JACOMY, VENTURINI et al. 2014) qui permet de positionner les nœuds d'un graphe afin de faciliter sa visualisation en utilisant un système de forces appliqué entre les nœuds et les arcs. Cet algorithme a la particularité d'être dirigé par la force (*Force-directed algorithms*), en d'autres termes, il simule un système physique afin de spatialiser un réseau. Les nœuds se repoussent comme des particules chargées, tandis que les arcs attirent leurs nœuds. L'attraction et la répulsion conjointe de ces forces permettent de créer un mouvement qui converge vers un état d'équilibre. La figure 2.18 présente une visualisation du graphe généré à partir des articles extraits de la revue *Corpus*. Nous précisons que les références bibliographiques communes ont été représentées visuellement par *Gephi* sous la forme de nœuds or dans les propriétés de représentation des données que nous avons définies plus haut ce sont bel et bien des attributs propres à chaque nœud.

Par le biais des propriétés du modèle de construction du graphe, nous pou-

60. http://www.touchgraph.com/
61. http://mrvar.fdv.uni-lj.si/pajek/

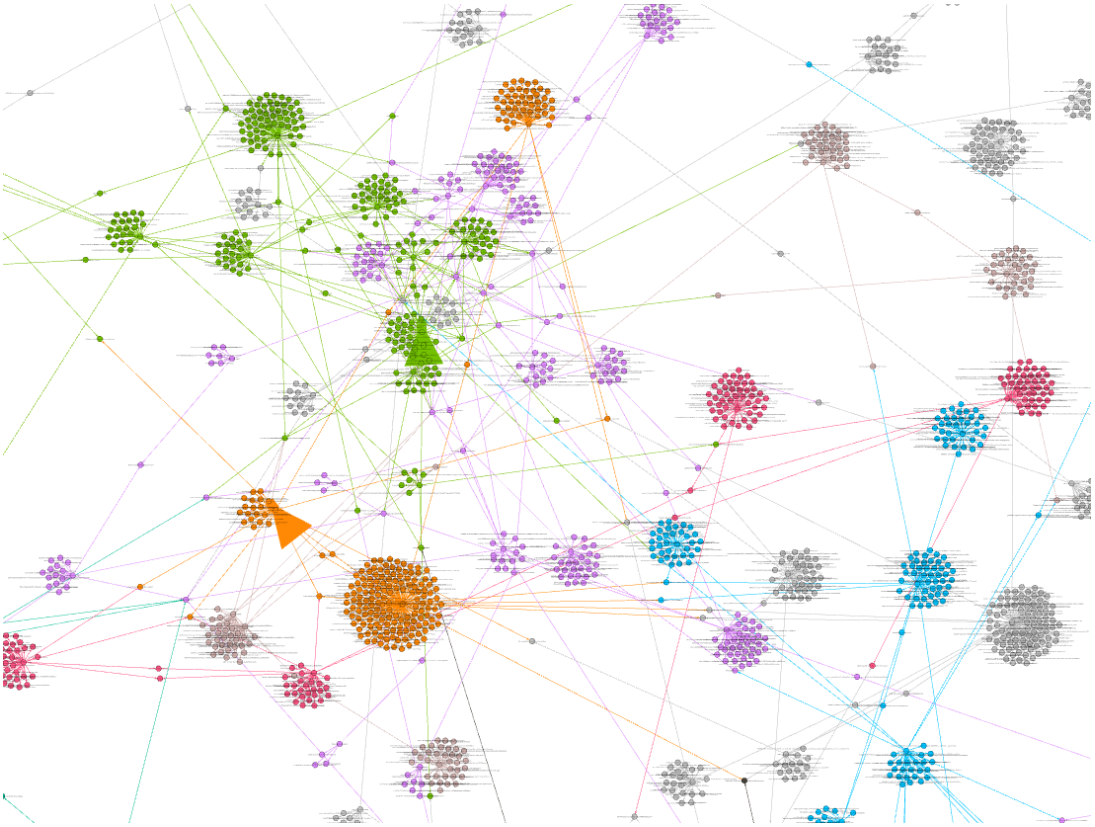


Figure 2.18. – Visualisation du graphe « *Directed Graph of Citations* » (DGC)

vons observer l'émergence de nombreux regroupements entre nœuds. Au vu des propriétés présentées précédemment, ces regroupements s'expliquent par la mobilisation de certaines références par de nombreux articles. *Via* l'algorithme de spacialisation utilisé, nous pouvons observer des nœuds plus épars à partir desquels s'établissent des interconnexions entre des regroupements de nœuds. Ces derniers réfèrent à l'utilisation de références communes entre différents articles. D'autres nœuds sont fortement spatialisés et ne présentent, par ailleurs, aucun arc vers d'autres regroupements. Ce phénomène s'explique par l'absence d'interconnectivité avec d'autres nœuds. La majorité de ces nœuds renvoie à des articles dont les références bibliographiques n'ont été citées nulle part ailleurs. Nous avons également noté la présence de nœuds uniques qui réfèrent à des articles ne possédant pas de zones dédiées aux références bibliographiques. En effet, nous avons pu observer que certains articles dépourvus de zone bibliographiques utilisent les notes afin d'identifier les références tout au long du document. Cette visualisation nous permet d'attester de la connectivité qui réside entre une référence bibliographique et différents articles provenant d'une même revue.

Bien que nos intentions premières ne sont pas orientées vers une analyse des

communautés issues du graphe DGC, nous avons voulu tester l'algorithme de modularité proposé dans *Gephi* (BLONDEL, GUILLAUME et al. 2008) dont les résultats sont également présentés dans la figure 2.18. *Via* cet algorithme, il est possible de décomposer un réseaux en sous-unités ou en communautés par le biais d'ensembles de nœuds fortement interconnectés (FORTUNATO et CASTELLANO 2012), ici chaque couleur correspond à une communauté. En d'autres termes, la modularité est décrite comme la proportion des arcs incidents sur une classe donnée moins la valeur qu'aurait été cette même proportion si les arcs étaient disposés au hasard entre les nœuds du graphe. Dans le cadre de l'algorithme proposé dans *Gephi*, la modularité ΔQ obtenue *via* le déplacement d'un nœud isolé i dans une communauté C est calculée comme suit :

$$\Delta Q = \left[\frac{\sum_{in}+K_{i,in}}{2m} - \left(\frac{\sum_{tot}+K_i}{2m}\right)^2\right] - \left[\frac{\sum_{in}}{2m} - \left(\frac{\sum_{tot}}{2m}\right)^2 - \left(\frac{k_i}{2m}\right)^2\right] \tag{2.12}$$

Où $\sum_{in}$ est la somme des poids des liens à l'intérieur de C, $\sum_{tot}$ est la somme des poids des liens incidents aux nœuds dans la communauté C, k_i est la somme des poids des liens incidents au nœud i, $k_{i,in}$ est la somme des poids des liens de i au nœuds dans la communauté C et m est la somme des poids de tous les liens du réseau.

Comme le présente la figure 2.19, au total 34 communautés ont été identifées au sein du graphe DGC comprenant pour chacune un nombre de nœuds très variable. Nous n'avons pas procédé à une expertise concernant la pertinence des communautés générées mais nous émettons l'hypothèse que ces communautés ont été formées par le biais de rapprochements thématiques. *Via* l'application de cet algorithme, des connexions entre des groupes de nœuds d'une même communauté mais également des connexions entre différentes communautés ont été établies. Ces phénomènes nous permettent d'émettre l'hypothèse de l'existence de références à la fois, propres à une thématique spécifique (*via* son appartenance à une même communauté) et au confluent de plusieurs thématiques spécifiques (*via* son utilisation au sein de différentes communautés). L'utilisation de cet algorithme de création de communautés nous permet d'imaginer, dans le cadre de la réalisation d'un système de recommandation de lectures, l'exploitation, par le biais de références communes, de documents liés (thématiquement) au document courant.

La figure 2.20 illustre l'utilisation d'une référence bibliographique extraite de deux des communautés générées par l'algorithme de modularité. Cette figure

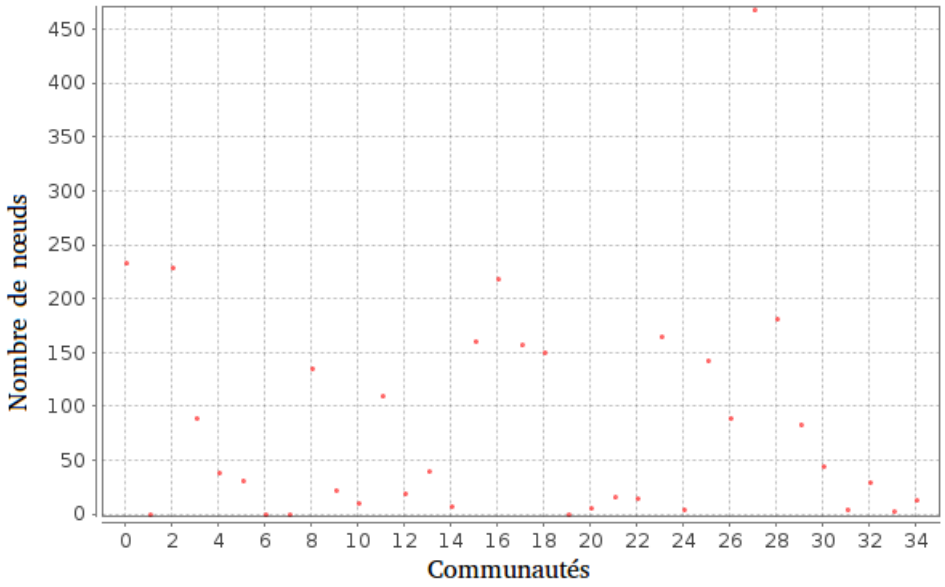


Figure 2.19. – Distribution des tailles de noeuds en fonction des classes

nous montre que la référence au centre en bleu référant à l'article « *Le problème épistémologique du contexte et le statut de l'interprétation dans les sciences du langage* » de F. Rastier est utilisée dans 3 articles différents provenant de deux communautés différentes (bleue et verte). La problématique de cet article s'articule autour de la définition du concept de contexte. Le tableau 2.16 présente une description succincte des articles citant cette référence.

Nous pouvons noter la présence de thématiques connexes entre les articles provenant de la communauté bleue orientée sur des problèmatiques liées à la contextulisation. Concernant l'appariement de cette référence à la communauté verte, il se trouve que dans l'article de F. Rastier, bien qu'orienté autour des phénomènes de contextualisation, une réflexion est posée sur la notion d'intertexte, ce phénomène explique donc les jointures entre communautés. Cette analyse nous permet donc d'envisager l'exploitation, par le biais de références communes, de documents liés (thématiquement) au document courant. Nous pouvons même envisager, à partir de l'exploitation des communautés extraire, à la fois, des documents dont les thématiques sont fortement liées et des documents proposant des thématiques connexes. Bien évidemment, le fait de n'exploiter des articles propres qu'à une revue nous permet de nous concentrer sur une affiliation disciplinaire majoritaire ce qui augmente les chances d'obtenir des interconnexions. Un axe de travail intéressant serait, par ailleurs, d'estimer le degré des interconnexions entre des disciplines différentes. À ce stade, ces

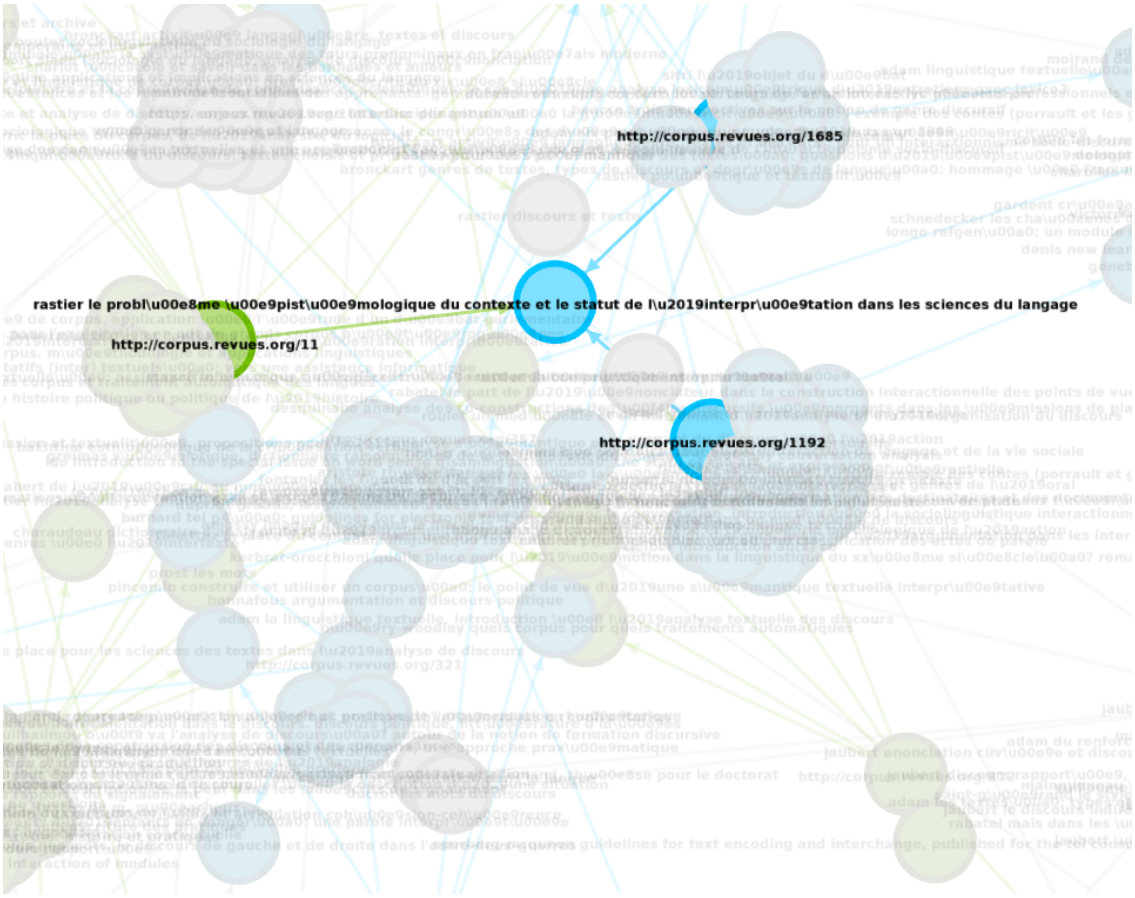


Figure 2.20. – Visualisation détaillée d'une référence intervenant auprès de plusieurs communautés thématiques de la revue *Corpus*

visualisations nous permettent d'ores et déjà d'illustrer l'impact que peut avoir l'utilisation des références comme vecteur d'informations dans la recommandation de lectures.

Pour conclure, dans cette section nous avons pu constater que l'exploitation des références bibliographiques permet de générer des liens entre différents contenus. L'étude du graphe DGC, nous a permis d'observer, par le biais des propriétés du modèle de construction du graphe, des regroupements entre contenus thématiquement liés. Ces travaux nous confortent dans la suite de nos recherches qui envisage d'exploiter en plus des références présentes dans les zones bibliographiques les références allusives comme vecteurs d'informations dédiés à la proposition de recommandation de lectures. En effet, comme nous le présentons dans le chapitre suivant les références allusives vont permettre, à la fois, d'établir des liens entre contenus mais également d'établir des indicateurs d'impact.

Titre	Résumé	Couleur de la communauté
Les corpus réflexifs : entre architextualité et hypertextualité [62]	Un des enjeux actuels du traitement sémantique des corpus textuels concerne la nécessaire tentative de contrôle et d'objectivation de l'intertexte.	Verte
Intertexte générique et interprétation des actes de parole dans un corpus d'émissions de plateaux télévisées [63]	Cet article propose deux mises à l'épreuve d'une modélisation du rôle du contexte dans l'interprétation des actes de parole.	Bleue
Les sphères de contextualisation. Réflexion méthodologique sur les passages de texte à texte(s) et la constitution des corpus [64]	Cet article propose d'interroger la notion de contextualisation à partir d'une perspective d'analyse textuelle des discours.	Bleue

Tableau 2.16. – Description des articles selon leur appartenance à une communauté

2.8. Conclusion

Ce chapitre nous a permis de constater que l'identification des références bibliographiques, et plus particulièrement des références bibliographiques allusives, n'est pas une tâche complètement résolue. En effet, bien que les références bibliographiques puissent apparaître comme des éléments structurés, de nombreux facteurs influencent sur leur structure ce qui rend la généralisation des traitements plus complexe. La multiplication des conventions stylistiques engendre, entre autres, de nombreuses variations impliquant de nombreux schémas de compositions. Nous avons vu que des spécifications à la fois, sur le formatage général ainsi que sur le formatage des champs, impliquent de nombreuses variations sur la structure et la composition des références.

L'étude plus détaillée des références allusives présentes dans les données d'OpenEdition nous a permis d'identifier des types de structure différents fortement liés au degré d'implicite des références. Afin de pallier les problèmes, nous avons proposé une méthode consistant, d'une part, à identifier les paragraphes qui contiennent des références *via* un processus de classification supervisée et, d'autre part, dans l'application de CCRF afin de détecter plus précisément les zones bibliographiques et d'annoter leurs contenus. L'utilisation d'un modèle SVM s'est faite en vue d'établir un pré-filtrage, au vu du nombre important de paragraphes ne comportant pas de références bibliographiques, afin d'identifier les paragraphes en contenant potentiellement. Quant à notre choix, porté sur l'emploi du mo-

dèle CCRF, nous avons estimé que ce type de modèle permettait de tenir compte distinctement de deux facteurs influençant les performances d'analyse des références, à savoir, le positionnement et la composition des références.

Les expérimentations sur les données d'OpenEdition ont permis de rendre compte de résultats satisfaisants concernant l'exécution individuelle de chacun des procédés et ainsi permettre d'attester de sa robustesse. L'éxécution consécutive de tous les procédés a révélé des résultats plus mitigés induits par des erreurs générées au cours de chaque processus. Cependant, les résultats restent encourageants et des solutions d'amélioration ont été proposées afin d'éviter la propagation des erreurs et ainsi augmenter la pertinence des résultats. Notre volonté d'évaluer la portabilité de cette approche sur des données aux conventions stylistiques et à la langue différentes aux données d'OpenEdition par le biais de notre participation à la campagne CLEF SBS 2016, a permis de nous conforter dans sa capacité d'adaptation *via* l'obtention de la meilleure précision. Les caractéristiques choisies ont démontré, au cours de ces expérimentations, leur robustesse ainsi qu'une capacité d'adaptation intéressante au vu des données hétérognènes traitées.

La suite de nos travaux s'est ensuite orientée sur une exploitation plus poussée des références bibliographiques *via* leur utilisation en tant qu'outil de liaison entre contenus. La modélisation sous la forme d'un graphe orienté, nous a permis d'attester de l'interconnectivité qui réside entre une référence bibliographique et différents articles provenant d'une même revue. Nous avons pu notamment observer des regroupements entre contenus thématiquement liés. Ces travaux, nous ont conforté dans la poursuite de nos recherches en vu d'exploiter en plus des références présentes dans les zones bibliographiques les références allusives comme vecteurs d'informations quantitatives dédiés à la proposition de recommandation de lectures.

Dans le chapitre suivant, nous allons développer cet axe *via* l'exploitation des références allusives. En effet, nous avons pu constater que les références bibliographiques permettent d'établir des hyperliens entre contenus. Nous avons choisi d'étendre ces travaux afin d'établir des indicateurs d'impact par le biais de l'exploitation des références allusives. Notre objectif est d'établir, *via* différents facteurs comme la fréquence d'apparition ainsi que la granularité entre chacune des apparitions d'une référence bibliographique, l'influence de cette dernière sur un document. Son utilisation dans le cadre d'un système de recommandation de lectures permettra de mettre en perspective une nouvelle approche basée sur des listes de recommandation classées selon le degré d'influence des références bibliographiques. De plus, en établissant un parallèle avec les travaux menés dans le domaine de la bibliométrie, l'originalité de cet indicateur est d'être basé sur l'analyse du contenu des documents et non plus sur l'exploitation des tradi-

tionnelles bases de données. L'exploitation d'informations comme la fréquence d'apparition et la répartion des références allusives peut permettre d'initier de nouvelles perspectives dans l'évaluation des travaux de recherche.

3. Exploitation des références allusives comme indicateur d'impact au sein d'un système de recommandation

3.1. Introduction

L'étude de l'écrit scientifique est aujourd'hui en plein essor : elle vise en particulier à mieux cerner ce qui fait la spécificité du langage scientifique et à comparer les rhétoriques disciplinaires (TUTIN et GROSSMANN 2013). Au centre des investigations menées, il y a la manière dont sont mises en scène, dans le lexique et le discours, les procédures de découverte et de validation des connaissances. Dans ce contexte d'analyse des stratégies rhétoriques, les citations se sont avérées prépondérantes *via* leur valeur, à la fois, rhétorique et argumentative (FLOREZ 2013). En effet, les références bibliographiques font partie intégrante du discours scientifique et permettent de véhiculer diverses informations relatives aux champs de recherche ou encore à la construction du point de vue de l'auteur (SWALES 1990). Au-delà de l'analyse de l'écrit scientifique, l'exploitation des références bibliographiques suscite un vif intérêt auprès des communautés scientifiques et industrielles. En effet, des domaines de recherche, tels que la bibliométrie, se sont concentrés sur leur analyse afin de rendre compte de l'activité scientifique ou technique par le biais d'études quantitatives (BELTER 2015). Ce type d'analyse traditionnellement basé sur l'exploitation de bases de données n'exploite pas le contenu et notamment l'analyse de l'emploi des références bibliographiques au sein des documents. Les cas d'exploitation de base de données les plus connus sont l'Institut d'information scientifique [1] (ISI) ainsi que Thomson-Reuters [2] et plus récemment Clarivate Analytics [3], en France, nous pouvons citer celle de l'Institut de l'information scientifique et technique [4] (INIST). Ces constatations nous mènent au premier objectif de ce chapitre qui est l'intégration d'une analyse du contenu, *via* l'exploitation des références bibliographiques allusives (cf. : chapitre 2), en vue d'évaluer leurs influences sur l'argumentaire de l'auteur au sein du document courant. Nous proposons de construire un indicateur d'impact à partir de ces dernières, dérivé des critères d'évaluation basés sur des mesures quantitatives « objectives » (OKUBO 1997). En effet, à partir d'obser-

1. http://isithomsonreuters.org/
2. https://www.thomsonreuters.com/en.html
3. https://clarivate.com/products/web-of-science/databases/
4. http://www.inist.fr/?-Databases-&lang=fr

vations menées sur l'utilisation des références bibliographiques allusives au sein d'articles scientifiques, nous avons établi des facteurs d'impact basés sur la fréquence d'apparition ainsi que la granularité de distribution entre chacune des apparitions d'une référence bibliographique. Par le biais de cet indicateur, nous proposons d'estimer le degré d'influence sur l'argumentaire de l'auteur de chaque référence bibliographique au sein du document courant.

Depuis de nombreuses années une interconnexion s'est établie entre la RI et la bibliométrie *via* l'application de méthodes bibliométriques dédiées à la recherche en RI et vice versa. Cette relation a permis entre autres d'évoluer vers la modélisation bibliométrique des processus des systèmes de RI à l'exploitation des relations de citation (MAYR, FROMMHOLZ et al. 2016). Influencée par cet axe de recherche, qui depuis les quarante dernières années s'intéresse à l'expansion des capacités de navigation, nous proposons d'exploiter cette intersection entre la RI et la bibliométrie afin d'élaborer un système de recommandation de lectures, ce qui constitue le second objectif de ce chapitre. Motivée par l'exploitation des relations entre citations afin d'obtenir des capacités de navigation étendues dans le but d'identifier des documents potentiellement pertinents (H. WHITE et MCCAIN 1998 ; GLÄNZEL et SCHUBERT 2004 ; WOLFRAM 2016), nous proposons d'étendre ces travaux *via* l'exploitation des références bibliographiques allusives pour la réalisation d'un système de recommandation basé sur des mesures bibliométriques. À partir de la construction de l'indicateur d'impact évoqué ci-dessus, nous proposons d'établir une liste de recommandations ordonnée en fonction du degré d'impact mesuré de chaque référence du document ciblé.

Ce chapitre est structuré comme suit : dans la section 3.2, nous effectuons une synthèse des travaux relatifs aux interactions relevées entre la RI et la bibliométrie. Ensuite, nous présentons un aperçu des travaux réalisés sur les mesures bibliométriques ainsi qu'au niveau des systèmes de recommandation. Dans la section 3.3, nous exposons l'approche dédiée à la construction de l'indicateur d'impact construit à partir des références bibliographiques allusives. La section 3.4 introduit les processus d'intégration de l'indicateur d'impact dans le cadre de la réalisation d'un système de recommandation de lectures. Dans la section 3.5, nous exposons la méthode mise en place afin d'évaluer le système de recommandation proposé. La section 3.6 présente, en premier lieu, les résultats obtenus dans le cadre de la comparaison des suggestions fournies par notre système de recommandation de lectures et celles fournies par le moteur de recherche *Search OpenEdition*. En second lieu, nous proposons une étude qualitative des documents suggérés par ces deux systèmes.

3.2. État de l'art

Dans cette section, nous présentons, tout d'abord, une revue des interactions relevées au sein de la littérature entre la bibliométrie et la RI. Ensuite, nous proposons un état des lieux sur deux domaines connexes aux travaux de recherche présentés dans ce chapitre, à savoir, la bibliométrie et les modèles de recommandation.

3.2.1. Les interactions entre la recherche d'information et la bibliométrie

La RI et la bibliométrie représentent deux axes de recherche fondamentaux au sein des sciences de l'information. La RI traite des problèmes liés à la collecte, à la représentation, au stockage, à l'indexation ou encore à la récupération de contenus documentaires, sous forme textuelle ou autre (MANNING, RAGHAVAN et al. 2008). La bibliométrie et ses disciplines connexes (infométrie [5], scientométrie [6], cybermétrie [7], webométrie [8]) examinent, quant à elles, quantitativement la production et l'utilisation du discours scientifique (C. WILSON 1999). Chacun de ces domaines a participé à l'amélioration de notre compréhension dans la façon dont l'information est créée, stockée, organisée, récupérée et utilisée. Jusqu'à récemment, les chercheurs ont traité chacun de ces domaines comme des domaines de recherche distincts, n'effectuant que peu de chevauchements entre les sujets de recherche et n'initiant que peu de collaborations entre les chercheurs provenant de ces deux domaines. Or, la RI et la bibliométrie sont étroitement liées.

Les chercheurs en bibliométrie reconnaissent depuis longtemps que des régularités ou des modèles empiriques existent dans la façon dont les informations sont produites et utilisées. Ces régularités s'étendent au contenu des systèmes de RI et à la manière dont les systèmes sont utilisés. Par exemple, des modèles basés sur les interactions des utilisateurs avec les systèmes de RI, peuvent s'avé-

5. L'infométrie est l'étude des aspects quantitatifs de l'information. Cela comprend la production, la diffusion et l'utilisation de toutes les formes d'information, quelle que soit sa forme ou son origine. L'infométrie englobe, par ailleurs, la scientométrie et la bibliométrie.

6. La scientométrie désigne la science de la mesure et l'analyse de la science. Elle est souvent en partie liée avec la bibliométrie cependant à la différence de cette dernière elle n'applique les techniques bibliométriques qu'au champ des études de la science et de la technologie, en comptabilisant les publications scientifiques et elle n'analyse pas seulement les publications mais également des financements, ressources humaines, brevets, etc.

7. La cybermétrie est une discipline comprenant toutes les méthodes et techniques de mesure appliquées au cyberespace et à la population qui l'occupe, principalement les internautes.

8. La webométrie est une discipline spécialisée dans l'analyse des pages et sites Web (et plus précisément des liens hypertextes) par le biais de métriques issues de la bibliométrie, la scientométrie et l'infométrie.

rer utiles au cours de la conception, l'utilisation et l'évaluation de systèmes de RI (WOLFRAM 2003). Dans cette optique, de nombreux aspects se sont prêtés à la modélisation bibliométrique comme la distribution de la fréquence des termes au sein d'un index, les répartitions de fréquence de co-occurrence de termes et plus récemment les aspects basés sur le Web tels que les distributions de fréquences de liens entrants/liens sortans (WOLFRAM 2016). À l'inverse, des techniques de visualisation, propres aux environnements de RI, intialement utilisées pour identifier les documents et leurs relations ont été appliquées aux auteurs, aux groupes de recherche, aux domaines d'études ou aux zones géographiques afin de mieux comprendre les dynamiques complexes inhérentes à la production et à l'utilisation des connaissances (BOLLEN, VAN DE SOMPEL et al. 2009 ; BÖRNER 2010).

La relation mutuellement bénéfique est évidente dans l'application de méthodes bibliométriques dédiées à la recherche en RI et vice versa. Un exemple majeur est le développement et l'utilisation de l'algorithme PageRank (PAGE, BRIN et al. 1999). Cette technique influencée par des idées provenant de l'analyse des citations des années 50-60 initiée par les travaux d'Eugene Garfield a été développée et popularisée par *Google* dans le but de fournir un classement des pages Web (BENSMAN 2013). Elle a depuis été appliquée dans la recherche de métriques évaluatives (DING, YAN et al. 2009 ; DING 2011). La reconnaissance de cette relation mutuellement bénéfique entre ces deux axes de recherche a augmenté au cours des 15 dernières années via l'apparition d'une littérature traitant spécifiquement de ce sujet (par exemple : WOLFRAM 2003 ; MAYR et SCHARNHORST 2015) et la création récente d'ateliers de récupération de l'information améliorée par le biais de la bibliométrie (MAYR, FROMMHOLZ et al. 2016). Un atelier se concentrant sur cette tâche est *Bibliometric-enhanced Information Retrieval*[9] (BIR). Parmi les dernières avancées présentées lors de cet atelier, nous pouvons citer : l'utilisation de l'indice *h* afin de procéder au classement de résultats de recherche (BAR-ILAN et LEVENE 2015) ou encore la proposition d'une méthode d'indexation sémantique dédiée à la RI et l'analyse bibliométrique (BAR-ILAN, JOHN et al. 2016). Plus récemment, la création de l'atelier BIRNDL[10] a permis de mener des réflexions autour des interactions possibles entre la RI, la bibliométrie et le TAL et ce, bien que ces applications linguistiques soient encore relativement nouvelles dans les contextes bibliométriques (CABANAC, CHANDRASEKARAN et al. 2016).

