THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par : *l'Université Toulouse 3 Paul Sabatier (UT3 Paul Sabatier)*

Présentée et soutenue le *19/09/2017* **par :**

Baptiste de La Robertie

Modèles d'Autorité et d'Expertise par les Graphes Hétérogènes
Application aux plateformes collaboratives et librairies digitales

---

JURY

Yoann Pitarch	Encadrant	Toulouse 3
Pascal Poncelet	Rapporteur	Montpellier
Celine Robardet	Examinatrice	INSA Lyon
Olivier Teste	Directeur	Toulouse 3
Julien Velcin	Rapporteur	Lyon 2
Bernard Dousset	Président	Toulouse 3

École doctorale et spécialité :
MITT : Image, Information, Hypermédia
Unité de Recherche :
Institut de Recherche en Informatique de Toulouse (IRIT)
Directeur(s) de Thèse :
Olivier Teste et *Yoann Pitarch (encadrant)*
Rapporteurs :
Pascal Poncelet et *Julien Velcin*

Abstract

With the advent of the Internet and the growth of Web technologies during the 90's, people have quickly changed the way they interact with each others. The multitude of search tools has shot up the informative landscape to create an horizontal dissemination system which has now difficulty in ranking informations published by internet users. The bundle of rumors arisen on social networks after the terrorist attacks in Paris in 2015 reminds us the difficulty in monitoring the spread of information on Web. According to the World Economic Forum, that massive distribution of false informations represents a significant threat to our societies.

To meet these challenges and expectations today, the scientific community has given some guidance for modelling the notions of authority and quality within networks. In particular, structural-based approaches, which use mathematical representations of networks exploiting the graph theory, constitute one of the most extensively studied and known solutions. However, if the scientific interest for modelling authority, quality and expertise is nothing new, current solutions still have various limitations. Firstly, since current structural approaches are generally developed in a very data specific way, they suffer from a lack of genericity, compromising both the generalizability and the extensibility of these approaches. Consequently, solutions are rarely transposable from one platform to another and are hardly adaptable to the evolution of the platforms. Secondly, if some state-of-the-art structural algorithms allow the integration of a quantitative knowledge, no solution allows the use of a qualitative knowledge about the quality of some particular entities.

In this manuscript, we propose to adress these limitations shared by numerous structural-based approaches. In particular, based on an heterogeneous graph representation of the platforms, a generic parametric model of influence is proposed allowing both the propagation of authority and quality scores in the graph and the correction of influence parameters with respect to an integrated qualitative knowledge. Furthermore, a definition of domain authority is proposed, allowing the characterization of both the notions of authority and expertise of the individuals in a single unified representation. The various contributions are evaluated over three collaborative sites Wikipedia, Stack Exchange and Academic Search Microsoft. Results, compared with state-of-the-art algorithms, outline the interest of the propositions and open numerous research questions.

Résumé

L'apparition d'internet dans les années 1990 a rapidement transformé la façon dont les individus s'informent et intéragissent. La multiplicité des outils de recherche et de veille, bousculant le paysage informationnel en fondant un système de diffusion horizontal, peine à hiérarchiser les informations et les contenus aujourd'hui publiés par les internautes eux-même. L'avalanche de rumeurs survenue sur les réseaux sociaux au lendemain des attentats de Paris en 2015 est un exemple qui rappelle la difficulté à contrôler la diffusion des informations sur le web. Dans ce contexte, et d'après le World Economic Forum, cette diffusion massive de fausses informations constituerait l'une des menaces les plus importante auxquelles nos sociétés sont confrontées.

Face à ces enjeux majeurs, la communauté scientifique apporte certains éléments de réponses notamment concernant la modélisation des notions d'autorité et de qualité au sein des réseaux. Dans ce contexte, les approches structurelles, basées sur une représentation mathématique des réseaux fondée sur la théorie des graphes, constituent une des familles de solutions la plus étudiée notamment pour le rôle historique tenue pendant le développement du moteur de recherche de Google. Pourtant, si l'intérêt scientifique pour les modèles d'autorité, de qualité et d'expertise n'est pas nouveau, les solutions actuelles partagent encore de nombreuses limitations. D'abord, les approches structurelles de la littérature sont généralement développées d'une façon très spécifique aux données utilisées. Celles-ci sont alors rarement transposables d'une plateforme à l'autre et difficilement adaptables à leurs évolutions. De plus, à notre connaissance, il n'existe pas dans la littérature d'approche structurelle permettant l'intégration de connaissances *a priori* présentées de façon qualitative, c'est à dire sous la forme d'une échelle ordinale.

Pour répondre à ces limitations, nous proposons dans ce manuscrit un modèle d'influence générique, basé sur l'exploitation d'une représentation par les graphes hétérogènes des plateformes, permettant une correction des paramètres d'influence guidée par un ensemble de connaissances *a priori* qualitatives. Par ailleurs, nous apportons une définition d'autorité de domaine, permettant la caractérisation conjointe des notions d'autorité et d'expertise des individus en exploitant à la fois certains indices de structure et de contenu. Les différentes contributions sont évaluées sur trois sites collaboratifs populaires Wikipédia, Stack Exchange et Microsoft Academic Search. Les résultats, comparés à l'état de l'art, soulignent l'intérêt de nos choix de représentations et ouvrent de nombreuses perspectives de recherche.

Remerciements

Mes remerciements vont d'abord à Olivier Teste pour avoir su réunir autour de cette thèse toutes les conditions nécessaires à sa réalisation (et dans les trois ans !). En plus de la confiance et des libertés accordées, Olivier s'est toujours placé en fin psychologue et a mené intelligemment nos échanges. Je remercie bien entendu aussi Yoann Pitarch pour avoir accompagné, soutenu et orienté ce travail durant ces trois années. Son suivi, la qualité de ses nombreux conseils et nombreuses corrections ont très clairement été une grande source d'apprentissage. Mon premier article n'aurait jamais vu le jour sans ton aide ! Mes remerciements vont aussi à Atsuhiro Takasu, et à tous ceux qui, d'une manière ou d'une autre, m'ont permis de réaliser un bout de cette thèse au NII. Je remercie aussi Pierre Fremaux pour avoir partagé le temps d'un article un jeu de données Babellio ainsi que Liana Ermakova pour son temps, son expertise, ses nombreux conseils et développements réalisés. Enfin, je remercie mes parents, non pour la thèse en particulier, mais bien pour tout le reste.

Publications

Les contributions présentées dans ce manuscrit ont fait l'objet de cinq publications internationales, une publication nationale ainsi qu'un chapitre d'ouvrage.

Chapitres d'ouvrage

- De la Robertie Baptiste, Pitarch Yoann et Teste Olivier. *Structure-Based Features for Predicting the Quality of Articles in Wikipedia.* Prediction and Inference from Social Networks and Social Media. Pages 115-140. Springer. 2017.

Articles de conférences internationales

- De la Robertie Baptiste, Ermakova Liana, Pitarch Yoann, Takasu Atsuhiro and Teste Olivier. *A Unified Approach for Learning Expertise and Authority in Digital Libraries.* International Conference on Database Systems for Advanced Applications (DASFAA), Suzhou, China, March 27 - 30, 2017. Regular paper.
- De la Robertie Baptiste, Pitarch Yoann, Takasu Atsuhiro and Teste Olivier. *Identifying Authoritative Researchers in Digital Libraries using External a Priori Knowledge.* Symposium on Applied Computing (SAC), Marrakech, Morocco, April 4 - 6, 2017. ACM. Regular paper.
- De La Robertie Baptiste. *Assessing the Authors of Online Books in Digital Libraries using Users Affinity.* Asia Information Retrieval Societies Conference (AIRS), Beijing, China, November 30 - December 02, 2016. Springer. Short paper.
- De la Robertie Baptiste, Pitarch Yoann and Teste Olivier. *Quality Prediction in Collaborative Platforms : A Generic Approach by Heterogeneous Graphs.* International Conference in Database and Expert Systems Applications (DEXA). Porto, Portugal, September 5-8, 2016. Springer. Short paper.
- De la Robertie Baptiste, Pitarch Yoann and Teste Olivier. *Measuring article quality in wikipedia using the collaboration network.* International Conference on Advances in Social Networks Analysis and Mining (ASONAM), Paris, France, August 25-28, 2015. ACM. Regular paper.

Articles de conférences nationales

- De la Robertie Baptiste, Pitarch Yoann et Teste Olivier. *Prédiction de la qualité dans les plateformes collaboratives : une approche générique*

par les graphes hétérogènes. Extraction et gestion des connaissances (EGC), Reims, France, 18-22 Janvier 2016. Papier court.

Table des matières

1 Introduction générale **13**
1 Introduction 13
2 Contributions de la thèse 15

2 Synthèse des travaux de l'état de l'art **19**
1 Indices et modèles de qualité et d'autorité 20
1.1 Indices de surface 20
1.2 Indices de contenu 23
1.3 Indices temporels 24
1.4 Indices de structure 25
2 Modèles de qualité et d'autorité 26
2.1 Approches inductives 27
2.2 Approches transductives 28
2.2.1 Graphes simples 28
2.2.2 Graphes bipartis 30
2.2.3 Graphes tri-partis 34
2.3 Discussion 36
3 Modèles d'expertise 37
3.1 Modèles centrés utilisateurs 39
3.2 Modèles centrés documents 40
3.3 Discussion 41
4 Conclusion 42

3 Modèle générique par les graphes hétérogènes **45**
1 Modélisation 46
1.1 Notations 46

1.2 Fonctions et graphe d'influence 48
1.3 Exercice des influences 49
1.3.1 Cas particulier 49
1.3.2 Cas général 50
1.4 Le principe de renforcement mutuel 51
2 Exemples d'applications 52
3 Algorithme 1 54
4 Intégration de connaissances 55
4.1 Modélisation des connaissances 56
4.2 Problème d'optimisation 56
4.3 Correction des paramètres 56
4.4 Algorithme 2 57
5 Discussion 64
6 Conclusion 67

4 Intégration de signaux temporels 69
1 La réactivité des utilisateurs sur les sites de Q&R 70
1.1 Motivations 70
1.2 Représentation 71
1.3 Modélisation 72
2 Le processus de révision des Wikis collaboratifs 73
2.1 Définitions 74
2.2 Modèle 75
2.3 Définition des influences 76
2.4 Discussion 78
3 Conclusion 79

5 Intégration de données textuelles 81
1 Motivations 82
2 Représentations 83
3 Modèle 85
4 Calculs 85
5 Modèle d'ordonnancement 86
6 Discussion 87
7 Conclusion 88

6 Validation des propositions et expérimentations 91
1 Jeux de données 92
1.1 Wikipedia 92
1.1.1 Description des données 93
1.1.2 Vérité terrain 93
1.1.3 Prétraitements 94
1.2 Stack Exchange 95
1.2.1 Description des données 95
1.2.2 Vérité terrain 95
1.3 Microsoft Academic Search 96
1.3.1 Description des données 97
1.3.2 Vérité terrain 97
1.3.3 Connaissance *a priori* 98
1.3.4 Prétraitements 98
2 Métriques d'évaluation 99
3 Validation des propositions 101
3.1 Microsoft Academic Search 101
3.1.1 Compétiteurs 102
3.1.2 Résultats 104
3.2 Stack Exchange 109
3.2.1 Compétiteurs 109
3.2.2 Résultats 110
3.3 Wikipedia 111
3.3.1 Compétiteurs 112
3.3.2 Résultats 113
4 Expérimentations 114
4.1 Ensemble de connaissances *a priori* 114
4.2 Coefficient pour l'expertise de domaine 116
4.3 Paramètre de co-éditions 117
4.4 Evaluation des signaux de structure 118
5 Conclusion 120

7 Conclusion générale 123
1 Synthèse des contributions 123
2 Perspectives de recherche 124

Chapitre 1

Introduction générale

1 Introduction

La révolution numérique, initiée dans les années 1990, a favorisé l'émergence d'un nouveau paradigme économique. Exploitant les notions de partage ou de co-création, le nombre d'entreprises comme BlablaCar [1] ou Uber [2] dans le transport, Airbnb [3] dans l'immobilier ou Kickstarter [4] dans le financement participatif a explosé. Si celles-ci partagaient en 2015 un marché d'environ 26 milliards de dollars, celui-ci devrait franchir, d'après le cabinet Forester [5], la barre des 100 milliards de dollars d'ici 2018. Au centre de cette nouvelle valeur économique réside bien une nouvelle figure commune, celle de l'utilisateur/producteur (Forestier et al., 2012).

Dans ce contexte, le domaine du collaboratif gratuit est tout aussi touché et accompagné d'une explosion d'informations produite par les internautes eux même (Forestier et al., 2011; Anokhin et al., 2012). Partageant l'idée d'un savoir accessible à tous, gratuitement et librement, la communauté mondiale formée par les Wikipédians constitue sans aucun doute la figure de référence du modèle collaboratif gratuit sur internet. Grace à la libre participation de 55 millions de contributeurs, 35 millions d'articles ont été co-écrits et publiés dans 280 langues sur la plateforme [6]. Le fonctionnement du site représente 10 millions de modifications d'articles par mois, soit environ une dizaine de modifications par seconde. De

1. https ://www.blablacar.fr/
2. https ://www.uber.com/fr/
3. https ://www.airbnb.fr/
4. https ://www.kickstarter.com/
5. https ://go.forrester.com/
6. https ://en.wikipedia.org/wiki/Wikipedia :Statistics

plus, Wikipedia serait aujourd'hui classé en 9e position des sites les plus consultés depuis la France et générant plus de 15 millions de visiteurs par mois [7]. Les sites de Questions-Réponses (Q&R) comme Stack Exchange [8] ou Quora [9] participent tout autant à la création d'une nouvelle figure du modèle participatif où les utilisateurs deviennent, tour à tour, producteurs et consommateurs d'informations. En 2016, le site Quora recencait plus de 100 millions d'utilisateurs actifs par mois et plus de 13 millions de questions auraient été publiées durant cette seule année. Son analogue Stackoverflow, spécialisé dans les questions techniques liées aux métiers du développement informatique, publierait quant à lui, grace au concours de ses utilisateurs, une nouvelle question toutes les 8 secondes [10].

L'usager est aujourd'hui devenu le premier producteur d'informations, exposant ces plateformes à de nombreux risques essentiellement liés à la qualité, véracité ou objectivité des productions. Si le manque d'expertise d'un internaute influe sur la qualité des informations qu'il publie, c'est plus généralement l'absence de filtre, le manque de légitimation éditoriale (Serres, 2013), qui permet une si grande variance de qualité des informations disponibles. Et même le contrôle des informations par les pairs sur Wikipédia, et donc la règle de la moyenne des points de vue, ne converge pas toujours, du moins dans de nombreux cas d'articles litigieux. De surcroît, devenus de stratégiques vecteurs d'informations, l'usage de ces sites est bien souvent dévoyé, souvent à des fins publicitaires voire dans des objectifs de propagande ou de désinformation. Les actes de vandalisme, notamment sur Wikipédia, dégradent la qualité des articles. Et les articles pris pour cible par les détracteurs ne sont pas toujours identifiés à temps par la communauté. Par conséquent, l'utilisateur non averti peut participer, à son insu, à la propagation de la fausse information. C'est le cas par exemple d'un article publié en 2005, concernant l'assassinat de Kennedy, sur lequel une "fake news", incriminant le journaliste J. Seigenthaler, est apparu sur sa page personelle Wikipedia. L'information erronée, accessible par tous, est restée en ligne pendant cinq mois avant d'être propagée et publiée par d'autres sites internet. La cadence accélérée du nombre de publications sur ces plateformes collaboratives et les réseaux sociaux incite alors à penser des méthodes automatiques pour se prémunir contre ces menaces.

7. www.mediametrie.fr
8. https ://stackexchange.com/
9. https ://www.quora.com/
10. https ://insights.stackoverflow.com/survey/2016

Dans ce manuscrit, nous nous intéressons à la représentation des notions d'autorité, de qualité et d'expertise sur les plateformes collaboratives, et à leur utilisation, en vue d'améliorer le processus d'identification de contenus de qualité ou d'utilisateurs faisant autorité sur ces plateformes. Nous basons notre travail sur une représentation par les graphes des plateformes, et plus particulièrement sur une représentation par les graphes hétérogènes de celles-ci, de façon à combiner de façon générique différentes représentations modélisant à la fois tous les types d'acteurs ou de contenus ainsi que les différentes relations ou intéractions entre eux. Nous évaluons nos propositions sur les quatre tâches d'identification de contenu suivantes :

1. une tâche d'identification des articles de qualité sur Wikipédia ;
2. une tâche d'identification des réponses de qualité sur les sites de Questions-Réponses ;
3. une tâche d'identification des chercheurs faisant autorité dans les librairies digitales ;
4. une tâche de recherche d'experts, c'est à dire d'identification des chercheurs faisant autorité dans un domaine d'expertise particulier.