Les nombreuses relations relevées entre ces deux domaines au sein de la littérature, nous ont motivée à orienter nos recherches en vue de mettre à profit les interactions bénéfiques que nous pouvions obtenir via l'association d'un système de RI et de mesures bibliométriques. Notre objectif étant non pas de réaliser des modélisations bibliométriques au sein d'un système de RI mais bel et bien

9. http://ceur-ws.org/Vol-1823/, http://ceur-ws.org/Vol-1610/
10. http://wing.comp.nus.edu.sg/birndl-sigir2017/

d'utiliser les informations issues de mesures bibliométriques dans le cadre de la réalisation d'un système de recommandation de lectures. En effet, comme nous l'avons déjà souligné, grâce à la fonctionnalité de représentation et de recherche de contenus trouvée dans des environnements RI et *via* les hyperliens qui imitent des relations semblables à des citations, ces domaines se prêtent à une utilisation combinée en vue de proposer des approches orientées sur la réalisation de systèmes de recommandation de lectures. De plus, l'exploitation d'une bibliothèque numérique d'articles scientifiques représente un environnement idéal dans lequel étudier l'intersection entre ces deux domaines.

3.2.2. La bibliométrie

Les méthodes bibliométriques sont maintenant solidement établies et font partie intégrante de la méthode d'évaluation de la recherche. Ces méthodes sont de plus en plus utilisées dans l'étude de divers aspects de la science et dans la manière dont les institutions et les universités sont classées dans le monde entier. Dans les sections qui suivent, nous fournissons un aperçu de ce qu'est la bibliométrie ainsi que de son application dans le domaine des SHS.

3.2.2.1. Bibliométrie : un aperçu

La bibliométrie peut être définie comme « l'application des mathématiques et des méthodes statistiques aux livres, articles et autres moyens de communication » (PRITCHARD 1969). Une définition plus contemporaine, associant bibliométrie, scientométrie et infométrie, serait en termes d'« analyse quantitative de l'activité et des réseaux scientifiques » (BORNMANN et MARX 2014). En effet, ce domaine scientifique englobe la mesure des « propriétés des documents et des processus liés aux documents » (BORGMAN et FURNER 2002). La gamme de techniques bibliométriques comprend l'analyse des fréquences de mots, l'analyse des citations, l'analyse des mots-clés et le comptage simple de documents, tels que le nombre de publications d'un auteur, d'un groupe de recherche ou d'un pays.

Bien que des techniques bibliométriques reconnaissables aient été appliquées depuis au moins un siècle, l'émergence de la bibliométrie comme domaine scientifique a été déclenchée (dans les années 1960) par l'ISI suite au développement du *Science Citation Index* (SCI) [11] d'Eugene Garfield (GARFIELD 1979). Le SCI a été créé en tant que base de données des références faite par les auteurs, dans des articles antérieurs, dans leurs articles publiés dans les revues scientifiques les plus importantes, initialement axée sur la science générale et la génétique. Plus tard, cette base de données s'est étendue *via* la création de l'index des cita-

11. http://ip-science.thomsonreuters.com/

tions en siences sociales [12] (SSCI) et en celui des arts et humanités [13] (AHCI). Par le biais de cette base de données, deux types d'application bibliométrique sont apparus : évaluatif et analytique (BORGMAN et FURNER 2002). La bibliométrie évaluative vise principalement à produire, éprouver et appliquer des indicateurs pour évaluer la production scientifique du niveau macro (des pays) au niveau micro (un scientifique). La bibliométrie analytique, quant à elle, vise à questionner, décrire et comprendre des phénomènes liés à la production et à l'exploitation des connaissances scientifiques.

Un certain nombre d'indicateurs ont été élaborés et utilisés par des bases de données comme l'ISI afin de compiler les résultats sur la performance et la productivité des chercheurs. Ces indicateurs se divisent en trois groupes : les indicateurs de production, les indicateurs d'impact et les indicateurs composites. Ces indicateurs peuvent s'appliquer à différentes échelles : micro (un chercheur, un groupe), meso (un département, une université) ou macro (une région, un pays, un continent). Les indicateurs de production font référence aux mesures relatives au calcul du nombre d'articles publiés dont les indices classiques sont les mesures de volume, de « parts de marché » sur une référence donnée (nationale, mondiale, etc.). Les indicateurs d'impact se fondent essentiellement sur des indices basés sur les citations entre articles. La base théorique de ces indicateurs découle des travaux de Robert Merton (MERTON 1957) relatifs aux comportements de citation et plus particulièrement à leur utilisation comme une marque, de la part des scientifiques, d'un travail influant antérieur. Sur cette base, une hypothèse a été émise concernant l'utilisation du comptage des citations comme un indice permettant d'estimer la valeur scientifique. L'un des indices les plus connus, formulé autour de cette hypothèse, est le facteur d'impact (FI) (BENSMAN 2007). Le FI se calcule en tenant compte du nombre de citations reçu au cours d'une année y par des articles publiés dans une même revue au cours des deux années précédentes, divisé par le nombre d'articles dits « citables » [14] publiés dans cette même revue au cours des deux années précédentes :

$$FI_y = \frac{Citations_{y-1} + Citations_{y-2}}{Publications_{y-1} + Publications_{y-2}} \tag{3.1}$$

Les indicateurs composites, également appelés indicateurs synthétiques, correspondent, quant à eux, à un agrégat d'indicateurs individuels. L'indicateur composite le plus connu est l'indice h (ou indice de Hirsch) (HIRSCH 2005). Par le biais de cet indice, il est possible de quantifier la productivité scientifique

12. https://www.thomsonreuters.com/social-sciences-citation-index/
13. https://www.thomsonreuters.com/arts_humanities_citation_index/
14. Les articles dits « citables » (*citable items*) sont le sous-ensemble des publications d'une revue qui contribuent à la documentation scientifique (MCVEIGH et MANN 2009).

et l'impact d'un scientifique en fonction du niveau de citation de ses publications. Formellement, un scientifique a un indice h si h de ses papiers N_p ont au moins h citations chacun et les autres $(N_p - h)$ papiers ont moins de $\leq h$ citations chacun. En d'autres termes, un indice h de 40 signifie, par exemple, qu'un scientifique a publié 40 articles qui ont chacun au moins 40 citations, ainsi que potentiellement d'autres articles avec moins de 40 citations.

Les indices d'impact fondés sur les citations sont toujours en place, mais sont maintenant complétés par une gamme de techniques complémentaires initiée par la mise à disposition de nouvelles sources d'informations sur la communication scientifique, telles que les pages Web et les statistiques d'utilisation des bibliothèques numériques (THELWALL 2008). Les mesures alternatives d'impact, ou *altmetrics* [15], viennent également compléter les indicateurs traditionnels. Les mesures alternatives d'impact visent à faire ressortir l'utilisation de résultats de recherche sur internet, en utilisant, entre autres, les partages sur *Twitter*, *Facebook* ou autres médias sociaux, les téléchargements sur des plates-formes telles que *Mendeley* [16] et les mentions dans les blogs, les wikis et les revues (HAUSTEIN, PETERS et al. 2013 ; COSTAS, ZAHEDI et al. 2014). De nouvelles bases de données alternatives ont également vu le jour permettant l'exploitation d'informations utiles en bibliométrie telles que *Web of Science* [17] (WoS), *Scopus* [18] ou *Google Scholar* [19] ainsi que des bibliothèques et des archives numériques spécifiques telles que *CiteSeer* [20] pour les sciences de l'informatique ou encore les archives *arXiv* [21] dédiées aux domaines de la physique, l'astrophysique, des mathématiques, de l'informatique, de la biologie, etc. (J. LI, BURNHAM et al. 2010). Des initiatives sont même actuellement menées par *OpenCitations* [22], en vue d'établir un consensus autour d'un formatage des citations permettant un traitement simplifié pour les machines, ainsi que la publication des données de citation par les maisons d'édition (actuellement *OpenCitations* concerne 45 % des articles possédant un DOI enregistré auprès de Crossref).

L'analyse bibliométrique suscite aujourd'hui de nombreux intérêts auprès de la communauté scientifique mais aussi parmi les milieux industriels. En effet du côté de la communauté scientifique, les statistiques établies à partir de l'analyse bibliométrique sont des critères importants dans les prises de décisions politiques concernant l'avenir de la recherche financée par l'État. De ce fait, des enjeux majeurs en découlent notamment au sujet du développement d'indicateurs ro-

15. http://altmetrics.org/manifesto/
16. https://www.mendeley.com/
17. https://login.webofknowledge.com
18. http://www.scopus.com/
19. https://scholar.google.fr/
20. http://citeseerx.ist.psu.edu
21. https://arxiv.org/
22. https://i4oc.org/

bustes et fiables. Du côté des entreprises, la bibliométrie est un outil permettant de mettre en œuvre des activités d'intelligence économique, de veille stratégique ou de veille technologique.

3.2.2.2. La bibliométrie en SHS

Dans le cadre de la réalisation de notre indicateur, nous continuons à exploiter les données fournies par OpenEdition provenant d'article en SHS. Suite à l'étude de la littérature, nous nous sommes aperçue que de nombreux problèmes ont été relevés concernant le traitement de ce type de données en bibliométrie. En effet, historiquement de nombreuses distinctions ont été établies entre les SHS et les sciences dures concernant les comportements liés aux citations (ARDANUY 2013). En effet, dans les domaines des sciences dures, les méthodes quantitatives sont devenues une partie intégrante de l'évaluation de la recherche tandis que dans le cas des SHS les méthodes quantitatives d'évaluation de la recherche étaient, ces dernières années, encore peu répandues (WALTMAN 2016). Ce phénomène s'explique principalement suite à des constats concernant la couverture insuffisante des publications par les chercheurs en SHS dans les bases de données de citations disponibles. En effet, des études ont démontré que ce domaine n'utilise pas les caractéristiques bibliographiques appropriées pour l'analyse bibliométrique telle qu'elle est actuellement pratiquée (BORNMANN, THOR et al. 2016). En SHS les résultats de la recherche sont diffusés au travers d'un éventail beaucoup plus large de médias que dans les sciences dures. Cela se reflète notamment dans le rôle plus important joué par les monographies, les documents de conférence ainsi que la littérature grise (TORRES-SALINAS, ROBINSON-GARCIA et al. 2014). Si nous prenons l'exemple des bases de données WoS et *Scopus*, la recherche en bibliométrie dédiée aux SHS se heurte à des problèmes de couverture principalement pour deux raisons : une proportion importante de revues n'est pas incluse dans les bases de données et une grande part des contributions sont faites par le biais de livres et de monographies. En effet, les sujets de recherche en SHS sont parfois plus localement orientés et, par conséquent, le lectorat ciblé est plus souvent limité à un pays ou une région. De ce fait, les universitaires en SHS publient plus souvent dans la langue vernaculaire et non dans les revues (internationales) incluses dans WoS ou *Scopus* (BORNMANN, THOR et al. 2016). Dans les sciences dures, les résultats de recherche sont plus souvent publiés sous forme d'articles au sein de revues spécialisées, dont WoS et *Scopus* présentent une large couverture (TORRES-SALINAS, ROBINSON-GARCIA et al. 2014).

Cependant, les recherches bibliométriques actuelles tendent à s'adapter aux SHS en tenant compte de leurs spécificités. Notamment, *via* des recherches orientées sur une analyse des champs de recherches spécifiques plutôt que sur de vastes collections hétérogènes de disciplines regroupées sous le label « SHS ». Les pratiques actuelles tendent non plus, comme nous venons de le souligner

précédemment, à effectuer des distinctions entre les SHS et les sciences dures mais plutôt vers des approches bibliométriques sensibles à l'organisation des champs de recherche en SHS (HAMMARFELT 2016). Des exemples de telles initiatives comprennent l'inclusion des « éléments non-sources » au sein de bases de données telles que WoS (HAMMARFELT 2011 ; LINMANS 2010) mais aussi par l'introduction de nouveaux services, qui ont par ailleurs influencé un grand nombre d'études (KOUSHA, THELWALL et al. 2011 ; GORRAIZ, PURNELL et al. 2013), tels que *Google Book Search* [23], *Google Scholar* [24] et *The Book Citation Index* [25]. Récemment, les possibilités offertes par des mesures alternatives dédiées aux SHS ont également été étudiées (HOLMBERG et THELWALL 2014 ; HAMMARFELT 2014). Dans cette même optique des projets de recherche ont vu le jour. Parmi ces projets, nous pouvons citer le projet RHECITAS (TANGUY, LALLEMAN et al. 2009) porté par le CNRS, et plus particulièrement par TGE-ADONIS [26]. Par le biais de ce projet, l'objectif principal est de promouvoir les humanités numériques au travers de l'analyse de citations dans le domaine des SHS via des procédés issus de la linguistique. Ce projet s'oriente notamment sur la caractérisation des rôles joués par les citations (importantes, de fond, superficielles) ainsi que sur l'extraction d'informations relatives aux intentions de l'auteur au travers des citances.

Des problèmes restent cependant inhérents aux pratiques utilisées en SHS comme l'utilisation d'objets de recherche anciens. Par exemple, une fenêtre de trois à cinq ans à partir du temps de publication est recommandée comme le choix idéal pour les citations dans les sciences naturelles et l'ingénierie tandis que les chercheurs en arts et en sciences humaines ont tendance à citer des ouvrages plus anciens (CHANG 2013). Or, la durabilité est rarement mesurée dans les exercices d'évaluation de la recherche. De plus, la construction d'indicateurs généraux est entravée par la nature hétérogène de la recherche. En effet, des études démontrent qu'il est difficile d'effectuer une analyse bibliométrique en SHS avec une méthode, à la fois, unique et polyvalente en raison du large éventail de disciplines (BORNMANN, THOR et al. 2016). À ces constations s'ajoutent les difficultés concernant les tentatives d'identification d'indicateurs de qualité dans les SHS pour lesquels plusieurs normes contradictoires ont été trouvées (OCHSNER, HUG et al. 2013).

Cette synthèse des travaux portant sur la bibliométrie, nous a permis de fournir un aperçu des problématiques relatives à ce domaine. Nous avons pu relever que de nombreuses difficultés se posent autour de la création mais aussi de l'application de mesures bibliométriques. De plus, nous avons pu noter que la portabilité de ces mesures est limitée et que, bien que l'émergence de nouvelles bases de

23. https://books.google.fr/
24. https://scholar.google.fr/
25. http://wokinfo.com/products_tools/multidisciplinary/bookcitationindex/
26. https://leo.hypotheses.org/2456

données et mesures alternatives offre de nouvelles perspectives, des difficultés persistent autour de la mise en place de procédés permettant le traitement de données hétérogènes tels que nous avons pu le relever au cours de la section 3.2.2.2. Suite à ces constatations, nous nous sommes penchée sur la conception d'un indicateur permettant d'associer une application bibliométrique à la fois évaluative et analytique. En effet, *via* cet indicateur nous proposons d'estimer l'impact d'une référence précise au sein d'un document particulier ce qui par la même occasion nous permet de qualifier les relations entre les citations au sein d'un même document. De plus, cet indicateur se veut, à la fois, unique et polyvalant car il ne tient pas compte des aspects tels que la discipline d'application ou le format de publication.

Dans la section qui suit nous présentons un aperçu, à la fois des approches de recommandation existantes ainsi que des systèmes réalisés à partir de ces approches.

3.2.3. Les modèles de recommandation

Il est nécessaire de filtrer et de classer par ordre de priorité les documents afin d'établir des recommandations pertinentes. Les systèmes de recommandation tentent de résoudre ce problème en parvenant à extraire des informations à partir d'un grand volume de données. En effet, définis comme « *une stratégie de prise de décision pour les utilisateurs dans des environnements d'information complexes* » (RASHID, ALBERT et al. 2002), ces systèmes permettent d'atténuer le problème de surchage d'information en fournissant des recommandations de contenu ainsi que des services personnalisés. Ils sont devenus de plus en plus populaires et sont utilisés dans une variété de domaines comme les films, la musique ou encore les articles de recherche (DE NART et TASSO 2014). Il existe également des systèmes de recommandation dédiés à des experts de domaine (H.-H. CHEN, ORORBIA II et al. 2015), à la recherche de collaborateurs (H.-H. CHEN, GOU et al. 2011 ; CABANAC 2011) ou pour les fils de *Twitter* (Pankaj GUPTA, GOEL et al. 2013).

Parmi les méthodes les plus couramment utilisées nous pouvons distinguer trois types d'approches : les systèmes à base de filtrage collaboratif, les systèmes de filtrage basés sur le contenu et les systèmes de filtrage hybride.

Les systèmes à base de filtrage collaboratif sont actuellement les plus matures et les plus couramment mis en œuvre (ISINKAYE, FOLAJIMI et al. 2015). Ces systèmes tentent de prédire l'intérêt des items pour un utilisateur en fonction des items précédemment évalués par d'autres usagers. Plus formellement, l'utilité de l'item pour l'utilisateur est estimée sur la base des utilités attribuées à ce même item par les usagers qui ont des goûts similaires à l'utilisateur actif. Par exemple, dans le cadre d'une recommandation de films, ces techniques vont essayer de

trouver des utilisateurs aux goûts semblables par le biais de calculs de similarité entre profils utilisateurs ou encore *via* les évaluations des autres utilisateurs. De nombreux systèmes collaboratifs ont été mis au point dans le milieu universitaire et dans l'industrie. Le système *Grundy* (RICH 1979), par exemple, a été l'un des premiers systèmes de recommandation à proposer d'utiliser des stéréotypes afin de construire un modèle utilisateur. Plus tard, *Video Recommender* (HILL, STEAD et al. 1995), *GroupLens* (KONSTAN, MILLER et al. 1997) et *Ringo* (L.-S. CHEN, HSU et al. 2008) furent les premiers systèmes à utiliser le filtrage collaboratif afin d'automatiser leurs prédictions. D'autres exemples de systèmes de recommandation collaboratifs incluent le système de recommandation de livres d'*Amazon*, le système *PHOAKS* qui aide les usagers à trouver des informations pertinentes sur le Web (TERVEEN, HILL et al. 1997), et le système *Jester* qui recommande des blagues (GOLDBERG, ROEDER et al. 2001).

Ce type de système utilise principalement les informations contenues dans la matrice d'usages comme donnée d'entrée. Cette matrice est construite à partir des comportements des utilisateurs ou de leurs avis sur les items qu'ils connaissent déjà. Il existe deux types principaux de filtrage collaboratif. Le filtrage collaboratif passif qui repose sur une analyse des comportements utilisateurs faite en « arrière-plan » de manière implicite (les achats effectués, les pages visitées, etc.) et le filtrage collaboratif actif qui repose sur du déclaratif (notes, commentaires) explicite de la part des utilisateurs. Le filtrage collaboratif passif ne nécessite aucune participation de la part du consommateur. Au lieu de cela, des informations sont collectées sur tous les utilisateurs lorsqu'ils naviguent sur un site ou des ressources. Un exemple courant de cette technique, extrait d'*Amazon*, est la rubrique « produits récemment consultés » qui reflète l'historique de navigation de l'utilisateur sur le site. Un autre dispositif est également disponible *via* la rubrique « les utilisateurs qui ont acheté x ont aussi acheté y » dans laquelle l'utilisateur dispose d'une liste des articles similaires au produit recherché à l'origine. Contrairement au filtrage passif, le filtrage collaboratif actif nécessite une participation active des internautes qui offrent leurs opinions. Un exemple classique de cette méthode est les « Commentaires clients » proposés également par *Amazon* dans lesquels les utilisateurs, anonymement ou non, sont autorisés à donner des avis sur les produits *via* un barème de notation allant d'une étoile à cinq étoiles

Les systèmes de filtrage basés sur le contenu, aussi appelés filtrages cognitifs, recommandent des items en se basant sur une comparaison entre le contenu des items et un profil utilisateur (DUNJA 1999). Le contenu de chaque item est représenté *via* un ensemble de descripteurs ou de termes, typiquement les mots qui apparaissent dans le document ou les métadonnées. Le profil utilisateur est, quant à lui, représenté avec les mêmes termes et construit en analysant le contenu des items qui ont été lus ou achetés par ce dernier. Ces techniques basent leurs prédictions sur les informations de l'utilisateur actif sans tenir compte des

contributions des autres usagers. Autrement dit, la sélection de documents se base sur une comparaison des thèmes abordés dans les documents par rapport aux thèmes intéressant l'utilisateur. Par exemple, cette technique peut être utilisée pour filtrer les résultats de recherche en décidant si un utilisateur est intéressé par une page Web ou non et, dans le cas négatif, empêcher qu'elle soit affichée. Ces techniques se tiennent à l'intersection de plusieurs domaines informatiques, notamment de la RI et de la fouille de données (LOPS, DE GEMMIS et al. 2011).

Dans les systèmes de RI, l'utilisateur exprime un besoin ponctuel d'information en donnant une requête (habituellement une liste de mots-clés), tandis que dans les systèmes de filtrage d'information, les besoins d'information de l'utilisateur sont représentés par son propre profil. Les items à recommander peuvent être de nature très différente tout comme les attributs utilisés pour les décrire. Généralement basés sur des valeurs connues il se peut que ces derniers n'aient pas de valeurs biens définies, dans ce cas, des techniques de modélisation provenant de la RI sont utilisées. La tâche de recommandation peut aussi s'apparenter à un problème de classification. En effet, à partir de l'historique de navigation de l'utilisateur ces techniques apprennent un profil. Cela implique généralement l'application de techniques d'apprentissage automatique, dont l'objectif est d'apprendre à catégoriser de nouveaux éléments sur la base d'informations précédemment vues qui ont été explicitement ou implicitement étiquetées comme intéressantes ou non par l'utilisateur.

De nombreuses applications basées sur ces techniques ont été élaborées, parmi elles, nous pouvons citer : *Movielens*[27] et *Pandora Radio*[28]. *Movielens* recommande des films en fonction des évaluations effectuées par l'utilisateur. Quant à *Pandora Radio*, ce populaire système de recommandation de musique, choisit des chansons ayant des caractéristiques similaires à la chanson fournie par l'utilisateur. Il existe bien d'autres domaines pour lesquels des systèmes de filtrage basés sur le contenu ont été élaborés notamment des systèmes de recommandation cinématographiques comme *Rotten Tomatoes*[29], *Internet Movie Database*[30], *Jinni*[31], *TiVo Corporation*[32] et *Jaman*[33].

Les systèmes de filtrage hybride proposent une alternative. En effet, malgré le succès des systèmes de filtrage collaboratif et ceux basés sur le contenu,

27. https://movielens.org/
28. http://www.pandora.com/
29. https://www.rottentomatoes.com/
30. http://www.imdb.com/
31. http://www.jinni.com/
32. https://business.tivo.com/
33. https://jaman.com/

plusieurs limites ont été identifiées. Pour les systèmes de filtrage basés sur le contenu des difficultés ont été identifiées sur des aspects tels que l'analyse limitée du contenu, la sur-spécialisation et la rareté des données (LOPS, DE GEMMIS et al. 2011). Concernant les méthodes de filtrage collaboratif, elles présentent des problèmes de démarrage à froid, d'incomplétude et d'évolutivité (ADOMAVICIUS et TUZHILIN 2005). C'est donc afin de pallier ces limitations que des approches de filtrages hybrides ont été proposées (GÖKSEDEF et GÜNDÜZ-ÖĞÜDÜCÜ 2010). Majoritairement basées sur la combinaison des deux méthodes présentées précédemment, plusieurs types de combinaisons ont été identifiés. Le tableau 3.1 extrait de (BURKE 2002) présente les différents types d'hybridation réalisés.

Filtrage Hybride	Description
Pondéré	Les scores de plusieurs techniques de recommandation sont combinés afin de ne produire qu'une seule prédiction.
Commutation	Le système bascule entre les techniques de recommandation en fonction de la situation actuelle.
Mixé	Les recommandations de plusieurs systèmes différents sont présentées en même temps.
Combinaison de caractéristiques	Les caractéristiques de différentes sources de données de recommandation sont regroupées.
Cascade	Un système va affiner les recommandations fournies par un autre système.
Augmentation des caractéristiques	La sortie d'un système est utilisée comme caractéristique d'entrée pour un autre.
Méta-niveau	Le modèle appris par un système de recommandation est utilisé comme entrée pour un autre.

Tableau 3.1. – Différentes méthodes d'hybridation

Une recommandation hybride pondérée consiste à combiner les scores d'un item produits par différents systèmes. Le Système *P-Tango* (CLAYPOOL, GOKHALE et al. 1999) utilise ce type d'hybridation *via* la combinaison des scores obtenus suite à l'utilisation de techniques de filtrage collaboratif et basées sur le contenu. Dans les systèmes hybrides basés sur des commutations un critère est utilisé pour basculer entre les différentes techniques de recommandation. Par exemple, le système *DailyLearner* (T. TRAN et R. COHEN 2000) utilise un hybride contenu/-collaboratif dans lequel une méthode de recommandation basée sur le contenu est utilisée en premier. Si le système basé sur le contenu ne peut pas faire une recommandation avec suffisamment de confiance, une recommandation basée sur du filtrage collaboratif est alors tentée. Lorsqu'il est pratique de faire un grand

nombre de recommandations simultanément, il est possible d'utiliser une hybridation, dites « mixée », où les recommandations de plusieurs techniques sont présentées ensemble. Par exemple, le système *PTV* (SMYTH et COTTER 2000) utilise, d'une part, des techniques basées sur le contenu fondées sur des descriptions textuelles d'émissions télévisées et, d'autre part, des informations issues du filtrage collaboratif sur les préférences d'autres utilisateurs. Les recommandations des deux techniques sont ensuite fusionnées. Un autre moyen de réaliser la fusion entre les méthodes de filtrage contenu/collaboratif est d'utiliser les informations collaboratives comme des données sur des caractéristiques supplémentaires associées à chaque item et des techniques basées sur le contenu sur cet ensemble de données augmenté (BASU, HIRSH et al. 1998). Dans les systèmes hybrides en cascade, une technique de recommandation est utilisée pour produire un classement grossier des candidats et une seconde technique affine la recommandation parmi l'ensemble des candidats. Dans le cas des systèmes hybrides basés sur une augmentation des caractéristiques, une technique est employée pour produire une classification d'un item et cette information est ensuite incorporée dans le traitement de la technique de recommandation suivante (MOONEY et ROY 2000). Par exemple, *GroupLens* travaille avec le filtrage d'actualité de *Usenet*[34] et s'en sert afin d'augmenter ses caractéristiques. Une autre façon de combiner deux techniques de recommandation peut être d'utiliser le modèle généré par un système comme entrée pour un autre : appelée modèle hybride méta-niveau, cette approche diffère de l'augmentation de caractéristiques dans laquelle un modèle appris pour générer des caractéristiques d'entrée à un deuxième algorithme est utilisé tandis que dans un hybride méta-niveau, le modèle entier devient l'entrée (SCHWAB, KOBSA et al. 2001).

Plus récemment, des services tels que *Netflix*[35] et *Babelio*[36] se sont servi de systèmes de filtrage hybride. *Netflix* effectue sa recommandation en comparant les habitudes de visionnage et de recherche des utilisateurs similaires à celui de l'utilisateur ciblé (filtrage collaboratif) et il propose également des films qui partagent les mêmes caractéristiques que les films notés par l'utilisateur ciblé (filtrage basé sur le contenu). *Babelio*, quant à lui, propose de cataloguer (en notant) des artistes et des œuvres pour définir sa bibliothèque de goûts à partir de laquelle l'algorithme (basé sur du filtrage collaboratif) propose de nouvelles choses à découvrir et propose un accès vers des voisins (membres ayant les mêmes centres d'intérêt). Le site, contributif, permet à ses membres d'enrichir la base de données.

Une grande partie des recommandations sont basées sur l'exploitation de requêtes ou de profils utilisateurs. Or, notre cadre applicatif, qui est orienté sur

34. https://en.usenet.nl/
35. https://www.netflix.com/fr/
36. http://www.babelio.com/

l'exploitation des données issues de la plateforme *OpenEdition*, permet des accès sans connexion. Par conséquent, nous ne sommes pas en mesure d'utiliser des profils utilisateurs. De ce fait, nous proposons de nous focaliser sur l'analyse du contenu des documents. Pour ce faire, nous proposons une approche reprenant les fondements des systèmes de filtrage basés sur le contenu, à savoir, proposer des items relatifs aux thèmes abordés dans les documents par rapport aux thèmes intéressant l'utilisateur. La différence notoire du système que nous proposons, comparativement aux systèmes de filtrage basés sur le contenu, se situe au niveau de l'exploitation du profil de l'utilisateur actif. En effet, au sein du système que nous proposons nous n'effectuons pas de comparaison entre des items et un profil utilisateur mais nous proposons de déterminer les intérêts de l'utilisateur actif à partir de l'analyse du contenu de l'article sélectionné par ce dernier. Afin d'extraire des thématiques connexes à ce document, et donc relatives aux intérêts de l'utilisateur, nous proposons d'établir un système basé sur des mesures bibliométriques. Nous proposons, *via* la génération d'un indicateur d'impact propre à chaque référence bibliographique, une liste de recommandations de lectures ordonnée selon leurs impacts sur le document ciblé. Par le biais de cette approche, nous suggérons uniquement des documents cités au sein du document ciblé. De ce fait, nous obtiendrons des documents que nous supposons thématiquement liés au document courant.

Au cours de la section suivante, nous introduisons la méthode de mesure de l'indicateur d'impact ainsi que les facteurs dont nous avons tenu compte afin de réaliser nos mesures.