Toutes les approches proposées dans ce manuscrit se basent sur le postulat d'un renforcement mutuel entre les notions d'autorité et de qualité. Autrement dit, nous supposons que les utilisateurs faisant autorité sur les plateformes sont généralement à l'origine de productions de qualité, et, inversement, les contenus de qualité sont généralement produits par les utilisateurs faisant autorité. Par ailleurs, nous proposons d'injecter, dans ce processus de renforcement mutuel, une connaissance *a priori* qualitative sur l'autorité ou la qualité de certaines des entités représentées. En particulier, nous montrons qu'injecter une connaissance *a priori* sur l'autorité de certaines conférences scientifiques au sein de l'approche structurelle proposée permet d'améliorer l'identification des chercheurs faisant autorité.

2 Contributions de la thèse

Nos contributions sont les suivantes :

1. Nous proposons un modèle générique basé sur une représentation par les graphes hétérogènes des plateformes. Cette première proposition permet notamment la redéfinition et l'exploitation du principe de renforcement mutuel entre un nombre arbitraire de familles d'entités afin d'améliorer l'identification de contenus de qualité et d'utilisateurs faisant autorité sur les plateformes collaboratives et librairies digitales.

La généricité de l'approche est démontrée en utilisant la formalisation introduite sur trois cas d'applications différents. En particulier, nous illustrons notre proposition sur les plateformes Wikipédia, Stack Exchange et Microsoft Academic Search pour résoudre les trois tâches (a) d'identification d'articles de qualité sur Wikipédia, (b) d'identification de réponses de qualité sur les sites de Questions-Réponses et (c) d'identification de chercheurs faisant autorité dans les librairies digitales. La proposition, évaluée sur ces trois jeux de données réels, souligne l'importance d'étendre les représentations de l'état de l'art, notamment en utilisant un graphe d'influence entre les familles d'entités ;

2. Nous étendons notre première contribution et proposons une solution paramétrique permettant l'intégration de connaissances exterieurs présentées de façon qualitative à l'algorithme. Cette proposition permet la correction, pendant la phase de propagation des scores dans le graphe, des paramètres du modèle de façon à respecter au mieux une relation de préférence traduisant un ordre sur la qualité de certaines entités modélisant la connaissance *a priori*. Par ailleurs, nous validons empiriquement notre proposition via l'intégration de connaissances extérieurs à propos de la qualité de certaines conférences scientifiques dans une tâche d'identification de chercheurs faisant autorité sur la librairie Microsoft Academic Search ;

3. Nous proposons deux instances différentes de notre contribution exploitant un ensemble d'indices temporels d'autorité et de qualité sur les deux plateformes Stack Exchange et Wikipédia. Sur cette dernière plateforme, nous proposons d'exploiter l'historique des articles et plus particulièrement le processus des révisions des articles pour définir une fonction de qualité des éléments textuels à partir de l'autorité des différents réviseurs. Nous montrons que la proposition permet d'une part la reformulation de certains travaux de la littérature et, d'autre part, d'étendre les représentations actuelles permettant la considération d'une combinaison linéaire de l'autorité des réviseurs d'un élément textuel. Sur Stack Exchange, nous motivons l'utilisation d'un graphe de réactivité dans notre approche générique pour améliorer la caractérisation de l'autorité des utilisateurs et donc l'identification des réponses de qualité sur les sites de type Questions-Réponses ;

4. Nous proposons une représentation vectorielle unifiant les notions d'autorité et d'expertise pour la caractérisation de l'autorité des utilisateurs dans un domaine d'expertise particulier. En se basant à nouveau sur une représentation par les graphes des librairies digitales faisant en plus intervenir le contenu des productions, nous formulons un

problème d'optimisation pour l'apprentissage des représentations proposées. En particulier, nous proposons de regrouper deux fonctions objectifs, l'une associée à un calcul d'autorité, l'autre à un calcul d'expertise, et réalisons son optimisation via un algorithme de descente de gradient. Par ailleurs, une série d'évaluations quantitative et qualitative montre des améliorations par rapport aux modèles de l'état de l'art sur la tâche de la recherche d'experts réalisée sur le jeu de données Microsoft Academic Search.

Le manuscrit est organisé de la façon suivante. L'état de l'art concernant les modèles d'autorité, de qualité et d'expertise, principalement basés sur une représentation par les graphes des données, est discuté dans le chapitre 2. Ensuite, dans le chapitre 3, nous définissons notre approche générique par les graphes hétérogène ainsi que les deux algorithmes de propagations associés. Le chapitre 4 est consacré à la définition et à l'intégration de signaux temporels dans notre proposition. Puis, dans le chapitre 5, est présenté le modèle d'autorité de domaine ainsi que le modèle d'ordonnancement associé utilisé dans le cadre d'une recherche d'experts. Enfin, les résultats aux expérimentations conduites sur les jeux de données Wikipédia, Stack Exchange et Microsoft Academic Search sont discutés dans le chapitre 6. Le chapitre 7 dresse les conclusions de la thèse et suggère quelques perspectives de recherche.

Chapitre 2

Synthèse des travaux de l'état de l'art

> "La seule autorité possible est fondée sur la compétence."
>
> Michel Serres

Introduction

Les travaux de recherche présentés dans ce manuscrit s'intéressent à la modélisation des notions d'autorité, de qualité et d'expertise sur les plateformes collaboratives et librairies digitales. Nous nous sommes précisément intéressés à l'étude d'une famille particulière d'approche, celle des approches structurelles, exploitant des représentations par les graphes des données comme support de formalisation des modèles. Aussi, l'état de l'art réalisé dans ce chapitre recouvre essentiellement cette famille particulière d'approches en détaillant certains modèles de références caractérisant l'autorité et l'expertise des contributeurs et la qualité de leurs contributions sur les sites dont le contenu est créé par l'utilisateur.

Ce chapitre est divisé en trois sections, les deux premières regroupant les indices et modèles d'autorité et de qualité, la dernière se consacrant aux modèles d'expertise. Si cette distinction est volontairement réalisée notamment pour s'aligner avec les travaux de la littérature, elle motive surtout une volonté de regrouper ces notions au sein d'un même modèle.

Dans un premier temps, dans la Section 1, les différentes familles d'indices de qualité et d'autorité utilisés pour nourrir les modèles de prédictions classiques sont introduites. Dans un second temps, dans la Section 2, est proposé un état de l'art des modèles d'autorité et de

qualité, détaillant plus particulièrement les approches structurelles en liens avec notre proposition. Dans cette section, sera soulevé le manque de généricité de ces approches et leur incapacité à considérer un ensemble de connaissances extérieures représentées sous la forme d'une préférence utilisateur. Enfin, dans la Section 3, est réalisé l'état de l'art des modèles d'expertise appliqués à la tâche de recherche d'experts sur les librairies digitales. On notera l'absence, dans la littérature, de modèle rapprochant les notions d'autorité et d'expertise au sein d'une même représentation.

1 Indices et modèles de qualité et d'autorité

Les indices de qualité issus des plateformes collaboratives étudiées dans ce manuscrit sont regroupés en quatre familles en fonction des caractéristiques des données :

1. *Les indices de surface* sont des statistiques généralement liées à un processus de comptage directement associés aux contenu des productions. Par exemple, le nombre de mots d'un article Wikipédia ou le nombre de votes associés à une réponse sur une plateforme commme Stack Exchange sont des indices de surface utilisés pour mesurer la qualité des contenus ;

2. *Les indices de contenu* sont des statistiques résumant les informations textuelles et sémantiques des productions. Par exemple, l'analyse des parties du discours ou certains indices de lisibilité sont des indices de contenus qui peuvent être utilisés pour caractériser la qualité des productions textuelles ;

3. *Les indices temporels* sont extraits des flux de productions ou des processus de révisions des plateformes à partir d'informations temporelles sur les données produites. Par exemple, l'âge des révisions d'un article sur Wikipédia ou le délai de réponse des utilisateurs sur un forum constituent des indices temporels de qualité et d'autorité ;

4. *Les indices de structures* sont extraits à partir d'une représentation par les graphes des données modélisant les relations entre les acteurs d'une plateforme. Par exemple, la centralité d'un nœud dans certaines représentations par les graphes peut constituer un indice structurel pour quantifier certaines notions de qualité ou d'autorité.

1.1 Indices de surface

Les indices de surface sont extraits des données brutes de la plateforme et nécessitent peu de traitements. Ils sont souvent le résultat

d'un processus de comptage. Sur Wikipédia par exemple, le nombre de caractères, de mots ou de phrases dans un article, une section ou un paragraphe sont des indices de surfaces de qualité, comme le sont le nombre de sections, d'images, de références ou le nombre de vues d'un article (Lih, 2004). La taille d'un article, et précisément le nombre de mots qu'il contient, est un bon indicateur de la qualité des articles sur Wikipédia (Blumenstock, 2008). La Figure 2.1 retrace les deux distributions du nombre de mots pour deux classes d'articles différentes. A gauche, est tracée la distribution associée à des articles choisis aléatoirement dans le corpus. A droite, est tracée la distribution associée à des articles labellisés "Featured Articles" et donc considérés comme de grande qualité. L'indicateur est clairement corrélé avec l'attribution du statut "Featured Articles" (FA) ce qui rend le nombre de mots d'un article un bon indicateur de qualité. En pratique, les performances d'un classifieur binaire basé sur ce seul critère sont très proches, du moins pour des articles écrits en Anglais, des modèles inductifs classiques entrainés sur une dizaine de caractéristiques de surface et de contenu (Blumenstock, 2008).

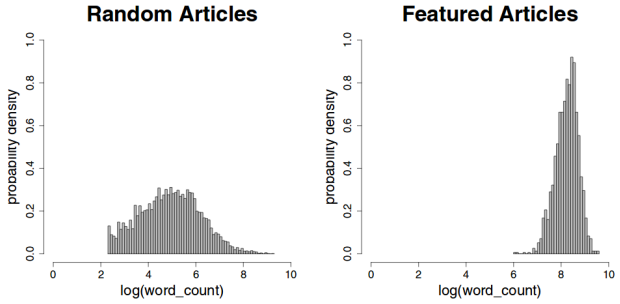

FIGURE 2.1 – Distributions du nombre de mots pour deux classes d'articles Wikipédia extraites du travail de (Blumenstock, 2008).

Toujours sur Wikipédia, le nombre d'éditeurs et le nombre d'éditions d'un article sont aussi deux variables corrélées avec l'attribution du statut "FA" aux articles. Une étude réalisée sur plus de 1,5 millions d'articles Wikipédia suggère que deux articles tirés aléatoirement, l'un à partir d'un ensemble d'article labellisés "Featured Articles" et l'autre à partir d'un ensemble hétérogène "Non Featured Articles", ont statistiquement aucune chance d'avoir le même nombre d'éditeurs ou d'éditions (Wilkinson and Huberman, 2007). Ces résultats sont d'ailleurs similaires lorsque d'autres variables de surfaces sont considérées. Par exemple, le nombre de commentaires générés par les contributeurs sur les "Talk page" est un indice utilisé. Notons que ces approches considèrent un

problème de classification binaire simplifiant la réalité. En effet, au moins 6 catégories d'articles de qualité croissante ont été définie par l'équipe éditoriale du site Wikipédia [1].

Connaissant la qualité de certains articles, l'efficacité d'un contributeur peut être mesurée par le ratio entre le nombre de ses contributions réalisées dans des articles labellisés "Featured Articles" et le nombre global de ses contributions (Betancourt et al., 2016). Formellement :

$$\text{Efficacité}(u) = \frac{\sum_{a \in FA} n(u,a)}{\sum_{a} n(u,a)} \tag{2.1}$$

où $n(u,a)$ mesure, en nombre de mots ou en nombre d'éditions réalisées, la quantité d'information que l'auteur u a apporté à l'article a. Par ailleurs, le fait que les contributions proviennent d'utilisateurs enregistrés ou anonymes peuvent renseigner sur le niveau de qualité des révisions des articles (Twidale, 2005).

Sur les sites de types Questions-Réponses, les badges attribués par certaines de ces plateformes, et plus précisément la variable quantitative associée au nombre de badges que possède un utilisateur, est un bon indicateur d'autorité (Dalip et al., 2013). Puisque le seul moyen pour un membre enregistré de gagner en autorité est de publier des questions, des réponses ou des commentaires, les statistiques relatives à son activité sont, d'une façon générale, de bons indices d'autorité (Shah and Pomerantz, 2010). D'ailleurs, réponses et questions semblent ne pas être apportées par les même membres de la communautés et répondre aux questions permettrait de davantage gagner en autorité (Anderson et al., 2012).

Une majorité des sites de Questions-Réponses permettent à leurs membres de faire valider les réponses publiées. Ainsi, l'acception d'une réponse constitue un très bon indicateur de qualité, créditant dans le même temps l'autorité de son auteur. D'une façon générale, les indices issus de la communauté utilisateurs, comme les votes sur la pertinence des réponses ou les signes de préférences sont largement utilisés comme indices de qualité et vérité terrain notamment pour calibrer les modèles (Agichtein et al., 2008; Wang et al., 2013).

Certains modèles proposent de séparer le calcul d'autorité d'un utilisateur en deux, en se basant, d'une part, sur les questions posées, et,

1. https //en.wikipedia.org/wiki/Template Grading_scheme

d'autre part, sur les réponses apportées (Zhang et al., 2007; Bian et al., 2009; Suryanto et al., 2009). Par exemple, la métrique *zscore*, formulée par l'Équation (2.2), calcule le rapport entre le nombre q de questions posées et le nombre a de réponses apportées par un utilisateur (Zhang et al., 2007).

$$z = \frac{aq}{\sqrt{a+q}} \tag{2.2}$$

On note que l'autorité de l'auteur des questions auxquelles répond l'utilisateur u n'est pas considérée dans le calcul de z. Autrement dit, ce qui pourrait refléter la difficulté de la question à laquelle répond un utilisateur n'est pas considéré.

1.2 Indices de contenu

Les indices de contenu regroupent essentiellement les caractéristiques lexicales ou sémantiques associées aux différentes productions textuelles. Ils sont le résultat de traitements spécifiques du langage naturel afin d'extraire certaines caractéristiques concernant les noms, pronoms ou adjectifs, ou plus généralement, des parties du discours d'un texte. Cette famille regroupe aussi les statistiques résultants de l'application de mesures de lisibilité.

Ainsi, les questions ou réponses publiées sur les sites de Q&R peuvent, par exemple, être jugées de mauvaise qualité par la communauté parce que non conformes à un certain formalisme d'écriture ou à certaines règles typographiques attendues (Shah and Pomerantz, 2010). Certains travaux intègrent donc dans leur modèle des indices comme Coleman-Liau (Smith et al., 1967), Flesch (Flesch, 1948) ou Gunning Fog (Gunning, 1968) pour identifier les réponses mal formulées. Certaines caractéristiques basées sur l'analyse des parties du discours (Agichtein et al., 2008), sur le nombre de mots en capitale ou sur la densité de ponctuation (Shah and Pomerantz, 2010) permettent aussi d'identifier certains contenus de qualité douteuse. Enfin des indices issus de la théorie de l'information (Li et al., 2012) peuvent être aussi utilisés pour caractériser la lisibilité d'un texte. On note qu'en se basant uniquement sur cette famille d'indices issus du contenu, il a été montré que les articles de classe "Featured Articles" du site Wikipédia semblent avoir été écrits dans une langue aussi formelle que ceux publiés par l'Encyclopedia Britannica, la plus ancienne encyclopédie de la langue anglaise (Emigh and Herring, 2005).

1.3 Indices temporels

Le processus de révision des articles sur Wikipédia est beaucoup exploité dans la littérature. Par exemple, le simple nombre de révisions d'un article est un bon indicateur de qualité d'un article (Hasan Dalip et al., 2009). Mais la majorité des travaux s'intéressent à la définition d'une mesure de qualité d'un mot ou d'une portion de texte à partir de son ancienneté dans le système de révisions. Plus une phrase est ancienne, vis-à-vis des autres phrases de l'article, plus sa qualité est créditée. Les modèles supposent donc que si la qualité d'une phrase n'était pas bonne, celle-ci aurait déjà été supprimée par un des contributeurs de l'article. Par conséquent, l'ancienneté d'une phrase ou d'une portion de texte devient un indice de qualité (ou un indice de confiance de qualité) qui croît avec le nombre des éditions. Différentes variantes de cet indicateur sont proposées. Par exemple, le rapport entre la quantité d'information ajoutée à l'article par un utilisateur et la quantité d'information ajoutée par lui mais supprimée ensuite par d'autres permet de caractériser l'autorité d'un contributeur (Adler and de Alfaro, 2007; Javanmardi and Lopes, 2010; Suzuki, 2015). Basé sur le même principe, le nombre d'éditeurs ayant effectué un changement dans l'article n'affectant pas les contributions d'un utilisateur constitue un indice d'autorité (Halfaker et al., 2009). Ces différentes métriques, basées sur le principe de révisions de Wikipédia, ont d'ailleurs fait l'objet de validations expérimentales via la plateforme Amazon Mechanical Turk (Biancani, 2014). Selon les auteurs, environ 10% des contributeurs de Wikipédia produiraient à eux seul plus de 90% de la valeur du site (Priedhorsky et al., 2007).