3.3. Les références allusives pour la construction des indicateurs d'impact

Dans le cadre de nos travaux, nous avons décidé de dériver l'utilisation traditionnelle des indicateurs d'impact (cf : section 3.2.2) *via* la génération d'un indicateur d'impact réalisé par le biais d'analyses quantitatives menées sur les références allusives. Dans les sections qui suivent, nous présentons la méthode globale de création de l'indicateur d'impact ainsi que les facteurs dont nous avons tenu compte afin d'établir nos mesures.

3.3.1. Méthode globale de construction de l'indicateur d'impact basée sur l'analyse des références bibliographiques allusives

À la différence des mesures bibliométriques traditionnellement basées sur l'exploitation des bases de données telles que l'ISI ou BIOSIS dédiée aux Sciences

de la vie, l'indicateur que nous proposons se base sur l'analyse du contenu des documents. Nous avons considéré plusieurs facteurs d'impact issus de l'analyse des références bibliographiques allusives au sein du discours scientifique. Nous nous sommes basée sur des facteurs issus d'analyses quantitatives basés principalement sur deux considérations : la fréquence d'apparition et la granularité de distribution de chacune des références allusives au sein d'un même document. Le premier facteur va permettre d'estimer le nombre d'occurences global d'une référence et le second va permettre, quant à lui, de qualifier leur répartition sur l'ensemble du document. La figure 3.1 présente en détail la procédure globale de création de cet indicateur d'impact.

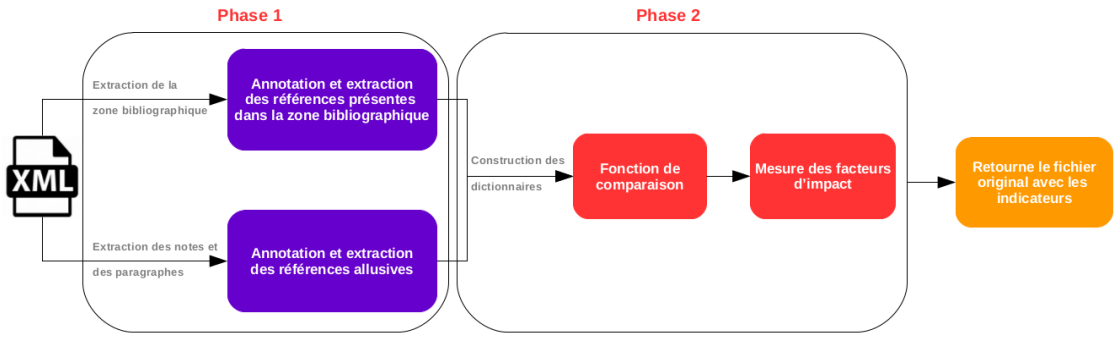


Figure 3.1. – Procédure globale de création des indicateurs d'impact. La couleur violette renvoie aux processus permettant l'application des modèles dédiés à l'annotation des références bibliographiques ainsi qu'à leur extraction, la couleur rouge à l'application des processus relatifs à la construction de l'indicateur d'impact et la couleur orange au(x) produit(s) résultant de la chaîne de traitement.

Précisément, nous prenons en entrée un document XML formaté selon la TEI. Au cours de la phase 1, à partir de ce document et des annotations fournies par la TEI, nous extrayons simultanément les références présentes dans les zones bibliographiques ainsi que celles présentes dans les notes et le corps du texte (références allusives). Les références présentes dans la zone bibliographique sont extraites afin de pouvoir établir une mesure propre à chaque référence. Nous avons estimé que, de par leur formalisme, les références présentes dans ces zones peuvent permettre de faire converger en un seul point toutes les références allusives relatives à une seule unité documentaire. Les références bibliographiques présentes dans la zone bibliographique ainsi que celles présentes dans les notes et le corps du texte sont ensuite présentées au logiciel Bilbo afin d'obtenir l'annotation des références selon leur zone d'apparition dans le document. Ces références sont annotées et extraites *via* les traitements présentés au cours de la section 2.4.

Au cours de la phase 2, plusieurs dictionnaires sont construits pour chacune des catégories de référence. Le dictionnaire dédié aux références présentes dans la zone bibliographique contient chacune des références de la liste ainsi que son indice d'apparition dans cette même liste. Concernant le dictionnaire relatif aux références allusives, ce dernier contient chacune des références allusives associée à des bornes de début et de fin correspondantes à leur position dans le document ainsi que le numéro de leur paragraphe d'apparition. La conservation de ces bornes et du numéro de paragraphe est destinée à établir les différents appariements entre chaque référence bibliographique se rapportant à une même unité documentaire effectués *via* l'utilisation des fonctions de correspondance appliquées lors de l'étape suivante. Ces fonctions de correspondance, dont nous détaillons le fonctionnement dans la section 3.3.2, sont employées afin de permettre d'effectuer différents appariements entre les références présentes au sein des dictionnaires. Suite à ces appariements nous procédons à la mesure des facteurs d'impact basée sur la fréquence d'apparition et la granularité de distribution de chacune des références allusives, dont nous expliquons le principe dans la section 3.3.3. Suite à ces processus, nous obtenons, en sortie, le document XML original auquel est agrégé l'indicateur d'impact propre à chaque référence présente dans la zone bibliographique.

Dans la section suivante, nous détaillons les fonctions de correspondance permettant de réaliser les appariements entre les références bibliographiques se rapportant à une même unité documentaire.

3.3.2. Fonctions de correspondance

Comme nous l'avons énoncé au cours de la section 3.3.1, l'utilisation de fonctions de correspondance nous permet d'effectuer une comparaison entre les dictionnaires établis suite à l'extraction des listes de références. Ces fonctions permettent d'établir des correspondances entre le dictionnaire relatif aux références allusives et celui relatif aux références présentes dans la zone bibliographique afin d'identifier les références allusives se rapportant à une même unité documentaire. L'objectif *via* ces correspondances est de permettre la mise en place du facteur dédié à la fréquence d'apparition présenté dans la section 3.3.3.1. Ces fonctions permettent également d'établir une comparaison entre les différentes références allusives relatives à une même unité documentaire présentes au sein d'un même dictionnaire. À partir de ces correspondances, nous établissons les facteurs de granularité de distribution présentés dans la section 3.3.3.2. Au total, deux fonctions ont été réalisées : une fonction effectuant une correspondance stricte et une fonction basée sur une distance de Levenshtein.

La fonction effectuant une correspondance stricte se base sur la recherche

d'une correspondance complète entre le contenu des champs bibliographiques relevé au sein d'une référence et le contenu de ces mêmes champs relevé au sein d'une autre référence et ce, même si les champs ne sont pas présents dans le même ordre. Les exemples présentés dans les figures 3.1 et 3.2 fournissent un cas de figure dans lequel la fonction de correspondance stricte est appliquée.

```
<bibl>
       <author><surname>Amossy</surname></author> <date>1999</date>
</bibl>
```

Listing 3.1 – Référence bibliographique allusive

```
<bibl>
   <author><surname>Amossy</surname>, <forename>Ruth</forename></author>. <date
       >1999</date>. <title>Images de soi dans le discours. La construction de l
       'ethos </title>(<pubPlace>Lausanne</pubPlace>: <publisher>Delachaux &amp;
        Niestlé</publisher>)
</bibl>
```

Listing 3.2 – Référence présente dans la zone bibliographique

La fonction basée sur une mesure de similarité permet quant à elle d'effectuer des correspondances entre des références dont le contenu des champs bibliographiques est sensiblement différent. En effet, suite à l'étude du corpus de test nous avons pu constater la présence de variations entre le contenu de certains champs, nous avons donc choisi de pallier ce problème *via* l'utilisation d'une mesure de similarité. Afin d'établir cette mesure de distance entre mots, nous avons opté pour une distance de Levenshtein. Notre choix s'est porté sur cette distance suite aux résultats satisfaisants que nous avons obtenus (section 2.6). Dans les exemples des figures 3.3 et 3.4, nous présentons un cas dans lequel intervient la fonction de correspondance basée sur une distance de Levenshtein.

```
<bibl>
       <author><surname>Watzlawick</surname></author> <abbr>et al.</abbr> <date
           >1972</date>
</bibl>
```

Listing 3.3 – Référence bibliographique allusive

```
<bibl>
   <author><surname>Watzlawick</surname>, <forename>Paul</forename></author>, <
       author><forename>Janet</forename> <surname>Helmick</surname> <forename>
       Beavin</forename></author> &amp; <author><surname>Don</surname> <forename
       >D.</forename> <forename>Jackson</forename></author>. <date>1972-1979</
       date>. <title>Une logique de la communication</title> (<pubPlace>Paris</
       pubPlace>: <publisher>Seuil</publisher>)
</bibl>
```

Listing 3.4 – Référence présente dans la zone bibliographique

Comme nous pouvons l'observer dans cet exemple entre ces deux références le contenu des balises <date> est différent, l'utilisation de la fonction basée sur une distance de Levenshtein, dans ce cas, permet tout de même d'établir une correspondance entre les deux références. Afin de mieux appréhender ces fonctions nous allons définir ci-dessous leur fonctionnement.

Considérons R comme l'ensemble des références bibliographiques allusives, tel que :

$$R = \{r_1, r_2, ..., r_n\} \tag{3.2}$$

Considérons R' comme l'ensemble des références avec lequel nous souhaitons établir une correspondance (R peut être égal à R'), tel que :

$$R' = \{r'_1, r'_2, ..., r'_m\} \tag{3.3}$$

Definition 1. (**Fonction de correspondance stricte**) Cette fonction recherche si l'ensemble des champs bibliographiques de la référence r_i ($i \in [1..n]$) et la référence r'_j ($j \in [1..m]$) correspondent. r_i est segmentée en mot (w_k) et r'_j est segmentée en mot (w'_k), avec $k = min(len(r_i), len(r'_j))$.

$$correspondance_{stricte}(r_i, r'_j) = \begin{cases} 1 \text{ si } [w_1\ w_2 ... w_k] = [w'_1\ w'_2 ... w'_k] \\ 0 \end{cases} \tag{3.4}$$

Definition 2. (**Fonction de correspondance à base de similarité**) $sim_{\theta,\delta}$ correspond à l'application de la distance de Levenshtein. θ réfère au vecteur $(w_1..w_k)$ de tokens dans r_i et δ correspond au vecteur $(w'_1..w'_k)$ de tokens extrait de r'_j, avec $k = min(len(r_i), len(r'_j))$.

$$correspondance_{similarité}(r_i, r'_j) = \begin{cases} 1 \text{ si } sim_{\theta,\delta} \geqslant \omega \\ 0 \end{cases} \tag{3.5}$$

Concernant l'application de ces fonctions, des difficultés lors de l'appariement des références allusives vers la référence présente dans la zone bibliographique correspondante sont observées. Les principales difficultés concernent les erreurs

d'annotation générées au cours de l'identification des différents champs bibliographiques, la variation du contenu d'un même champ bibliographique entre des références se rapportant à une même unité documentaire et l'occultation de ces derniers.

Concernant les erreurs d'annotation, seules les références présentes dans la section bibliographique de fin d'article ont été annotées par l'outil Bilbo. Afin d'estimer l'existence d'un impact réel *via* l'exploitation des références bibliographiques allusives dans le cadre d'un système de recommandation de lectures nous avons utilisé des articles pour lesquels les références allusives ont été préalablement annotées. Cependant, ces corpus n'étant pas initialement prévus pour le traitement des références présentes dans les zones bibliographiques, ils ne sont pas pouvus d'annotations à ce niveau. De ce fait, bien que des résultats satisfaisants aient été observés au cours de précédents travaux (Y.-M. KIM, BELLOT et al. 2011 ; Y.-M. KIM, BELLOT, TAVERNIER et al. 2012), des erreurs d'annotation ont été notifiées. Les références présentes dans les zones bibliographiques bien que structurées ne possèdent pas un seul et même formalisme (cf : section 2.3.1.1) ce qui engendre des erreurs d'annotation comme l'illustre la figure 3.5 dans laquelle l'auteur n'a pas été correctement identifié.

```
<bibl>
       <title>Clauzon G. (1982)</title> - <title>Le canyon messinien du Rhône:
           une preuve décisive du "Desiccated deep-basin model</title>" [<
           booktitle>Hsu</booktitle>, <pubPlace>Cita, Ryan</pubPlace>, <date>
           1973</date>]. <booktitle>Bulletin de la Société Géologique de France
           ,</booktitle> 24, 3, 597-610.
</bibl>
```

Listing 3.5 – Exemple d'une référence incorrectement annotée

Ces erreurs d'annotation engendrent des difficultés d'appariement. En effet dans ce cas précis, si les références allusives se rapportant à cette unité documentaire ne comportent pas d'informations aussi discriminantes que le titre de l'article, sans le nom de l'auteur il est difficile de procéder à une mise en correspondance. Concernant les variations de contenu observées entre un même champ bibliographique se rapportant à une même unité documentaire, nous avons pu noter des distinctions dans l'énonciation des éléments. Comme le démontre cet exemple nous pouvons observer que le nom de cet auteur est tronqué lors de son utilisation au sein des références allusives.

```
<!-- Référence présente dans la section bibliographique -->
<bibl>
       <author><surname>Choderlos de Laclos</surname></author>. <date>1951</
           date>. <title>{\OE}uvres complètes </title>(éd. <author><forename>
           Allem</forename></author>) (<pubPlace>Paris</pubPlace>: <publisher>
           Gallimard</publisher>)
```

```
</bibl>
<!-- Référence allusive extraite du corps du texte -->
<bibl>
       <author><surname>Laclos</surname></author> <date>1951</date>: <biblScope
           >19</biblScope>
</bibl>
```

Listing 3.6 – Références dont les champs ne correspondent pas dans leur intégralité

Ce type de variation engendre également des difficultés d'appariement. En effet, bien que nous ayons mis en place une fonction de correspondance basée sur une mesure de similarité il est difficile de réaliser, selon le degré de variation entre deux chaînes, un appariement. Dans ce cas précis, la fonction de correspondance basée sur une distance de Levenshtein établie à partir du nombre d'insertions, suppression ou substitutions de caractère permettant de transformer l'une des chaînes en l'autre, est en dessous du seuil que nous avons fixé afin de permettre la mise en correspondance des deux chaînes. Concernant l'occultation de certains champs bibliographiques, nous avons pu observer que certains champs présents dans la composition des références allusives ne le sont pas dans la composition des références présentes dans la zone bibliographique. Comme nous l'avons relevé au cours de la section 2.3.1.1, certains éléments bibliographiques susceptibles de composer le schéma d'un type de document peuvent être facultatifs ou obligatoires selon la convention stylistique utilisée. De ce fait, des éléments correspondant à une zone d'information tels que le nombre de volumes ou le nombre de pages sont parfois omis dans les références présentes dans la zone bibliographique car leur renseignement est facultatif (KYHENG 2003). A contrario, ces zones d'information peuvent se trouver renseignées au sein des références allusives bien qu'elles ne soient pas présentes dans la référence présente dans la zone bibliographique correspondante. L'exemple qui suit fournit une illustration de ce phénomène *via* l'occultation, au sein de la référence présente dans la zone bibliographique, des informations relatives aux numéros de pages identifées par le biais de la balise <biblScope>.

```
<!-- Référence présente dans la section bibliographique -->
<bibl>
   <author><surname>Rousseau</surname>, <forename>Jean-Jacques</forename></
       author>. <date>1959-1961</date>. <title>{\OE}uvres complètes 1-2</title>
       (éd. Gagnebin &amp;amp; <author><forename>Raynaud</forename></author>) (<
       pubPlace>Paris</pubPlace>: <publisher>Gallimard</publisher>)
</bibl>

<!-- Référence allusive extraite du corps du texte -->
<bibl>
       <author><surname>Rousseau</surname></author> <date>1960</date>: <
           biblScope>15</biblScope>
</bibl>
```

Listing 3.7 – Références dont des champs sont occultés

Ce type de configuration peut s'avérer problématique lors de la gestion des différents appariements. En effet dans ce cas précis, nous pouvons noter que la balise <biblScope> n'est pas présente lors de la description de l'unité documentaire au sein de la zone bibliographique. De plus, les balises <date> ne correspondent pas. De ce fait, nous ne possédons que le nom de l'auteur qui peut s'avérer insuffisant pour permettre un appariement sur l'unité documentaire correspondante dans le cas ou plusieurs travaux de ce même auteur sont cités. Dans l'exemple qui suit nous avons un autre cas de ce que nous pouvons apparenter à un problème lié à l'occultation de certains champs bibliographiques mais cette fois non pas engendrée par la composition du schéma imposée d'un type de document mais par un formalisme dépendant de la convention stylistique employée. En effet, dans ce cas de figure les travaux d'un même auteur présentés successivement sont représentés sous la forme de tiret qui permettent de ne pas répéter plusieurs fois le nom de ce dernier.

```
 <bibl>
       <author><surname>Geoffroy</surname> <forename>Éric</forename></author>,
           <date>2003</date>, <title>Initiation au soufisme</title>, <pubPlace>
           Paris</pubPlace>, <publisher>Fayard</publisher>.
</bibl>
<bibl>
       -, "<title>Le soufisme d'Occident dans le miroir du soufisme d'Orient</
           title>", <edition>in www.religioperennis.org (non daté</edition>).
</bibl>
```

Listing 3.8 – Exemple de références ne répétant pas le nom des collectifs d'auteur

Ce type de structure ne permet pas d'apparier correctement les références allusives qui correspondent à cette unité documentaire dans le cas ou ces dernières ne possèdent que le nom de l'auteur et une date par exemple. Concernant l'application des fonctions de correspondance entre les références allusives se rapportant à une même unité documentaire, la principale difficulté rencontrée réside dans l'identification de la source initiale des *ibidem*. En effet, les *ibidem* sont fréquemment utilisés pour éviter les répétitions afin de signifier que le mot ou le passage cité fait référence à la même source que la citation précédente. L'exemple présenté dans la figure 3.2 illustre l'utilisation des *ibidem* (abrégés sous la forme *ibid.*) au sein du discours.

L'utilisation des *ibidem* étant important au sein du corpus de test (1,5 *ibidem* en moyenne par document), nous avons dû pallier ce problème afin d'éviter une perte d'information trop importante. Le traitement de ces derniers a nécessité l'implémentation de traitements spécifiques permettant l'appariement des *ibidem* à leur source initiale. Chaque *ibidem* est remplacé au sein du dictionnaire des références allusives par le nom de la source initiale permettant ainsi l'application des fonctions de correspondances. L'algorithme 2 présente la chaîne de traitement proposée afin de traiter les *ibidem*.

```
<p>Tout jeune, il faisait déjà preuve d'une merveilleuse compréhension de la musique, en accumulait de
telles réserves dans son cerveau et possédait un don si exceptionnel de virtuose, qu'il me fallut
suivre le vieux conseil : enseigner un savant n'aboutit qu'à le déformer (<bibl><author>
<surname>Neuhaus</surname></author>, <date>1971</date> <biblScope>: 181</biblScope></bibl>).</p>
<p>Un même scepticisme s'empare de l'éminent professeur lorsqu'un autre élève d'exception,
Emil Guilels, joue la Rhapsodie espagnole de Liszt :</p>
<p>J'ai souvent pensé que, n'étant pas en mesure de jouer les octaves aussi vite, aussi fort et d'une
façon aussi brillante que mon élève, il serait plus raisonnable de lui trouver un
 professeur digne de lui, c'est-à-dire capable de résoudre encore plus brillamment ces difficultés
( <bibl>ibid.: <biblScope>183</biblScope></bibl> ).</p>
<p>Plus tard, après avoir beaucoup travaillé sur la musique et sur lui-même, Guilels me disait que le
plus beau concerto du monde était celui de Schumann ( <bibl>ibid.: <biblScope>217</biblScope></bibl> ).</p>
```

Figure 3.2. – Exemple de références sous la forme d'*ibidem*

Algorithme 2 Identification et remplacement des ibidem

Prérequis: Liste des références allusives $List_R$. Expression régulière correspondante aux abbréviations des ibidem rencontrées $ibid_{pattern}$

Garantie: Document XML dont les références bibliographiques sont annotées.

1: $indice \leftarrow 0$
2: **Tant que** $indice$à$(longueur$ de $List_R)$ **faire**
3: $\quad référence_{courante} = List_R[indice]$
4: $\quad$ **Si** $indice$à$(longueurde List_R - 1)$ **alors**
5: $\quad\quad référence_{suivante} = List_R[indice + 1]$
6: $\quad$ **Si** $référence_{suivante} == ibid_{pattern}$ **alors**
7: $\quad\quad référence_{suivante} = référence_{courante}$
8: $\quad indice \leftarrow indice + 1$

Comme nous venons de le soulever de nombreuses difficultés majoritairement induites par les variations, à la fois, structurelles et compositionnelles des références peuvent causer des pertes d'informations. Afin de pallier certaines de ces difficultés, plusieurs axes de travail sont envisagés notamment du côté des seuils de mise en correspondance. Une des pistes envisagées est l'attribution d'un poids plus important à certains champs comme le champ auteur afin de ne plus considérer l'ensemble des résultats retournés par la mesure de Levenshtein sur l'ensemble des champs bibliographiques. Certaines limites sont cependant difficiles à dépasser notamment en ce qui concerne l'occultation de certains champs. En effet, ceux-ci peuvent s'avérer discriminants et permettre d'établir des correspondances comme dans l'exemple de la figure 3.7.

Dans la section qui suit nous détaillons le fonctionnement des facteurs d'impact que nous avons établis.

3.3.3. Construction des facteurs d'impact

Comme défini dans la littérature, le rôle des références bibliographiques, dans le cadre de publications scientifiques, est de permettre de donner une valeur

rhétorique et argumentative à l'écrit scientifique au travers de leur utilisation [37]. Ces valeurs peuvent être véhiculées au sein du discours par le biais de positionnement des travaux de recherche d'un chercheur vis-à-vis de ses pairs (Mitra 1970) ou encore dans le but de convaincre son lectorat afin d'être accepté en tant que membre d'une communauté de discours (Swales 1990). De ce fait, les références bibliographiques font partie intégrante du discours et sont donc omniprésentes. Comme nous avons pu le soulever au cours du chapitre 2, les références bibliographiques peuvent se retrouver à différents niveaux du document et notamment sous la forme de références que nous avons qualifié d'allusives. À partir de l'analyse de ces références allusives, dont la caractéristique principale est d'être disséminées dans le corps du texte, nous avons pu constater que ces références sont utilisées afin de ponctuer l'argumentaire de l'auteur, elles font partie intégrante de la syntaxe du discours. De ce fait, nous avons émis l'hypothèse qu'extraire des informations à partir de ces références peut permettre d'estimer l'impact d'une référence sur l'argumentaire de l'auteur au sein d'un document. Afin d'établir nos facteurs d'impact, nous nous sommes donc basée sur des observations en rapport avec la construction de l'argumentaire de l'auteur. Nous avons établi, à partir de facteurs fondés sur des mesures quantitatives telles que la fréquence d'apparition et la granularité de distribution de chacune des références allusives, un indicateur d'impact propre à chaque unité documentaire.

Les sections suivantes définissent les différents facteurs que nous avons mis en place ainsi que leur fonctionnement.

3.3.3.1. Le facteur de fréquence d'apparition

Ce facteur se mesure à partir de l'estimation de la valeur relative au nombre de répétitions d'une référence se rapportant à une même unité documentaire au sein d'un même document. De ce fait, des références allusives présentant des structures différentes mais se rapportant à la même unité documentaire sont considérées comme les éléments d'un même ensemble. L'hypothèse relative à la construction de ce facteur est la suivante : plus une référence bibliographique est citée au sein du document plus son impact est fort dans l'argumentaire du document ciblé. Afin de mesurer ce facteur, nous nous sommes basée sur des comparaisons effectuées *via* la mise en correspondance des références allusives et des références présentes dans la zone bibliographique. En effet, de par leur formalisme plus conventionnel mais aussi de par l'exhaustivité de leurs zones d'informations relatives aux différents champs bibliographiques, les références présentes dans les zones bibliographiques sont un moyen de faire converger en

37. Au cours de cette thèse, nous reconnaissons l'existance de l'utilisation de références dites « superficielles » qui est une problématique importante pour l'analyse des citations (Zhao, Cappello et al. 2017). Cependant, nous n'effectuerons pour le moment aucune distinction sur les fonctions des citations.

un seul point toutes les références allusives relatives à une seule unité documentaire. La figure 3.3 propose un exemple d'une référence présente à la fois dans les paragraphes et les notes à laquelle s'applique le facteur de fréquence.

```
<p>L'entrave temporaire à anneaux articulés répondrait davantage à des besoins ponctuels. Le principe de
l'entrave articulée serait effectivement répandue en Celtique ( Thompson 1993, p. 145-149), en particulier
 le type de Chalon (Bourgogne)<note>De nombreuses entraves articulées sont mentionnées pour les régions du
 Doubs et de la Saône ( Audin &amp;amp; Armand-Calliat 1962 ). Elles sont présentes sous la forme de stock pour
 la pratique de l'asservissement, connue chez les Eduens ( Daubigney &amp;amp; Guillaumet 1985, 175 ). À Sanzeno
 même sont attestés d'autres types d'entraves articulées ( Thompson 1993, 73, 93 ), comme les entraves de cou
 et des menottes ( Nothdurfter 1979 ). Ainsi, la forme simple a dû coexister avec différents spécimens
articulés</note>.</p>
 <p>Une variante du type riveté</p>
 <p>Dans une tombe de Selca (ancien Pelion, en Thessalie), datée de la seconde moitié du IIIe s. av. J.-C.,
 le squelette portait une entrave fermée (conservée au National Historical Museum de Tirana, Albanie ;
Thompson 1993, p. 133 ). Il s'agit d'un type ancien d'entraves rivetées, jointes distinctement par deux ou plusieurs
 anneaux à un anneau central, comme une chaîne. Le principe de rivetage définitif est apparenté aux pièces du Vallon
 du Fou : le jonc est en deux parties, donc mobile pour l'installer sur la cheville, mais le deuxième rivet est mis en
 place par matage, l'entrave une fois posée sur l'individu. Ce type est archéologiquement attesté pour les esclaves des
 mines du Laurion de la péninsule grecque ( Thompson 1993, p. 131 ), mais il est également cité comme préférable pour
les travaux agricoles, limitant les mouvements du captif tout en lui laissant les mains libres.</p>
```

Figure 3.3. – Exemple d'une référence à laquelle s'applique le facteur de fréquence

L'algorithme 3 décrit la chaîne de traitement proposée afin d'incrémenter le facteur de fréquence d'apparition. Nous commençons par initialiser un tableau $valeurs$ à 0 afin de pouvoir stocker les résultats obtenus lors des tests réalisés sur les fonctions de correspondance. Chaque indice du tableau $valeurs$ correspond à un des index de positionnement des références présentes dans la zone bibliographique r'_j de l'ensemble R'. Ensuite pour chaque référence allusive de l'ensemble R nous effectuons des tests de correspondance avec les références de l'ensemble R' *via* les fonctions $correspondance_{stricte}(r_i, r'_j)$ et $correspondance_{similarité}(r_i, r'_j)$. Si les conditions sont remplies, $valeur[i]$ est incrémenté selon la valeur attribuée à α. α renvoie au paramètre relatif au poids attribué à ce facteur.

Algorithme 3 Calcul du facteur de fréquence d'apparition

Pour i de 0 à (longueur de $valeurs$) **faire**
    $valeurs[i] \leftarrow 0$
Pour i de 0 à (longueur de R') **faire**
    Pour j de 0 à (longueur de R) **faire**
        Si $correspondance_{stricte}(r_i, r'_j) = 1$
 ou $correspondance_{similarité}(r_i, r'_j) = 1$ **alors**
            $valeurs[i] \leftarrow valeurs[i] + \alpha$

La section suivante détaille les deux facteurs de granularité de distribution que nous avons mis en place afin d'établir le degré de distribution d'une référence au sein d'un document.

3.3.3.2. Les facteurs de granularité de distribution

Ces facteurs se mesurent en tenant compte de différents niveaux de granularité de distribution que nous avons mis en place, à savoir, une granularité fine et une granularité large. Le principe de ces facteurs est d'attribuer un poids supplémentaire à la référence présente dans la zone bibliographique dont les références allusives remplissent les conditions de distribution.

La **granularité fine** se mesure en tenant compte des références allusives renvoyant à une même unité documentaire et ce, au sein d'un même paragraphe. Afin d'établir si la granularité entre deux références allusives est fine, au préalable, nous comptons le nombre de mots qui sépare chaque référence se rapportant à un même document seulement dans le cas où elles sont utilisées au sein d'un même paragraphe. Ensuite, nous attribuons un poids uniquement si le nombre de mots entre deux références allusives se rapportant à la même unité documentaire est inférieur à la moyenne des écarts. Afin de mieux appréhender l'utilisation de ce facteur nous allons détailler ci-dessous son estimation.

Premièrement, nous définissons $countp()$ comme une fonction permettant de retourner le nombre de paragraphes.

Considérons Ref comme l'ensemble des références $\{r_0, .., r_k\}$ dans la zone bibliographique et P un paragraphe. Nous définissons $Allu_P$ comme un ensemble ordonné de références allusives extraites d'un même paragraphe, tel que :

$$Allu_P = \{i_1, i_2, ..., i_n\} \tag{3.6}$$

Tel que $i_n = (r'_n, a_n, b_n)$ avec $r'_n \in Ref$, $a_n \in [0..size(P)]$, $b_n \in [0..size(P)]$. a_n est la i_n position de départ dans le paragraphe P et b_n est sa position de fin dans le paragraphe P. Chaque i_n est ordonné selon sa position d'apparition dans P.