Les indices temporels pour les sites de Q&R sont moins exploités mais semblent pourtant de bons indicateurs d'autorité ou de qualité. En effet, la plupart des experts sur les plateformes de type Q&R semblent répondre plus rapidement que les autres aux nouvelles questions (Anderson et al., 2012; Movshovitz-Attias et al., 2013). Ainsi, des métriques basées sur les délais d'apparition entre la publication des questions et des réponses caractérisant la réactivité des utilisateurs sont proposées (Li et al., 2012). On note que, du point de vue des utilisateurs en attente d'une réponse, le fait que les réponses arrivent par odre de qualité décroissante constitue une caractéristique intéressante. D'ailleurs, 80% des réponses seraient apportées dans les 24 heures suivant la date de publication d'une question (Yeniterzi and Callan, 2014).

D'autres indices directement associés aux utilisateurs, comme la date

d'inscription à la plateforme ou la date de dernière connexion, sont particulièrement discriminants (Dalip et al., 2013). L'activité horaire, en particulier le nombre de publications à l'heure, est utilisé, notamment pour estimer les périodes durant lesquelles un utilisateur u est le plus susceptible de répondre aux questions (Chang and Pal, 2013) :

$$\frac{\#\text{ Réponses publiées par } u \text{ à l'heure } h}{\#\text{ Réponses publiées par } u} \tag{2.3}$$

L'utilisation de fonctions d'amortissement permet de pondérer les métriques usuelles et, par exemple, de donner davantage de poids aux publications récentes (Yeniterzi and Calla, 2015). La mesure du *zscore* introduite en Section 1.1 est alors reformulée par une agrégation des valeurs z_i sur toutes les périodes de temps considérées :

$$z = \sum_i z_i e^{-\Delta t_i} \tag{2.4}$$

où $\Delta t_i = d(t_q) - d(t_i)$ mesure le délai, par exemple sur un site de Questions-Réponses, entre la date de publication d'une question t_q et celle d'une réponse t_i.

1.4 Indices de structure

Les indices de structure, lorsqu'ils ne sont pas le résultat d'approches structurelles complexes décrites dans la Section 2.2, sont extraits directement de représentations par les graphes des plateformes. Certaines de ces métriques, et en particulier les métriques calculées localement dans le graphe, peuvent exprimer les mêmes informations que les métriques de surface ou de contenu. Par exemple, sur une plateforme de type Questions-Réponses, l'activité d'un utilisateur peut être exprimée directement avec un processus de comptage du nombre de ses publications (indice de surface), ou bien à partir du graphe représentant la plateforme, en comptant le nombre de liens associés à une relation de publication. Ainsi, le degré des nœuds d'un graphe modélisant tous les éléments d'un site de Questions-Réponses est un indice de structure d'autorité pour les utilisateurs (Anderson et al., 2012).

Dans ce contexte, les graphes de citations sont particulièrement utilisés pour mesurer l'importance relative des articles scientifiques dans les librairies digitales (Mimno and McCallum, 2007; Chen et al., 2007) ou des articles Wikipédia (Bellomi and Bonato, 2005). Le degré des nœuds, le degré d'*assortativité* (coefficient de Pearson pour le degré des noeuds connectés), les coefficients de clustering ou de réciprocité sont

alors autant d'indices de structure largement corrélés avec la qualité des articles Wikipédia (Hasan Dalip et al., 2009). Les graphes de collaborations (Newman, 2004) peuvent aussi informer de l'autorité des chercheurs. D'ailleurs, sur les sites de Questions-Réponses comme Quora ou Stack Overflow, certaines études suggèrent (Anderson et al., 2012; Wang et al., 2013) que les liens d'affinités entre les utilisateurs participent à une création de valeur. Les ensembles d'utilisateurs connectés via des liens d'affinités auraient en effet tendance à produire, ensemble, des réponses de meilleure qualité que des groupes d'utilisateurs non connectés. De la même manière, sur Wikipédia, les contributeurs faisant autorité semblent s'organiser pour contribuer ensemble à l'élaboration d'articles de qualité.

D'une façon générale, les résultats obtenus sur des tâches de prédiction de qualité regroupant de nombreuses familles d'indices suggèrent que les indices d'origine structurelle sont ceux apportant le plus d'information aux modèles (Flekova et al., 2014).

2 Modèles de qualité et d'autorité

Une première famille d'approches, dont certaines sont citées en Section 2.1, regroupe les différents modèles d'apprentissage à partir de cas. Les indices d'autorité et de qualité sont alors utilisés pour décrire les producteurs et productions des plateformes et construire des modèles descriptifs ou prédictifs. De nombreux modèles ont été proposés dans la littérature pour mieux identifier les règles et corrélations entre ces variables et seuls les plus représentatifs sont présentés dans ce chapitre. Une seconde famille d'approches regroupe les modèles structurels transductifs (Güttler et al., 2016), c'est à dire les solutions qui reposent sur une représentation par les graphes des données utilisées et exploitent les relations pour propager les scores de qualité et d'autorité. Compte tenu de la proximité de ces travaux avec la problématique générale de cette thèse, cette famille d'approches est plus largement détaillée. Par ailleurs, les approches sont organisées en fonction du type de graphe utilisé pour la construction des modèles. Nous montrons alors que toutes les solutions de l'état de l'art sont développées d'une façon très spécifique à la plateforme utilisée. Par ailleurs, nous constatons qu'aucun modèle semble permettre l'intégration de connaissance *a priori* qualitative sur la qualité ou l'autorité de certaines entités.

2.1 Approches inductives

Les modèles inductifs, largement majoritaires dans notre contexte, regroupent l'ensemble des solutions d'apprentissage automatique à partir de cas. Les indices introduits dans la Section 1 sont alors utilisés pour caractériser les productions et optimiser les paramètres des modèles.

Les modèles à base de mélanges gaussiens, par exemple, sont utilisés pour caractériser la qualité des articles Wikipédia en exploitant un ensemble d'indices de surface (Wilkinson and Huberman, 2007; De la Calzada and Dekhtyar, 2010). La qualité Q_a d'un article a est alors définie par la somme des fonctions de densité associées à chacune des composantes i estimées sur un corpus d'entrainement :

$$Q_a = \sum_i q_{ia}, \text{ avec } q_{ia} = \frac{1}{\sigma_i \sqrt{2\pi}} ^{-\frac{(x_a - \mu_i)^2}{2\sigma_i^2}} \tag{2.5}$$

Les machines à vecteurs de supports sont aussi appliqués à Wikipédia pour discriminer les articles de qualité (Flekova et al., 2014). Les résultats montrent que les gains d'information associés aux indices temporels et de surface sont plus importants que ceux associés aux indices de contenu pour identifier les articles de qualité. Les modèles de régression permettent de considérer plusieurs niveaux de qualité des articles (Hasan Dalip et al., 2009). Sur *Yahoo ! Answers*, les modèles de régression bayesiens (Wang et al., 2009) et logistiques (Shah and Pomerantz, 2010) sont appliqués pour identifier les meilleures réponses aux questions.

Plus récemment, une solution basée sur l'utilisation des réseaux de neurones est proposée pour répondre à cette même tâche de classification (Dang and Ignat, 2016). Les résultats aux expérimentations conduites sur une représentation latente des mots, apprise via le modèle Word2vec (Mikolov et al., 2013), semblent égaler les performances des modèles de prédiction utilisés jusqu'alors entrainés sur une vingtaine de caractéristiques de l'état de l'art. Les modèles de forêts aléatoires sont aussi utilisés pour identifier les experts d'une plateforme de type Questions-Réponses (Movshovitz-Attias et al., 2013; Dalip et al., 2013). Les arbres de décisions sont aussi utilisés (Agichtein et al., 2008), en particulier, pour étudier l'impact de différentes familles d'indices issus à la fois du contenu des productions et des indices de structure de graphes de collaborations.

Si les approches inductives permettent la construction de règles pour prédire la qualité de nouvelles observations étant donné un vecteur de

caractéristiques, elles ne permettent pas la création d'indices. Les solutions présentées dans la section suivante ont l'intérêt de construire des indicateurs de qualité, à partir d'un choix de représentation des plateformes, qui, à leur tour, pourront nourir un vecteur de caractéristiques et améliorer les prédictions.

2.2 Approches transductives

Les modèles transductifs, lorsqu'ils sont développés sur des données de graphe, sont des cas particuliers d'algorithmes de propagation de scores ou d'étiquettes (Güttler et al., 2016). Dans notre contexte, ils reposent sur un principe de renforcement mutuel entre l'autorité des utilisateurs et la qualité des productions associées. Nous regroupons ces approches selon le nombre et le type d'entités et de relations considérées. La Section 2.2.1 introduit les approches classiques développées sur une représentation par les graphes simples des contenus Web. Les Section 2.2.2 et Section 2.2.3 détaillent les solutions basées sur des graphes bipartis et tri-partis respectivement. Enfin, la Section 2.3 discute les limitations des approches présentées.

2.2.1 Graphes simples

Les approches basées sur les graphes simples ne modélisent qu'un seul type de nœuds et, généralement, un seul type de relation. Dans notre contexte, il s'agit généralement d'appliquer de légères variantes des algorithmes PageRank (Page et al., 1999) et HITS (Kleinberg, 1999) historiquement appliqués aux réseaux de pages Web.

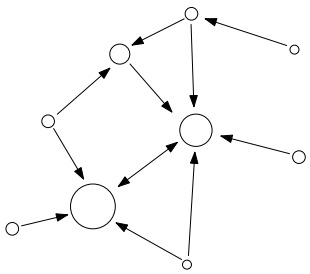

FIGURE 2.2 – La notion d'autorité structurelle de l'algorithme PageRank. La taille d'un sommet du graphe reflète son autorité, fonction de l'autorité des sommets voisins.

L'algorithme PageRank (Page et al., 1999), qui a joué un rôle crucial dans le développement du moteur de recherche de Google, est un

exemple phare d'approche structurelle largement utilisée pour les calculs d'autorité [2]. Sa version originelle permet le calcul d'un score d'autorité d'une page Web en ne tenant compte que de la structure des liens hypertextes. Autrement dit, l'emplacement d'une page Web ou d'un site dans l'organisation générale du Web est un gage d'autorité. Le principe de renforcement mutuel est donc ici exploité, sans distinction faite sur la nature des nœuds. PageRank est donc une méthode plus sophistiquée que le simple degré des nœuds puisque le score d'autorité d'une page Web est fonction des scores d'autorité des pages pointant vers elle. La Figure 2.2 illustre ce fonctionnement. Dans la pratique, l'hypothèse est simple : une page Web citée par un site faisant autorité comme *LeMonde.fr* n'hérite pas du même crédit que si elle est pointée par un blog isolé ou peu connu. Formellement, l'autorité d'un sommet u_i d'un graphe simple est calculé de la façon suivante :

$$Autorite(u_i) = \frac{1}{Z} \sum_{u_j \to u_i} Autorite(u_j) \tag{2.6}$$

où Z est un facteur de normalisation qui assure généralement qu'un nœud u_j distribue à part égale son autorité à ses voisins. L'équation est récursive, mais sous certaines conditions, son calcul pour tous les sommets du graphe converge en un temps fini vers une solution stable. Si A est la matrice d'adjacence du graphe G représentant l'ensemble des pages Web reliées entre elles par des liens hypertextes et $\mathbf{x} \in \mathbb{R}^n$ le vecteur d'autorité des nœuds ($x_i \in \mathbb{R}$ est le score d'autorité du sommet i dans G), alors l'Équation (2.6) peut être reformulée de la façon suivante :

$$x_i = \frac{A_i \cdot \mathbf{x}}{|A_i|_1} \tag{2.7}$$

où A_i est la ligne i de la matrice A. L'algorithme converge vers un vecteur d'autorité $\mathbf{x}$ satisfaisant l'égalité

$$\mathbf{x} = A \cdot \mathbf{x} \tag{2.8}$$

D'un point de vue algébrique, le vecteur $\mathbf{x}$ est donc le premier vecteur propre de l'application engendrée par la matrice A. D'un point de vue probabiliste, $\mathbf{x}$ définit la probabilité stationnaire de la chaine de markov engendrée par A.

De nombreux graphes simples sont exploités dans la littérature pour caractériser la qualité ou l'autorité des entités. Largement exploités dans

2. Il serait plus précis de parler de popularité, l'algorithme privilégiant le quantitatif au qualitatif.

les librairies digitales pour mesurer l'impact des membres, les graphes de collaborations, construits par exemple à partir des productions communes entre les utilisateurs, sont des graphes simples ne faisant intervenir qu'un seul type d'entités (Egghe et al., 2000; Farkas et al., 2002; Liu et al., 2005; Anokhin et al., 2012). Sur les sites de Questions-Réponses, le flux des réponses permet la création d'un graphe de co-occurences mesurant le nombre de fois qu'un utilisateur répond aux questions d'un autre utilisateur (Zhang et al., 2007). Les calculs de l'algorithme PageRank peuvent être reformulés à partir des collaborateurs d'un utilisateur. Formellement, à un facteur de normalisation près, on a :

$$\text{Authorité}(u) = \sum_{v \in \text{Collaborateurs}(u)} \text{Authorité}(v) \tag{2.9}$$

Puisque les productions ne sont pas toujours représentées dans ces modèles, la qualité de celles-ci est souvent assimilée à l'autorité de leur(s) auteur(s).

2.2.2 Graphes bipartis

Le principe de renforcement mutuel entre la qualité des productions et l'autorité des producteurs d'information peut être représenté par un graphe biparti défini, d'une part, sur l'ensemble des productions, et d'autre part, sur l'ensemble des producteurs. De nombreuses variantes ou reformulations des algorithmes PageRank (Page et al., 1999) ou HITS (Kleinberg, 1999) ont alors été proposées pour calculer et propager les indices d'autorité et de qualité dans le graphe (Hu et al., 2007; Suzuki and Yoshikawa, 2013; Li et al., 2015).

Sur Wikipédia, il est fréquent de considérer un graphe biparti qui associe l'ensemble des contributeurs à l'ensemble de leurs articles en cours de développement (Wilkinson and Huberman, 2007; Vydiswaran et al., 2011; Suzuki and Yoshikawa, 2013; Li et al., 2015). L'hypothèse sous-jacente affirme que l'autorité d'un contributeur est fonction de la qualité de ses productions, et inversement, que la qualité d'une production est fonction de l'autorité des différents contributeurs. Ainsi, lorsque les arcs de la relation d'auteur sont pondérés par une agrégation de différents indices de surface, ils deviennent à la fois vecteur et porteur d'information (Li et al., 2015).

Par exemple, le modèle Basic (Hu et al., 2007) propose de pondérer la relation d'auteur avec le nombre de mots qu'un utilisateur a écrit dans un article. Le graphe biparti entre les contributeurs et les articles est illustré dans la Figure 2.3. En distinguant l'autorité A_j d'un contributeur j de la qualité Q_i d'un article i, le principe de renforcement

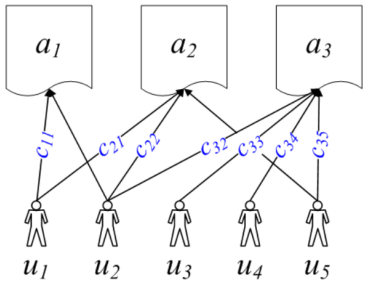

FIGURE 2.3 – Graphe biparti sur Wikipédia.

mutuel entre A et Q est alors formulé de la façon suivante :

$$\begin{aligned} Q_i &= \sum_j c_{ij} A_j \\ A_j &= \sum_i c_{ij} Q_i \end{aligned} \tag{2.10}$$

Autrement dit, la qualité d'un article i est calculée à partir de l'autorité de ses contributeurs, pondérée par la quantité d'information c_{ij} apportée à l'article. Inversement, l'autorité d'un contributeur j est calculée à partir de la qualité de l'ensemble des articles auxquels il a participé. La résolution du système formulé par l'Équation (2.10) se fait itérativement en calculant alternativement, jusqu'à stabilité des scores, les quantités Q_i et A_j pour chacun des articles et des utilisateurs.

Ce modèle est étendu afin d'y intégrer des indices de qualité liés au processus de révision des articles. Plus spécifiquement, le modèle PEER REVIEW (Hu et al., 2007) crédite davantage les mots les plus anciens avec l'autorité des réviseurs, c'est à dire avec l'autorité des utilisateurs ayant contribué au moins une fois à l'article en laissant sur place ces mots ci. Le modèle, défini par l'Équation (2.11), permet alors d'exploiter le principe de renforcement mutuel aini que cet indice temporel de qualité issu du processus de révision des articles, propageant les scores de qualité dans le graphe de proche en proche.

$$\begin{aligned} q_{ik} &= \sum_{j \to w_{ik}} A_j \\ A_j &= \sum_{w_{ik} \leftarrow j} q_{ik} \end{aligned} \tag{2.11}$$

Formellement, la qualité q_{ik} d'un mot w_{ik} apparaissant dans l'article i est définie à partir de l'autorité A_j des réviseurs $j \to w_{ik}$ du mot, c'est

à dire l'ensemble des utilisateurs ayant participé au moins une fois à l'élaboration de l'article i *après* l'ajout du mot w_{ik}. Inversement, l'autorité A_j d'un contributeur j est définie à partir de la qualité q_{ik} de l'ensemble $w_{ik} \leftarrow j$ des mots révisés, c'est à dire l'ensemble des mots qui ont été ajoutés à l'article avant la plus récente de ses contributions. Notons la proximité de la formulation de la quantité q_{ik} avec les indices temporels introduits en Section 1.3. En effet, un mot w_{ik} est ancien si le nombre de révisions suivant w_{ik} est grand, impliquant donc potentiellement davantage de réviseurs. La qualité Q_i d'un article est alors exprimée à partir de la qualité de l'ensemble des mots qu'il contient, $Q_i = \sum_k q_{ik}$.

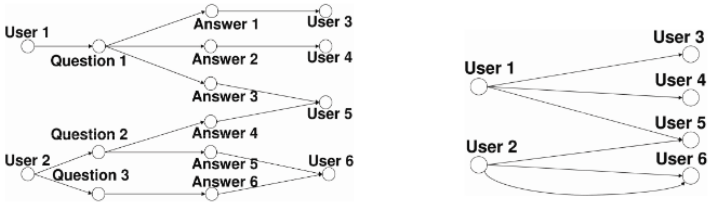

FIGURE 2.4 – Construction d'un graphe biparti. A gauche, une représentation immédiate des données issues de la plateforme Yahoo ! Answers. A droite, une nouvelle représentation construite à partir de la représentation immédiate. L'auteur d'une réponse est associée à l'auteur de la question.

L'usage des graphes bipartis permet aussi la caractérisation de l'autorité des utilisateurs sur les sites de Questions-Réponses. Par exemple, un graphe biparti, illustré en Figure 2.4, peut être construit en considérant d'une part les auteurs des questions et d'autre part les auteurs des réponses. Un sommet du premier ensemble est relié à un sommet du second ensemble si ce dernier est l'auteur d'une réponse associée à au moins une question du premier (Jurczyk and Agichtein, 2007). L'application de l'algorithme HITS sur cette représentation permet de formuler le modèle suivant, à un facteur de normalisation près :

$$H(u) = \sum_{u \sim v} A(v) \text{ et } A(v) = \sum_{v \sim u} H(u) \tag{2.12}$$

où H et A sont les scores d'autorité des utilisateurs associés aux deux ensembles représentés et $u \sim v$ est une notation simplifiée indiquant la présence d'un arc entre deux nœuds u et v. On note ici l'absence de pondération.

Toujours sur les sites de Questions-Réponses, le modèle EXHITS (Suryanto et al., 2009), basé aussi sur un graphe biparti, décompose l'autorité d'un utilisateur en deux. Un premier calcul d'autorité y_u^q caractérise l'autorité de l'utilisateur u à partir des questions q posées. Un second calcul d'autorité y_u^a caractérise l'autorité d'un utilisateur u à partir des réponses a apportées. Le score d'autorité global d'un utilisateur est alors défini par une combinaison linéaire des deux critères introduits :

$$x_a = \alpha y_u^q + (1 - \alpha) y_u^a \tag{2.13}$$

Les deux scores d'autorité partiels sont eux calculés en exploitant le principe de renforcement mutuel entre l'autorité des auteurs des questions et l'autorité des auteurs des réponses, formellement :

$$\begin{aligned} y_u^a &= \sum_{u \sim u'} y_{u'}^q \\ y_u^q &= \sum_{u' \sim u} y_{u'}^a \end{aligned} \tag{2.14}$$

Un autre type de graphe biparti associant l'auteur des questions aux différentes questions posées est utilisé pour capturer le principe de renforcement mutuel entre la qualité Q des questions et l'autorité U des utilisateurs sur la plateforme Yahoo ! Answers (Li et al., 2012). En outre, les auteurs postulent que les questions ayant des caractéristiques similaires devraient avoir des scores de qualité similaires, tout comme les utilisateurs avec des profils similaires devraient avoir des scores de qualité similaires. Cette hypothèse ou contrainte posée sur les scores de nœuds voisins similaires repose sur le principe sociologique d'homophilie, largement étudié dans la littérature, qui postule que les acteurs similaires d'un réseau tendent naturellement à se rassembler structurellement (Kossinets and Watts, 2009; Boucher, 2012; Currarini et al., 2016). L'algorithme de renforcement mutuel proposé, MRLP, intègre alors deux matrices de similarité M et N associées aux utilisateurs et aux questions respectivement de façon à favoriser, pendant la phase de propagation des scores, le transfert de qualité entre entités similaires (via le poids des matrices de similarité). En particulier, si $\mathbf{x}_i$ et $\mathbf{x}_j$ sont deux vecteurs de caractéristiques associées à deux questions q_i et q_j, alors ces matrices de similarité sont calculées à partir de la formule suivante :

$$sim(q_i, q_j) = \exp(-\frac{||\mathbf{x}_i - \mathbf{x}_j||^2}{\lambda_q^2}) \tag{2.15}$$

où λ_q est un paramètre utilisateur. Une combinaison linéaire entre les scores issus des matrices de similarité et les scores issus de la matrice

d'adjacence du graphe est utilisée pour pondérer les deux sources de scores. Formellement, si E est la matrice d'adjacence du graphe biparti, la proposition est modélisée par :

$$\begin{aligned} U &= \alpha \cdot M \cdot U + (1-\alpha) \cdot E \cdot Q \\ Q &= \beta \cdot N \cdot Q + (1-\beta) \cdot E^T \cdot U \end{aligned} \tag{2.16}$$

Il est aussi naturel de représenter les librairies digitales par des graphes bipartis, en associant l'ensemble des chercheurs à l'ensemble de leurs articles publiés. Dans ce contexte, biaiser l'algorithme PageRank par un vecteur exprimant la pertinence thématique d'une requête q pour les documents d modélisés permet l'identification de nœuds faisant autorité dans un domaine d'expertise particulier (Serdyukov et al., 2008). Ainsi, le principe de renforcement mutuel est exploité tout en privilégiant, pendant la propagation des scores, les documents et utilisateurs similaires à la requête. Formellement, le modèle est le suivant :

$$\begin{aligned} P(d) &= \alpha P(d|q) + (1-\alpha) \sum_{e \sim d} P(d|e) P(e) \\ P(e) &= \alpha P(e|q) + (1-\alpha) \sum_{d \sim e} P(e|d) P(d) \end{aligned} \tag{2.17}$$

où α pondère l'importance du critère de pertinence thématique et $P(d|q)$ (resp. $P(e|q)$), mesure la similarité entre d et q (resp. entre e et q). Une extension du modèle, formalisé par l'Équation 2.18, permet la prise en compte des relations inter-documents et inter-utilisateurs.

$$\begin{aligned} P(d) &= \alpha P(d|q) + (1-\alpha)(\beta \sum_{e \sim d} P(d|e) P(e) (1-\beta) \sum_{d' \sim d} P(d|d') P(d')) \\ P(e) &= \alpha P(e|q) + (1-\alpha)(\beta \sum_{d \sim e} P(e|d) P(d) (1-\beta) \sum_{e' \sim e} P(e|e') P(e')) \end{aligned} \tag{2.18}$$

Dans les deux cas, l'algorithme calcule alternativement les quantités $P(d)$ et $P(e)$ jusqu'à convergence. Les utilisateurs sont alors ordonnés par score décroissant donnés par $P(e)$.

2.2.3 Graphes tri-partis

A notre conaissance, aucune modélisation par les graphes tri-partis n'a été proposée pour la représentation des plateformes de type Wikis. En revanche, cette modélisation se prête bien aux données des sites de Questions-Réponses et certains modèles étendent le principe de renforcement mutuel pour ce type de représentation. En effet, si les utilisateurs faisant autorité posent de bonnes questions ou répondent juste

aux questions posées, il semblerait que les bonnes questions attirent de bonnes réponses et, qu'inversement, les mauvaises questions attirent les mauvaises réponses (Agichtein et al., 2008). Ainsi, certains travaux proposent d'utiliser des graphes hétérogènes contenant exactement trois ensembles d'entités (les utilisateurs, les questions et les réponses du site) ainsi que les relations d'auteurs et de liaisons entres ces éléments (Bian et al., 2009; Zhang et al., 2014).

Une solution séparant les calculs d'autorité en fonction du type de publication réalisée est proposée (Bian et al., 2009). Plus précisemment, l'autorité y_u^q d'un utilisateur u calculée à partir de la qualité y_q des questions q posées est différenciée de l'autorité y_u^a d'un utilisateur u calculée à partir de la qualité y_a des réponses a données. Ainsi, si M_{UA} est la matrice d'adjacence associée à la relation d'auteur des réponses et M_{UQ} la matrice d'adjacence associée à la relation d'auteur des questions, le calcul des scores d'autorité partielle des utilisateurs est le suivant :

$$\begin{aligned} y_u^q &= M_{UQ} y_q \\ y_u^a &= M_{UA} y_a \end{aligned} \tag{2.19}$$

Les scores de qualité des réponses et des questions sont calculés selon deux combinaisons linéaires. En particulier, la qualité y_a d'une réponse est définie à partir de l'autorité de l'auteur de la réponse et de l'autorité de l'auteur de la question. De la même manière, la qualité d'une question y_q est défine à partir de l'autorité de l'auteur de la question et de la qualité des réponses associées. Formellement :

$$\begin{aligned} y_a &= \alpha M_{UA}^T y_u^a + (1-\alpha) M_{QA}^T y_q \\ y_q &= \gamma M_{UQ}^T y_u^q + (1-\gamma) M_{QA}^T y_a \end{aligned} \tag{2.20}$$

La résolution du système étant réalisée en calculant alternativement et jusqu'à convergence, les quantités exprimées par l'Équation (2.19) et l'Équation (2.20).

Récemment, le modèle NCR (Zhang et al., 2014) s'est basé sur une représentation similaire, illustrée en Figure 2.5. Le principe de renforcement mutuel est formulé par un système d'équations reliant trois variables de popularité, de qualité et d'autorité associées aux types d'entités différentes que sont les questions, les réponses et les utilisateurs. Les auteurs postulent que :

- la popularité P_q d'une question q est fonction de :
- l'autorité A_u de l'auteur u de la question q ;
- la qualité Q_a de toutes les réponses associées à q.

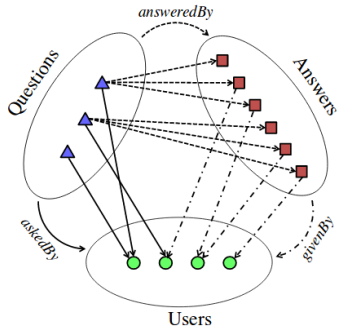

FIGURE 2.5 – Graphe hétérogène du modèle NCR (Zhang et al., 2014).

- la qualité Q_a d'une réponse a est fonction de :
- l'autorité A_u de l'auteur u de la réponse ;
- la popularité P_q de la question associée.
- l'autorité A_u d'un utilisateur u est fonction de :
- la qualité Q_a de l'ensemble des réponses publiées par u ;
- la popularité P_q de l'ensemble des questions publiées par u.

Formellement, le modèle est exprimé par le système de l'Équation (2.21).

$$\forall u, p, q \quad \begin{cases} A_u = \sum_{a \sim u} Q_a + \sum_{q \sim u} P_q \\ P_q = A_u + \sum_{a \sim q} Q_a \\ Q_a = A_u + P_q \end{cases} \tag{2.21}$$

2.3 Discussion

Si la majorité des propositions présentées dans la section précédente conservent la sémantique des nœuds, celles-ci sont en revanche confrontées aux trois limitations suivantes :

1. *Manque d'évolutivité.* La plupart des propositions structurelles modélisent les plateformes digitales par des graphes ne représentant jamais plus de trois types d'entités différents. Or, pour ne citer que l'exemple des sites de Questions-Réponses, quatre ensembles d'entités devraient être représentés (utilisateurs, questions, réponses *et commentaires*). Notons qu'une prise en compte de nouvelles entités par les approches existantes nécessiterait une refonte des modèles ;

2. *Manque de généricité.* En outre, ces solutions sont formulées pour des plateformes particulières. Il paraît difficile d'appliquer directement une

des approche de l'état de l'art développée pour Wikipédia sur une plateforme de type Questions-Réponses. Pourtant, certaines caractéristiques communes de qualité ou d'autorité peuvent être exploitées de la même manière en utilisant une approche générique par les graphes hétérogène. Une telle approche, basée sur un nombre arbitraire de familles d'entités, permettrait d'ailleurs de pallier au problème lié à l'évolution des plateformes ;

3. *Absence de connaissance qualitative.* Finalement, on constate que les approches structurelles ne permettent pas l'intégration de connaissance *a priori* qualitative sur la qualité des entités. Si ce genre d'intégration a déjà été formalisée via la prise en compte d'un vecteur de biais *réel* et *fixé* avant la propagatation (pour privilégier les nœuds dont le contenu textuel est en lien avec une requête utilisateur (Serdyukov et al., 2008)), elle n'a, à notre connaissance, jamais été formalisée de façon à permettre l'apprentissage de ces biais, de façon à ce que les scores de qualité et d'autorité calculés par l'algorithme respectent aux mieux une préférence utilisateur. Par exemple, il paraît intéressant d'injecter dans un algorithme de propagation de scores le fait qu'une conférence scientifique soit réputée de meilleure qualité qu'une autre, sans pour autant mesurer, de façon quantitative donc, cette connaissance. Un encodage ordinal, basé uniquement sur l'ordre induit par la connaissance, nous pourait plus judicieux.

Les différentes approches structurelles détaillées dans cette section sont regroupées dans le Tableau 2.1. Notons que le travail de (Li et al., 2015) est générique dans le sens où l'algorithme PageRank est appliqué sur un graphe simple uniformisé. Si les auteurs prennent en compte un ensemble de connaissances *a priori*, celui-ci est représenté sous la forme d'un vecteur quantitatif fixé avant l'exécution de l'algorithme destiné à favoriser la propagation des scores depuis certains noeuds particuliers.

3 Modèles d'expertise

Les modèles d'expertise étendent les modèles classiques de recherche d'information à l'identification de personnes associées à un ensemble de productions généralement textuelles. Les modèles proposés ont alors pour objectif d'identifier, dans un corpus exhibant des informations sur au moins deux types d'entités (les producteurs et leurs productions), les individus ayant une certaine autorité dans un domaine d'expertise particulier. Les travaux de la littérature s'intéressent aux deux problématiques suivantes, souvent attaquées de façon conjointe :

	Générique	A priori
(Campbell et al., 2003)	✓	
(Hu et al., 2007)		
(Jurczyk and Agichtein, 2007)	✓	
(Agichtein et al., 2008)		
(Bian et al., 2009)	✓	
(Wang et al., 2009)		
(Javanmardi and Lopes, 2010)		
(Suryanto et al., 2009)		
(Gollapalli et al., 2011)	✓	
(Li et al., 2012)		
(Movshovitz-Attias et al., 2013)	✓	
(Yeniterzi and Calla, 2014)	✓	
(Zhang et al., 2014)		
(Li et al., 2015)	✓	✓

TABLE 2.1 – Classification des approches structurelles de la littérature appliquées à une tâche d'identification de contenus de qualité sur les plateformes collaboratives et librairies digitales.

- *Modélisation* du niveau d'expertise d'un individu dans un domaine d'expertise particulier ;
- *Recherche* des individus dans un domaine d'expertise particulier formulé par une requête textuelle ;

Généralement, les phases principales de conception d'un moteur de recherche d'experts sont les suivantes :

1. *Aquisition des données.* C'est la première étape associée au recueil des données regroupant les preuves d'expertise (Craswell et al., 2001). Par exemple, les articles de Wikipédia, les publications sur les sites de Questions-Réponses ou les articles scientifiques sur les librairies digitales sont les éléments textuels contenant les preuves d'expertise des producteurs d'informations ;

2. *Prétraitement et indexation.* Il s'agit notamment d'une étape d'extraction de caractéristiques pour construire une représentation des profils des utilisateurs. Les méthodes de réduction de dimensions que sont les modèles LDA (Blei et al., 2003), LSI (Deerwester et al., 1990) ou

SVD (Lathauwer et al., 2000) sont souvent utilisées pour extraire les thématiques des documents associés aux utilisateurs ;

3. *Modélisation et recherche.* Modélisation de l'association entre les candidats et une requête utilisateur pour induire un ordre sur les candidats en fonction de la force de l'association et des niveaux d'expertise des individus (Balog et al., 2006; Tang et al., 2011) ;

4. *Visualisation.* Présentation des résultats sous la forme de graphiques, indices ou mots clefs.