Nous appellons $Allu_{P,d} \subset Allu_P$, le sous-ensemble ordonné de références allusives présentes dans le même paragraphe et correspondantes à un document d. Nous signifions que :

$$\forall (i_u, i_v) \in Allu_{P,d} \times Allu_{P,d}, \\ correspondance\_stricte(r'_u, r'_v) = 1 \vee \\ correspondance\_similarité(r'_u, r'_v) = 1$$

Ensuite, nous définissons $Avg_{P,d}$ comme la moyenne des distances en mot

entre les références allusives d'un document d dans le paragraphe P, tel que :

$$Avg_{P,d} = \frac{\sum_{u=1}^{|Allu_{P,d}|} \sum_{v=u+1}^{|Allu_{P,d}|} |a_v - b_u|}{|Allu_{P,d}|} \tag{3.7}$$

Où u et v sont les indices d'apparition de deux éléments i_u et i_v de $Allu_{P,d}$.

Après, étant donné une fonction $count\_allu(P)$ donnant le nombre d'ensembles $Allu_{P,d}$ générés pour chaque paragraphe P (nous considérons tous les sous-ensembles de tous les documents référencés dans le paragraphe P), nous avons :

$$Avg_{Allu} = \frac{\sum Avg_{P,d}}{\sum count\_allu(P)} \ , P \in [1, countp()] \tag{3.8}$$

Avg_{Allu} est donc la moyenne de toutes les moyennes des distances calculées précédemment.

À partir de ces informations, étant donné $i_u \in Allu_{P,d}$ et $i_v \in Allu_{P,d}$ référant à un même document d dans le même paragraphe P, la fonction de granularité fine $granularité_{fine}(i_u, i_v)$ peut être calculée comme suit :

$$granularité_{fine}(i_u, i_v) = \begin{cases} 1 \text{ si } a_v - b_u < Avg_{Allu} \\ 0 \end{cases} \tag{3.9}$$

Concernant la mesure d'une **granularité large**, nous tenons compte des références allusives renvoyant à une même unité documentaire sur l'ensemble du document. Afin d'établir si la granularité entre deux références allusives est large, nous reprenons le même principe que celui appliqué pour la mesure d'une granularité fine seulement nous nous positionnons cette fois sur le nombre de paragraphes qui séparent deux références allusives se rapportant à la même unité documentaire. En d'autres termes, nous comptons le nombre de paragraphes qui séparent chaque référence renvoyant à une même unité documentaire tout au long du document. Ensuite, nous attribuons un poids uniquement si le nombre de paragraphes entre deux références allusives se rapportant au même document est inférieur à la moyenne des écarts. Afin de mieux appréhender l'utilisation de ce facteur nous allons détailler ci-dessous son estimation.

Considérons $Allu_{P,d}$, un ensemble de références allusives correspondant à un même document d provenant d'un paragraphe P, et $Allu_{Q,d}$ un ensemble de références allusives correspondantes à un même document d provenant d'un paragraphe Q (nous considérons que les fonctions $correspondance\_stricte$ et $correspondance\_similarité$ sont remplies), nous avons :

$$Avg_d = \frac{\sum_{P=1}^{countp()} \sum_{Q=P+1}^{countp()} (Q - P) \times |Allu_{P,d}| \times |Allu_{Q,d}|}{\sum_{R=1}^{countp()} |Allu_{R,d}|} \tag{3.10}$$

Avg_d est donc la moyenne des distances en paragraphe qui sépare deux références allusives correspondantes au même document d.

Ensuite, si n est le nombre de documents différents d, nous définissons :

$$Avg'_{Allu} = \frac{\sum Avg_d}{n}, d \in [1, n] \tag{3.11}$$

Avg'_{Allu} est donc la moyenne de toutes les moyennes des distances calculées précédemment.

À partir de ces informations, étant donné P et Q, deux paragraphes tel que $Q \geq P$, et $i_u \in Allu_{P,d}$ et $i_v \in Allu_{Q,d}$ référant au même document d, la fonction de granularité large $granularité_{large}(i_u, i_v)$ peut être calculée comme suit :

$$granularité_{large}(i_u, i_v) = \begin{cases} 1 \text{ si } Q - P < Avg'_{Allu} \\ 0 \end{cases} \tag{3.12}$$

L'algorithme 4 présente la chaîne de traitements proposée afin d'incrémenter les facteurs de granularité de distribution. Nous reproduisons sensiblement le même cheminement que lors de la présentation de l'algorithme 3 permettant l'incrémentation du facteur de fréquence d'apparition.

Algorithme 4 Calcul des facteurs de granularité de distribution

Pour u de 0 à (longueur de R) **faire**
  Pour v de 0 à (longueur de R) **faire**
    Si $u \neq v$ **alors**
      Si $correspondance_{stricte}(i_u.r', i_v.r') = 1$ **alors**
        Si $granularité_{fine}(i_u, i_v) = 1$
ou $granularité_{large}(i_u, i_v) = 1$ **alors**
          $values[i_u.r'] \leftarrow values[i_u.r'] + \beta$
      Sinon Si $correspondance_{similarité}(i_u.r', i_v.r') = 1$ **alors**
        Si $granularité_{fine}(i_u, i_v) = 1$
ou $granularité_{large}(i_u, i_v) = 1$ **alors**
          $values[i_u.r'] \leftarrow values[i_u.r'] + \gamma$

Pour chaque couple (i_u, i_v) de l'ensemble $Allu$ de références allusives, nous effectuons des tests de correspondance entre les deux références avec les fonctions $correspondance_{stricte}(i_u.r', i_v.r')$ et $correspondance_{similarité}(i_u.r', i_v.r')$. Si les conditions sont vérifiées, nous effectuons d'autres tests basés sur la mesure des facteurs de granularité de distribution. Si les conditions pour l'application des fonctions $granularité_{fine}(i_u, i_v)$ et $granularité_{large}(i_u, i_v)$ sont remplies, $values[i_u.r']$ est incrémenté selon la valeur assignée à β ou γ. β et γ correspondent aux poids

attribués à chacun de ces paramètres. Notons que la taille de $values$ est la même que celle de Ref.

La figure 3.4 présente un exemple d'une référence pour laquelle les deux facteurs de granularité ont été appliqués. Dans ce cas-ci, la moyenne des écarts établie pour ce document dans le cadre de la granularité fine est de 55 mots et de 7 pour la granularité large ce qui fait que l'indicateur d'impact de cette référence est influencé par les facteurs de granularité de distribution.

```
      <p>The Aven d'Orgnac karst, between the Ardèche and Cèze Gorges (fig. 2), provides a unique record
 of rising hydrological base level during the Pliocene. Two karst planation surfaces have been recognised
 (400 m and 260 m a.s.l.). The first formed before the Messinian salinity crisis (Clauzon, 1982 ); the
second is connected to the end of the Pliocene continental aggradation (Mocochain, 2007). These
sub-horizontal features, which indicate a low gravitational energy, are in sharp contrast with the deep
incision of Ardèche and Cèze Gorges. However, studies of the Orgnac underground system have shown that
both planation and gorge formation do not tell the entire story.</p>
      <p>Fig. 2 - Diagrammatic cross section showing the karst evolution of the Aven d'Orgnac (Ardèche):
subterranean karst records of Mio-Pliocene eustatic variations. Variation in the level of the Rhône
(Clauzon, 1982;  Mocochain, 2007). Inset map showing the Orgnac area.<lb/>Fig. 2 - Coupe synthétique de
l'évolution karstogénique de l'aven d'Orgnac (Ardèche) : les enregistrements endokarstiques des variations
eustatiques mio-pliocènes. Variation du niveau du Rhône d'après  Clauzon (1982) et Mocochain (2007). Encart
 montrant la localisation d'Orgnac.</p>
```

Figure 3.4. – Exemple d'une référence à laquelle s'applique les facteurs de granularité fine et large

Via la mesure de ces facteurs, il est possible de cerner plus précisément le degré d'impact d'une référence au sein du document. En effet, nous émettons l'hypothèse qu'une référence citée tout au long du document, donc possédant un facteur de granularité large plus important, a été utilisée afin de construire l'argumentaire de l'auteur dans sa globalité. *A contrario*, la présence plus concentrée de références allusives au sein de zones à la densité textuelle plus faible tend vers une construction non plus globale mais partielle de l'argumentaire de l'auteur.

3.3.3.3. Combinaison linéaire des facteurs d'impact

Afin de procéder au calcul de l'indicateur d'impact propre à chaque référence bibliographique, nous effectuons à une combinaison linéaire des facteurs d'impact présentés précédemment. Pour chaque référence r_i de l'ensemble R correspondant aux références présentes dans la zone bibliographique, nous procédons à la factorisation des différents facteurs d'impact, tel que :

$$indicateur(r_i) = \alpha freq_{fréquence} + \beta freq_{Granularité_{fine}} + \gamma freq_{Granularité_{large}} \quad (3.13)$$

Où $freq_{fréquence}$, $freq_{Granularité_{fine}}$, $freq_{Granularité_{large}}$ correspondent respectivement à la fréquence d'évaluations positives des conditions relatives à l'application des fonctions correspondant à la mesure des fréquences d'apparition et des granularités de distribution. Chacune des variables $freq$ est ici multipliée par le coefficient correspondant à la valeur attribuée aux paramètres de chaque fonction soit, α, β et γ. La somme de ces coefficients est égale à 1.

La section qui suit présente l'utilisation de l'indicateur d'impact que nous proposons au sein d'un système de recommandation basé sur des mesures bibliométriques.

3.4. Ordonnancement des indicateurs d'impact au sein d'un système de recommandation basé sur des critères bibliométriques

Comme nous avons pu le constater au cours de la section 3.2.3, les modèles de recommandation traditionnels sont basés sur l'exploitation de requêtes ou de profils utilisateurs. Cependant notre cadre applicatif nous permet d'envisager d'autres types de recommandation *via* l'exploitation d'informations issues de l'analyse des références bibliographiques allusives. Suite à la production de l'indicateur que nous avons présenté lors de la section 3.3 basée sur des mesures quantitatives, nous avons pu obtenir des informations sur l'impact d'une référence sur l'argumentaire de l'auteur au sein d'un document. À partir de ces informations, nous proposons d'utiliser cet indicateur d'impact afin de fournir une liste de recommandations ordonnée en fonction du degré d'impact de chaque référence du document ciblé.

Nous proposons pour un article donné, dont nous présentons un exemple dans la figure 3.5, de fournir à partir des références bibliographiques allusives présentes au sein de ce dernier une recommandation de lectures.

Nous nous sommes basée uniquement sur les indicateurs que nous avons pu extraire à partir de la mesure des facteurs. Nous obtenons *via* la mesure de ces facteurs des indicateurs propres à chaque unité documentaire présente dans le document. Suite aux calculs des indicateurs d'impact, nous obtenons sur les 21 références présentes dans la zone bibliographique de l'article *La défunte aux entraves* de S. Duval une liste de recommandations dont nous présentons un extrait dans la figure 3.6.

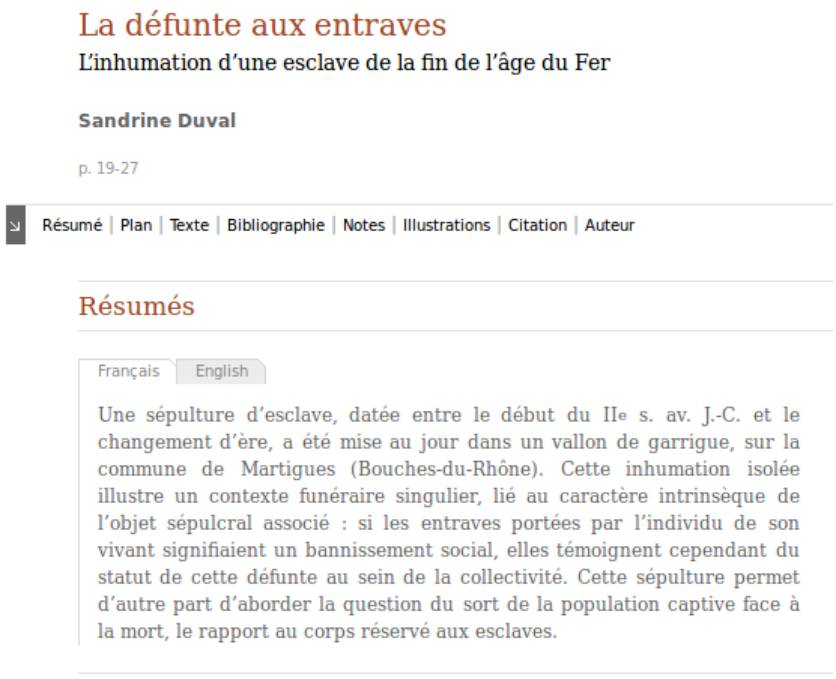


Figure 3.5. – Extrait de l'article *La défunte aux entraves* issu de la revue Préhistoires Méditerranéennes

```
1 Thompson 1993 Iron Age and Roman Slave-Shackles, The Archaeological Journal London
2 Dumont 1987 La mort de l'esclave La Mort les morts et l'au-delà dans le monde romain actes du Colloque de Caen 20-22 novembre
3 Chabot 2004 L'oppidum de la Cloche (Les Pennes-Mirabeau, Bouches-du-Rhône Protohistoire européenne
4 Audin Armand-Calliat 1962 Entraves en Bourgogne et dans le Lyonnais Revue archéologique de l'Est
5 Daubigney 1985 L'entrave de Glanon (Côte d'Or), les Eduens et l'esclavage Revue archéologique de l'Est et du Centre Est
```

Figure 3.6. – Extrait de la liste de recommandations fournie pour l'article *La défunte aux entraves*

3.4.1. Analyse et perspectives

Afin de réaliser la liste de recommandations proposée, nous procédons à un ordonnancement basé sur un tri des indicateurs par ordre décroissant afin de mettre en évidence les références les plus plébiscitées par l'auteur. *Via* l'utilisation de ce type d'ordonnancement nous avons pu constater certaines limites notamment en présence d'unités documentaires possédant des indicateurs aux degrés d'impact similaires, comme nous pouvons l'observer dans la figure 3.7 présentant un extrait de la liste de recommandations fournie pour l'article *Le karst : des archives paléogéographiques aux indicateurs de l'environnement*[38] (les indicateurs d'impact sont soulignés en rouge).

En effet, en présence de ce type de configuration, nous ne disposons pas de ressources supplémentaires permettant d'établir une distinction plus marquée entre des unités documentaires. Une des pistes envisageables est l'inclusion d'informa-

38. http://geomorphologie.revues.org/7520

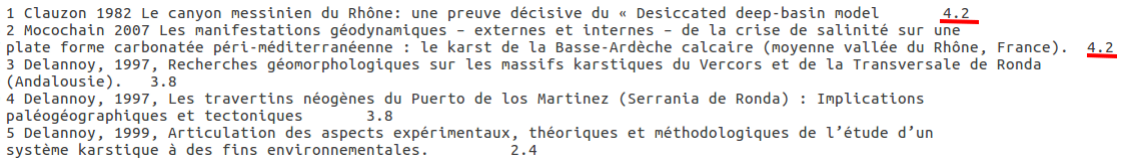


Figure 3.7. – Extrait de la liste de recommandations fournie pour l'article *Le karst : des archives paléogéographiques aux indicateurs de l'environnement*

tions sociales telles que le comportement des utilisateurs (navigation, clics, etc.) mais aussi des informations issues de comptes rendus de lecture. L'exploitation d'informations sociales peut nous permettre d'intégrer des informations quantitatives *via* le nombre de clics à partir duquel nous pouvons obtenir un facteur d'impact relatif à la popularité. L'exploitation des comptes rendus de lecture, quant à elle, peut nous permettre d'intégrer des informations reflétant les opinions relatives à l'article courant.

Une autre constatation, cette fois plus générale est que nous ne sommes pas en mesure de juger dans quel cadre est employée une référence bibliographique. En effet, nous nous sommes basée sur des mesures quantitatives. Or, il se peut que, bien qu'une référence ait un indicateur d'impact important, son utilisation au sein du discours ne soit pas faite de façon positive. *Via* cette constation, plusieurs pistes de recherche se dessinent à partir desquelles nous pouvons envisager d'établir une analyse des motivations de citation ou encore des fonctions de citation (analyse de sentiment).

Au cours de la section suivante, nous présentons la méthode d'évaluation que nous avons mise en place.

3.5. Méthode d'évaluation

Nous avons proposé à des utilisateurs d'évaluer les recommandations proposées par notre système [39] ainsi que les recommandations fournies lors de l'interrogation du moteur de recherche d'OpenEdition *Search OpenEdition* [40]. Suite à l'étude de la littérature, nous n'avons pu trouver de systèmes de recommandation ouverts basés sur des mesures bibliométriques avec lesquels nous aurions pu envisager d'effectuer une comparaison des performances.

Dans les sections qui suivent, nous présentons brièvement le processus d'interrogation de *Search OpenEdition* ainsi que la plateforme mise en place afin d'évaluer les recommandations fournies par les deux systèmes.

39. http://grapheval.openeditionlab.org/
40. https://search.openedition.org/

3.5.1. Processus d'interrogation de *Search OpenEdition*

Afin de pouvoir établir une comparaison des recommandations fournies par le système de recommandation basé sur des mesures bibliométriques que nous proposons, nous avons procédé à l'interrogation du moteur de recherche *Search OpenEdition*.

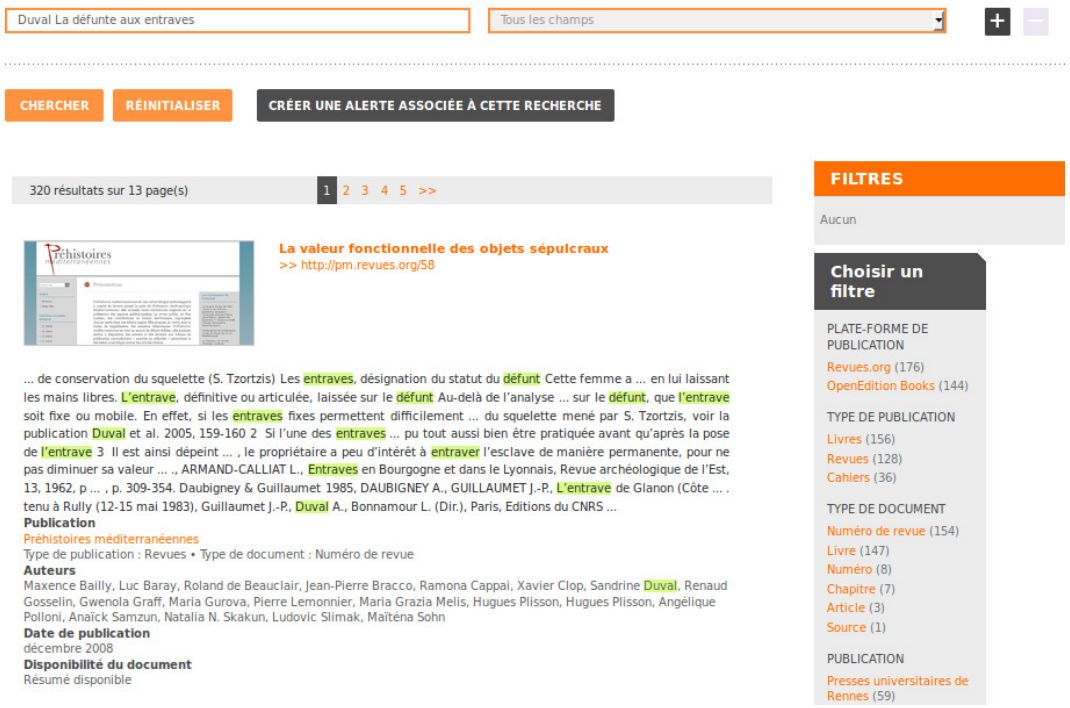


Figure 3.8. – Plateforme Search OpenEdition

Comme le présente la figure 3.8, l'interrogation de *Search OpenEdition* s'effectue sous la forme d'une requête à partir de laquelle nous obtenons une liste des documents les plus pertinents en fonction de la requête posée. De manière générale, les requêtes soumises sur le portail d'OpenEdition (composées en moyenne de 2,1 mots) sont essentiellement liées aux contenus des plates-formes, se référant à des ouvrages précis, des titres, des auteurs et plus généralement à des éléments relatifs aux contenus proposés (LEVA 2011). Une fois la requête formulée, le moteur de recherche interroge les contenus des quatre plateformes d'OpenEdition, et les données communes aux revues partagées avec les portails Cairn [41] et Persée [42]. *Search OpenEdition*, basé sur la plateforme logicielle de moteur de recherche SolR [43], permet différents modes d'interrogation :

41. https://www.cairn.info/
42. http://www.persee.fr/
43. http://lucene.apache.org/solr/

- le mode principal d'interrogation qui consiste à effectuer une recherche générale en soumettant directement une requête *via* l'utilisation de champs de recherche liés à des caractéristiques des documents (titre, auteur, bibliographie, notes, etc.) reliés par un opérateur booléens (et, ou, sauf)
- le mode de recherche avancée qui permet l'application de filtres ou l'interrogation par facettes. Ce mode permet de raffiner la sélection des documents *via* une sélection précise de la plateforme de publication, du type de publication, de la date, etc.

Concernant le processus d'indexation des données d'OpenEdition, chacun des documents a été représenté sous la forme d'un ensemble de champs auxquels sont associés des valeurs. Les champs utilisés par cet index peuvent aller des plus communs, tels que les champs titre et auteur, aux plus spécifiques, tels que des champs propres aux revues et aux livres comme des champs relatifs à l'identifiant DOI ou à la bibliographie. À partir de cet index et d'une requête formulée par un utilisateur, une liste des documents les plus pertinents est générée. Par le biais de l'analyseur *ExtendedDisMax Query Parsers*, chaque terme de la requête est comparé aux termes présents dans les différents champs. Cet analyseur a la particularité de traiter des phrases simples entrées par l'utilisateur *via* le moteur de recherche et de rechercher chacun des mots de la requête au sein de plusieurs champs. Pour chaque champ, des poids ont été préalablement attribués afin d'augmenter la pertinence des correspondances entre ces champs et les termes de la requête. Suite à l'étude des contextes de recherche effectuée sur les requêtes formulées sur *Search OpenEdition* (LEVA 2013), 22 champs différents ont été pourvus de poids dont les champs correspondant au titre (*naked_titre*) et aux noms d'auteur (*contributeur_auteur*). La liste de ces champs pouvus de poids est présentée dans la figure 3.9.

```
'naked_titre' => 8.0,
'parts_titre' => 8.0,
'contributeur_auteur' => 4.0,
'naked_soustitre' => 2.0,
'parts_soustitre' => 2.0,
'entree' => 0.6,
'naked_texte' => 0.5,
'parts_texte' => 0.5,
'resume_fr' => 0.5,
'resume_en' => 0.5,
'resume_es' => 0.5,
'resume_de' => 0.5,
'parts_resume' => 0.5,
'naked_introduction' => 0.5,
'entity_isbn' => 0.5,
'naked_notesbaspage' => 0.4,
'parts_notesbaspage' => 0.4,
'naked_bibliographie' => 0.4,
'parts_bibliographie' => 0.4,
'naked_annexes' => 0.4,
'naked_addenda' => 0.2,
'site' => 0.1
```

Figure 3.9. – Liste des poids de confiance attribués à chaque champ

Suite au cumul des scores produit pour chacun des termes présent dans la requête correspondant à des termes présents dans les différents champs des documents de l'index, *Search OpenEdition* retourne une liste des documents les plus pertinents.

Dans le cadre de nos travaux, nous avons interrogé *Search OpenEdition* sans spécifier de champs de recherche liés à des caractéristiques des documents (titre, auteur, bibliographie, notes, etc.) et ni d'opérateurs booléens. Cependant, nous avons eu recours au mode de recherche avancée *via* l'application d'un filtre permettant une interrogation uniquement de la plateforme Revues.org. Ainsi l'application de ce filtre, nous permet de concentrer les recherches sur les mêmes données que celles exploitées par notre système de recommandation également extraites de cette même plateforme. Afin de nous rapprocher au plus près des traitements effectués lors de la création des indicateurs, soit l'exploitation des références bibliographiques allusives, les requêtes proposées au moteur de recherche, lors de l'évaluation, ont été formulées à partir du nom du premier auteur suivi du titre de l'article (cf. : section 2.3.2.2.b).

Dans la section suivante, nous présentons la plateforme dédiée à l'évaluation des deux systèmes de recommandation que nous avons mise en place.

3.5.2. Plateforme d'évaluation des systèmes de recommandation

Afin de permettre à un panel d'utilisateurs provenant d'horizons disciplinaires divers d'évaluer la pertinence des retours de chaque système de recommandation proposé, une interface graphique a été réalisée.

Nous proposons tout d'abord aux utilisateurs de s'identifier par le biais d'un formulaire afin d'obtenir des informations telles que le nom, le prénom et le domaine d'activité. Suite à leur identification, les utilisateurs ont accès à la liste des articles à évaluer dont un exemple est présenté dans la figure 3.10.

Via cette liste, les utilisateurs ont la possibilité de sélectionner un article. Une fois l'article sélectionné deux listes de cinq articles sont proposées pour chaque système, comme l'illustre la figure 3.11. Seulement les cinq premiers articles ont été sélectionnés pour chaque liste proposée afin d'éviter de rendre cette évaluation trop fastidieuse. De plus, comme l'illustre cet exemple, pour certaines propositions, notamment celles de *Search OpenEdition*, les retours sont parfois trop faibles et ne permettent pas de proposer cinq articles. Nous supposons que ce phénomène est induit par la configuration par défaut de l'analyseur *Extended-DisMax Query Parsers* utilisé pour lequel la totalité des termes de la requête doit figurer dans les différents champs des documents indexés.

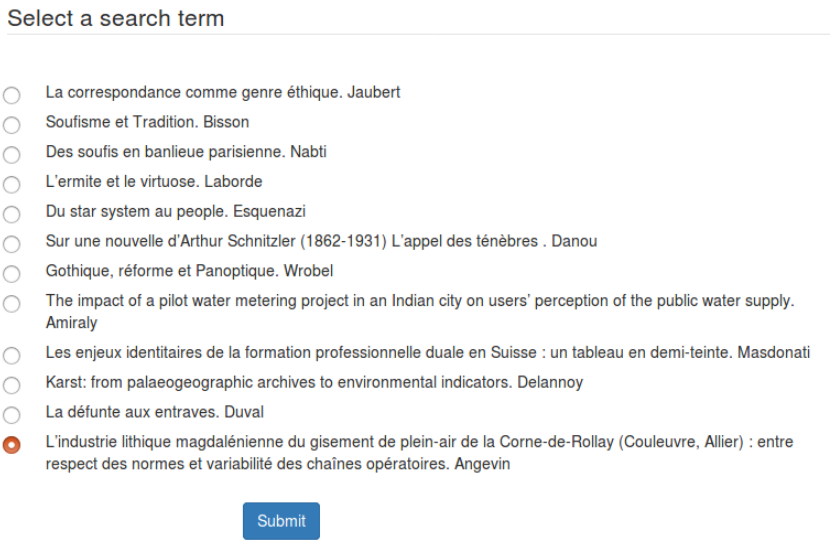


Figure 3.10. – Liste des thématiques de recherche

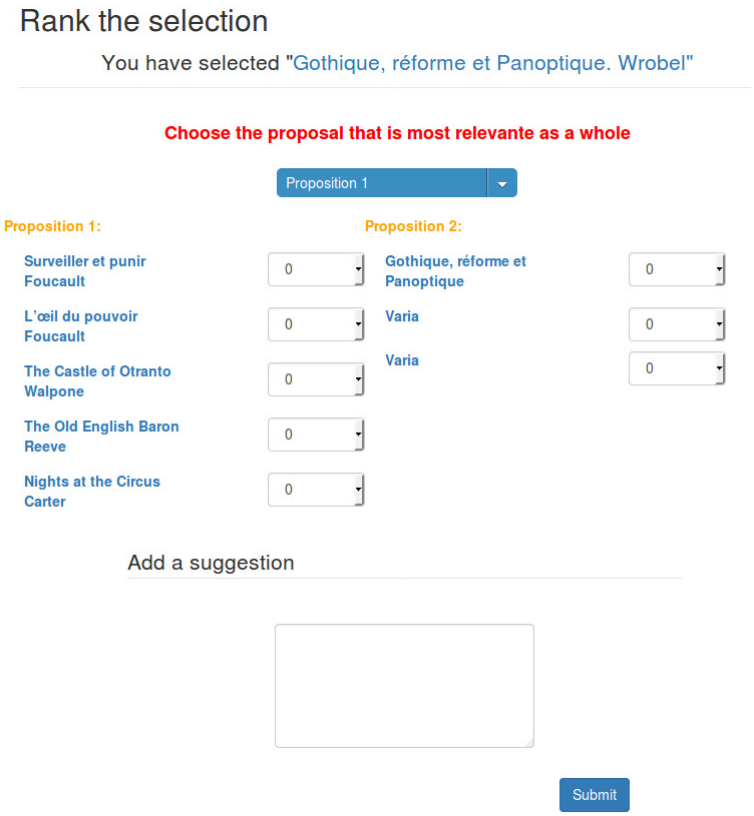


Figure 3.11. – Formulaire d'évaluation des articles

Par le biais de ce formulaire, il suffit aux utilisateurs de signifier leur choix

via la sélection d'une de ces deux listes. Lors de l'évaluation, la provenance des suggestions n'est pas signifiée afin de ne pas orienter le choix des utilisateurs (dans la figure 3.11, la proposition 1 correspond aux suggestions fournies par le système de recommandation que nous proposons et la proposition 2 à celles fournies par *Search OpenEdition*). Il est également possible d'affiner leur évaluation en émettant une appréciation allant de 0 à 5 (0 signifiant que l'article est hors sujet et 5 signifiant que l'article est totalement en accord avec la thématique abordée). La majorité des documents proposés dispose d'un lien cliquable qui permet d'obtenir la version originale du document ou une description. Ceci permet aux utilisateurs d'obtenir un aperçu du contenu de l'article courant mais aussi des articles à évaluer. Un champ « suggestion » est également disponible afin de permettre aux utilisateurs d'exprimer plus concrètement un avis ou des remarques sur les recommandations présentées.