Les modèles de recherche d'experts, essentiellement basés sur les modèles de langue (Serdyukov et al., 2008; Balog et al., 2012), peuvent être regroupés en deux groupes. D'une part, les modèles centrés utilisateurs, qui modélisent d'abord l'expertise des utilisateurs à partir de toutes les preuves d'expertise associées avant d'appliquer différents modèles de recherche d'information classique au sein de la base de profils créée. D'autre part, les modèles centrés documents, qui se proposent d'identifier d'abord les preuves d'expertise pertinentes à la requête avant de retrouver, sur la base des seuls utilisateurs associés aux preuves d'expertises pertinentes, les experts les plus probables. Dans les deux cas, étant donnée une requête q, les candidats, notés e, sont triés selon la probabilité $P(e|q)$ traduisant la similarité ou le rapprochement entre la requête et l'expert. Le théorème de Bayes est largement utilisé pour estimer cette probabilité étant donné l'ensemble des profils d'expertise (Balog et al., 2006). En particulier, en supposant $P(q)$ fixe :

$$\begin{aligned} P(e|q) &= \frac{P(q|e)P(e)}{P(q)} \\ &\simeq P(q|e)P(e) \end{aligned} \tag{2.22}$$

Ainsi, sans connaissance *a priori* sur e, les modèles de l'état de l'art s'intéressent à l'estimation de la quantité $P(q|e)$.

Ces deux grandes familles d'approches de la litérature proposant une modélisation la quantité $P(q|e)$ sont présentées dans les Section 3.1 et Section 3.2 suivantes.

3.1 Modèles centrés utilisateurs

Les modèles centrés utilisateurs construisent une représentation vectorielle ou probabiliste des utilisateurs à partir de leurs productions et autres preuves d'expertise. La représentation, indépendante de la requête et souvent assimilée au profil d'expertise, est alors utilisée pour caractériser les compétences d'un utilisateur dans les thématiques traitées

dans les documents. Ainsi, étant donnée une requête, les candidats, associés à ces profils d'expertises, sont ordonnés selon les scores résultant d'un processus de réconciliation entre deux représentations (profil et requête).

Généralement, le profil d'un candidat est représenté par une distribution de probabilités multinomiales sur un vocabulaire de termes. La représentation de ce profil, notée θ_e, est alors calculée pour chaque candidat, à partir des documents associés à e, induisant une probabilité $P(t|\theta_e)$ pour chaque terme t. Ainsi, en supposant une indépendance des termes de la requête, les candidats peuvent être ordonnés selon la probabilité $P(q|\theta_e)$:

$$P(q|\theta_e) = \prod_{t \in q} P(t|\theta_e)^{n(t,q)} \tag{2.23}$$

où $n(t, q)$ est le nombre d'occurences du terme t dans la requête q. Le calcul de la probabilité $P(t|\theta_e)$ est réalisé à partir de l'ensemble des documents associés aux utilisateurs. Plus spécifiquement, si $P(t|d)$ est la probabilité d'un terme t dans un document d, et $P(d|e)$ la probabilité que le document d soit associé au canditat e, alors le modèle classique centré utilisateur est formellement exprimé par l'Équation suivante :

$$P(q|\theta_e) = \prod_{t \in q} \left(\sum_d P(t|d)P(d|e) \right)^{n(t,q)} \tag{2.24}$$

En pratique, la fonction de lissage suivante est souvent appliquée pour rendre le calcul de $P(t|\theta_e)$ plus robuste :

$$P(t|\theta_e) = (1 - \lambda)P(t|\theta_e) + \lambda P(t) \tag{2.25}$$

où $P(t)$ est la probabilité du terme t dans le corpus et λ le paramètre de lissage. La formulation générale du modèle devient alors :

$$P(e|\theta_q) = \prod_{t \in q} \left((1 - \lambda) \sum_d P(t|d)P(d|e) + \lambda P(t) \right)^{n(t,q)} \tag{2.26}$$

On note que le poids $P(d|e)$ de la relation entre un utilisateur et un document pondère chaque terme de la requête appartenant au document.

3.2 Modèles centrés documents

Les modèles centrés documents, dépendant de la requête, appliquent dans un premier temps un processus de réconciliation classique entre

la requête et l'ensemble des productions textuelles associées aux utilisateurs avant d'identifier, parmi les utilisateurs associés à un sous-ensemble de documents pertinents, les candidats les plus susceptibles de répondre au besoin en information formulé par la requête. Autrement dit, les candidats ne sont pas directement modélisés ou représentés hors ligne mais utilisés pour faire le lien avec les documents pertinents à la requête. Formellement, on suppose :

$$P(q|e) = \sum_{d} P(q|d,e)P(d|e) \tag{2.27}$$

Dans cette nouvelle expression de $P(q|e)$, $P(q|d,e)$ mesure combien le document d est une preuve de l'autorité du candidat e sur q, et $P(d|e)$ mesure l'importance de la relation entre le document d et le candidat e. Encore une fois, en supposant une indépendance des termes de la requête, le calcul de la probabilité d'une requête sachant un candidat et un document est formulé de la façon suivante :

$$P(q|d,e) = \prod_{t \in q} P(t|d,e)^{n(t,q)} \tag{2.28}$$

En supposant $P(t|d,e) = P(t|\theta_d)$, on obtient finalement :

$$P(q|e) = \sum_{d} \left(\prod_{t \in q} P(t|\theta_d)^{n(t,q)} \right) P(d|e) \tag{2.29}$$

On note que le poids de la relation entre un utilisateur et un document, formalisée ici par la probabilité $P(d|e)$, intervient après que le rapprochement entre le document et la requête, réalisé ici au niveau des termes par le calcul $\prod_{t \in q} P(t|\theta_d)^{n(t,q)}$, ait été effectué.

3.3 Discussion

Les propositions de la littérature sont essentiellement consacrées à l'application de modèles de recherche d'information classiques. Or, ces modèles ne considèrent pas d'autres critères de pertinence que celui lié à la thématique abordée dans les documents liés aux producteurs d'informations. Il paraît crucial de prendre en compte, en plus du niveau d'expertise caractérisé par le contenu des productions textuelles associées à un utilisateur, un niveau de qualité attaché à ses productions. Si certains travaux (Deng et al., 2008) intégrent un *a priori fixe* sur la qualité des documents via la probabilité $P(e)$ de l'Équation (2.22), deux problèmes différents sont résolus. A notre connaissance, il semblerait qu'aucun modèle de recherche d'experts ne permette de réaliser,

conjointement, les deux calcluls liés à l'extraction des domaines d'expertises d'un corpus de documents associés aux utilisateurs et ceux liés à l'extraction des indices structurels de qualité associés aux documents et/ou aux utilisateurs. Une telle proposition sera alors présentée dans le Chapitre 5.

4 Conclusion

Dans ce chapitre, nous avons étudié les principales approches permettant la caractérisation des notions d'autorité, de qualité et d'expertise dans les plateformes collaboratives et librairies digitales.

L'état de l'art concernant les approches structurelles utilisées pour modéliser la qualité et l'autorité des éléments au sein des plateformes collaboratives montre que les formulations manquent globalement d'uniformité. Pourtant très similaires, aussi bien au niveau des représentations qu'au niveau des techniques de propagation, ces approches gagneraient à être présentées sous une forme commune. Une formulation générique permettrait une homogénéisation des différents modèles et une rapide adaptation face à l'apparition de nouveaux éléments. Par ailleurs, nous avons constaté qu'aucune approche n'intègre de connaissance *qualitative a priori* concernant la qualité des éléments. Pourtant, il nous paraît intéressant de modéliser la connaissance extérieure que l'on peut avoir sur la qualité ou l'autorité des acteurs modélisés le long d'une échelle ordinale.

L'état de l'art réalisé sur les modèles d'identification d'experts montre une lacune immédiate. En effet, le seul contenu textuel des documents produits par les utilisateurs est utilisé pour caractériser l'expertise. La qualité des productions n'est donc pas prise en compte, limitant automatiquement l'expressivité des modèles. L'ordre induit par les moteurs de recherche d'experts sur les candidats peut certainement être amélioré en considérant certains indices de qualité, au niveau des documents associés à un utilisateurs ou au niveau de l'utilisateur lui-même, pour privilégier les auteurs de contenus de qualité.

Dans la suite, nous nous intéressons à ces différentes limitations soulevées par l'étude de l'existant. Nous proposerons une solution pour faire face à la généricité des approches structurelles liées aux calculs d'autorité et de qualité ainsi qu'à l'intégration de connaissance *a priori* qualitative. De plus, nous proposerons une modélisation de l'expertise considérant à la fois le contenu des productions et un ensemble d'indices de qualité des

productions lié à la structure de la représentation de la plateforme. Les résultats aux expérimentations de ces différentes propositions, discutés dans le Chapitre 6, montrent d'ailleurs l'intérêt de nos deux principales contributions et ouvrent de nombreuses perspectives de recherche.

Chapitre 3

Modèle générique par les graphes hétérogènes

Introduction

Afin de répondre aux limitations des solutions présentées dans le Chapitre 2, et notamment à celle de généricité, nous proposons dans ce chapitre une approche basée sur une représentation des plateformes par les graphes hétérogènes. La proposition, généralisant le principe de renforcement mutuel aux graphes contenant un nombre arbitraire de familles d'entités et de relations, permet de généraliser et d'étendre certaines approches existantes de l'état de l'art. Le modèle proposé, basé sur les notions de fonctions et de graphe d'influence, permet de préciser avec davantage de flexibilité que les méthodes actuelles les différentes influences entre les familles d'entités. L'intérêt de préciser non pas les relations entre les familles d'entités mais bien les influences entre celles-ci sera d'ailleurs démontré.

Par ailleurs, une extension de la proposition, caractérisée par la paramétrisation des fonctions d'influence est proposée. Plus spécifiquement, nous proposons dans ce chapitre d'intégrer dans l'algorithme de propagation de scores une phase de correction des paramètres d'influence, guidée par le respect d'une relation de préférence sur la qualité des entités. Ainsi, nous privilégions la propagation des scores de qualité dans le graphe depuis les sources faisant autorité sans avoir à connaître ou spécifier numériquement ce signal d'autorité. L'apprentissage automatique de ces biais d'influence, réalisé de façon conjointe avec la phase de propagation des scores, permet alors de créditer la qualité de certaines entités et, de proche en proche, la qualité des nœuds voisins.

Les contributions de ce chapitre sont les suivantes :

1. *Approche structurelle générique.* L'approche étend et redéfinit certains travaux de la littérature. Elle permet d'être utilisée sur différentes plateformes ;

2. *Définition d'un graphe d'influence.* La modélisation des plateformes par un ensemble de fonctions d'influence permet une specification précise de la façon dont les scores se propagent. En outre, elle permet d'ajuster ou, au contraire, de ne pas considérer certaines relations indésirables ;

3. *Optimisation des paramètres d'influence.* La correction des paramètres d'influence en fonction d'une relation d'ordre sur la qualité de certaines entités permet de propager une connaissance extérieure qualitative dans le graphe.

Le chapitre est organisé de la façon suivante. La Section 1, après avoir introduit les notations, présente et illustre notre modèle structurel. En s'appuyant sur trois plateformes collaboratives différentes, la Section 2 illustre le caractère générique de l'approche. Un premier algorithme de calcul des scores de qualité est décrit en Section 3. Ensuite, une extension de l'approche motivée par la prise en compte de connaissances extérieures est détaillée en Section 4. La Section 5 discute et motive l'intérêt de la proposition, et en particulier, celui de l'exploitation d'un graphe d'influence. Finalement, la Section 6 dresse les conclusions du chapitre.

1 Modélisation

Cette première section est consacrée à la formalisation de l'approche structurelle par les graphes hétérogènes. La Section 1.1 introduit d'abord les notations utilisées pour formuler le modèle. Les Section 1.2 et Section 1.3 présentent les notions de fonctions d'influence et de graphes d'influence. La Section 1.4 exploite les concepts introduits pour proposer une version générique du principe de renforcement mutuel.

1.1 Notations

Un graphe hétérogène dirigé est utilisé pour représenter une plateforme collaborative ou une librairie digitale. Ainsi, à chaque nœud et arc du graphe est associé un type, une catégorie ou une sémantique. Nous notons $\mathcal{T}_V = 1, 2, ..., T_V$ l'ensemble des T_V types de nœuds de la plateforme et $\mathcal{T}_E = 1, 2, ..., T_E$ l'ensemble des T_E types de relations. Le graphe hétérogène $G = (H, E)$ modélisant les données est défini sur un ensemble $H = \{V_k\}_{1 \leq k \leq T_V}$ de T_V familles de nœuds et sur un ensemble

$E \subseteq H \times H$ de T_E relations binaires entre ces familles. Notons que les relations considérées étant binaires, on a $T_E \leq (T_V - 1)T_V$. On note n_k le nombre de nœuds ou entités de la famille V_k, i.e., $n_k = |V_k|$. Le nœud $j \leq n_k$ de type $k \in \mathcal{T}_V$ appartenant à la famille V_k est noté v_k^j. Le nombre total de nœuds dans G est noté N et vérifie $\sum_{k \leq T_V} n_k = N$. Soit $(V_i, V_j) \in E$ une paire de familles d'entités de G. On note E_{ij} la relation binaire définit sur $V_i \times V_j$, i.e., $E_{ij} \subseteq V_i \times V_j$ et $A_{ij} \in M_{n_i n_j}(\mathbb{R})$ la matrice d'adjacence associée. En particulier, $(v_i^a, v_j^b) \notin E_{ij} \Leftrightarrow A_{ij}(v_i^a, v_j^b) = 0$. Finalement, on note $\mathbf{x}_k \in \mathbb{R}^{n_k}$ le vecteur de scores de qualité associé à la famille V_k. En particulier, $x_k^j \in \mathbb{R}$ est le score de qualité associé au nœud $v_k^j \in V_k$. Un résumé des notations utilisées dans le manuscrit est donné dans le tableau 3.1. Par ailleurs, l'Exemple 1.1 illustre l'ensemble de ces notations à partir du graphe présenté en Figure 3.1.

Notation	Définition
$\mathcal{T}_V$	L'ensemble des types de nœuds
$\mathcal{T}_E$	L'ensemble des types de relations
T_V	Le nombre distinct de types de nœuds
T_E	Le nombre distinct de types de relations
G	Le graphe hétérogène
H	L'ensemble des familles de nœuds de G
E	L'ensemble des familles de relations de G
N	Le nombre distinct de nœuds dans G
V_k	L'ensemble de nœuds de G de type $k \in \mathcal{T}_V$
n_k	Le nombre de nœud de la famille V_k
v_k^j	Le nœud j de type $k \in \mathcal{T}_V$ de G
$E_{ij} \subseteq V_i \times V_j$	La relation binaire entre deux familles V_i et V_j
$A_{ij} \in M_{n_i n_j}(\mathbb{R})$	La matrice d'adjacence associée à E_{ij}
$\theta_{ij} \in M_{n_i n_j}(\mathbb{R})$	La matrice de paramètres associée à $f_{\theta_{ij}}$
$\mathbf{x}_k \in \mathbb{R}^{n_k}$	Le vecteur de scores de la famille V_k
$x_k^j \in \mathbb{R}$	Le score du nœud v_k^j de V_k

TABLE 3.1 – Résumé des notations utilisées dans ce manuscrit.

Exemple 1.1. Considérons un site de type Q&R faisant intervenir $T_V = 3$ types de contenus dont 3 utilisateurs, 3 questions et 2 réponses ($N = 8$). Le graphe hétérogène $G = (H, E)$, illustré en Figure 3.1, est défini sur $T_E = 3$ types de relations (utilisateurs/questions, utilisateurs/réponses, questions/réponses).

Par ailleurs, $H = \{V_1, V_2, V_3\}$, avec $V_1 = \{v_1^1, v_1^2, v_1^3\}$, $V_2 =$

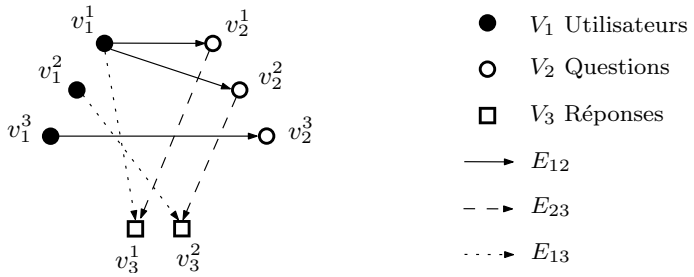

FIGURE 3.1 – Graphe hétérogène associé à l'Exemple 1.1.