Une fois l'évaluation d'un article terminée les utilisateurs soumettent leurs réponses qui sont stockées dans une base de données MySQL. Ils sont ensuite renvoyés à la page d'accueil permettant la sélection d'une nouvelle thématique de recherche afin de procéder à une nouvelle évaluation.

La section suivante présente les expérimentations que nous avons menées d'une part, sur la modulation du poids des facteurs d'impact et d'autre part, sur les retours des utilisateurs concernant la pertinence des recommandations fournies par les systèmes de recommandation présentés précédemment.

3.6. Expérimentations

Au cours de cette section, nous présentons dans un premier temps les données de tests. Dans un second temps, nous proposons d'analyser le poids accordé à chacun des facteurs d'impact afin d'estimer si l'ordonnancement des propositions de la liste de recommandations est influencé par cet aspect. Dans un troisième temps, nous présentons une étude des retours utilisateurs.

3.6.1. Données de tests

Comme nous avons pu le signifier au cours de la section 3.3, afin d'établir la mesure des facteurs d'impact nous exploitons des articles scientifiques pourvus d'une zone dédiée à la bibliographie. Or, nous avons pu constater au cours du chapitre 2 qu'il est d'usage en SHS d'utiliser les notes de bas de page afin d'identifier les références tout au long du document ce qui engendre des articles dépourvus de zones bibliographiques. De ce fait, nous avons dû, à partir du corpus présenté dans la section 2.5.1 dédiée à l'apprentissage des références allusives, faire une sélection des seuls articles possédant une zone bibliographique spé-

cifique. Nous avons choisi d'exploiter à nouveau les articles de ce corpus pour lesquels les références allusives sont déjà annotées afin d'estimer si les indicateurs d'impact mesurés à partir de références correctement annotées permettent de fournir des recommandations de lectures pertinentes. Sur les 42 articles de revues initialement annotés pour l'apprentissage des références allusives, nous n'avons pu en conserver que 12, les 30 articles restants n'étant pas pourvus de zones bibliographiques. Bien évidemment nous sommes conscients de la perte qu'occasionne l'utilisation uniquement des articles possédant une zone dédiée aux références bibliographiques de ce fait, nous proposerons au cours de nos perspectives d'étendre ces travaux afin de pouvoir générer des indicateurs à partir des références présentes dans les notes de bas de page.

Pour rappel, les articles composant ce corpus ont été sélectionnés aléatoirement indépendamment de l'affiliation disciplinaire des revues. La figure 3.12 présente un extrait d'un des articles conservés avec les références allusives en gras et leurs références bibliographiques associées.

Les médias, comme le détaillera **Edgar Morin dans *L'Esprit du temps* (1964)**, seront non seulement réceptifs mais plus encore les véritables propagandistes de l'individualité starisée et de ses formes de vie. Ce sont les textes médiatiques (**Dyer, 1979**) qui édifient l'image des stars comme ils construisent l'image d'une existence emplie de divertissements.
De ce point de vue, l'éclosion du star system et son épanouissement peuvent être mis en relation avec de grandes transformations sociales qu'ont mises en évidence par exemple **David Riesman (1964)** dans son ouvrage ***La foule solitaire*** ou **Antoine Prost (1999)** à travers sa contribution à la monumentale ***Histoire de la vie privée***.

MORIN, EDGAR (1984), *LES STARS*, PARIS, ÉDITIONS GALILÉE.

DYER, RICHARD (1979), *STARS*, LONDRES, BRITISH FILM INSTITUTE.

RIESMAN, DAVID (1964), *LA FOULE SOLITAIRE*, PARIS, ÉDITIONS B. ARTHAUD.

PROST, ANTOINE (1997), « FRONTIÈRES ET ESPACES DU PRIVÉ », DANS ANTOINE PROST ET G. VINCENT (DIR.), *HISTOIRE DE LA VIE PRIVÉE TOME V*, PARIS, ÉDITIONS DU SEUIL, P. 61-113.

Figure 3.12. – Extrait de l'article *Du star system au people : l'extension d'une logique économique*

La section suivante présente une analyse fondée sur la modulation du poids accordé à chacun des facteurs d'impact.

3.6.2. Expérimentations sur la modulation du poids des facteurs de granularité de distribution

Nous avons proposé de tenir compte de plusieurs facteurs d'impact dont les facteurs de granularité de distribution. Ces facteurs mesurés à partir de l'analyse du comportement des références allusives au sein d'un même document tiennent compte de deux types de granularité : une granularité fine et une granularité large (cf : section 3.3.3.2). Les hypothèses émises autour de la création de ces facteurs sont, pour la granularité large, qu'une référence citée tout au long du

document est utilisée afin de construire l'argumentaire de l'auteur dans sa globalité tandis que pour la granularité fine, l'utilisation d'une référence dans ce cas tend vers une construction non plus globale mais partielle de l'argumentaire de l'auteur. *Via* les expérimentations sur la modulation du poids attribué aux facteurs de granularité de distribution, nous avons voulu observer si l'ordonnancement des propositions de la liste de recommandations générée par le système de recommandation que nous proposons est influencé par cet aspect. En d'autres termes, au travers de ces expérimentations nous voulons observer si, *via* la mesure de ces facteurs et selon le poids accordé à chacun, nous pouvons identifier les références bibliographiques utilisées dans le but de construire l'argumentaire global ou partiel de l'auteur. Nous précisons qu'au cours de cette expérimentation nous n'avons pas établi une étude qualitative des articles recommandés mais uniquement une étude sur l'ordonnacement des références bibliographiques. Nous précisons également que seuls les extraits des recommandations proposées sont issus de la liste de recommandations générée pour l'article *Le karst : des archives paléogéographiques aux indicateurs de l'environnement* [44] mais que nous avons procédé à l'étude de l'ensemble des articles. Pour chacune des expérimentations effectuée, nous avons comparé les listes de recommandations obtenues, premièrement, *via* les modulations des poids effectuées sur les facteurs de granularité de distribution et secondement, *via* l'utilisation d'un poids unique pour chacun des facteurs mesuré.

Pour cette première expérimentation, nous avons choisi d'attribuer un poids plus important aux références possédant une granularité large tout en respectant un seuil de 1 concernant la somme des coefficients (cf : section 3.3.3.3). Pour rappel *via* cette mesure, nous attribuons un poids uniquement si le nombre de paragraphes entre deux références allusives se rapportant au même document est inférieur à la moyenne des écarts. Concernant la configuration de γ et β qui correspondent respectivement à la valeur attribuée aux paramètres de la granularité large et celle attribuée à la granularité fine, nous avons attribué un poids plus important pour la granularité large ($\gamma = 0{,}4$) et poids plus faible à la granularité plus fine ($\beta = 0{,}1$). Nous avons effectué une comparaison entre les cinq premières références bibliographiques proposées *via* cette modulation et les cinq premières références établies sur un paramétrage unique de 0,5 pour chaque facteur.

Les tableaux 3.2 et 3.3 présentent, respectivement, les recommandations proposées suite à l'augmentation du poids attribuée au facteur de granularité large et les recommandations proposées suite à l'attribution d'un poids unique pour chaque facteur. Comme nous pouvons l'observer sur ces extraits, très peu de

44. http://geomorphologie.revues.org/7520

Référence bibliographique recommandée	Indicateur d'impact
Clauzon, 1982, Le canyon messinien du Rhône : une preuve décisive du « Desiccated deep-basin model »	4,2
Mocochain, 2007, Les manifestations géodynamiques – externes et internes – de la crise de salinité sur une plate forme carbonatée péri-méditerranéenne : le karst de la Basse-Ardèche calcaire (moyenne vallée du Rhône, France).	4,2
Delannoy, 1997, Recherches géomorphologiques sur les massifs karstiques du Vercors et de la Transversale de Ronda (Andalousie).	3,8
Delannoy, 1997, Les travertins néogènes du Puerto de los Martinez (Serrania de Ronda) : Implications paléogéographiques et tectoniques	3,8
Delannoy, 1999, Articulation des aspects expérimentaux, théoriques et méthodologiques de l'étude d'un système karstique à des fins environnementales.	2,4

Tableau 3.2. – Résultats des recommandations obtenues suite à l'augmentation du poids de la granularité large

changements sont constatés suite à la modulation du poids attribué au facteur de granularité large. *Via* ces tableaux, nous observons que seul l'ordonnancement est quelque peu modifié avec l'inclusion d'une nouvelle référence. Suite à l'étude des recommandations proposées, nous avons pu notifier que l'inclusion de cette nouvelle référence bibliographique est due à son utilisation à trois reprises sur l'ensemble du document. *A contrario* la référence substituée dans le tableau 3.3 est utilisée aussi à plusieurs reprises mais sur un espace textuel réduit. Ces résultats nous permettent d'affirmer que la modulation du poids d'impact attribué au facteur de granularité large permet de faire ressortir les références utilisées afin de construire l'argumentaire global de l'auteur.

Suite à l'analyse de l'ensemble du corpus, les principaux phénomènes constatés sont des modifications de l'ordonnancement et pour seulement quatre articles l'inclusion d'une nouvelle référence bibliographique.

L'expérimentation qui suit, dont les résultats sont présentés dans le tableau 3.4, a été effectuée suite à l'attribution d'un poids plus important aux références possédant une granularité fine, pour rappel *via* cette mesure, nous attribuons un poids uniquement si le nombre de mots entre deux références allusives se rapportant à la même unité documentaire est inférieur à la moyenne des écarts.

Référence bibliographique recommandée	Indicateur d'impact
Clauzon, 1982, Le canyon messinien du Rhône : une preuve décisive du «Desiccated deep-basin model»	4,5
Mocochain, 2007, Les manifestations géodynamiques – externes et internes – de la crise de salinité sur une plate forme carbonatée péri-méditerranéenne : le karst de la Basse-Ardèche calcaire (moyenne vallée du Rhône, France).	4,5
Delannoy, 1997, Recherches géomorphologiques sur les massifs karstiques du Vercors et de la Transversale de Ronda (Andalousie).	4
Delannoy, 1997, Les travertins néogènes du Puerto de los Martinez (Serrania de Ronda) : Implications paléogéographiques et tectoniques	4
Mocochain, 2006, La grotte de Saint-Marcel (Ardèche) : un référentiel pour l'évolution des endokarsts méditerranéens depuis 6 Ma.	2

Tableau 3.3. – Résultats des recommandations obtenues suite à l'utilisation d'un poids unique pour chaque facteur d'impact

Concernant la configuration des paramètres γ et β, nous avons attribué un poids plus important pour la granularité fine (β = 0,4) et poids plus faible à la granularité plus large (γ = 0,1). Nous avons également effectué une comparaison entre les cinq premières références bibliographiques proposées *via* cette modulation et les cinq premières références établies sur un paramétrage unique de chaque facteur présenté dans le tableau 3.3 (somme des coefficients égale à 1).

Suite aux résultats obtenus *via* l'augmentation des poids attribuée à la granularité fine présentés dans le tableau 3.4, nous pouvons observer un ordonnancement similaire à celui présenté dans le tableau 3.3. Ce phénomène est en corrélation avec les observations que nous avons effectuées lors de la comparaison entre le tableau 3.2 et 3.3 suite au réordonnancement des références bibliographiques. En effet, la référence substituée lors de l'augmentation des poids attribuée à la granularité large se trouve, dans le cas inverse, à la même position. Cela est dû au fait que cette référence est utilisée également trois fois mais sur un espace textuel réduit. Le fait d'augmenter le poids attribué à la granularité fine permet donc de mettre en exergue des références utilisées afin de construire l'argumentaire partiel de l'auteur. Suite à l'analyse de l'ensemble du corpus, nous constatons les mêmes phénomènes que lors de l'augmentation des poids attribuée à la granu-

Référence bibliographique recommandée	**Indicateur d'impact**
Clauzon, 1982, Le canyon messinien du Rhône : une preuve décisive du «Desiccated deep-basin model»	4,2
Mocochain, 2007, Les manifestations géodynamiques – externes et internes – de la crise de salinité sur une plate forme carbonatée péri-méditerranéenne : le karst de la Basse-Ardèche calcaire (moyenne vallée du Rhône, France).	4,2
Delannoy, 1997, Recherches géomorphologiques sur les massifs karstiques du Vercors et de la Transversale de Ronda (Andalousie).	3,8
Delannoy, 1997, Les travertins néogènes du Puerto de los Martinez (Serrania de Ronda) : Implications paléo-géographiques et tectoniques	3,8
Mocochain, 2006, La grotte de Saint-Marcel (Ardèche) : un référentiel pour l'évolution des endokarsts méditerranéens depuis 6 Ma.	2,4

Tableau 3.4. – Résultats des recommandations obtenues suite à l'augmentation du poids de la granularité fine

larité large soit, quelques modifications de l'ordonnancement et pour seulement deux articles l'inclusion d'une nouvelle référence bibliographique.

Via ces extraits, nous avons pu observer, lors de modulation des poids des facteurs de granularité de distribution, des modifications d'ordonnancement uniquement au niveau de la dernière proposition. Nous avons pu noter que ces références bibliographiques obtiennent le même indicateur d'impact lors de l'application d'un poids unique. Or, comme nous l'avons souligné au cours de la section 3.4, nous ne disposons pas de ressources supplémentaires permettant d'établir une distinction plus marquée entre des unités documentaires possédant un indicateur d'impact similaire. Cependant, *via* la modulation des poids des facteurs de granularité nous pouvons observer une réelle démarcation de ces références.

Concernant la stabilité de l'ordonnancement et ce, malgré la modulation des poids des facteurs de la granularité de distribution, nous la supposons majoritairement induite par l'application du facteur de fréquence d'apparition. En effet, ce facteur consiste uniquement à dénombrer les occurences des références allusives se rapportant à une seule unité documentaire tandis que les facteurs de granularité de distribution sont tenus par des conditions d'application *via* le respect de la moyenne des écarts. De ce fait, le facteur de fréquence d'apparition est plus largement employé. Ces expérimentations nous ont permis d'établir que

certaines références se distinguent de par leur utilisation au sein du discours. *Via* les constatations que nous avons effectuées précédemment des perspectives de travail se dessinent à partir desquelles nous pouvons envisager d'établir une analyse des motivations de citation, des fonctions de citation ou encore de sentiment orientée sur l'analyse du contexte d'apparition des références au sein du discours et dont les facteurs de granularité pourraient permettre de dégager les motivations, fonctions ou sentiments autour d'une unité documentaire selon sa distribution.

Dans la section suivante, nous effectuons une étude sur les retours des utilisateurs concernant la pertinence du système de recommandation basé sur des mesures bibliométriques que nous proposons et celui basé sur les retours de *Search OpenEdition*.

3.6.3. Étude des retours utilisateurs sur la pertinence des systèmes de recommandation

Afin de procéder à l'évaluation de notre système de recommandation, nous avons proposé à des utilisateurs, *via* la plateforme présentée dans la section 3.5.2, d'évaluer les recommandations proposées par notre système (MB [45] recommandation) ainsi que les recommandations fournies lors de l'interrogation du moteur de recherche *Search OpenEdition* (SoE recommandation). Par le biais de cette interface graphique, nous proposons aux utilisateurs d'estimer quelle liste de recommandations est, selon l'article concerné, la plus pertinente. Au total, 12 articles, issus du corpus présenté dans la section 3.6.1, ont été soumis à cette évaluation. Cette évaluation s'est déroulée sur un période de 1 mois durant laquelle nous avons sollicité des membres de la communauté OpenEdition provenant d'horizons disciplinaires divers. Sur ce laps de temps, nous avons dénombré 31 participants avec un ratio de 2,3 articles évalués en moyenne par personne.

Concernant la configuration du système de recommandation que nous proposons, les indicateurs établis pour chacune des propositions ont été mesurés *via* l'intégration de tous les facteurs présentés dans la section 3.3.3, à savoir, le facteur de fréquence d'apparition et les facteurs de granularité de distribution. Nous avons paramétré le poids de chacun des facteurs d'impact de façon à obtenir la somme des coefficients égale à 1, notre but n'étant pas d'observer l'influence de certaines références bibliographiques sur l'argumentaire de l'auteur mais plutôt d'évaluer leur impact global en tenant compte de chacun des facteurs de façon égale.

Concernant le processus d'interrogation de la plateforme *Search OpenEdition*,

45. MB correspond à l'acronyme : Mesures Bibliographiques

nous avons utilisé la même configuration que celle présentée au cours de la section 3.5.1, à savoir, l'application d'un filtre permettant une interrogation uniquement de la plateforme Revues.org ainsi que des requêtes formulées à partir du nom du premier auteur suivi du titre de l'article. Le tableau 3.5 recense les résultats obtenus pour chacun des articles proposé à l'évaluation selon la liste de recommandations choisie par les utilisateurs. Les valeurs présentées pour chaque article correspondent au nombre d'utilisateurs ayant sélectionné les recommandations fournies soit par MB recommandation ou SoE recommandation.

Article	**MB recommandation**	**SoE recommandation**
Jaubert - La correspondance comme genre éthique	7	5
Bisson - Souffisme et Tradition	8	0
Nabti - Des soufis en banlieu parisienne	5	0
Laborde - L'ermite et le virtuose	6	0
Esquenazi - Du star system au people	4	4
Danou - Sur une nouvelle d'Arthur Schnitzler (1862-1931) L'appel des ténèbres	2	1
Wrobel - Gothique, réforme et Panoptique	3	0
Amiraly - The impact of a pilot water metering project in an Indian city on users perception of the public water supply	4	0
Masdonati - Les enjeux identitaires de la formation professionnelle duale en Suisse : un tableau en demi-teinte	1	5
Delannoy - Karst : from palaeogeographic archives to environmental indicators	1	0
Duval - La défunte aux entraves	2	2
Angevin - L'industrie lithique magdalénienne du gisement de plein-air de la Corne-de-Rollay (Couleuvre, Allier) : entre respect des normes et variabilité des chaînes opératoires	2	0

Tableau 3.5. – Résultats des recommandations obtenues par les systèmes MB recommandation et SoE recommandation

Comme nous pouvons l'observer *via* le tableau 3.5, les utilisateurs ont sélectionné à 45 reprises les propositions de recommandation issues du système de

recommandation que nous proposons comme étant les plus pertinentes tandis que les propositions issues de l'interrogation de *Search OpenEdition* ont été sélectionnées à 17 reprises. Cependant, si nous observons les performances en détail, nous pouvons noter la présence de performances très contrastées selon les articles. Au total, 8 articles se démarquent clairement par un clivage en faveur du système de recommandation que nous proposons tandis que le reste obtient des performances substanciellement semblables voir similaires. Grâce aux commentaires fournis par les utilisateurs, nous avons pu interpréter cela. Le ressenti général est que les deux propositions fournies offrent des lectures pertinentes mais, sur l'ensemble, celles proposées par notre système le sont plus. Les propositions fournies lors de l'interrogation de *Search OpenEdition* sont liées à la thématique de l'article ciblé mais sur des aspects scientifiques plus génériques.

Concernant les limitations du système proposé, les principaux retours dénotent des propositions fortement liées à l'article ciblé qui nécessitent selon les références une connaissance aiguisée du domaine. Or, les utilisateurs sollicités ne sont pas forcément spécialisés sur l'ensemble des domaines. De ce fait, ils ont dû effectuer des recherches afin de prendre connaissance du contexte de chaque discipline. En effet, selon les références il est parfois difficile pour les utilisateurs de parvenir à faire un lien avec leur appartenance à la thématique générale de l'article de par leur très forte spécificité. Certaines références sont, en effet, utilisées non pas dans le but de construire le cadre général de l'argumentaire mais comme moyen d'étayer certains aspects scientifiques pointus. Nous pouvons par ailleurs constater que certains articles ont été très peu évalués à cause de la complexité du sujet abordé tel que l'article *Karst : from palaeogeographic archives to environmental indicators*.

Concernant *SearchOpenedition*, les utilisateurs ont dénoté la présence d'articles, bien que liés à la thématique de l'article ciblé, référant à des aspects trop génériques. Cela est induit par le fait que les propositions sont souvent extraites de la même revue que celle à laquelle est affilié l'article ciblé. En effet, les revues sont souvent spécialisées dans un domaine précis et chaque parution se focalise sur une thématique spécifique. De ce fait, les propositions fournies par *SearchOpenedition* proviennent bien du même domaine mais sans pour autant s'intéresser aux mêmes aspects précis que l'article ciblé.

Bien évidemment le fait que chaque utilisateur (faute de temps) n'ait pas évalué l'ensemble des articles ne nous permet d'avoir un recul suffisant. De plus, certaines recommandations ont également été difficiles à évaluer car les utilisateurs ne bénéficiaient d'aucune information accessible concernant l'article ciblé (souvent des articles anciens n'ayant probablement pas été numérisés). Il est évident que l'évaluation de chacun de ces articles nécessiterait l'expertise de spécialistes du domaine. Or, nous ne disposions pas des ressources nécessaires afin

de mobiliser des comités scientifiques des revues. Cependant, cette expérimentation nous permet malgré tout, *via* le panel utilisateur mobilisé, d'estimer que sur l'ensemble des documents les utilisateurs ont tendance à trouver les recommandations fournies par le système que nous proposons plus pertinentes. Nous pouvons envisager de fusionner les propositions fournies par les deux systèmes afin de produire une liste de recommandations permettant de pallier les limitations que nous avons soulevées précédemment.

Dans la section suivante, nous proposons une analyse des évaluations fournies sur deux articles.

3.6.4. Analyse des documents suggérés par les systèmes de recommandation

Dans cette section, nous proposons une analyse des propositions suggérées par chacun des systèmes de recommandation présenté précédemment pour deux des articles évalués. Nous avons pris deux articles obtenant des résultats différents. En effet, nous avons choisi d'étudier les principaux cas de figure que nous avons relevés au cours de la section 3.6.3, à savoir, un article pour lequel un clivage des évaluations est constaté et un autre pour lequel les résultats relevés sont similaires. Les listes de recommandations analysées sont celles fournies pour l'article *Souffisme et Tradition* [46] et celles fournies pour l'article *Du star system au people* [47]. Nous souhaitons, dans un premier temps, établir une analyse qualitative au cours de laquelle nous allons évaluer si les propositions de chacun des systèmes correspondent à la thématique abordée par l'article ciblé. Dans un second temps, nous allons tenir compte des appréciations fournies par le panel d'utilisateurs afin d'établir s'il existe une corrélation entre les résultats obtenus par chacune des propositions et le choix des utilisateurs concernant la liste de recommandations la plus pertinente. Afin d'établir cette étude, nous nous sommes basée sur les appréciations attribuées par chacun des utilisateurs sur chaque proposition au cours de l'évaluation des systèmes de recommandation. Pour rappel, chaque utilisateur, *via* l'interface d'évaluation, a la possibilité d'affiner son évaluation en émettant une appréciation allant de 0 à 5 pour chaque article suggéré (0 signifiant que la proposition est hors-sujet et 5 signifiant que la proposition est totalement en accord avec la thématique abordée). À partir de ces informations, nous avons réalisé la moyenne des appréciations données que nous avons reportée au sein des tableaux 3.6 et 3.7 et ce, pour chaque article provenant de chaque proposition. Nous voulons seulement estimer si les propositions fournies par chacun des systèmes sont thématiquement liées à l'article ciblé.

46. http://assr.revues.org/11343
47. http://communication.revues.org/1247

Proposition	Article	Appréciation moyenne
MB recommandation	La fonction de René Guénon et le sort de l'Occident	1
MB recommandation	La lecture de Yahyâ Guénon à travers l'opérativité spirituelle d'une communauté d'intellectuels musulmans européens	2,3
MB recommandation	Initiation au soufisme	3
MB recommandation	Orient et Occident	2,8
MB recommandation	Le Symbolisme de la Croix	2,2
SoE recommandation	Varia	1,5
SoE recommandation	L'ésotérisme. Thèmes, motifs et acteurs d'une culture	0
SoE recommandation	Le religieux interrogé par les chercheurs	0
SoE recommandation	Musiques rituelles	5
SoE recommandation	21 | Gland – Hadjarien	3,2

Tableau 3.6. – Moyenne des appréciations données pour l'article *Souffisme et Tradition*

Le premier article pour lequel nous avons réalisé une analyse qualitative des suggestions proposées par chacun des systèmes est *Souffisme et Tradition*. Les appréciations établies par les utilisateurs pour chacun des articles suggérés sont présentées dans le tableau 3.6. Cet article rédigé par David Bisson est dédié à l'influence de l'intellectuel René Guénon sur l'islam soufi européen.

Les deux premiers articles relatent les pensées de René Guénon sur la question occidentale et sur l'étude d'une communauté d'intellectuels musulmans européens. La troisième proposition refère à un ouvrage dédié au souffisme. La quatrième et la dernière proprosition renvoient à des ouvrages relatifs au protagoniste de l'article ciblé, à savoir, René Guénon aussi connu sous le nom de Abd al-Wâhid Yahyâ. Cette brève analyse, nous permet d'ores et déjà de constater l'existence de liens thématiques entre chacune des propositions du système que nous proposons. En effet, nous pouvons observer que pour chacune des proposi-

tions au moins une thématique relative à l'article ciblé est présente, à savoir, le souffisme et/ou René Guénon.

Concernant les propositions fournies par *Search OpenEdition*, la première proposition correspond à la revue dédiée aux confréries soufies en métropoles de laquelle est extraite l'article ciblé. La seconde proposition renvoie à un article du même auteur que l'article ciblé mais cette fois relatif à l'ésotérisme. La troisième proposition se réfère à un article dans lequel les auteurs examinent comment la recherche, dans différentes disciplines, appréhende et construit l'objet religieux. La quatrième proposition correspond à une revue proposant divers articles dédiés à l'analyse des musiques utilisées au cours de rites. La dernière proposition, quant à elle, renvoie à une encyclopédie relative au monde berbère.

Suite à cette analyse, nous pouvons constater que les propositions fournies par notre système sont plus étroitement liées à la thématique de l'article ciblé que celles de *Search OpenEdition* qui sont plus généralistes. *Search OpenEdition* propose des articles orientés sur différents aspects de la religion. Ces constatations nous permettent d'établir une liaison avec les constations émisent par les utilisateurs que nous avons relayées au cours la section 3.6.3, à savoir, une tendance pour le système que nous proposons à fournir des propositions étroitement liées à la thématique de l'article ciblé tandis que, les propositions suggérées par *Search OpenEdition* sont orientées sur des aspects de la thématique plus généralistes.

Via ce même tableau, nous pouvons observer qu'en moyenne les appréciations fournies par les utilisateurs sont de 2,3 pour les articles proposés par notre système de recommandation et de 1,9 pour les articles suggérés par *Search OpenEdition*. Cette moyenne corrèle avec le fait que les 8 utilisateurs ont choisi la liste de recommandations proposée par notre système de recommandation. Cependant, des articles issus de la liste de recommandations fournie par *Search OpenEdition* obtiennent des appréciations bien supérieures à celles du système que nous proposons comme l'article *Musiques rituelles* [48]. A contrario, bien que des articles proposés par *Search OpenEdition* aient reçu de bonnes appréciations, deux d'entre eux obtiennent une appréciation de 0 (hors-sujet). Nous supposons que cela a incité les utilisateurs à choisir le système que nous proposons.

Le second article pour lequel nous avons réalisé une analyse qualitative est *Du star system au people*. Les appréciations établies par les utilisateurs pour chacun des articles suggérés sont présentées dans le tableau 3.7. Cet article rédigé par Jean-Pierre Esquenazi est dédié à l'étude des stratégies marketing mises en place *via* le « star system ».

48. http://ethnomusicologie.revues.org/2421

Proposition	**Article**	**Appréciation moyenne**
MB recommandation	L'économie du star system	4
MB recommandation	La distinction	2,9
MB recommandation	L'esprit du temps	2
MB recommandation	Frontières et espaces du privé	2,5
MB recommandation	La foule solitaire	2
SoE recommandation	L'information-people	2
SoE recommandation	Le cas de l'information-people en Suisse romande	3,5
SoE recommandation	Vol. 27/1	1
SoE recommandation	Du populaire au populisme ?	1
SoE recommandation	Recherches au féminin en Sciences de l'Information	4

Tableau 3.7. – Moyenne des appréciations données pour l'article *Du star system au people*

Concernant les propositions fournies par le système de recommandation que nous proposons, la première proposition fait référence à un article orienté sur l'exploitation économique de la notoriété. La seconde proposition renvoie à un livre dans lequel il est question de la consommation culturelle et les styles de vie. La troisième référence correspond à une étude réalisée sur la culture de masse. La quatrième proposition réfère à un ouvrage dans lequel il est question de l'évolution de la vie privée de la première guerre à nos jours. La dernière proposition, quant à elle, renvoie à un livre portant sur l'ère de la consommation de masse.

Concernant les propositions fournies par *Search OpenEdition*, la première proposition renvoie à un article orienté sur le phénomène de peopolisation. La seconde proposition renvoie à un article relatif à la construction de l'objet people et de ses spécificités en Romandie. La troisième proposition correspond à la revue dédiée à l'information people de laquelle est extrait l'article ciblé. La quatrième proposition correspond à un article référant à l'idéologie et la négociation des

valeurs dans la presse people française. La dernière proposition, quant à elle, renvoie à une revue proposant divers articles dédiés aux chercheures et les recherches au féminin en Sciences de l'Information et de la Communication.

Comme nous avons pu également le relever suite à l'étude du précédent article, chacune des propositions fournie par notre système sur cet article possède un lien thématique avec l'article ciblé. Chacune des propositions réfère à au moins une des thématiques de l'article ciblé, à savoir, l'industrie culturelle et/ou le vedettariat. Nous pouvons également constater qu'à la différence des articles proposés pour l'article *Souffisme et Tradition*, les propositions de *Search OpenEdition* sont cette fois plus étroitement liées à la thématique générale de l'article ciblé. En effet, nous pouvons observer que la plupart des articles proposés sont orientés sur l'objet people qui est l'une des thématiques principales.