$\{v_2^1, v_2^2, v_2^3\}$ et $V_3 = \{v_3^1, v_3^2\}$. De plus, $E = \{(V_1, V_2), (V_1, V_3), (V_2, V_3)\}$, avec $E_{12} = \{(v_1^1, v_2^1), (v_1^1, v_2^2), (v_1^3, v_2^3)\}$, $E_{13} = \{(v_1^1, v_3^1), (v_1^2, v_3^2\}$ et $E_{23} = \{(v_2^1, v_3^1), (v_2^2, v_3^2\}$. Enfin, les 3 matrices d'adjacences sont :

$$A_{12} = \begin{array}{c} \\ v_1^1 \\ v_1^2 \\ v_1^3 \end{array} \begin{array}{c} \begin{array}{ccc} v_2^1 & v_2^2 & v_2^3 \end{array} \\ \begin{pmatrix} 1 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \end{array} A_{13} = \begin{array}{c} \\ v_1^1 \\ v_1^2 \\ v_1^3 \end{array} \begin{array}{c} \begin{array}{cc} v_3^1 & v_3^2 \end{array} \\ \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{pmatrix} \end{array} A_{23} = \begin{array}{c} \\ v_2^1 \\ v_2^2 \\ v_2^3 \end{array} \begin{array}{c} \begin{array}{cc} v_3^1 & v_3^2 \end{array} \\ \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{pmatrix} \end{array}$$

1.2 Fonctions et graphe d'influence

Le graphe d'influence permet de formuler de façon générique l'ensemble des différentes influences au sein d'une plateforme. Autrement dit, il nous permet d'exprimer et de mesurer à quel point l'autorité ou la qualité de certaines familles d'entités influe sur la qualité des nœuds des autres familles. L'intérêt d'exploiter ce graphe d'influence plutôt qu'une représentation immédiate basée sur le graphe hétérogène construit sur l'ensemble des relations sera montré en Section 5.

Définition 1.1. (*Fonction d'influence*) Soit $G = (H, E)$ un graphe hétérogène d'ordre T_V et $(V_i, V_j) \in E$ une paire de familles d'entités. L'influence qu'exerce la famille V_i sur V_j est donnée par la fonction $f_{\theta_{ij}}$

$$f_{\theta_{ij}} : \mathbb{R}^{n_i} \times \mathbb{R}^{n_j} \to \mathbb{R}^{n_j} \tag{3.1}$$

appellée *fonction d'influence*, dont les paramètres sont spécifiés par le vecteur $\theta_{ij} \in \mathbb{R}^{n_i n_j}$.

Définition 1.2. (*Graphe des influences*) Soit $G = (H, E)$ un graphe hétérogène d'ordre T_V. Le *graphe d'influence* est le graphe orienté $G' = (H, E')$ dont les *sommets* sont les familles de G. De plus, il existe un arc dirigé du noeud V_i vers V_j si la fonction d'influence $f_{\theta_{ij}}$ n'est pas nulle, i.e., $(V_i, V_j) \in E' \Leftrightarrow f_{\theta_{ij}} \neq 0$.

Une représentation graphique d'un graphe d'influence construit sur un graphe hétérogène d'ordre 2 et défini à partir d'une unique fonction d'influence $f_{\theta_{ij}}$ est donnée en Figure 3.2.

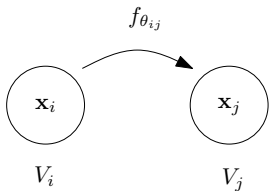

FIGURE 3.2 – Influence d'une famille V_i sur V_j.

1.3 Exercice des influences

1.3.1 Cas particulier

Etant donnée une fonction $f_{\theta_{ij}}$ décrivant l'influence d'une famille V_i sur une famille V_j, l'exercice de l'influence de V_i sur V_j est le calcul des scores de qualité des nœuds de la famille V_j à partir de l'état des variables $\mathbf{x}_i$, $\mathbf{x}_j$ et θ_{ij}. Une définition formelle de la notion d'exercice d'influence est donnée en Définition 1.3.

Définition 1.3. (*Exercice d'influence*) Soit $G = (H, E)$ un graphe hétérogène d'ordre T_V. Soit $(V_i, V_j) \in E$ une paire de familles d'entités et $f_{\theta_{ij}}$ une fonction d'influence non nulle. Si $\mathbf{x}_j^{(t)}$ et $\mathbf{x}_j^{(t+1)}$ sont les vecteurs de scores de la famille V_j avant et après l'application de $f_{\theta_{ij}}$ respectivement, alors :

$$\mathbf{x}_j^{(t+1)} = f_{\theta_{ij}}(\mathbf{x}_i^{(t)}, \mathbf{x}_j^{(t)}) \tag{3.2}$$

L'application de la fonction $f_{\theta_{ij}}$ sur les vecteurs de scores $\mathbf{x}_j^{(t)}$ et $\mathbf{x}_i^{(t)}$ est appelée *exercice d'influence*.

L'Exemple 1.2 détaille les calculs liés à l'exercice d'une unique fonction d'influence sur la famille d'entités V_1 de la Figure 3.2.

Exemple 1.2. Considérons les deux ensembles V_1 (utilisateurs) et V_3 (réponses) de l'Exemple 1.1 en relation par E_{13}. Si $f_{\theta_{13}}$ mesure combien l'autorité $\mathbf{x}_1$ des utilisateurs influe sur la qualité $\mathbf{x}_3$ des réponses apportées, on a :

$$\mathbf{x}_3^{(t+1)} = f_{\theta_{13}}(\mathbf{x}_1^{(t)}, \mathbf{x}_3^{(t)}) \tag{3.3}$$

En considérant une influence linéaire des scores d'autorité $\mathbf{x}_1^{(t)}$ des utilisateurs, et en posant $f_{\theta_{13}} = A_{13}^T \circ \theta_{13}$, avec

$$\theta_{13} = \begin{pmatrix} \theta_{13}^{11} & \theta_{13}^{12} & \theta_{13}^{13} \\ \theta_{13}^{21} & \theta_{13}^{22} & \theta_{13}^{23} \end{pmatrix} \in M_{23}(\mathbb{R}) \tag{3.4}$$

on a

$$\mathbf{x}_3 = \begin{pmatrix} \theta_{13}^{11} & 0 & 0 \\ 0 & \theta_{13}^{22} & 0 \end{pmatrix} \begin{pmatrix} x_1^1 \\ x_1^2 \\ x_1^3 \end{pmatrix} = (\theta_{13}^{11} x_1^1, \theta_{13}^{22} x_1^2) \tag{3.5}$$

La qualité des réponses associées aux nœuds v_3^1 et v_3^2 est alors exprimée par $\theta_{13}^{11} x_1^1$ et $\theta_{13}^{22} x_1^2$ respectivement, c'est à dire linéairement par rapport à l'autorité des auteurs.

1.3.2 Cas général

Lorsque plusieurs familles d'entités exercent en même temps leur influence sur une famille V_k, une aggrégation des résultats associés aux différentes fonctions d'influence doit être considérée. Une formulation générique est donnée par l'Équation (3.6) :

$$\forall k \in \mathcal{T}_V, \mathbf{x}_k^{(t+1)} = g_k(\mathbf{x}_k^{(t)}, f_{\theta_{1k}}(\mathbf{x}_1^{(t)}), f_{\theta_{2k}}(\mathbf{x}_2^{(t)}), \ldots, f_{\theta_{T_V k}}(\mathbf{x}_{T_V}^{(t)})) \tag{3.6}$$

avec $g_k : M_{(k+1)T_V}(\mathbb{R}) \to \mathbb{R}^k$ une fonction d'aggrégation des différents vecteurs de scores calculés par l'ensemble des fonctions d'influences.

Dans ce manuscrit, nous avons considéré une aggrégation linéaire des différentes influences. Formellement, l'exercice de l'ensemble des fonctions d'influence sur une famille V_k est calculé de la façon suivante :

$$\forall k \in \mathcal{T}_V, \mathbf{x}_k^{(t+1)} = \sum_{j \in \mathcal{T}_V} f_{\theta_{jk}}(\mathbf{x}_j^{(t)}, \mathbf{x}_k^{(t)}) \tag{3.7}$$

L'Exemple 1.3 détaille les calculs liés à l'exercice de plusieurs fonctions d'influence sur une même famille d'entités.

Exemple 1.3. Considérons les trois ensembles V_1 (utilisateurs), V_2 (questions) et V_3 (réponses) de l'Exemple 1.1. Si $f_{\theta_{21}}$ et $f_{\theta_{31}}$ mesurent combien la qualité $\mathbf{x}_2$ des questions et la qualité $\mathbf{x}_3$ des réponses influent sur l'autorité $\mathbf{x}_1$ des utilisateurs, alors on a :

$$\begin{cases} \mathbf{x}_1^{(t+1)} & = f_{\theta_{21}}(\mathbf{x}_1^{(t)}, \mathbf{x}_2^{(t)}) \\ \mathbf{x}_1^{(t+1)} & = f_{\theta_{31}}(\mathbf{x}_1^{(t)}, \mathbf{x}_3^{(t)}) \end{cases} \tag{3.8}$$

Ainsi, d'après l'Equation 3.6, on a :

$$\mathbf{x}_1^{(t+1)} = g_1(f_{\theta_{21}}(\mathbf{x}_1^{(t)}, \mathbf{x}_2^{(t)}), f_{\theta_{31}}(\mathbf{x}_1^{(t)}, \mathbf{x}_3^{(t)})) \tag{3.9}$$

En considérant une agrégation linéaire des deux fonctions d'influences $f_{\theta_{21}}$ et $f_{\theta_{31}}$, finalement :

$$\mathbf{x}_1^{(t+1)} = f_{\theta_{21}}(\mathbf{x}_1^{(t)}, \mathbf{x}_2^{(t)}) + f_{\theta_{31}}(\mathbf{x}_1^{(t)}, \mathbf{x}_3^{(t)})) \tag{3.10}$$

1.4 Le principe de renforcement mutuel

Le principe de renforcement mutuel permet d'exprimer une relation d'influence circulaire entre la qualité de deux familles d'entités. Par exemple, dans l'exemple illustré en Figure 3.1, le principe de renforcement mutuel permet d'exprimer que l'autorité des utilisateurs influe sur la qualité de leurs réponses, et qu'inversement, la qualité des réponses influe sur l'autorité de leurs auteurs. Cette relation circulaire mène nécessairement à un système d'équations défini comme suit.

Définition 1.4. (*Principe de renforcement mutuel*) Soit $G = (H, E)$ un graphe hétérogène d'ordre T_V. Le *principe de renforcement mutuel* défini sur une paire de familles d'entités $(V_i, V_j) \in E$ est exprimé par la paire de fonctions d'influence $(f_{\theta_{ij}}, f_{\theta_{ji}})$. Son exercice est exprimé par le système d'équations suivant :

$$\begin{cases} \mathbf{x}_j^{(t+1)} & = f_{\theta_{ij}}(\mathbf{x}_i^{(t)}, \mathbf{x}_j^{(t)}) \\ \mathbf{x}_i^{(t+1)} & = f_{\theta_{ji}}(\mathbf{x}_j^{(t)}, \mathbf{x}_i^{(t)}) \end{cases} \tag{3.11}$$

La Figure 3.3 illustre ce principe de renforcement mutuel entre deux familles d'entités.

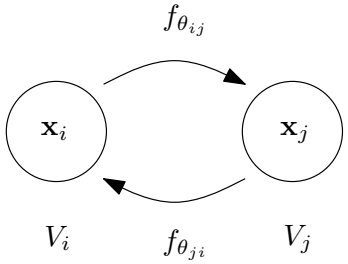

FIGURE 3.3 – Principe de renforcement mutuel entre V_i et V_j.

2 Exemples d'applications

Le caractère générique de la modélisation introduite en Section 1 est appuyé en considérant trois instances différentes associées aux plateformes Wikipédia [1], Stack Exchange [2] et Microsoft Academic Search [3]. Les modèles proposés, nommés Wiki 1, Mas 1 et Ste 1 sont formellement définis par les Définition 2.1, Définition 2.2 et Définition 2.3 respectivement.

Définition 2.1. (Wiki 1) L'instance Wiki 1 est représentée par un graphe hétérogène $G = (H, E)$ d'ordre 2 défini sur V_1 l'ensemble des contributeurs, et sur V_2 l'ensemble des articles. Par ailleurs, $T_E = 2$; l'ensemble des relations $E = \{(V_1, V_1), (V_1, V_2)\}$ définit une relation de collaboration E_{11} et une relation d'édition E_{12}. Le modèle Wiki 1 est exprimé par :

$$\text{Wiki 1} \begin{cases} \mathbf{x}_1^{(t+1)} & = f_{\theta_{11}}(\mathbf{x}_1^{(t)}) + f_{\theta_{21}}(\mathbf{x}_2^{(t)}) \\ \mathbf{x}_2^{(t+1)} & = f_{\theta_{12}}(\mathbf{x}_1^{(t)}) \end{cases} \tag{3.12}$$

où les trois fonctions d'influences expriment l'influence

- $f_{\theta_{11}}$ des co-auteurs sur l'autorité d'un utilisateur ;
- $f_{\theta_{21}}$ des articles sur l'autorité d'un utilisateur ;
- $f_{\theta_{12}}$ des auteurs sur la qualité d'un article.

Une représentation graphique du graphe d'influence associée au modèle Wiki 1 est donnée en Figure 3.4.

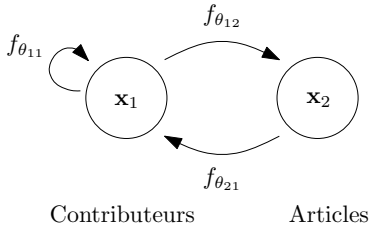

Figure 3.4 – Graphe d'influence du modèle Wiki 1.

Définition 2.2. (Mas 1) L'instance Mas 1 est représentée par un graphe hétérogène $G = (H, E)$ d'ordre 3 définit sur V_1 l'ensemble des chercheurs, V_2 l'ensemble des articles et V_3 l'ensemble des conférences.

1. https ://fr.wikipedia.org/
2. https ://stackexchange.com/
3. http ://academic.research.microsoft.com/

De plus, $T_E = 4$, avec $E = \{(V_1, V_1), (V_1, V_2), (V_2, V_2), (V_2, V_3)\}$ définissant E_{11} une relation de collaboration, E_{12} une relation d'auteur, E_{22} une relation de citation et E_{23} une relation de publication. Le modèle MAS 1 est exprimé par :

$$\text{MAS 1} \begin{cases} \mathbf{x}_1^{(t+1)} & = f_{\theta_{11}}(\mathbf{x}_1^{(t)}) + f_{\theta_{21}}(\mathbf{x}_2^{(t)}) \\ \mathbf{x}_2^{(t+1)} & = f_{\theta_{12}}(\mathbf{x}_1^{(t)}) + f_{\theta_{22}}(\mathbf{x}_2^{(t)}) + f_{\theta_{32}}(\mathbf{x}_3^{(t)}) \\ \mathbf{x}_3^{(t+1)} & = f_{\theta_{23}}(\mathbf{x}_2^{(t)}) \end{cases} \tag{3.13}$$

où les six fonctions d'influence expriment l'influence

- $f_{\theta_{11}}$ des collaborateurs sur l'autorité d'un chercheur ;
- $f_{\theta_{21}}$ des articles sur l'autorité d'un chercheur ;
- $f_{\theta_{12}}$ des auteurs sur la qualité d'un article ;
- $f_{\theta_{22}}$ des articles (citant) sur la qualité d'un article (cités) ;
- $f_{\theta_{23}}$ de la conférence sur la qualité d'un article ;
- $f_{\theta_{32}}$ des articles sur la qualité d'une conférence.

Une représentation graphique du modèle MAS 1 est donnée en Figure 3.5.

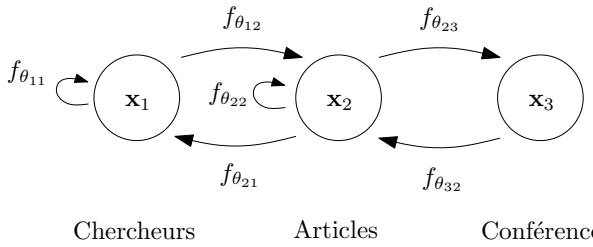

FIGURE 3.5 – Graphe d'influence du modèle MAS 1.

Définition 2.3. (STE 1) L'instance STE 1 est représentée par un graphe hétérogène $G = (H, E)$ d'ordre 4 définit sur V_1 l'ensemble des utilisateurs, V_2 l'ensemble des questions, V_3 l'ensemble des réponses et V_4 l'ensemble des commentaires. De plus, $T_E = 6$ avec E_{12}, E_{13} et E_{14} des relations d'auteur, E_{23}, E_{24} et E_{34} des relations de liaisons. Le modèle STE 1 est exprimé par :

$$\text{STE 1} \begin{cases} \mathbf{x}_1^{(t+1)} & = f_{\theta_{21}}(\mathbf{x}_2^{(t)}) + f_{\theta_{31}}(\mathbf{x}_3^{(t)}) + f_{\theta_{41}}(\mathbf{x}_4^{(t)}) \\ \mathbf{x}_2^{(t+1)} & = f_{\theta_{12}}(\mathbf{x}_1^{(t)}) + f_{\theta_{32}}(\mathbf{x}_3^{(t)}) \\ \mathbf{x}_3^{(t+1)} & = f_{\theta_{13}}(\mathbf{x}_1^{(t)}) \\ \mathbf{x}_4^{(t+1)} & = f_{\theta_{14}}(\mathbf{x}_1^{(t)}) \end{cases} \tag{3.14}$$

où les sept fonctions d'influence expriment l'influence

- $f_{\theta_{21}}$ des questions sur l'autorité d'un utilisateur ;
- $f_{\theta_{31}}$ des réponses sur l'autorité d'un utilisateur ;
- $f_{\theta_{41}}$ des commentaires sur l'autorité d'un utilisateur ;
- $f_{\theta_{12}}$ d'un utilisateur sur la qualité d'une question ;
- $f_{\theta_{13}}$ de l'auteur sur la qualité d'une réponse ;
- $f_{\theta_{14}}$ de l'auteur sur la qualité d'un commentaire ;
- $f_{\theta_{32}}$ des réponses sur la qualité d'une question ;

Une représentation graphique du modèle STE 1 est donnée en Figure 3.6.

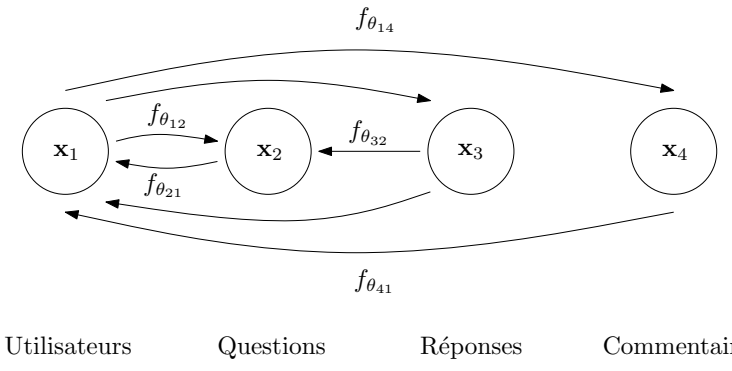

FIGURE 3.6 – Graphe d'influence du modèle STE 1.

3 Algorithme 1

Un premier algorithme de propagation de scores dans le graphe représentant la plateforme est proposé. Etant donné un graphe hétérogène $G = (H, V)$ d'ordre T_V, un ensemble de fonctions d'influence $\{f_{\theta_{ij}}\}$ et un ensemble de vecteurs de scores d'initialisation $\{\mathbf{x}_i^{(0)}\}_{i \leq T_V}$, l'algorithme consiste en l'exercice successif des différentes fonctions d'influence jusqu'à stabilité des scores. Les phases principales de l'algorithme sont les suivantes :

1. *Initialisation* des vecteurs de scores $\mathbf{x}_i \in \mathbb{R}^{n_i}$ associés à chaque famille d'entités ;
2. *Propagation* des scores par l'exercice des différentes fonctions d'influences ;
3. *Normalisation* des vecteurs de scores par famille d'entités ;
4. *Vérification* du critère de convergence.

Le critère de stabilité des scores dans le graphe est calculé à partir de la différence entre les valeurs des vecteurs de scores avant et après

Algorithme 1 $(\mathcal{T}_V, \{f_{\theta_{ij}}\}, \lambda)$

/** Initialisation **/
$t \leftarrow 1$
pour tout $k \in \mathcal{T}_V$ **faire**
    $\mathbf{x}_k^{(t)} \leftarrow$ Aléatoire($\mathbb{R}^{n_k}$)
fin pour
/** Critère de convergence **/
tant que $\sum_{k \in \mathcal{T}_V} ||\mathbf{x}_k^{(t+1)} - \mathbf{x}_k^t||_2 > \lambda$ **faire**
    /** Propagation **/
    pour tout $k \in \mathcal{T}_V$ **faire**
        $\mathbf{x}_k^{(t+1)} \leftarrow 0$
        pour tout $k' \in \mathcal{T}_V$ **faire**
            $\mathbf{x}_k^{(t+1)} \leftarrow \mathbf{x}_k^{(t+1)} + f_{\theta_{k'k}}(\mathbf{x}_{k'}^{(t)}, \mathbf{x}_k^{(t)})$
        fin pour
    fin pour
    /** Normalisation **/
    pour tout $k \in \mathcal{T}_V$ **faire**
        $\mathbf{x}_k^{(t+1)} \leftarrow \frac{\mathbf{x}_k^{(t+1)}}{||\mathbf{x}_k^{(t+1)}||_2}$
    fin pour
    $t \leftarrow t + 1$
fin tant que
Retourner $\{\mathbf{x}_k^t\}_{k \leq T_V}$

l'exercice des fonctions d'influence. Formellement, pour un $\lambda \in \mathbb{R}$ fixé, l'algorithme converge lorsque l'inégalité formulée par l'Equation (3.15) est vérifiée.

$$\sum_{i \in \mathcal{T}_V} ||\mathbf{x}_i^{t+1} - \mathbf{x}_i^t||_2 \leq \lambda \tag{3.15}$$

L'Algorithme 1 détaille les principales étapes de la proposition.

4 Intégration de connaissances

Dans cette section est détaillée la proposition permettant l'intégration de connaissances extérieures dans le modèle structurel présenté. La Section 4.1 s'intéresse à la représentation des connaissances *a priori*. La Section 4.2 formalise le problème sous la forme d'un problème d'optimisation. La Section 4.3 s'intéresse à la résolution du problème d'optimisation et décrit la phase de correction des paramètres associée. Enfin, une extension de l'algorithme 1 est proposée en Section 4.4.

4.1 Modélisation des connaissances

Soit $G = (H, V)$ un graphe hétérogène représentant une plateforme. Nous supposons dans la suite que les ensembles $\{V_k\}_{k \leq T_V}$ sont des ensembles partiellement ordonnés. Autrement dit, chaque ensemble est muni d'une relation d'ordre indiquant, pour certains couples d'éléments $(v_k^i, v_k^j) \in V_k^2$, si l'un est préféré à l'autre. On note alors $v_k^i \succ v_k^j$ lorsque l'entité associée au nœud v_k^i est préférée à v_k^j. Ces relations d'ordre ou de préférences utilisateur sur les différents ensembles V_k permettent la représentation des connaissances extérieures sur la qualité des entités. Dans la suite, nous noterons $\mathcal{P}_k = \{(v_k^i, v_k^j) \in V_k^2 : v_k^i \succ v_k^j \vee v_k^i \prec v_k^j\}$ l'ensemble des connaissances présentées de façon qualitative sur V_k.

4.2 Problème d'optimisation

Soit $y_k^{ij} \in \{-1, +1\}$ le label associé à la paire $(v_k^i, v_k^j) \in \mathcal{P}_k$ traduisant la préférence utilisateur sur un ensemble V_k, vérifiant :

$$\forall k \in \mathcal{T}_V : \mathcal{P}_k \neq \emptyset \begin{cases} y_k^{ij} = & +1 \text{ si } v_k^i \succ v_k^j, \\ y_k^{ij} = & -1 \text{ si } v_k^i \prec v_k^j \end{cases} \tag{3.16}$$

Etant supposé fixés les scores de qualité des différentes familles d'entités, la phase de correction des paramètres des fonctions d'influence est guidée par la minimisation du risque empirique $\mathcal{L}$ formulé par l'Équation (3.17) :

$$\mathcal{L} = \sum_{k \in \mathcal{T}_V} \sum_{(v_k^i, v_k^j) \in \mathcal{P}_k} \Delta(y_k^{ij}, x_k^{ij}) \tag{3.17}$$

avec $x_k^{ij} = x_k^i - x_k^j$ et $\Delta : \mathbb{R}^2 \to \mathbb{R}$ la fonction de coût de Hinge (Xia et al., 2008) définie par :

$$\Delta(y_k^{ij}, x_k^{ij}) = \max(0, 1 - y_k^{ij} x_k^{ij}) \tag{3.18}$$

En particulier, soit $(v_k^i, v_k^j) \in \mathcal{P}_k$ une paire de nœuds du graphe tels que $v_k^i \succ v_k^j$. Le modèle $\boldsymbol{\theta} = \{\theta_{ij}\}_{i,j \leq \mathcal{T}_V}$ est en accord avec la préférence utilisateur sur la paire (v_k^i, v_k^j) si le score de qualité renvoyé par le modèle pour le nœud v_k^i est supérieur (de plus de 1) à celui du nœud v_k^j, formellement lorsque $1 - x_k^{ij} y_k^{ij} \leq 0$. A l'inverse, le modèle est en désaccord avec la préférence utilisateur sur la paire (v_k^i, v_k^j) lorsque $1 - x_k^{ij} y_k^{ij} > 0$.

4.3 Correction des paramètres

Si $f^i_{\theta_{k'k}}$ est l'influence d'une famille d'entité $V_{k'}$ sur un élément v_k^i de V_k, alors dans le cas d'influences linéaires, le score de qualité x_k^i associé

peut s'exprimer par :

$$x_k^i = \sum_{k' \in \mathcal{T}_V} f^i_{\theta_{k'k}}(\mathbf{x}_{k'}) = \sum_{k'=1}^{T_V} \sum_{l=1}^{n_{k'}} A^{il}_{k'k} \theta^{il}_{k'k} x^l_{k'} \tag{3.19}$$

Aussi, le gradient de la fonction d'erreur Δ en tout point $\theta^{cl}_{k'k}$ s'exprime :

$$\frac{\partial \Delta(y_k^{ij}, x_k^{ij})}{\partial \theta^{cl}_{k'k}} = \begin{cases} -A^{cl}_{k'k} x^l_{k'} y^{ij}_k & \text{si } c = i \text{ et } 1 - y^{ij}_k x^{ij}_k \geq 0, \\ +A^{cl}_{k'k} x^l_{k'} y^{ij}_k & \text{si } c = j \text{ et } 1 - y^{ij}_k x^{ij}_k \geq 0, \\ 0 & \text{sinon.} \end{cases} \tag{3.20}$$

L'ajustement des paramètres des fonctions d'influence consiste donc à modifier la valeur de $\boldsymbol{\theta}$ dans le sens opposé au gradient lorsque, pour chaque paire de nœuds $(v_k^i, v_k^j) \in V_k$ tels que $v_k^i \succ v_k^j$, la prédiction donnée par le modèle est incorrecte. En l'occurence, pour un taux d'apprentissage ϵ fixé par l'utilisateur, les modifications suivantes sur le modèle sont réalisées pour toute paire (y_k^{ij}, x_k^{ij}) vérifiant $\Delta(y_k^{ij}, x_k^{ij}) > 0$:

$$\theta^{il}_{k'k} = \theta^{il}_{k'k} \pm \epsilon \cdot A^{il}_{k'k} \cdot x^l_{k'} y^{ij}_k \tag{3.21}$$

Notons que la solution est *conservative* dans le sens où aucune modification des paramètres n'est réalisée si les prédictions respectent les préférences utilisateurs. L'inscription de cette étape de correction dans l'algorithme est détaillée dans la Section 4.4.

4.4 Algorithme 2

L'algorithme présenté dans cette section étend l'Algorithme 1 en intégrant la phase de correction des paramètres d'influence. Les nouvelles phases de l'algorithme sont les suivantes :

1. *Initialisation* des vecteurs de scores ;
2. *Propagation* des scores par l'exercice des fonctions d'influence ;
3. *Correction* des paramètres des fonctions d'influence ;
4. *Normalisation* des vecteurs de scores ;
5. *Vérification* du critère de convergence.

En particulier, les étapes de propagation et de correction sont réalisées alternativement de façon à calculer d'une part, les vecteurs de scores, et d'autre part, les paramètres d'influence. L'Algorithme 2 détaille les principales étapes de la proposition. L'Exemple 4.1 illustre le fonctionnement de l'algorithme en ne considérant, par simplicité, qu'une seule famille d'entités.

Algorithme 2 $(\mathcal{T}_V, \mathcal{P}_k, \{f_{\theta_{ij}}\}_{i,j \leq T_V}, \epsilon, \lambda)$

$t \leftarrow 1$
pour tout $k \in \mathcal{T}_V$ **faire**
  $\mathbf{x}_k^{(t)} \leftarrow \text{Random}(\mathbb{R}^{n_k})$
fin pour
tant que $\sum_{k \in \mathcal{T}_V} ||\mathbf{x}_k^{(t+1)} - \mathbf{x}_k^t||_2 \geq \lambda$ **faire**
  pour tout $k \in \mathcal{T}_V$ **faire**
    $\mathbf{x}_k^{(t+1)} \leftarrow 0$
    pour tout $k' \in \mathcal{T}_V$ **faire**
      $\mathbf{x}_k^{(t+1)} \leftarrow \mathbf{x}_k^{(t+1)} + f_{\theta_{k'k}}(\mathbf{x}_{k'}^{(t)}, \mathbf{x}_k^{(t)})$
    fin pour
  fin pour
  pour tout $k \in \mathcal{T}_V$ **faire**
    pour tout $(v_k^i, v_k^j) \in \mathcal{P}_k$ **faire**
      si $\Delta(y_k^{ij}, x_k^{ij}) > 0$ **alors**
        pour tout $k' \in \mathcal{T}_V : (V_{k'}, V_k) \in E$ **faire**
          pour tout $l \leq n_{k'}$ **faire**
            $\theta_{k'k}^{il} \leftarrow \theta_{k'k}^{il} + \epsilon \cdot A_{k'k}^{il} \cdot x_{k'}^l y_k^{ij}$
            $\theta_{k'k}^{jl} \leftarrow \theta_{k'k}^{il} - \epsilon \cdot A_{k'k}^{jl} \cdot x_{k'}^l y_k^{ij}$
          fin pour
        fin pour
      fin si
    fin pour
  fin pour
  pour tout $k \in \mathcal{T}_V$ **faire**
    $\mathbf{x}_k^{(t+1)} \leftarrow \frac{\mathbf{x}_k^{(t+1)}}{||\mathbf{x}_k^{(t+1)}||_2}$
  fin pour
  $t \leftarrow t + 1$
fin tant que
Retourner $\{\mathbf{x}_k^t\}_{k \leq T_V}$

Exemple 4.1. Soit le graphe $G = (H, E)$ définie sur $T_V = 1$, $H = \{V_1\}$, avec $V_1 = \{v_1^1, v_1^2, v_1^3, v_1^4\}$ et $E = \{(V_1, V_1)\}$ avec $E_{11} = \{(v_1^1, v_1^3), (v_1^2, v_1^3), (v_1^3, v_1^4)\}$. Le graphe est illustré en Figure 3.7. On considère $\mathcal{P}_k = \{(v_1^3, v_1^4)\}$ avec $y_1^{34} = 1$. La fonction d'erreur s'exprime donc :

$$\mathcal{L} = \max(0, 1 - y_1^{34}(x_1^3 - x_1^4))$$

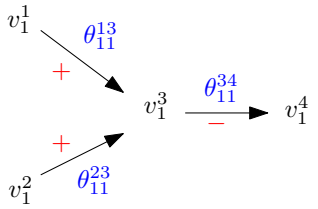

FIGURE 3.7 – Graphe associé à l'Exemple 4.1. Les symboles $\{+, -\}$ associés aux arcs indiquent comment l'étape de correction met à jour les valeurs des paramètres du modèle pour que la contrainte $v_1^3 \succ v_1^4$ soit respectée.