Via ce même tableau, nous pouvons observer qu'en moyenne les appréciations fournies par les utilisateurs pour cet article sont de 2,7 pour les articles proposés par notre système de recommandation et de 2,3 pour les articles suggérés par *Search OpenEdition*. Ces appréciations sont moins contrastées que celles obtenues pour l'article précédent ce qui corrèle avec le fait que les performances obtenues par chacune des listes de recommandations soient similaires. En effet, nous avons pu observer suite à l'analyse qualitative des propositions fournies par chacun des systèmes que les articles suggérés sont de, part et d'autre, étroitement liés à la thématique de l'article ciblé. De ce fait, chacune des propositions fournies suggère une recommandation pertinente.

3.7. Conclusion

Comme nous avons pu le constater au cours de ce chapitre les références bibliographiques servent plusieurs desseins. Leur exploitation suscite un vif intérêt autant auprès des communautés scientifiques qu'industrielles. Dans le cadre de ce chapitre, nous avons voulu effectuer un parallèle avec les travaux menés en bibliométrie *via* l'exploitation des références bibliographiques présentes au sein du discours scientifique que nous avons qualifiées de références bibliographiques allusives. À partir de ces références, nous avons proposé de construire un indicateur d'impact dont l'originalité est de se baser sur l'analyse du contenu des documents et non plus sur l'exploitation des traditionnelles bases de données. Nous avons choisi d'extraire des informations à partir de ces références dans le but d'estimer l'impact d'une référence sur l'argumentaire de l'auteur au sein d'un document. Afin d'établir nos facteurs d'impact, nous nous sommes fondée sur des mesures quantitatives telles que la fréquence d'apparition et la granularité de distribution de chacune des références allusives afin de construire un indica-

teur d'impact propre à chaque unité documentaire. *Via* ces facteurs nous avons proposé, au cours de la section 3.3.1, une méthode permettant l'implémentation de ces facteurs. Suite à la construction de cet indicateur, nous avons réalisé un système de recommandation de lectures basé sur des mesures bibliométriques. Ce système de recommandation dont nous avons présenté l'implémentation au cours de la section 3.4, nous permet d'obtenir une liste de recommandations ordonnée en fonction du degré d'impact mesuré de chaque référence du document ciblé. Les expérimentations menées ont permis d'obtenir des résultats plus que satisfaisants avec 72,8% des utilisateurs ayant opté pour les listes de recommandations proposées par notre système comme étant plus pertinentes que celles de *Search OpenEdition*. Les retours plus détaillés émis par les utilisateurs ont également permis de mettre en avant une réelle corrélation entre les thématiques de l'article ciblé et celles des articles recommandés.

Les recherches effectuées au cours de ce chapitre ouvrent la voie à d'autres projets de recherche. Du côté des indicateurs d'impact, nous envisageons d'étendre la couverture de l'indicateur aux références bilbiographiques présentes dans les notes de bas de page afin de pallier la perte d'information qu'induit leur non traitement. Sur l'aspect RI, nos perspectives s'orientent vers l'intégration de travaux issus de l'analyse de citations (DING, G. ZHANG et al. 2014 ; C. LU, DING et al. 2017) mais aussi de sentiments (T. WILSON, WIEBE et al. 2005 ; HTAIT, FOURNIER et al. 2017). L'idée est d'établir dans quels contextes sont employées les références allusives afin de pouvoir caractériser plus finement l'impact (motivation, fonction, polarité) d'une référence dans le cadre d'un article donné.

Du côté du système de recommandation basé sur des mesures bibliométriques que nous proposons, et plus particulièrement dans le cadre de son application aux données d'OpenEdition, nous nous orientons vers un enrichissement des retours de *Search OpenEdition via* l'inclusion de recommandations proposées par notre système. L'intégration d'informations sociales (clics, etc.) provenant des logs d'OpenEdition mais aussi d'informations issues des comptes rendus de lecture disponibles sur la plateforme *Review of Books*[49] est également une piste envisagée. En effet, l'intégration de ces informations a notamment pour objectif de nous permettre d'établir une distinction plus marquée entre des unités documentaires possédant un indicateur d'impact similaire.

Sur des aspects plus génériques liés à la recommandation, nous avons relevé le fait que l'approche proposée ne fournit, pour le moment, que des recommandations de lectures issues du document ciblé. De ce fait, nos recommandations sont figées car elles ne sollicitent aucune ressource externe mais elles sont également restreintes de par leur affiliation à une fenêtre temporelle (un article

49. http://reviewofbooks.openeditionlab.org/

publié en 2013 cite nécessairement des articles antérieurs ou provenant de cette même année). Afin de pallier les limitations de notre système, nous proposons de nous orienter vers une modélisation de nos données sous la forme d'un graphe afin d'exploiter des algorithmes dédiés au parcours de ce dernier (BENKOUSSAS 2016). L'application de ces algorithmes, dans le cadre de la réalisation d'un graphe de recommandation, nous permettra d'envisager d'établir des parcours permettant d'intégrer des ressources externes au document ciblé lors de la recommandation mais aussi de ternir compte d'aspect tel que la nouveauté. À l'image des travaux présentés dans l'article H. D. TRAN, CABANAC et al. 2017, nous pouvons envisager au travers d'un graphe de qualifier les relations entre les références et ainsi évaluer leur proximité. Par exemple, nous pourrions privilégier, pour un document donné, les références « citant-cité », leur affiliation à un domaine scientifique similaire (l'appartenance à une même revue), des collectifs d'auteur similaires, etc. Nous envisageons également d'utiliser les profils utilisateurs afin d'obtenir une recommandation personnalisée *via* l'analyse des historiques de navigation (LIU, DOLAN et al. 2010 ; JANNACH et LUDEWIG 2017).

Conclusion générale

Nous avons présenté dans ce manuscrit les travaux que nous avons menés au cours de ces trois années de thèse. Nos contributions se situent dans le cadre de la réalisation d'un système de recommandation de lectures. Dans ce dernier chapitre, nous commençons par dresser une synthèse de nos contributions, incluant un rappel des méthodes proposées et des résultats obtenus, avant de discuter de nos perspectives de recherche.

Synthèse des contributions

Cette thèse réalisée dans le cadre de l'équipement d'excellence DILOH part du constat que les usages des technologies numériques au sein des SHS sont inégaux. Des études ont montré que ce domaine pâtit notamment de restrictions technologiques d'accès aux contenus (DACOS et MOUNIER 2014). Dans ce contexte, cette thèse s'est alors orientée sur la création d'outils exploitant les possibilités du numérique en matière d'accès à l'information scientifique.

Nous avons pu relever que le *continuum* itératif de la production scientifique engendre un besoin de ventilation et de croisement spécifique des éléments trouvés. Afin de répondre à cette problématique, nous nous sommes penchée sur la réalisation d'un système de recommandation de lectures à partir duquel il est possible d'établir des suggestions permettant de guider les utilisateurs dans ce flux d'information mais également d'engendrer un croisement des ressources *via* l'exploitation d'un très riche corpus en SHS associant contenus et métadonnées normalisées.

Au cours du chapitre 1 et 2, nous nous sommes tout d'abord intéressée à deux aspects inhérents aux systèmes de recommandation : l'enrichissement de la représentation des requêtes et la représentation des documents.

Concernant l'enrichissement de la représentation des requêtes, et plus particulièrement des requêtes de recherche de livres, nous avons pu notifier la présence de verbosité au sein de ce type de requêtes. Au travers de cette verbosité, les utilisateurs expriment des informations plus complexes sur leurs besoins. Notre objectif était donc de parvenir à capturer les informations pertinentes au sein de ce type de requêtes afin de satisfaire au mieux le besoin d'information des utilisateurs. Nous nous sommes alors penchée sur la qualification de ces besoins *via* la réalisation d'une taxonomie propre au traitement des requêtes de recherche de livres. Suite à ces travaux, nous avons dénombré 5 taxons permettant d'identifier les besoins informationnels des utilisateurs. Nous nous sommes donc ensuite

intéressée à des solutions permettant le traitement de ces requêtes selon la qualification d'un besoin particulier. À des fins expérimentales, nous nous sommes concentrée sur un type de requêtes bien particulier à savoir les requêtes analogues. Nous avons présenté une approche de recommandation fondée sur l'analyse de ces requêtes basée sur des procédés issus de la classification supervisée et du TAL. Nous avons pu observer que la méthode de classification des requêtes selon la taxonomie que nous avons établie offre des résultats plus que satisfaisants avec une précision moyenne relevée pour nos deux algorithmes à plus de 90% sur les classes : analogue et non analogue. Concernant l'utilisation d'un analyseur en dépendance, nous avons pu constater que son utilisation dans l'analyse des requêtes analogues nous permet de faire un pas vers une meilleure interprétation des besoins exprimés par les utilisateurs. En effet, l'analyse plus fine des livres retournés pour une requête nous a permis de constater que l'apport d'informations supplémentaires par expansion permet d'améliorer la pertinence des résultats.

Concernant la représentation des documents par le biais des références bibliographiques, nous rappelons que notre but n'était pas d'améliorer la compréhension des documents mais bien d'exploiter les références bibliographiques présentes afin de mettre en perspective de nouveaux liens thématiquement liés au document courant. Comme nous avons pu le souligner, des études portées sur l'écrit scientifique ont permis d'identifier plusieurs types de références. Parmi elles, certaines sont des références explicites à l'image des références que nous pouvons trouver à la fin des articles ou des livres, tandis que d'autres références, que nous avons qualifié d'allusives, sont disséminées dans le corps du texte. Au cours de nos recherches, nous nous sommes particulièrement intéressée à l'identification automatique de ce type de références. Pour ce faire, nous avons proposé une méthode qui consiste d'une part, à identifier les paragraphes qui contiennent des références via un processus de classification supervisée et d'autre part, dans l'application de CCRFs afin de détecter plus précisément les zones bibliographiques et d'annoter leurs contenus. Nous avons pu constater que l'identification des références bibliographiques, et plus particulièrement des références bibliographiques allusives, n'est pas une tâche complètement résolue. En effet, bien que les références bibliographiques puissent apparaître comme des éléments structurés de nombreux facteurs influencent sur leur structuration ce qui rend la généralisation des traitements plus complexe. Les expérimentations que nous avons réalisées sur les données d'OpenEdition ont permis de rendre compte de résultats satisfaisants concernant l'exécution individuelle de chacun des procédés et ainsi permettre d'attester de sa robustesse. L'éxécution consécutive de tous les procédés a révélé, quant à elle, des résultats plus mitigés induits par des erreurs générées au cours de chaque processus. Parallèlement, l'application de la méthode que nous avons proposée sur les données issues de la campagne d'évaluation CLEF SBS 2016, a permis de nous conforter dans sa ca-

pacité d'adaptation *via* l'obtention de la meilleure précision. Les caractéristiques choisies ont démontré, aux cours de ces expérimentations, leur robustesse ainsi qu'une capacité d'adaptation intéressante au vu des données hétérognènes traitées.

La suite de nos travaux s'est ensuite orientée sur une exploitation plus poussée des références bibliographiques *via* leur utilisation en tant qu'outil de liaison entre contenus. L'étude des travaux dédiés à l'analyse des écrits scientifiques nous a permis d'établir que l'utilisation de références bibliographiques est un élément crucial dans la création et la diffusion de l'information. Dans le cadre d'une bibliothèque d'articles scientifiques, les références bibliographiques peuvent s'avérer être une source de liens majeure. En effet, à partir des références bibliographiques, il est possible d'obtenir des informations permettant l'identification d'un document en tant qu'unité documentaire. De ce fait, nous avons proposé d'exploiter les références bibliographiques afin d'établir des liens entre contenus. La modélisation sous la forme d'un graphe orienté, nous a permis d'attester de l'interconnectivité qui réside entre une référence bibliographique et différents articles provenant d'une même revue. Nous avons pu notamment observer des regroupements entre contenus thématiquement liés.

Au cours du chapitre 3, nous avons souhaité poursuivre ces travaux en proposant un système de recommandation de lectures basé sur l'exploitation des références bibliographiques. En effet, nous avons pu constater que les références bibliographiques avaient la capacité de faire émerger des liens permettant d'établir des liens entre contenus. Nous avons choisi d'étendre ces travaux afin de construire un indicateur d'impact, à partir des références bibliographiques allusives, dérivé des critères d'évaluation traditionnels basés sur des mesures quantitatives « objectives ». À partir d'observations menées sur l'utilisation des références bibliographiques allusives au sein d'articles scientifiques, nous avons établi des facteurs d'impact basés sur la fréquence d'apparition ainsi que la granularité de distribution entre chacune des apparitions d'une référence bibliographique. Par le biais cet indicateur, nous avons proposé d'estimer le degré d'influence propre à chaque référence bibliographique au sein d'un document. À travers ces facteurs nous avons proposé une méthode permettant leur implémentation. Suite à la construction de cet indicateur, nous avons voulu mettre en avant les interactions déjà relevées au sein de la littérature entre la RI et la bibliometrie *via* la réalisation d'un système de recommandation de lectures basé sur des mesures bibliométriques. Ce système de recommandation nous a permis d'obtenir des listes de recommandations ordonnées en fonction du degré d'impact mesuré de chaque référence du document ciblé. Les expérimentations menées, suite à la réalisation de ce système, ont permis d'obtenir des résultats plus que satisfaisants avec 72,8% des utilisateurs ayant opté pour les listes de recommandations proposées par notre système comme étant les plus pertinentes.

Les retours plus détaillés émis par les utilisateurs sur les propositions fournies par le système que nous avons proposé ont également permis de mettre en avant une réelle corrélation entre les thématiques de l'article ciblés et celles des articles recommandés.

Perspectives

Les travaux réalisés au cours de cette thèse ont ouvert la voie à d'autres projets de recherche envisageables à court, moyen et long terme.

Concernant les travaux effectués autour l'enrichissement de la représentation des requêtes, les tests de significativité que nous avons menés, nous ont permis d'observer des degrés de significativité différents sur certaines requêtes. Ce phénomème nous a amené à penser que des sous-classes ou des facettes sont envisageables selon les caractéristiques de certaines requêtes analogues afin d'employer le modèle le plus adapté. De plus, l'analyse détaillée des résultats obtenus pour une requête analogue sur chacun de nos modèles nous a permis de corroborer cette hypothèse. À court terme, nous tenterons d'établir les caractéristiques des requêtes analogues en fonction des meilleures performances obtenues selon les modèles employés. À plus long terme, nous envisagerons de mettre en place une stratégie de représentation propre au type de requêtes en fonction de la classe prédite lors de la classification automatique. Par exemple, nous pensons extraire les termes qui viennent particulariser la requête pour les requêtes orientées.

Concernant la représentation des documents par le biais des références bibliographiques, nous avons proposé des solutions d'amélioration afin d'éviter la propagation des erreurs et ainsi augmenter la pertinence des résultats. Du côté de la classification supervisée, à court terme, nous nous pencherons sur une analyse approfondie des attributs sélectionnés afin d'identifier à la fois leur pertinence et leur portabilité. À moyen terme, la mise en place d'approches basées sur l'extraction de séquences ainsi que sur l'élaboration de motifs syntaxiques utilisés dans le traitement des requêtes verbeuses (ETTALEB, LATIRI et al. 2016) représentent des pistes envisagées au vu de la densité textuelle de nos données. Concernant les modèles utilisés par nos CRFs, à court terme, nous proposons d'étudier l'influence de chacune des caractéristiques lors de l'apposition d'une étiquette afin de détecter quelles sont les caractéristiques les plus discriminantes. Nous supposons que l'intégration des améliorations induites par ces investigations permettront d'améliorer chacun des procédés de l'approche proposée et ainsi les performances globales.

Du côté des indicateurs d'impact, à moyen terme, nous envisageons d'étendre

la couverture de l'indicateur proposé aux références bilbiographiques présentes dans les notes de bas de page afin de pallier la perte d'information qu'induit leur non traitement. Sur l'aspect RI, à long terme, nos perspectives s'orientent vers l'intégration de travaux issus de l'analyse de citations (DING, G. ZHANG et al. 2014 ; C. LU, DING et al. 2017) mais aussi de sentiments (T. WILSON, WIEBE et al. 2005 ; HTAIT, FOURNIER et al. 2017). L'idée est d'établir dans quels contextes sont employées les références allusives afin de pouvoir caractériser plus finement l'impact (motivation, fonction, polarité) d'une référence dans le cadre d'un article donné. Concernant les aspects bibliométriques, l'indicateur proposé, au-delà de son inclusion dans le cadre d'un système de recommandation, peut permettre d'initier de nouvelles perspectives dans l'évaluation des travaux de recherche. En effet, cet indicateur s'oriente sur une analyse des procédés de citation propre au document et non globale comme nous pouvons actuellement le constater en bibliométrie (BELTER 2015). De ce fait, son utilisation peut permettre de mettre en évidence, pour un document donné, l'influence de chacune des références. Cependant, cela n'empêche pas l'intégration des informations issues de cet indicateur en externe.

Du côté du système de recommandation que nous avons proposé, et plus particulièrement dans le cadre de son application aux données d'OpenEdition, à moyen terme, nous nous orientons vers un enrichissement des retours de *Search OpenEdition via* l'inclusion de recommandations proposées par notre système. L'intégration d'informations sociales (clics, etc.) provenant des logs d'OpenEdition mais aussi d'informations issues des comptes rendus de lectures disponibles sur la plateforme *Review of Books* [50] est également une piste envisagée à moyen terme. En effet, l'intégration de ces informations a notamment pour objectif de nous permettre d'établir une distinction plus marquée entre des unités documentaires possédant un indicateur d'impact similaire.

Comme nous avons pu le relever, le système de recommandation basé sur des mesures bibliométriques que nous proposons fournit des recommandations figées et restreintes à une fenêtre temporelle. Afin de pallier les limitations de notre système, nous proposons, à court terme, de nous orienter vers une modélisation de nos données sous la forme d'un graphe afin d'exploiter des algorithmes dédiés au parcours de ce dernier (BENKOUSSAS 2016). L'application de ces algorithmes, dans le cadre de la réalisation d'un graphe de recommandation, nous permettra d'envisager d'établir des parcours permettant d'intégrer des ressources externes au document ciblé lors de la recommandation mais aussi de ternir compte d'aspects tels que la nouveauté. Nous envisageons également, à long terme, d'utiliser les profils utilisateurs afin d'obtenir une recommandation personnalisée *via* l'analyse des historiques de navigation (LIU, DOLAN et al.

50. http://reviewofbooks.openeditionlab.org/

2010 ; JANNACH et LUDEWIG 2017).

Au-delà de ces perspectives qui concernent un cadre purement applicatif de nombreux questionnements inhérents aux systèmes de recommandation restent encore à investiguer. En effet, les systèmes que nous pouvons trouver à l'heure actuelle sont souvent centrés sur la recherche d'algorithmes de recommandation plus précis (BOBADILLA, ORTEGA et al. 2013). Cependant de nombreux aspects sont importants mais difficilement intégrables lors de la recommandation, tels que :

- diversité : les utilisateurs ont tendance à être plus satisfaits des recommandations quand il y a une plus grande diversité intra-liste, comme par exemple des articles de différents artistes (CASTELLS, VARGAS et al. 2011).
- recommandation persistante : comment arriver à valoriser les nouveaux contenus ?
- intimité : un certain nombre d'informations sont disponibles sur les utilisateurs mais la privatisation des données rend leur utilisation compliquée (SCHEIN, POPESCUL et al. 2002).
- robustesse : peut on être certain de la véracité des informations données ? (HERLOCKER, KONSTAN et al. 2004)
- surprendre : comment parvenir à surprendre l'utilisateur en restant borné sur les éléments fournis ? (L. ZHANG 2013)
- confiance : comment faire confiance à un système de recommandation ? (MONTANER, LÓPEZ et al. 2002)

Bibliographie

[1] Gediminas ADOMAVICIUS et Alexander TUZHILIN. « Toward the next generation of recommender systems : A survey of the state-of-the-art and possible extensions ». In : *IEEE Transactions on Knowledge and Data Engineering*. T. 17. 6. IEEE, 2005, p. 734–749 (cf. p. 153).

[2] Sam ANZAROOT et Andrew MCCALLUM. « A new dataset for fine-grained citation field extraction ». In : *ICML Workshop*. 2013 (cf. p. 32, 81, 86, 107).

[3] Denis APOTHÉLOZ. *Rôle et fonctionnement de l'anaphore dans la dynamique textuelle*. Librairie Droz, 1995 (cf. p. 83).

[4] Jordi ARDANUY. « Sixty years of citation analysis studies in the humanities (1951–2010) ». In : *Journal of the American Society for Information Science and Technology* 64.8 (2013), p. 1751–1755 (cf. p. 148).

[5] Jon Scott ARMSTRONG. *Principles of forecasting : a handbook for researchers and practitioners*. T. 30. Springer Science & Business Media, 2001 (cf. p. 28).

[6] Azin ASHKAN, Charles CLARKE, Eugene AGICHTEIN et Qi GUO. « Classifying and characterizing query intent ». In : *European Conference on Information Retrieval*. Springer. 2009, p. 578–586 (cf. p. 48).

[7] Laura AURIA et Rouslan MORO. « Support vector machines (SVM) as a technique for solvency analysis ». In : *DIW Berlin Discussion Paper*. German Institute for Economic Research. 2008 (cf. p. 73).

[8] Ricardo BAEZA-YATES, Berthier RIBEIRO-NETO et al. *Modern information retrieval*. T. 463. ACM press, 1999 (cf. p. 27).

[9] Peter BAILEY, Ryen WHITE, Han LIU et Giridhar KUMARAN. « Mining historic query trails to label long and rare search engine queries ». In : *ACM Transactions on the Web (TWEB)* 4.4 (2010), p. 15 (cf. p. 42).

[10] Niranjan BALASUBRAMANIAN, Giridhar KUMARAN et Vitor CARVALHO. « Exploring reductions for long web queries ». In : *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*. ACM. 2010, p. 571–578 (cf. p. 41).

[11] Johan BALTIÉ. « DataMining : ID3 et C4. 5 ». In : Epita SCIA, 2002 (cf. p. 68).

[12] Judit BAR-ILAN, Marcus JOHN, Rob KOOPMAN, Shenghui WANG, Philipp MAYR, Andrea SCHARNHORST et Dietmar WOLFRAM. « Bibliometrics and information retrieval : Creating knowledge through research synergies ». In : *Proceedings of the Association for Information Science and Technology* 53.1 (2016), p. 1–4 (cf. p. 144).

[13] Judit BAR-ILAN et Mark LEVENE. « The hw-rank : An h-index variant for ranking web pages ». In : *Scientometrics* 102.3 (2015), p. 2247–2253 (cf. p. 144).

[14] Chumki BASU, Haym HIRSH, William COHEN et al. « Recommendation as classification : Using social and content-based information in recommendation ». In : *Proceedings of Knowledge-Based Electronic Markets, Papers from the AAAI Workshop*. 1998 (cf. p. 154).

[15] Joeran BEEL, Stefan LANGER, Marcel GENZMEHR, Bela GIPP, Corinna BREITINGER et Andreas NÜRNBERGER. « Research paper recommender system evaluation : a quantitative literature survey ». In : *Proceedings of the International Workshop on Reproducibility and Replication in Recommender Systems Evaluation*. ACM. 2016, p. 15–22 (cf. p. 28).

[16] Christopher BELTER. « Bibliometric indicators : opportunities and limits ». In : *Journal of the Medical Library Association : JMLA* 103.4 (2015), p. 219 (cf. p. 82, 141, 197).

[17] Michael BENDERSKY et Bruce CROFT. « Analysis of long queries in a large scale search log ». In : *Proceedings of the 2009 workshop on Web Search Click Data*. ACM. 2009, p. 8–14 (cf. p. 40).

[18] Michael BENDERSKY et Bruce CROFT. « Modeling higher-order term dependencies in information retrieval using query hypergraphs ». In : *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*. ACM. 2012, p. 941–950 (cf. p. 42).

[19] Michael BENDERSKY, Bruce CROFT et David SMITH. « Joint annotation of search queries ». In : *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies-Volume 1*. Association for Computational Linguistics. 2011, p. 102–111 (cf. p. 43).

[20] Michael BENDERSKY, Donald METZLER et Bruce CROFT. « Parameterized concept weighting in verbose queries ». In : *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*. ACM. 2011, p. 605–614 (cf. p. 42).

[21] Chahinez BENKOUSSAS. « Approches non supervisées pour la recommandation de lectures et la mise en relation automatique de contenus au sein d'une bibliothèque numérique. » Thèse de doct. Université Aix-Marseille, 2016 (cf. p. 192, 197).

[22] Chahinez BENKOUSSAS et Patrice BELLOT. « Book recommendation based on social information ». In : *CLEF 2013 Evaluation Labs and Workshop Online Working Notes*. 2013 (cf. p. 45).

[23] Chahinez BENKOUSSAS, Hussam HAMDAN, Shereen ALBITAR, Anaïs OLLAGNIER et Patrice BELLOT. « Collaborative Filtering for Book Recommendation ». In : *Working Notes for CLEF 2014 Conference*. 2014 (cf. p. 62).

[24] Chahinez BENKOUSSAS, Anaïs OLLAGNIER et Patrice BELLOT. « Book Recommendation Using Information Retrieval Methods and Graph Analysis ». In : *Working Notes for CLEF 2015 Conference*. CLEF. 2015 (cf. p. 46).

[25] James BENNETT, Stan LANNING et al. « The netflix prize ». In : *Proceedings of KDD cup and workshop*. 2007, p. 37 (cf. p. 27).

[26] Stephen BENSMAN. « Garfield and the impact factor ». In : *Annual Review of Information Science and Technology* 41.1 (2007), p. 93–155 (cf. p. 146).

[27] Stephen BENSMAN. « Eugene Garfield, Francis Narin, and PageRank : The Theoretical Bases of the Google Search Engine ». In : *arXiv preprint arXiv :1312.3872* (2013) (cf. p. 144).

[28] Adam BERGER et John LAFFERTY. « Information retrieval as statistical translation ». In : *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*. ACM. 1999, p. 222–229 (cf. p. 44).

[29] Shane BERGSMA et Qin Iris WANG. « Learning Noun Phrase Query Segmentation. » In : *EMNLP-CoNLL*. T. 7. 2007, p. 819–826 (cf. p. 43).

[30] Marc BERTIN et Iana ATANASSOVA. « A study of lexical distribution in citation contexts through the IMRaD standard ». In : *PloS Negl. Trop. Dis* 1.200,920 (2014), p. 83–402 (cf. p. 102).

[31] Sumit BHATIA, Cliff BRUNK et Prasenjit MITRA. « Analysis and automatic classification of web search queries for diversification requirements ». In : *Proceedings of the American Society for Information Science and Technology* 49.1 (2012), p. 1–10 (cf. p. 48).

[32] Ismail BISKRI et Louis ROMPRE. « Using Associated Rules for Query Reformulation ». In : *Next Generation Search Engine : Advanced Models for Information Retrieval*. IGI-Global, 2012, p. 291–303 (cf. p. 38, 49).

[33] Vincent BLONDEL, Jean-Loup GUILLAUME, Renaud LAMBIOTTE et Etienne LEFEBVRE. « Fast unfolding of communities in large networks ». In : *Journal of statistical mechanics : theory and experiment* 2008.10 (2008), P10008 (cf. p. 135).

[34] Jesús BOBADILLA, Fernando ORTEGA, Antonio HERNANDO et Abraham GUTIÉRREZ. « Recommender systems survey ». In : *Knowledge-based systems* 46 (2013), p. 109–132 (cf. p. 198).

[35] Toine BOGERS, Iris HENDRICKX, Marijn KOOLEN et Suzan VERBERNE. « Overview of the SBS 2016 Mining Track ». In : *Ceur Workshop Proceedings*. 2016 (cf. p. 128).

[36] Johan BOLLEN, Herbert VAN DE SOMPEL, Aric HAGBERG, Luis BETTENCOURT, Ryan CHUTE, Marko RODRIGUEZ et Lyudmila BALAKIREVA. « Clickstream data yields high-resolution maps of science ». In : *PLoS One* 4.3 (2009), e4803 (cf. p. 144).

[37] Ludovic BONNEFOY, Romain DEVEAUD et Patrice BELLOT. « Do Social Information Help Book Search ? » In : *Workshop INEX*. 2012, p. 109 (cf. p. 45).

[38] Christine BORGMAN et Jonathan FURNER. « Scholarly communication and bibliometrics ». In : *Annual Review of Information Science and Technology* 36.1 (2002), p. 3–72 (cf. p. 145, 146).

[39] Katy BÖRNER. *Atlas of science : Visualizing What We Know*. MIT Press, 2010 (cf. p. 144).

[40] Lutz BORNMANN et Werner MARX. « How to evaluate individual researchers working in the natural and life sciences meaningfully ? A proposal of methods based on percentiles of citations ». In : *Scientometrics* 98.1 (2014), p. 487–509 (cf. p. 145).

[41] Lutz BORNMANN, Andreas THOR, Werner MARX et Hermann SCHIER. « The application of bibliometrics to research evaluation in the humanities and social sciences : An exploratory study using normalized Google Scholar data for the publications of a research institute ». In : *Journal of the Association for Information Science and Technology* 67.11 (2016), p. 2778–2789 (cf. p. 148, 149).

[42] Thorsten BRANTS. « Natural Language Processing in Information Retrieval ». In : *CLIN*. 2003 (cf. p. 27).

[43] Andrei BRODER. « A taxonomy of web search ». In : *ACM Sigir forum*. T. 36. 2. ACM. 2002, p. 3–10 (cf. p. 31, 47, 49, 54).

[44] Christopher BURGES. « A tutorial on support vector machines for pattern recognition ». In : *Data mining and knowledge discovery* 2.2 (1998), p. 121–167 (cf. p. 88).

[45] Robin BURKE. « Hybrid recommender systems : Survey and experiments ». In : t. 12. 4. Springer, 2002, p. 331–370 (cf. p. 153).

[46] Guillaume CABANAC. « Accuracy of inter-researcher similarity measures based on topical and social clues ». In : *Scientometrics* 87.3 (2011), p. 597–620 (cf. p. 150).

[47] Guillaume CABANAC. « Extracting and quantifying eponyms in full-text articles ». In : *Scientometrics* 98.3 (2014), p. 1631–1645 (cf. p. 86).

[48] Guillaume CABANAC, Muthu Kumar CHANDRASEKARAN, Ingo FROMMHOLZ, Kokil JAIDKA, Min-Yen KAN, Philipp MAYR et Dietmar WOLFRAM. « Joint workshop on bibliometric-enhanced information retrieval and natural language processing for digital libraries (BIRNDL 2016) ». In : *Digital Libraries (JCDL), 2016 IEEE/ACM Joint Conference on*. IEEE. 2016, p. 299–300 (cf. p. 144).

[49] Liliana CALDERON-BENAVIDES, Cristina GONZALES-CARO et Ricardo BAEZA-YATES. « Towards a deeper understanding of the user's query intent ». In : *Proceedings of ACM SIGIR Workshop on Query Representation and Understanding*. 2010, p. 21–24 (cf. p. 48).

[50] James CALLAN et Bruce CROFT. « An evaluation of query processing strategies using the TIPSTER collection ». In : *Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM. 1993, p. 347–355 (cf. p. 41).

[51] Huanhuan CAO, Derek HAO HU, Dou SHEN, Daxin JIANG, Jian-Tao SUN, Enhong CHEN et Qiang YANG. « Context-aware query classification ». In : *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*. ACM. 2009, p. 3–10 (cf. p. 48).

[52] Pablo CASTELLS, Miriam FERNANDEZ et David VALLET. « An adaptation of the vector-space model for ontology-based information retrieval ». In : *IEEE transactions on knowledge and data engineering* 19.2 (2007) (cf. p. 44).

[53] Pablo CASTELLS, Saúl VARGAS et Jun WANG. « Novelty and diversity metrics for recommender systems : choice, discovery and relevance ». In : (2011) (cf. p. 198).

[54] Messaoud CHAA, Patrice BELLOT et Omar NOUALI. « New Technique to Deal With Verbose Queries in Social Book Search ». In : *2017 IEEE/-WIC/ACM International Conference on Web Intelligence (WI)*. Leipzig, Germany, août 2017, Long paper (cf. p. 37).

[55] Yu-Wei CHANG. « A comparison of citation contexts between natural sciences and social sciences and humanities ». In : *Scientometrics* 96.2 (2013), p. 535–553 (cf. p. 149).

[56] Patrick CHARAUDEAU, Dominique MAINGUENEAU et D DICTIONNAIRE. *Analyse du discours*. T. 98. Éditions du Seuil, 1985 (cf. p. 84).

[57] Chien-Chih CHEN, Kai-Hsiang YANG, Hung-Yu KAO et Jan-Ming HO. « BibPro : A Citation Parser Based on Sequence Alignment Techniques ». In : *Proceedings of the 22nd International Conference on Advanced Information Networking and Applications*. 2008, p. 1175–1180 (cf. p. 91).

[58] Hung-Hsuan CHEN, Liang GOU, Xiaolong ZHANG et Clyde Lee GILES. « Collabseer : a search engine for collaboration discovery ». In : *Proceedings of the 11th annual international ACM/IEEE joint conference on Digital libraries*. ACM. 2011, p. 231–240 (cf. p. 150).

[59] Hung-Hsuan CHEN, Alexander ORORBIA II et Lee GILES. « ExpertSeer : a Keyphrase Based Expert Recommender for Digital Libraries ». In : *arXiv preprint arXiv :1511.02058* (2015) (cf. p. 150).

[60] Long-Sheng CHEN, Fei-Hao HSU, Mu-Chen CHEN et Yuan-Chia HSU. « Developing recommender systems with the consideration of product profitability for sellers ». In : *Information Sciences* 178.4 (2008), p. 1032–1048 (cf. p. 151).

[61] Mark CLAYPOOL, Anuja GOKHALE, Tim MIRANDA, Pavel MURNIKOV, Dmitry NETES et Matthew SARTIN. « Combining content-based and collaborative filters in an online newspaper ». In : *Proceedings of ACM SIGIR workshop on recommender systems*. T. 60. Citeseer. 1999 (cf. p. 153).

[62] Rodrigo COSTAS, Zohreh ZAHEDI et Paul WOUTERS. « Do altmetrics correlate with citations ? Extensive comparison of altmetric indicators with citations from a multidisciplinary perspective ». In : *CoRR* abs/1401.4321 (2014). URL : http://arxiv.org/abs/1401.4321 (cf. p. 147).

[63] Isaac COUNCILL, Lee GILES et Min-yen KAN. « ParsCit : An open-source CRF reference string parsing package ». In : *LREC*. European Language Resources Association, 2008 (cf. p. 86, 91, 112).

[64] Blaise CRONIN. *The Citation Process : The Role and Significance of Citations in Scientific Communication*. London : Taylor Graham, 1984 (cf. p. 81, 82).

[65] Marin DACOS et Pierre MOUNIER. *Humanités numériques – État des lieux et positionnement de la recherche française dans le contexte international*. Institut français, 2014 (cf. p. 193).

[66] Van DANG et Bruce CROFT. « Query reformulation using anchor text ». In : *Proceedings of the third ACM international conference on Web search and data mining*. ACM. 2010, p. 41–50 (cf. p. 42).

[67] Nicola DE BELLIS. *Bibliometrics and citation analysis : from the science citation index to cybermetrics*. Scarecrow Press, 2009 (cf. p. 85).

[68] Marie-Catherine DE MARNEFFE, Bill MACCARTNEY, Christopher MANNING et al. « Generating typed dependency parses from phrase structure parses ». In : *Proceeding of the 5th edition of the International Conference on Language Ressources and Evaluation*. LREC. 2006, p. 98–109 (cf. p. 62).

[69] Dario De Nart et Carlo Tasso. « A personalized concept-driven recommender system for scientific libraries ». In : *Procedia Computer Science* 38 (2014), p. 84–91 (cf. p. 150).

[70] Stéphane Deconinck. « Artificial intelligence a modern approach ». In : (2010) (cf. p. 32, 86).

[71] Adèle Désoyer, Frédéric Landragin, Isabelle Tellier, Anaïs Lefeuvre, Jean-Yves Antoine et Marco Dinarelli. « Coreference Resolution for French Oral Data : Machine Learning Experiments with ANCOR ». In : *17th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing'2016)*. 2016 (cf. p. 84).

[72] Emanuele Di Buccio, Massimo Melucci et Federica Moro. « Detecting verbose queries and improving information retrieval ». In : *Information Processing & Management* 50.2 (2014), p. 342–360 (cf. p. 37, 39).

[73] Mariella Di Giacomo, Dan Mahoney, Johan Bollen, Andres Monroy-Hernandez et Cesar Ruiz Meraz. « MyLibrary, A Personalization Service for Digital Library Environments. » In : 2001 (cf. p. 28).

[74] Ying Ding. « Topic-based PageRank on author cocitation networks ». In : *Journal of the Association for Information Science and Technology* 62.3 (2011), p. 449–466 (cf. p. 144).

[75] Ying Ding, Erjia Yan, Arthur Frazho et James Caverlee. « PageRank for ranking authors in co-citation networks ». In : *Journal of the American Society for Information Science and Technology* 60.11 (2009), p. 2229–2243 (cf. p. 144).

[76] Ying Ding, Guo Zhang, Tamy Chambers, Min Song, Xiaolong Wang et Chengxiang Zhai. « Content-based citation analysis : The next generation of citation analysis ». In : *Journal of the Association for Information Science and Technology* 65.9 (2014), p. 1820–1833 (cf. p. 191, 197).

[77] Mladenic Dunja. « Text-learning and related intelligentagents : a survey ». In : t. 14. 4. IEEE Computer Society, 1999, p. 44–54 (cf. p. 151).

[78] Mohamed Ettaleb, Chiraz Latiri, Brahim Douar et Patrice Bellot. « SBS 2016 Track mining : Classification with linguistic features for book search requests classification ». In : (2016), p. 1079–1088 (cf. p. 120, 196).

[79] Shao-Hui Feng, Bo-Wen Zhang, Zan-Xia Jin, Xu-Cheng Yin, Jian-Lin Jin, Jian-Wei W, Le-Le Zhang, Hao-Jie Pan, Fan Fang et Fang Zhou. « USTB at Social Book Search 2016 Suggestion Task : Active Books Set and Reranking ». In : (2016), p. 1089–1096 (cf. p. 46).

[80] Alejandro Figueroa. « Exploring effective features for recognizing the user intent behind web queries ». In : *Computers in Industry* 68 (2015), p. 162–169 (cf. p. 49).

[81] Magda FLOREZ. « la citation positionnée dans l'écrit scientifique ». In : *L'écrit scientifique : du lexique au discours* (2013), p. 67–84 (cf. p. 141).

[82] Santo FORTUNATO et Claudio CASTELLANO. « Community structure in graphs ». In : *Computational Complexity*. Springer, 2012, p. 490–512 (cf. p. 135).

[83] Edward FOX et Ricardo da SILVA TORRES. « Digital Library Technologies : Complex Objects, Annotation, Ontologies, Classification, Extraction, and Security ». In : *Synthesis Lectures on Information Concepts, Retrieval, and Services* (2014), p. 1–205 (cf. p. 110).

[84] Maria GÄDE, Mark HALL, Hugo HUURDEMAN, Jaap KAMPS, Marijn KOOLEN, Mette SKOV, Elaine TOMS et David WALSH. « Overview of the SBS 2015 Interactive Track ». In : *CLEF 2015 Evaluation Labs and Workshop Online Working Notes*. 2015 (cf. p. 46).

[85] Eugene GARFIELD. *Citation indexing : Its theory and application in science, technology, and humanities*. T. 8. Information sciences series. Isi Press, 1979 (cf. p. 145).

[86] Wolfgang GLÄNZEL et András SCHUBERT. « Analysing scientific networks through co-authorship ». In : *Handbook of quantitative science and technology research*. Springer, 2004, p. 257–276 (cf. p. 142).

[87] Murat GÖKSEDEF et Şule GÜNDÜZ-ÖĞÜDÜCÜ. « Combination of Web page recommender systems ». In : *Expert Systems with Applications* 37.4 (2010), p. 2911–2922 (cf. p. 153).

[88] Ken GOLDBERG, Theresa ROEDER, Dhruv GUPTA et Chris PERKINS. « Eigentaste : A constant time collaborative filtering algorithm ». In : *Information Retrieval* 4.2 (2001), p. 133–151 (cf. p. 151).

[89] Carlos GOMEZ-URIBE et Neil HUNT. « The netflix recommender system : Algorithms, business value, and innovation ». In : *ACM Transactions on Management Information Systems (TMIS)* 6.4 (2016), p. 13 (cf. p. 27).

[90] Juan GORRAIZ, Philip PURNELL et Wolfgang GLÄNZEL. « Opportunities for and limitations of the book citation index ». In : *Journal of the American Society for Information Science and Technology* 64.7 (2013), p. 1388–1398 (cf. p. 149).

[91] Pablo GRANITTO, Cesare FURLANELLO, Franco BIASIOLI et Flavia GASPERI. « Recursive feature elimination with random forest for PTR-MS analysis of agroindustrial products ». In : *Chemometrics and Intelligent Laboratory Systems* 83.2 (2006), p. 83–90 (cf. p. 119).

[92] Manish GUPTA et Michael BENDERSKY. « Information Retrieval with Verbose Queries ». In : *Proposal for a Tutorial at SIGIR'15 Conference*. 2015 (cf. p. 37).

[93] Pankaj GUPTA, Ashish GOEL, Jimmy LIN, Aneesh SHARMA, Dong WANG et Reza ZADEH. « Wtf : The who to follow service at twitter ». In : *Proceedings of the 22nd international conference on World Wide Web*. ACM. 2013, p. 505–514 (cf. p. 150).

[94] Isabelle GUYON, Jason WESTON, Stephen BARNHILL et Vladimir VAPNIK. « Gene selection for cancer classification using support vector machines ». In : *Machine learning* 46.1 (2002), p. 389–422 (cf. p. 88).

[95] Björn HAMMARFELT. « Interdisciplinarity and the intellectual base of literature studies : Citation analysis of highly cited monographs ». In : *Scientometrics* 86.3 (2011), p. 705–725 (cf. p. 149).

[96] Björn HAMMARFELT. « Using altmetrics for assessing research impact in the humanities ». In : *Scientometrics* 101.2 (2014), p. 1419–1430 (cf. p. 149).

[97] Björn HAMMARFELT. « Beyond Coverage : Toward a Bibliometrics for the Humanities ». In : *Research Assessment in the Humanities*. Springer, 2016, p. 115–131 (cf. p. 149).

[98] Li HANG. « A short introduction to learning to rank ». In : *IEICE TRANSACTIONS on Information and Systems* 94.10 (2011), p. 1854–1862 (cf. p. 45).

[99] Stefanie HAUSTEIN, Isabella PETERS, Cassidy SUGIMOTO, Mike THELWALL et Vincent LARIVIÈRE. « Tweeting biomedicine : an analysis of tweets and citations in the biomedical literature ». In : *CoRR* abs/1308.1838 (2013). URL : http://arxiv.org/abs/1308.1838 (cf. p. 147).

[100] Jonathan HERLOCKER, Joseph KONSTAN, Loren TERVEEN et John RIEDL. « Evaluating collaborative filtering recommender systems ». In : *ACM Transactions on Information Systems (TOIS)* 22.1 (2004), p. 5–53 (cf. p. 28, 198).

[101] Irazú HERNÁNDEZ, Parth GUPTA, Paolo ROSSO et Martha ROCHA. « A simple model for classifying web queries by user intent ». In : *Proceedings of the 2nd Spanish Conference on Information Retrieval*. CERI. 2012, p. 235–240 (cf. p. 48, 49).

[102] Will HILL, Larry STEAD, Mark ROSENSTEIN et George FURNAS. « Recommending and evaluating choices in a virtual community of use ». In : *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM Press/Addison-Wesley Publishing Co. 1995, p. 194–201 (cf. p. 151).

[103] Jorge HIRSCH. « An index to quantify an individual's scientific research output ». In : *Proceedings of the National academy of Sciences of the United States of America* (2005), p. 16569–16572 (cf. p. 146).

[104] Kim HOLMBERG et Mike THELWALL. « Disciplinary differences in Twitter scholarly communication ». In : *Scientometrics* 101.2 (2014), p. 1027–1042 (cf. p. 149).

[105] Amal HTAIT, Sébastien FOURNIER et Patrice BELLOT. « Identification Automatique de Mots-Germes pour l'Analyse de Sentiments et son Intensité ». In : *CORIA-RJCRI*. Marseille, France, 2017 (cf. p. 191, 197).

[106] I-Ane HUANG, Jan-Ming HO, Hung-Yu KAO et Wen-Chang LIN. « Extracting citation metadata from online publication lists using BLAST ». In : *Advances in Knowledge Discovery and Data Mining*. Springer, 2004, p. 539–548 (cf. p. 86).

[107] David HULL. « Using statistical testing in the evaluation of retrieval experiments ». In : *Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM. 1993, p. 329–338 (cf. p. 70).

[108] Samuel HUSTON et Bruce CROFT. « Evaluating verbose query processing techniques ». In : *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*. ACM. 2010, p. 291–298 (cf. p. 37, 41).

[109] Melanie IMHOF. « BM25 for Non-Textual Modalities in Social Book Search ». In : *Working Notes of CLEF 2016*. CLEF. 2016, p. 1123–1129 (cf. p. 46).

[110] Fo ISINKAYE, Yo FOLAJIMI et Ba OJOKOH. « Recommendation systems : Principles, methods and evaluation ». In : *Egyptian Informatics Journal* 16.3 (2015), p. 261–273 (cf. p. 150).

[111] Mathieu JACOMY, Tommaso VENTURINI, Sebastien HEYMANN et Mathieu BASTIAN. « ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software ». In : *PloS one* 9.6 (2014), e98679 (cf. p. 133).

[112] Dietmar JANNACH et Malte LUDEWIG. « Determining characteristics of successful recommendations from log data : a case study ». In : *Proceedings of the Symposium on Applied Computing*. ACM. 2017, p. 1643–1648 (cf. p. 192, 198).

[113] Bernard JANSEN et Danielle BOOTH. « Classifying web queries by topic and user intent ». In : *CHI'10 Extended Abstracts on Human Factors in Computing Systems*. ACM. 2010, p. 4285–4290 (cf. p. 48).

[114] Bernard JANSEN, Danielle BOOTH et Amanda SPINK. « Determining the informational, navigational, and transactional intent of Web queries ». In : *Information Processing & Management* 44.3 (2008), p. 1251–1266 (cf. p. 48, 49).

[115] Rosie JONES, Benjamin REY, Omid MADANI et Wiley GREINER. « Generating query substitutions ». In : *Proceedings of the 15th international conference on World Wide Web*. ACM. 2006, p. 387–396 (cf. p. 43).

[116] Ashish KATHURIA, Bernard JANSEN, Carolyn HAFERNIK et Amanda SPINK. « Classifying the user intent of web queries using k-means clustering ». In : *Internet Research* 20.5 (2010), p. 563–581 (cf. p. 48).

[117] Maxwell Mirton KESSLER. « Bibliographic coupling between scientific papers ». In : *Journal of the Association for Information Science and Technology* 14.1 (1963), p. 10–25 (cf. p. 132).

[118] Kyoung-Min KIM, Jin-Hyuk HONG et Sung-Bae CHO. « A semantic Bayesian network approach to retrieving information with intelligent conversational agents ». In : *Information Processing & Management* 43.1 (2007), p. 225–236 (cf. p. 44).

[119] Young-Min KIM, Patrice BELLOT, Elodie FAATH et Marin DACOS. « Automatic annotation of bibliographical references in digital humanities books, articles and blogs ». In : *Proceedings of the 4th ACM workshop on Online books, complementary social media and crowdsourcing*. ACM. 2011, p. 41–48 (cf. p. 81, 107, 110, 160).

[120] Young-Min KIM, Patrice BELLOT, Elodie FAATH et Marin DACOS. « Annotated Bibliographical Reference Corpora in Digital Humanities. » In : *LREC*. 2012, p. 329–340 (cf. p. 92).

[121] Young-Min KIM, Patrice BELLOT, Jade TAVERNIER, Elodie FAATH et Marin DACOS. « Evaluation of BILBO reference parsing in digital humanities via a comparison of different tools ». In : *Proceedings of the 2012 ACM symposium on Document engineering*. ACM. 2012, p. 209–212 (cf. p. 107, 160).

[122] Joseph KONSTAN, Bradley MILLER, David MALTZ, Jonathan HERLOCKER, Lee GORDON et John RIEDL. « GroupLens : applying collaborative filtering to Usenet news ». In : t. 40. 3. ACM, 1997, p. 77–87 (cf. p. 28, 151).

[123] Marijn KOOLEN, Toine BOGERS et Jaap KAMPS. « Overview of the SBS 2016 Suggestion Track ». In : (2016), p. 1039–1052 (cf. p. 34, 38, 123).

[124] Kayvan KOUSHA, Mike THELWALL et Somayeh REZAIE. « Assessing the citation impact of books : The role of Google Books, Google Scholar, and Scopus ». In : *Journal of the Association for Information Science and Technology* 62.11 (2011), p. 2147–2164 (cf. p. 149).

[125] Giridhar KUMARAN et James ALLAN. « Effective and efficient user interaction for long queries ». In : *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*. ACM. 2008, p. 11–18 (cf. p. 42).

[126] Giridhar KUMARAN et Vitor CARVALHO. « Reducing long queries using query quality predictors ». In : *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*. ACM. 2009, p. 564–571 (cf. p. 40, 41).

[127] Rossitza KYHENG. « La référence bibliographique : NORME ET PRAXIS À l'aide des spécialistes en sciences humaines et sociales ». In : (2003) (cf. p. 95, 161).

[128] John LAFFERTY, Andrew MCCALLUM et Fernando PEREIRA. « Conditional Random Fields : Probabilistic Models for Segmenting and Labeling Sequence Data ». In : *Proceedings of the Eighteenth International Conference on Machine Learning*. ICML '01. San Francisco, CA, USA : Morgan Kaufmann Publishers Inc., 2001, p. 282–289 (cf. p. 86, 87, 90).

[129] Thomas LAVERGNE, Olivier CAPPÉ et François YVON. « Practical very large scale CRFs ». In : *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics. 2010, p. 504–513 (cf. p. 86).

[130] Sarah LEROY. « Lecture de : Catherine Schnedecker, Nom propre et chaînes de référence ». In : *Cahiers de praxématique* 32 (1999), p. 226–228 (cf. p. 84).

[131] Simon LEVA. « Caractérisation linguistique des requêtes des utilisateurs d'un moteur de recherche : vers une détection automatique des types de besoins d'information ». In : *Mémoire de Master 2 TAL* (2011) (cf. p. 173).

[132] Simon LEVA. « Les sessions de recherche comme contexte des requêtes ». In : *13e Conférence Francophone sur l'Extraction et la Gestion des Connaissances*. 2013, p. 1–12 (cf. p. 174).

[133] Dirk LEWANDOWSKI, Jessica DRECHSLER et Sonja MACH. « Deriving query intents from web search engine queries ». In : *Journal of the American Society for Information Science and Technology* 63.9 (2012), p. 1773–1788 (cf. p. 48, 49).

[134] Dingcheng LI, Karin KIPPER-SCHULER et Guergana SAVOVA. « Conditional random fields and support vector machines for disorder named entity recognition in clinical texts ». In : *Proceedings of the workshop on current trends in biomedical natural language processing*. Association for Computational Linguistics. 2008, p. 94–95 (cf. p. 90).

[135] Jie LI, Judy BURNHAM, Trey LEMLEY et Robert BRITTON. « Citation analysis : Comparison of web of science®, scopus™, SciFinder®, and google scholar ». In : *Journal of electronic resources in medical libraries*. T. 7. 2010, p. 196–217 (cf. p. 147).

[136] AJM LINMANS. « Why with bibliometrics the humanities does not need to be the weakest link ». In : *Scientometrics* 83.2 (2010), p. 337–354 (cf. p. 149).

[137] Christina LIOMA et Roi BLANCO. « Part of speech based term weighting for information retrieval ». In : *Advances in information retrieval*. Springer, 2009, p. 412–423 (cf. p. 27).

[138] Émile LITTRÉ. « Dictionnaire de la langue française ». In : 1 (1874), p. 1408 (cf. p. 53).

[139] Jiahui LIU, Peter DOLAN et Elin Rønby PEDERSEN. « Personalized news recommendation based on click behavior ». In : *Proceedings of the 15th international conference on Intelligent user interfaces*. ACM. 2010, p. 31–40 (cf. p. 192, 197).

[140] Laurence LONGO. « Vers des moteurs de recherche «intelligents» : un outil de détection automatique de thèmes ». Thèse de doct. Université François Rabelais, 2013 (cf. p. 84).

[141] Patrice LOPEZ. « GROBID : combining automatic bibliographic data recognition and term extraction for scholarship publications ». In : *Proceedings of the 13th European conference on Research and advanced technology for digital libraries*. ECDL'09. Springer-Verlag, 2009, p. 473–474 (cf. p. 91).

[142] Pasquale LOPS, Marco DE GEMMIS et Giovanni SEMERARO. « Content-based recommender systems : State of the art and trends ». In : *Recommender systems handbook*. Springer, 2011, p. 73–105 (cf. p. 152, 153).

[143] Chao LU, Ying DING et Chengzhi ZHANG. « Understanding the impact change of a highly cited article : a content-based citation analysis ». In : *Scientometrics* (2017), p. 1–19 (cf. p. 191, 197).

[144] Jie LU, Dianshuang WU, Mingsong MAO, Wei WANG et Guangquan ZHANG. « Recommender system application developments : a survey ». In : *Decision Support Systems* 74 (2015), p. 12–32 (cf. p. 28).

[145] Louise-Noëlle MALCLÈS. *La Bibliographie*. PUF, 1977, p. 456 (cf. p. 93).

[146] Christopher MANNING, Prabhakar RAGHAVAN, Hinrich SCHÜTZE et al. *Introduction to information retrieval*. T. 1. 1. Cambridge university press Cambridge, 2008 (cf. p. 43, 143).

[147] Melvin Earl MARON et John KUHNS. « On relevance, probabilistic indexing and information retrieval ». In : *Journal of the ACM (JACM)* 7.3 (1960), p. 216–244 (cf. p. 44).

[148] Tamsin MAXWELL et Bruce CROFT. « Compact query term selection using topically related text ». In : *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*. ACM. 2013, p. 583–592 (cf. p. 41).

[149] Philipp MAYR, Ingo FROMMHOLZ et Guillaume CABANAC. « Bibliometric-Enhanced Information Retrieval : 3rd International BIR Workshop ». In : *European Conference on Information Retrieval*. Springer. 2016, p. 865–868 (cf. p. 142, 144).

[150] Philipp MAYR et Andrea SCHARNHORST. « Scientometrics and information retrieval : weak-links revitalized ». In : *Scientometrics* 102.3 (2015), p. 2193–2199 (cf. p. 144).

[151] Andrew MCCALLUM et Kamal NIGAM. « A comparison of event models for naive bayes text classification ». In : *AAAI-98 workshop on learning for text categorization*. Citeseer. 1998, p. 41–48 (cf. p. 66).

[152] Marie MCVEIGH et Stephen MANN. « The journal impact factor denominator : defining citable (counted) items ». In : *Jama* 302.10 (2009), p. 1107–1109 (cf. p. 146).

[153] Robert MERTON. « Priorities in scientific discovery : a chapter in the sociology of science ». In : *American sociological review* 22.6 (1957), p. 635–659 (cf. p. 146).

[154] AC MITRA. « The Bibliographical reference : a review of its role ». In : (1970) (cf. p. 164).

[155] Alaa MOHASSEB, Maged EL-SAYED et Khaled MAHAR. « Automated Identification of Web Queries using Search Type Patterns. » In : *WEBIST (2)*. 2014, p. 295–304 (cf. p. 48, 49).

[156] Miquel MONTANER, Beatriz LÓPEZ et Josep Lluís de la ROSA. « Developing trust in recommender agents ». In : *Proceedings of the first international joint conference on Autonomous agents and multiagent systems : part 1*. ACM. 2002, p. 304–305 (cf. p. 198).

[157] Raymond MOONEY et Loriene ROY. « Content-based book recommending using learning for text categorization ». In : *fifth ACM conference on Digital libraries*. ACM. 2000, p. 195–204 (cf. p. 154).

[158] Cristian MORAL, Angélica de ANTONIO, Ricardo IMBERT et Jaime RAMÍREZ. « A survey of stemming algorithms in information retrieval ». In : t. 19(1). 2014 (cf. p. 27).

[159] Judith MUZERELLE, Emmanuel SCHANG et Jean-Yves ANTOINE. « Annotations en chaînes de coréférences et anaphores dans un corpus de discours spontané en français ». In : *SHS Web of Conferences*. T. 1. EDP Sciences. 2012, p. 2497–2516 (cf. p. 84).

[160] Judith MUZERELLE, Emmanuel SCHANG et Jean-Yves ANTOINE. « Annotation en relations anaphoriques d'un corpus de discours oral spontané en français ». In : *Congrès Mondial de Linguistique Française, CMLF'2012*. 2013, 15–pp (cf. p. 84).

[161] James NORRIS. *Markov chains*. 2. Cambridge university press, 1998 (cf. p. 87).

[162] Michael OCHSNER, Sven HUG et Hans-Dieter DANIEL. « Four types of research in the humanities : Setting the stage for research quality criteria in the humanities ». In : *Research Evaluation* 22.2 (2013), p. 79–92 (cf. p. 149).

[163] Yoshiko OKUBO. *Indicateurs bibliometriques et analyse des systemes de recherche*. OECD Publishing, 1997 (cf. p. 141).

[164] Anaïs OLLAGNIER, Sébastien FOURNIER et Patrice BELLOT. « Cascade de CRFs et SVM pour la détection de références bibliographiques diffuses dans les articles scientifiques ». In : *CORIA*. Semaine du Document Numérique et de la Recherche d'Information : COnférence en Recherche d'Information et Applications, 2016 (cf. p. 90).

[165] Lawrence PAGE, Sergey BRIN, Rajeev MOTWANI et Terry WINOGRAD. *The PageRank citation ranking : Bringing order to the web.* Rapp. tech. Stanford InfoLab, 1999 (cf. p. 144).

[166] Jiaul PAIK et Douglas OARD. « A fixed-point method for weighting terms in verbose informational queries ». In : *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*. ACM. 2014, p. 131–140 (cf. p. 42).

[167] Bo PANG, Lillian LEE et al. « Opinion mining and sentiment analysis ». In : *Foundations and Trends® in Information Retrieval* 2.1–2 (2008), p. 1–135 (cf. p. 44).

[168] Nish PARIKH, Prasad SRIRAM et Mohammad AL HASAN. « On segmentation of ecommerce queries ». In : *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*. ACM. 2013, p. 1137–1146 (cf. p. 43).

[169] Alan PRITCHARD. « Statistical Bibliography or Bibliometrics ». In : t. 25. 4. 1969, p. 348–349 (cf. p. 145).

[170] Vahed QAZVINIAN et Dragomir RADEV. « Scientific paper summarization using citation summary networks ». In : *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*. Association for Computational Linguistics. 2008, p. 689–696 (cf. p. 85).

[171] Lawrence RABINER. « A tutorial on hidden Markov models and selected applications in speech recognition ». In : *Proceedings of the IEEE* 77.2 (1989), p. 257–286 (cf. p. 86).

[172] Al Mamunur RASHID, Istvan ALBERT, Dan COSLEY, Shyong LAM, Sean MCNEE, Joseph KONSTAN et John RIEDL. « Getting to know you : learning new user preferences in recommender systems ». In : *Proceedings of the 7th international conference on Intelligent user interfaces*. ACM. 2002, p. 127–134 (cf. p. 150).

[173] Paul RESNICK, Neophytos IACOVOU, Mitesh SUCHAK, Peter BERGSTROM et John RIEDL. « GroupLens : an open architecture for collaborative filtering of netnews ». In : *Proceedings of the 1994 ACM conference on Computer supported cooperative work*. ACM. 1994, p. 175–186 (cf. p. 27).

[174] Paul RESNICK et Hal VARIAN. « Recommender systems ». In : *Communications of the ACM* 40.3 (1997), p. 56–58 (cf. p. 27).

[175] Behnam REZAEI et Alice MUNTZ. *System and method for context-based knowledge search, tagging, collaboration, management, and advertisement*. 2013 (cf. p. 32, 86).

[176] Francesco RICCI, Lior ROKACH et Bracha SHAPIRA. « Introduction to recommender systems handbook ». In : *Recommender systems handbook*. Springer, 2011, p. 1–35 (cf. p. 27).

[177] Elaine RICH. « User modeling via stereotypes ». In : *Cognitive science* 3.4 (1979), p. 329–354 (cf. p. 27, 151).

[178] Anna RITCHIE, Stephen ROBERTSON et Simone TEUFEL. « Comparing citation contexts for information retrieval ». In : *Proceedings of the 17th ACM conference on Information and knowledge management*. ACM. 2008, p. 213–222 (cf. p. 85).

[179] Stephen ROBERTSON, Cornelis Joost VAN RIJSBERGEN et Martin PORTER. « Probabilistic Models of Indexing and Searching ». In : *SIGIR*. 1980, p. 35–56 (cf. p. 44, 63).

[180] Daniel ROSE et Danny LEVINSON. « Understanding user goals in web search ». In : *Proceedings of the 13th international conference on World Wide Web*. ACM. 2004, p. 13–19 (cf. p. 31, 48, 54).

[181] Gerard SALTON. « Recent studies in automatic text analysis and document retrieval ». In : *Journal of the ACM (JACM)* 20.2 (1973), p. 258–278 (cf. p. 43).

[182] Gerard SALTON. « Mathematics and information retrieval ». In : *Journal of Documentation* 35.1 (1979), p. 1–29 (cf. p. 44).

[183] Gerard SALTON. « Automatic text processing : The transformation, analysis, and retrieval of ». In : (1989) (cf. p. 28).

[184] Lawrence SAUL, Yair WEISS et Léon BOTTOU. *Advances in Neural Information Processing Systems 17*. 2005 (cf. p. 110).

[185] Andrew SCHEIN, Alexandrin POPESCUL, Lyle UNGAR et David PENNOCK. « Methods and metrics for cold-start recommendations ». In : *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM. 2002, p. 253–260 (cf. p. 198).

[186] Catherine SCHNEDECKER. *Nom propre et chaînes de référence*. Librairie KLINCKSIECK, 1997 (cf. p. 84).

[187] Martin SCHOLZ, George FORMAN et Rong PAN. *Collaborative filtering model having improved predictive performance*. US Patent 9,355,414. Mai 2016 (cf. p. 28).

[188] Ingo SCHWAB, Alfred KOBSA et Ivan KOYCHEV. « Learning user interests through positive examples using content analysis and collaborative filtering ». In : (2001) (cf. p. 154).

[189] Fei SHA et Fernando PEREIRA. « Shallow parsing with conditional random fields ». In : *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*. Association for Computational Linguistics. 2003, p. 134–141 (cf. p. 90).

[190] Shai SHALEV-SHWARTZ et Shai BEN-DAVID. *Understanding machine learning : From theory to algorithms*. Cambridge university press, 2014 (cf. p. 110).

[191] Daniel SHELDON, Milad SHOKOUHI, Martin SZUMMER et Nick CRASWELL. « LambdaMerge : merging the results of query reformulations ». In : *Proceedings of the fourth ACM international conference on Web search and data mining*. ACM. 2011, p. 795–804 (cf. p. 42).

[192] Barry SMYTH et Paul COTTER. « A personalised TV listings service for the digital TV age ». In : *Knowledge-Based Systems* 13.2 (2000), p. 53–59 (cf. p. 154).

[193] Léna SOLER. *Introduction à l'épistémologie*. Ellipses, 2009 (cf. p. 93).

[194] Luciana B SOLLACI et Mauricio G PEREIRA. « The introduction, methods, results, and discussion (IMRAD) structure : a fifty-year survey ». In : *Journal of the medical library association* 92.3 (2004), p. 364 (cf. p. 103).

[195] Charles SUTTON. *Efficient training methods for conditional random fields*. ProQuest, 2008 (cf. p. 90).

[196] Charles SUTTON et Andrew MCCALLUM. « An introduction to conditional random fields for relational learning ». In : *Introduction to statistical relational learning* (2006), p. 93–128 (cf. p. 90).

[197] Charles SUTTON et Andrew MCCALLUM. « An Introduction to Conditional Random Fields ». In : *Foundations and Trends in Machine Learning* 4 (4 2011), p. 267–373 (cf. p. 87).

[198] John SWALES. *Genre analysis : English in academic and research settings*. Cambridge University Press, 1990 (cf. p. 99, 141, 164).

[199] Michael SYMONDS, Peter BRUZA, Guido ZUCCON, Bevan KOOPMAN, Laurianne SITBON et Ian TURNER. « Automatic query expansion : A structural linguistic perspective ». In : *Journal of the Association for Information Science and Technology* 65.8 (2014), p. 1577–1596 (cf. p. 42).

[200] Bin TAN et Fuchun PENG. « Unsupervised query segmentation using generative language models and wikipedia ». In : *Proceedings of the 17th international conference on World Wide Web*. ACM. 2008, p. 347–356 (cf. p. 43).

[201] Buzhou TANG, Yudong FENG, Xiaolong WANG, Yonghui WU, Yaoyun ZHANG, Min JIANG, Jingqi WANG et Hua XU. « A comparison of conditional random fields and structured support vector machines for chemical entity recognition in biomedical literature ». In : *Journal of cheminformatics* 7.1 (2015), S8 (cf. p. 90).

[202] Ludovic TANGUY, Fanny LALLEMAN, Claire FRANÇOIS, Philippe MULLER et Patrick SÉGUÉLA. « RHECITAS : citation analysis of French humanities articles ». In : *Corpus Linguistics 2009*. 2009, http–ucrel (cf. p. 149).

[203] Loren TERVEEN, Will HILL, Brian AMENTO, David MCDONALD et Josh CRETER. « PHOAKS : A system for sharing recommendations ». In : *Communications of the ACM* 40.3 (1997), p. 59–62 (cf. p. 151).

[204] Mike THELWALL. « Bibliometrics to webometrics ». In : *Journal of information science* 34.4 (2008), p. 605–621 (cf. p. 147).

[205] Joachims THORSTEN. *Making large scale SVM learning practical*. Rapp. tech. Universität Dortmund, 1999 (cf. p. 119).

[206] Daniel TORRES-SALINAS, Nicolas ROBINSON-GARCIA, Juan MIGUEL CAMPANARIO et Emilio DELGADO LOPEZ-COZAR. « Coverage, field specialisation and the impact of scientific publishers indexed in the Book Citation Index ». In : *Online Information Review* 38.11 (2014), p. 24–42 (cf. p. 148).

[207] Hong Diep TRAN, Guillaume CABANAC et Gilles HUBERT. « Expert suggestion for conference program committees ». In : *Research Challenges in Information Science (RCIS), 2017 11th International Conference on*. IEEE. 2017, p. 221–232 (cf. p. 192).

[208] Thomas TRAN et Robin COHEN. « Hybrid recommender systems for electronic commerce ». In : *Proceedings of Knowledge-Based Electronic Markets, Papers from the AAAI Workshop*. 2000 (cf. p. 153).

[209] Howard TURTLE et Bruce CROFT. « Evaluation of an inference network-based retrieval model ». In : *ACM Transactions on Information Systems (TOIS)* 9.3 (1991), p. 187–222 (cf. p. 44).

[210] Agnès TUTIN. « Dans cet article, nous souhaitons montrer que... Lexique verbal et positionnement de l'auteur dans les articles en sciences humaines ». In : *Lidil. Revue de linguistique et de didactique des langues* 41 (2010), p. 15–40 (cf. p. 99).

[211] Agnès TUTIN et Francis GROSSMANN. *L'écrit scientifique : du lexique au discours*. Presses Universitaires de Rennes, 2013 (cf. p. 81, 141).

[212] Vladimir VAPNIK. « An overview of statistical learning theory ». In : *IEEE transactions on neural networks* 10.5 (1999), p. 988–999 (cf. p. 88).

[213] Vladimir Naumovich VAPNIK et Vlamimir VAPNIK. *Statistical learning theory*. T. 1. Wiley New York, 1995 (cf. p. 89).

[214] Ludo WALTMAN. « A review of the literature on citation impact indicators ». In : *Journal of Informetrics* 10.2 (2016), p. 365–391 (cf. p. 148).

[215] Xuanhui WANG et ChengXiang ZHAI. « Mining term association patterns from search logs for effective query reformulation ». In : *Proceedings of the 17th ACM conference on Information and knowledge management*. ACM. 2008, p. 479–488 (cf. p. 42).

[216] Howard WHITE et Katherine MCCAIN. « Visualizing a discipline : An author co-citation analysis of information science, 1972-1995 ». In : *Journal of the American society for information science* 49 (1998), p. 327–355 (cf. p. 142).

[217] Concepción WILSON. « Informetrics. » In : *Annual Review of Information Science and Technology (ARIST)* 34 (1999), p. 107–247 (cf. p. 143).

[218] Edwin WILSON. « Probable inference, the law of succession, and statistical inference ». In : *Journal of the American Statistical Association* 22.158 (1927), p. 209–212 (cf. p. 63).

[219] Theresa WILSON, Janyce WIEBE et Paul HOFFMANN. « Recognizing contextual polarity in phrase-level sentiment analysis ». In : *Proceedings of the conference on human language technology and empirical methods in natural language processing*. Association for Computational Linguistics. 2005, p. 347–354 (cf. p. 191, 197).

[220] Dietmar WOLFRAM. « Applications of informetrics to information retrieval research ». In : *Informing Science* 3 (2003), p. 77–82 (cf. p. 144).

[221] Dietmar WOLFRAM. « The symbiotic relationship between information retrieval and informetrics ». In : *Scientometrics* 102.3 (2015), p. 2201–2214 (cf. p. 33).

[222] Dietmar WOLFRAM. « Bibliometrics, Information Retrieval and Natural Language Processing : Natural Synergies to Support Digital Library Research. » In : 2016, p. 6–13 (cf. p. 142, 144).

[223] Paulus Franciscus WOUTERS et al. « The citation culture ». In : (1999) (cf. p. 32, 81).

[224] Dayong WU, Yu ZHANG, Shiqi ZHAO et Ting LIU. « Identification of Web Query Intent Based on Query Text and Web Knowledge ». In : *Pervasive Computing Signal Processing and Applications*. PCSPA. 2010, p. 128–131 (cf. p. 49).

[225] Shih-Hung WU, Yi-Hsiang HSIEH, Liang-Pu CHEN et Ping-Che YANG. « Query Expansion by Word Embedding in the Suggestion Track of CLEF 2016 Social Book Search Lab ». In : (2016), p. 1155–1165 (cf. p. 46, 60).

[226] Shih-Hung WU, Pei-Kai LIAO, Hua-Wei LIN, Li-Jen HSU, Wei-Lun XIAO, Liang-Pu CHEN, Tsun KU et Gwo-Dong CHEN. « Query Type Recognition and Result Filtering in INEX 2014 Social Book Search Track ». In : *CLEF 2014 Evaluation Labs and Workshop Online Working Notes*. 2014 (cf. p. 46, 60).

[227] Xiaobing XUE et Bruce CROFT. « Modeling subset distributions for verbose queries ». In : *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*. ACM. 2011, p. 1133–1134 (cf. p. 41).

[228] Xiaobing XUE et Bruce CROFT. « Generating reformulation trees for complex queries ». In : *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*. ACM. 2012, p. 525–534 (cf. p. 41).

[229] Xiaobing XUE, Jiwoon JEON et Bruce CROFT. « Retrieval models for question and answer archives ». In : *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*. ACM. 2008, p. 475–482 (cf. p. 42).

[230] Xiaobing XUE, Yu TAO, Daxin JIANG et Hang LI. « Automatically mining question reformulation patterns from search log data ». In : *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics : Short Papers-Volume 2*. Association for Computational Linguistics. 2012, p. 187–192 (cf. p. 42).

[231] Bishan YANG, Nish PARIKH, Gyanit SINGH et Neel SUNDARESAN. « A study of query term deletion using large-scale e-commerce search logs ». In : *European Conference on Information Retrieval*. Springer. 2014, p. 235–246 (cf. p. 41).

[232] Nazpar YAZDANFAR et Alex THOMO. « Link recommender : Collaborative-filtering for recommending urls to twitter users ». In : *Procedia Computer Science* 19 (2013), p. 412–419 (cf. p. 82).

[233] Jeonghe YI et Farzin MAGHOUL. « Mobile search pattern evolution : the trend and the impact of voice queries ». In : *Proceedings of the 20th international conference companion on World wide web*. ACM. 2011, p. 165–166 (cf. p. 40).

[234] Liang ZHANG. « The Definition of Novelty in Recommendation System. » In : *Journal of Engineering Science & Technology Review* 6.3 (2013) (cf. p. 198).

[235] Bo-Wen ZHANG, Xu-Cheng YIN, Xiao-Ping CUI, Jiao QU, Bin GENG, Fang ZHOU et Hong-Wei HAO. « USTB at INEX2014 : Social Book Search Track ». In : *CLEF 2014 Evaluation Labs and Workshop Online Working Notes*. 2014 (cf. p. 46).

[236] Dangzhi ZHAO, Alicia CAPPELLO et Lucinda JOHNSTON. « Functions of Uni-and Multi-citations : Implications for Weighted Citation Analysis ». In : *Journal of Data and Information Science* 2.1 (2017), p. 51–69 (cf. p. 164).

ANNEXES

A. Résultats de la classification supervisée des paragraphes contenant des références bibliographiques

Dans cette annexe, nous présentons les différentes expérimentations effectuées afin de trouver le paramétrage optimal permettant la classification supervisée des paragraphes contenant des références bibliographiques. Nous avons opté pour deux algorithmes couramment utilisés lors de la sélection d'attributs : *Info Gain Attribute* (IGA) et *Gain Ration Attribute* (GRA). Les expérimentations effectuées sur ces algorithmes se sont majoritairement concentrées sur la manipulation de la fréquence minimale d'apparition des termes que nous souhaitions inclure dans la liste des attributs. Une fois cette liste obtenue, nous avons fait varier sa taille via l'inclusion ou non des mots obtenant un score « Élimination récursive de caractéristiques » (RFE) égal à 0. Cette liste a été ensuite utilisée afin de modéliser les données d'entrée pour l'apprentissage des modèles SVM. Ces évaluations ont été faites sur un échantillon contenant 340 paragraphes soit 170 exemples par classe.

Algorithme	Fréquence termes	Score RFE	Exactitude	Précision	Rappel
IGA	1	Tous	83,33 %	81,40 %	84,85 %
IGA	1	Sans 0	85,67 %	84,52 %	86,06 %
GRA	1	Tous	84,12 %	81,61 %	86,59 %
GRA	1	Sans 0	83,24 %	83,65 %	81,10 %
IGA	**3**	**Tous**	**85,29 %**	**86,08 %**	**82,93 %**
IGA	3	Sans 0	82,94 %	83,54 %	80,49 %
GRA	3	Tous	84,71 %	85,44 %	82,32 %
GRA	3	Sans 0	85,29 %	83,53 %	86,59 %
IGA Stopword	3	Tous	83,82 %	85,16 %	80,49 %
IGA Stemmer	3	Tous	84,41 %	86,27 %	80,49 %

Tableau .8. – Résultats de la classification supervisée des paragraphes

B. Compte rendu sur les retours de l'index SoLR d'OpenEdition

Ce compte rendu présente l'analyse des retours de l'index SoLR d'OpenEdition effectuée sur des requêtes basées sur des références bibliographiques. Le corpus d'étude se compose de 20 documents extraits du corpus présenté dans la section

2.5.1. Ce corpus contient uniquement des articles comportant des zones bibliographiques présentes en fin de document. Les requêtes ont été faites à la fois sur la base de données de Books et Revues.org. Cette étude nous permet d'estimer le nombre de retours de l'index SoLR et ainsi d'établir l'impact de cet index dans la génération potentielle de nouveaux nœuds pour le graphe de recommandation. Au total 365 références ont été extraites. Les résultats sont interprétés selon un test de classification binaire basé sur l'interprétation du résultat comme étant un faux positif (FP) ou un vrai positif (VP). Pour une question de temps, nous n'avons pu établir de tests concernant les faux négatifs et les vrais négatifs. En fonction des informations les plus couramment disponibles au sein des références bibliographiques, nous avons fait deux expérimentations : une interrogation de SoLR *via* le titre du document et le nom de l'auteur principal et une seconde interrogation de SoLR basée uniquement sur le titre du document.

B.1. Interrogations basées sur Auteur + Titre

Revues	Requête envoyée	URL retournée	Résultats
formationemploi-1253	titre :"Les attitudes à l'égard de l'insertion professionnelle d'apprentis de l'enseignement supérieur" auteur :"Cohen*"	http://osp.revues.org/5194	VP
aad-985	titre :"Œuvres complètes" auteur :"M*"	http://studifrancesi.revues.org/2854	FP
assr-11343	titre :"Témoignages" auteur :"*"	http://hleno.revues.org/511	FP
osb-872	titre :"" auteur :"HM"	http://monderusse.revues.org/7270	FP
osb-872	titre :"Les partenariats public-privé" auteur :(Voisin*)	http://pmp.revues.org/6558	FP* [51]
droitcultures-2224	titre :"Colloque de Cerisy" auteur :(J*)	http://rfp.revues.org/2079	FP
droitcultures-2224	titre :"Les impressionnistes" auteur :(B*)	http://ml.revues.org/248	FP
rsa-241	titre :"La parenté" auteur :(Barry*)	rl{http://lhomme.revues.org/284}&FP*	

51. * : renvoie à un document qui n'est pas totalement faux l'auteur et le titre sont présents

rsa-241	titre :"L'apport des familles homoparentales dans le débat actuel sur la construction de la parenté" auteur :(Cadoret*)	http://lhomme.revues.org/9081	VP
rsa-241	titre :"L'anonymat des dons d'engendrement est-il vraiment éthique" auteur :(Th*)	http://revdh.revues.org/193	VP
racf-1421	titre :"Préhistoire et Ethnologie. Le geste retrouvé" auteur :(Geneste*)	http://tc.revues.org/680	FP*
racf-1421	titre :"Contribution à l'étude de la circulation sur de longues distances des matières premières lithiques au Paléolithique" auteur :(Alixx*)	http://paleo.revues.org/1537	VP
recherches psychanalyse-1696	titre :"" auteur :(L*)	http://books.openedition.org/editionsmsh/1321	FP
recherches psychanalyse-1696	titre :"" auteur :(Calame*)	http://books.openedition.org/editionsmsh/1696	FP*
corpus-1672	titre :"Récritures et variation : pour une génétique linguistique et textuelle" auteur :(Adam*)	http://ml.revues.org/332	VP
corpus-1672	titre :"Les corpus réflexifs : entre architextualité et hypertextualité" auteur :(Mayaffre*)	http://corpus.revues.org/11	VP
corpus-1672	titre :"D'une sémiotique de l'altération" auteur :(Peytard*)	http://semen.revues.org/4182	VP
corpus-1672	titre :"Quelle place pour les sciences des textes dans l'Analyse de Discours" auteur :(Viprey*)	http://semen.revues.org/1995	VP

pm-202	titre :"Une sépulture d'esclave à Martigues Bouches-du-Rhône" auteur :(R*)	http://dam.revues.org/588	VP
communication-1247	titre :"L'image" auteur :(K*)	http://books.openedition.org/oep/751	FP

Tableau .9. – Résultats des retours de SoLR sur les interrogations Auteur + Titre

Seulement 20 retours de SoLR et seulement 9 VP. Nous avons pu constater que certaines erreurs étaient dues :

- Parsing des noms
- Titre incomplet ou trop commun (un seul mot)
- Certains éléments sont vides

B.2. Interrogations basées uniquement sur le Titre

Revues	Requête envoyée	URL retournée	Résultats
formationemploi-1253	titre :"Les attitudes à l'égard de l'insertion professionnelle d'apprentis de l'enseignement supérieur"	http://osp.revues.org/5194	VP
formationemploi-1253	titre :"La crise des identités "	http://osp.revues.org/5231	FP*
formationemploi-1253	titre :"Identités"	http://books.openedition.org/editionsmsh/6516	FP
aad-985	titre :"L'Art épistolaire"	http://rhetorique.revues.org/427	FP*
aad-985	titre :"Œuvres complètes"	http://ahrf.revues.org/1856	FP*
aad-985	titre :"Correspondance"	http://books.openedition.org/pum/1297	FP*
aad-985	titre :"L'ordre de l'iinteraction"	http://philonsorbonne.revues.org/102	FP*

aad-985	titre :"La lecture pragmatique"	http://rde.revues.org/4815	FP
aad-985	titre :"Lettres"	http://books.openedition.org/pur/33265	FP
aad-985	titre :"Œuvres complètes"	http://ahrf.revues.org/1856	FP
assr-11343	titre :"l'islam"	http://books.openedition.org/editionscnrs/842	FP
assr-11343	titre :"Témoignages"	http://cemoti.revues.org/922	FP
osb-872	titre :""	http://books.openedition.org/editionsehess/562	FP
osb-872	titre :"Les partenariats public-privé"	http://aspd.revues.org/365	FP*
droitcultures-2224	titre :"Colloque de Cerisy"	http://sabix.revues.org/335	FP*
droitcultures-2224	titre :"Les impressionnistes"	http://bcrfj.revues.org/7059	FP*
droitcultures-2224	titre :"Magie et religion"	http://lhomme.revues.org/14052	FP*
droitcultures-2224	titre :"L'appel des ténèbres"	http://droitcultures.revues.org/2224	FP*
rsa-241	titre :"La parenté"	http://books.openedition.org/editionsmsh/1185	FP*
rsa-241	titre :"L'apport des familles homoparentales dans le débat actuel sur la construction de la parenté"	http://lhomme.revues.org/9081	VP
rsa-241	titre :"Kinship"	http://lhomme.revues.org/38	FP*
rsa-241	titre :"Sociologie et Anthropologie"	http://lectures.revues.org/8324	FP*
rsa-241	titre :"Dir"	http://encyclopedieberbere.revues.org/2264	FP*

rsa-241	titre :"L'anonymat des dons d'engendrement est-il vraiment éthique"	http://revdh.revues.org/193	VP
racf-1421	titre :"Préhistoire et Ethnologie. Le geste retrouvé"	http://tc.revues.org/680	FP*
racf-1421	titre :"l'étude de la circulation sur de longues distances des matières premières lithiques au Paléolithique"	http://paleo.revues.org/1537	VP
racf-1421	titre :"Tardiglaciaire"	http://rae.revues.org/5499	FP*
geomorphologie-7520	titre :"sociétés et archéologie"	http://cem.revues.org/12950	FP
recherches psychanalyse-1696	titre :"L'homosexualité féminine dans l'Antiquité grecque et romaine"	http://genrehistoire.revues.org/307	FP*
recherches psychanalyse-1696	titre :"Le sexe incertain. Androgynie et hermaphrodisme dans l'Antiquité gréco-romaine"	http://kernos.revues.org/1249	FP*
recherches psychanalyse-1696	titre :""	http://books.openedition.org/editionsehess/562	FP
recherches psychanalyse-1696	titre :""	http://books.openedition.org/editionsehess/562	FP
recherches psychanalyse-1696	titre :"L'Éros dans la Grèce antique"	http://kernos.revues.org/1248	FP*
recherches psychanalyse-1696	titre :"Histoire de la sexualité"	http://lectures.revues.org/4840	FP*
recherches psychanalyse-1696	titre :"Désir et contraintes en Grèce ancienne"	http://clio.revues.org/7519	FP*

corpus-1672	titre :"Sciences du texte et analyse de discours. Enjeux d'une interdisciplinarité"	http://mots.revues.org/831	FP*
corpus-1672	titre :"La linguistique textuelle. Introduction à l'analyse textuelle des discours"	http://alsic.revues.org/300	VP
corpus-1672	titre :"Récritures et variation : pour une génétique linguistique et textuelle"	http://ml.revues.org/332	VP
corpus-1672	titre :"Le texte littéraire. Pour une approche interdisciplinaire"	http://corpus.revues.org/1957	FP*
corpus-1672	titre :"Enjeux d'une interdisciplinarité"	http://questionsdecommunication.revues.org/2661	FP*
corpus-1672	titre :"Les corpus réflexifs : entre architextualité et hypertextualité"	http://corpus.revues.org/11	VP
corpus-1672	titre :"D'une sémiotique de l'altération"	http://semen.revues.org/4182	VP
corpus-1672	titre :"Quelle place pour les sciences des textes dans l'Analyse de Discours"	http://semen.revues.org/1995	VP
pm-202	titre :"Une sépulture d'esclave à Martigues Bouches-du-Rhône"	http://dam.revues.org/588	VP
pm-202	titte :"Vol"	http://books.openedition.org/ifea/6639	FP
pm-202	titre :"Une nécropole hellénistique à la Pointe de Vella Port-de-Bouc, Bouches-du-Rhône"	http://dam.revues.org/563	VP
pm-202	titre :"Zabern"	http://abstractairanica.revues.org/6099	FP*

pm-202	titre :"Les bronzes grecs et romains : recherches récentes"	http://inha.revues.org/3245	VP
communication-1247	titre :" L'image"	http://books.openedition.org/oep/751	FP*
communication-1247	titre :"La distinction"	http://osp.revues.org/1526	FP
communication-1247	titre :"Stars"	http://map.revues.org/1548	FP
communication-1247	titre :"Télévision et démocratie"	http://mots.revues.org/7373	FP*
communication-1247	titre :"La nouvelle vague"	http://books.openedition.org/editionscnrs/2683	FP*
communication-1247	titre :"L'esprit du temps"	http://communication.revues.org/2559	FP*
communication-1247	titre :"Les stars"	http://map.revues.org/1548	FP
communication-1247	titre :"L'élite journalistique et son pouvoir"	http://questionsdecommunication.revues.org/7413	FP*
communication-1247	titre :"La presse féminine"	http://e-migrinter.revues.org/466	FP*

Tableau .10. – Résultats des retours de SoLR sur les interrogations uniquement sur le Titre

Seulement 58 retours de SoLR et seulement 12 VP. Nombreux FP* sont des comptes rendus de lecture des ouvrages concernés. La plupart des erreurs sont dues à des titres trop courts ou composés de mots trop communs. La plupart des ulrs correctes retournées ont des titres longs et précis. Des améliorations sont à prévoir quant à la capture du nom des auteurs qui pour le moment ne correspond qu'à une simple expression régulière. Nous avons pu constater que dans le cas d'un titre court le nom de l'auteur permet de désambiguïser et ainsi affiner la recherche. A l'inverse n'utiliser que le titre permet d'élargir le champ des possibles et d'avoir, quand l'auteur est mal détecté, des VP.

B.3. Analyse des retours de la revue *Corpus*

Nous nous sommes penchés sur les requêtes extraites de la revue *Corpus* afin de savoir si les articles ne sont pas récupérés car les entrées des requêtes sont mauvaises ou si l'article n'est pas dans la base de données.

Titre	**URL retournée**	**Résultats**
U. Heidmann,Sciences du texte et analyse de discours. Enjeux d'une interdisciplinarité	http://mots.revues.org/831	compte rendu
Adam Jean-Michel,La linguistique textuelle. Introduction à l'analyse textuelle des discours	http://alsic.revues.org/300	VP
Adam Jean-Michel,Récritures et variation : pour une génétique linguistique et textuelle	http://ml.revues.org/332	VP
Heidmann Ute,Le texte littéraire. Pour une approche interdisciplinaire	http://corpus.revues.org/1957	compte rendu
Cerquiglini Bernard,Eloge de la variante. Histoire critique de la philologie	NONE	Pas dans la base
Gresillon Almuth,Eléments de critique génétique	NONE	Pas dans la base
U. Heidmann,Enjeux d'une interdisciplinarité	http://questions.decommunication.revues.org/2661	compte rendu
C. Calame,Epistémologie et pratique de la comparaison différentielle	NONE	Pas dans la base
Jeanneret Michel,Chantiers de la Renaissance. Les variations de l'imprimé au XVIe siècle	NONE	Pas dans la base
Legallois Dominique,Le texte et le problème de son et ses unités : propositions pour une déclinaison	NONE	Pas dans la base
Legallois Dominique,Des phrases entre elles : unité réticulaire du texte	NONE	Pas dans la base
Lever Maurice,Romanciers du grand siècle	NONE	Pas dans la base

Lussault Michel,Dictionnaire de la géographie et de lespace des sociétés	NONE	Pas dans la base
Mayaffre Damon,Les corpus réflexifs : entre architextualité et hypertextualité	http://corpus.revues.org/11	VP
P. M. Wetherill,Manuscrits littéraires : comparaisons et histoire littéraire	NONE	Pas dans la base
Peytard Jean,D'une sémiotique de l'altération	http://semen.revues.org/4182	VP
G. Willliams,Enjeux épistémologiques de la linguistique de corpus	NONE	Pas dans la base
Viprey Jean-Marie,Quelle place pour les sciences des textes dans l'Analyse de Discours	http://semen.revues.org/1995	VP
Viprey Jean-Marie,Structure non-séquentielle des textes	NONE	Pas dans la base
Viprey Jean-Marie,"Language, Discourse : textual analysis, computer and statistics"	NONE	Pas dans la base
Zumthor Paul,Essai de poétique médiévale	NONE	Pas dans la base

Tableau .11. – Résultats des retours de SoLR sur les interrogations Auteur + Titre

Nous pouvons constater que dans le cas de cette revue la majorité des FP sont dûs à des articles qui ne sont pas dans la base.