Ou, de façon équivalente, en faisant intervenir les paramètres des fonctions d'influence :

$$\mathcal{L} = \max(0, 1 - y_1^{34}((\theta_{11}^{13}x_1^1 + \theta_{11}^{23}x_1^2) - \theta_{11}^{34}x_1^3))$$

Soit $\epsilon = 10^{-1}$ le pas de gradient et $\lambda = 10^{-2}$ le seuil de convergence. Les étapes principales de l'Algorithme 2 appliqué à cet exemple, jusqu'à convergence, sont les suivantes :

- *Initialisation.*

$$\mathbf{x}_1^{(t_0)} = (1, 1, 1, 1), \text{ et } (\theta_{11}^{13}, \theta_{11}^{23}, \theta_{11}^{34})^{(t_0)} = (1, 1, 1)$$

- *Itération 1.*
  - (a) *Propagation*

$$\begin{cases} x_1^{1(t_1)} = 1 \\ x_1^{2(t_1)} = 1 \\ x_1^{3(t_1)} = \theta_{11}^{13}x_1^{1(t_0)} + \theta_{11}^{23}x_1^{2(t_0)} = 1 \cdot 1 + 1 \cdot 1 = 2 \\ x_1^{4(t_1)} = \theta_{11}^{34}x_1^{3(t_0)} = 1 \cdot 1 = 1 \end{cases}$$

  - (b) *Correction ?*

$$\mathcal{L}(\mathbf{x}_1^{t_1}, \boldsymbol{\theta}_{11}^{(t_1)}) = \max(0, 1 - 1(2 - 1)) = 0$$

  $\mathcal{L} \leq 0$: aucune correction des paramètres d'influence.

  - (c) *Convergence ?*

$$||\mathbf{x}_1^{(t_1)} - \mathbf{x}_1^{(t_0)}||_2 = 1 > \lambda$$

- *Itération 2.*

  - (a) *Propagation*

$$\begin{cases} x_1^{1(t_2)} = 1 \\ x_1^{2(t_2)} = 1 \\ x_1^{3(t_2)} = \theta_{11}^{13} x_1^{1(t_1)} + \theta_{11}^{23} x_1^{2(t_1)} = 1 \cdot 1 + 1 \cdot 1 = 2 \\ x_1^{4(t_2)} = \theta_{11}^{34} x_1^{3(t_1)} = 1 \cdot 2 = 2 \end{cases}$$

  - (b) *Correction ?*

$$\mathcal{L}(\mathbf{x}_1^{t_2}, \boldsymbol{\theta}_{11}^{(t_2)}) = \max(0, 1 - 1(2 - 2)) = 1$$

  $\mathcal{L} > 0$: correction des paramètres d'influence, avec :

$$\begin{cases} \dfrac{\partial \mathcal{L}^{(t_2)}}{\partial \theta_{11}^{13}} = -x_1^{1,(t_2)} = -1 \\ \dfrac{\partial \mathcal{L}^{(t_2)}}{\partial \theta_{11}^{23}} = -x_1^{2,(t_2)} = -1 \\ \dfrac{\partial \mathcal{L}^{(t_2)}}{\partial \theta_{11}^{34}} = x_1^{3,(t_2)} = 2 \end{cases}$$

  donc

$$\begin{cases} \theta_{11}^{13} = \theta_{11}^{13} + \epsilon = 1.1 \\ \theta_{11}^{23} = \theta_{11}^{23} + \epsilon = 1.1 \\ \theta_{11}^{34} = \theta_{11}^{34} - 2 \cdot \epsilon = 0.8 \end{cases}$$

  - (c) *Convergence ?*

$$||\mathbf{x}_1^{(t_2)} - \mathbf{x}_1^{(t_1)}||_2 = 1 > \lambda$$

- *Itération 3.*

  - (a) *Propagation*

$$\begin{cases} x_1^{1(t_3)} = 1 \\ x_1^{2(t_3)} = 1 \\ x_1^{3(t_3)} = \theta_{11}^{13} x_1^{1(t_2)} + \theta_{11}^{23} x_1^{2(t_2)} = 1.1 \cdot 1 + 1.1 \cdot 1 = 2.2 \\ x_1^{4(t_3)} = \theta_{11}^{34} x_1^{3(t_2)} = 0.8 \cdot 2 = 1.6 \end{cases}$$

- (b) *Correction* ?

$$\mathcal{L}(\mathbf{x}_1^{t_3}, \boldsymbol{\theta}_{11}^{(t_3)}) = \max(0, 1 - 1(2.2 - 1.6)) = 0.4$$

$\mathcal{L} > 0$: correction des paramètres d'influence, avec :

$$\begin{cases} \dfrac{\partial \mathcal{L}^{(t_3)}}{\partial \theta_{11}^{13}} = -x_1^{1,(t_3)} = -1 \\ \dfrac{\partial \mathcal{L}^{(t_3)}}{\partial \theta_{11}^{23}} = -x_1^{2,(t_3)} = -1 \\ \dfrac{\partial \mathcal{L}^{(t_3)}}{\partial \theta_{11}^{34}} = x_1^{3,(t_3)} = 2.2 \end{cases}$$

donc

$$\begin{cases} \theta_{11}^{13} = \theta_{11}^{13} + \epsilon = 1.2 \\ \theta_{11}^{23} = \theta_{11}^{23} + \epsilon = 1.2 \\ \theta_{11}^{34} = \theta_{11}^{34} - 2.2 \cdot \epsilon = 0.8 - 2.2 \cdot \epsilon = 0.66 \end{cases}$$

- (c) *Convergence* ?

$$||\mathbf{x}_1^{(t_3)} - \mathbf{x}_1^{(t_2)}||_2 = 0.2 > \lambda$$

- *Itération 4.*

- (a) *Propagation*

$$\begin{cases} x_1^{1(t_4)} = 1 \\ x_1^{2(t_4)} = 1 \\ x_1^{3(t_4)} = \theta_{11}^{13} x_1^{1(t_3)} + \theta_{11}^{23} x_1^{2(t_3)} = 1.2 \cdot 1 + 1.2 \cdot 1 = 2.4 \\ x_1^{4(t_4)} = \theta_{11}^{34} x_1^{3(t_3)} = 0.66 \cdot 2.2 = 1.452 \end{cases}$$

- (b) *Correction* ?

$$\mathcal{L}(\mathbf{x}_1^{t_4}, \boldsymbol{\theta}_{11}^{(t_4)}) = \max(0, 1 - 1(2.4 - 1.452)) = 0.052$$

$\mathcal{L} > 0$: correction des paramètres d'influence, avec :

$$\begin{cases} \dfrac{\partial \mathcal{L}^{(t_4)}}{\partial \theta_{11}^{13}} = -x_1^{1,(t_4)} = -1 \\ \dfrac{\partial \mathcal{L}^{(t_4)}}{\partial \theta_{11}^{23}} = -x_1^{2,(t_4)} = -1 \\ \dfrac{\partial \mathcal{L}^{(t_4)}}{\partial \theta_{11}^{34}} = x_1^{3,(t_4)} = 2.4 \end{cases}$$

donc

$$\begin{cases} \theta_{11}^{13} = \theta_{11}^{13} + \epsilon = 1.3 \\ \theta_{11}^{23} = \theta_{11}^{23} + \epsilon = 1.3 \\ \theta_{11}^{34} = \theta_{11}^{34} - 2 \cdot \epsilon = 0.66 - 2.4 \cdot \epsilon = 0.42 \end{cases}$$

- (c) *Convergence* ?

$$||\mathbf{x}_1^{(t_4)} - \mathbf{x}_1^{(t_3)}||_2 = 0.062 > \lambda$$

- *Itération 5.*

- (a) *Propagation*

$$\begin{cases} x_1^{1(t_5)} = 1 \\ x_1^{2(t_5)} = 1 \\ x_1^{3(t_5)} = \theta_{11}^{13} x_1^{1(t_4)} + \theta_{11}^{23} x_1^{2(t_4)} = 1.3 \cdot 1 + 1.3 \cdot 1 = 2.6 \\ x_1^{4(t_5)} = \theta_{11}^{34} x_1^{3(t_4)} = 0.42 \cdot 2.4 = 1.008 \end{cases}$$

- (b) *Correction* ?

$$\mathcal{L}(\mathbf{x}_1^{t_5}, \boldsymbol{\theta}_{11}^{(t_5)}) = \max(0, 1 - 1(2.6 - 1.008)) = 0$$

$\mathcal{L} < 0$: pas de correction.

- (c) *Convergence* ?

$$||\mathbf{x}_1^{(t_4)} - \mathbf{x}_1^{(t_3)}||_2 = 0.24 > \lambda$$

- *Itération 6.*

- (a) *Propagation*

$$\begin{cases} x_1^{1(t_6)} = 1 \\ x_1^{2(t_6)} = 1 \\ x_1^{3(t_6)} = \theta_{11}^{13} x_1^{1(t_5)} + \theta_{11}^{23} x_1^{2(t_5)} = 1.3 \cdot 1 + 1.3 \cdot 1 = 2.6 \\ x_1^{4(t_6)} = \theta_{11}^{34} x_1^{3(t_5)} = 0.42 \cdot 2.6 = 1.092 \end{cases}$$

- (b) *Correction*. Evaluation de la fonction de coût :

$$\mathcal{L}(\mathbf{x}_1^{t_6}, \boldsymbol{\theta}_{11}^{(t_6)}) = \max(0, 1 - 1(2.6 - 1.092)) = 0$$

$\mathcal{L} < 0$: pas de correction.

- (c) *Convergence* ?

$$||\mathbf{x}_1^{(t_5)} - \mathbf{x}_1^{(t_4)}||_2 = 7.056 \times 10^{-3} < \lambda$$

- (d) *Retourner* $\mathbf{x}_1^{(t_5)}$.

L'état des variables est résumé dans les matrices suivantes. En particulier, l'évolution des valeurs du vecteur $\mathbf{x}_1$ est

$$\mathbf{x}_1^{(t)} : \begin{array}{c} \\ x_1^1 \\ x_1^2 \\ x_1^3 \\ x_1^4 \end{array} \begin{array}{c} \begin{array}{ccccccc} t_0 & t_1 & t_2 & t_3 & t_4 & t_5 & t_6 \end{array} \\ \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 2 & 2.2 & 2.4 & 2.6 & 2.6 \\ 1 & 1 & 2 & 1.6 & 1.452 & 1.008 & 1.092 \end{pmatrix} \end{array}$$

L'évolution des paramètres d'influences :

$$\boldsymbol{\theta}_{11}^{(t)} : \begin{array}{c} \\ \theta_{11}^{13} \\ \theta_{11}^{23} \\ \theta_{11}^{34} \end{array} \begin{array}{c} \begin{array}{ccccccc} t_0 & t_1 & t_2 & t_3 & t_4 & t_5 & t_6 \end{array} \\ \begin{pmatrix} 1 & 1 & 1.1 & 1.2 & 1.3 & 1.3 & 1.3 \\ 1 & 1 & 1.1 & 1.2 & 1.3 & 1.3 & 1.3 \\ 1 & 1 & 0.8 & 0.66 & 0.42 & 0.42 & 0.42 \end{pmatrix} \end{array}$$

L'évolution des valeurs du critère de convergence :

$$||\mathbf{x}_1^{(t)} - \mathbf{x}_1^{(t-1)}||_2 : \begin{array}{c} \begin{array}{cccccc} t_1 & t_2 & t_3 & t_4 & t_5 & t_6 \end{array} \\ \begin{pmatrix} 1 & 1 & 0.2 & 0.062 & 0.24 & 0.007 \end{pmatrix} \end{array}$$

L'évolution des valeurs de la fonction de coût :

$$\mathcal{L}(\mathbf{x}_1^t, \boldsymbol{\theta}_{11}^{(t)}) : \begin{array}{cccccc} t_1 & t_2 & t_3 & t_4 & t_5 & t_6 \\ (0 & 1 & 0.4 & 0.052 & 0 & 0) \end{array}$$

On note que $x_1^3 > x_1^4$, respectant la préfèrence utilisateur introduite. Il convient de préciser que, sans correction des paramètres d'influence, nous aurions eu, à convergence, $x_1^3 < x_1^4$ (le noeud v_1^4 accumulant les scores de ses prédécesseurs).

5 Discussion

L'approche structurelle classique utilisée pour mesurer l'autorité d'un élement dans un réseau consiste à modéliser le réseau par un graphe simple puis à appliquer directement sur celui-ci une des méthodes de l'état de l'art homogénéisant les relations du graphe. Si cette solution est efficace lorsque l'homogénéité du graphe est désirée et lorsque la notion d'autorité peut être approchée de celle de popularité, elle présente des limites lorsqu'une ou l'autre de ces deux conditions n'est plus vérifiée. Dans cette section, on montre qu'une telle approche possède les deux limitations suivantes :

- elle ne permet pas la définition explicite des relations d'influence entre les entités du graphe ;
- elle biaise, par définition, les entités au coeur du réseau puisque possédant de nombreux liens.

Pour illustrer notre propos, nous développons les calculs effectués par l'algorithme HITS (Kleinberg, 1999) sur la représentation MAS 1 associée à une librairie digitale contenant chercheurs, articles et conférences.

Soit $G = (V, E)$ le graphe non orienté modélisant la librairie, avec $V = 1, 2, ..., N$ l'ensemble des N sommets représentant les différentes entités de la librairie et $E \subseteq V \times V$ représentant les différentes relations. L'ensemble V, homogénéisant les entités, modélise à la fois les conférences, les articles et les chercheurs. De la même manière, l'ensemble E homogénéise les quatres relations suivantes :

- *Relation d'auteur* : $(i, j) \in E$ si et seulement si le chercheur associé au noeud i est l'auteur de l'article j ;
- *Relation de co-auteur* : $(i, j) \in E$ si et seulement si les chercheurs associés aux nœuds i et j sont co-auteurs d'un article ;

- *Relation de publication* : $(i, j) \in E$ si et seulement si l'article associé au noeud i est publiée dans la conférence j ;
- *Relation de citation* : $(i, j) \in E$ si et seulement si l'article i est cité par j.

En organisant la matrice d'adjacence M du graphe G en lignes et en colonnes selon le type des entités, on obtient la formulation par bloc suivante :

$$M = \begin{array}{c} \\ R \\ A \\ C \end{array} \begin{array}{c} \begin{array}{ccc} R & A & C \end{array} \\ \begin{pmatrix} M_{RR} & M_{RA} & M_{RC} \\ M_{AR} & M_{AA} & M_{AC} \\ M_{CR} & M_{CA} & M_{CC} \end{pmatrix} \end{array} \tag{3.22}$$

où $R = \{r_i\}_{i \leq n}$, $A = \{a_i\}_{i \leq m}$ et $C = \{c_i\}_{i \leq p}$ sont les ensembles représentant les n chercheurs, m articles et p conférences respectivement, avec $n + m + p = N$, et $\{M_{ij}\}_{i,j \in \{R,A,C\}}$ sont les matrices d'adjacences associées à toutes les relations binaires possibles entre les trois familles d'entités R, A et C. Sans perte de généralité, on peut définir les matrices d'adjacence de la façon suivante :

- $M_{RA}(r_i, a_j) = 1$ si le chercheur r_i est l'auteur de l'article a_j ;
- $M_{RR}(r_i, r_j) = 1$ si les chercheurs r_i et r_j sont co-auteurs ;
- $M_{AC}(a_i, c_j) = 1$ si l'article a_i est publié dans la conférence c_j ;
- $M_{AA}(a_i, a_j) = 1$ si l'article a_i est cité par l'article a_j ;
- $M_{AR} = M_{RA}^T$;
- $M_{CA} = M_{AC}^T$;
- $M_{RC} = M_{RC}^T = 0$;
- $M_{CC} = 0$.

Par souci de simplicité, les relations entre chercheurs et conférences et les relations entre conférences ne sont pas considérées. La matrice M peut donc être reformulée de la façon suivante :

$$M = \begin{pmatrix} M_{RR} & M_{RA} & 0 \\ M_{RA}^T & M_{AA} & M_{AC} \\ 0 & M_{AC}^T & 0 \end{pmatrix} \tag{3.23}$$

Une des implémentations possible de l'algorithme HITS, connue sous le nom de la méthode des puissances (Golub and Van Loan, 2012), calcule le point fixe $\mathbf{v}^* \in \mathbb{R}^N$ de l'application engendrée par la matrice MM^T. Techniquement, l'algorithme calcule itérativement, jusqu'à convergence, l'image du vecteur $\mathbf{v}^k$ par l'application MM^T, en normalisant les résultats afin d'éviter d'obtenir des valeurs trop grandes. Formellement, la méthode est exprimée par la relation de récurrence suivante :

$$\mathbf{v}^{k+1} = \frac{MM^T\mathbf{v}^k}{||MM^T\mathbf{v}^k||_2} \tag{3.24}$$

De la même manière que nous avons reparti les blocs de la matrice M en fonction des types de nœuds du graphe, le vecteur $\mathbf{v}^k$ est décomposé en fonction des trois types d'entité :

$$\mathbf{v}^k = \begin{pmatrix} \mathbf{v}_R^k \\ \mathbf{v}_A^k \\ \mathbf{v}_C^k \end{pmatrix} \tag{3.25}$$

Dès lors, la relation de récurrence formulée par l'Équation (3.24) peut être décomposée et exprimée :

$$\begin{array}{rl} \mathbf{v}_R^{k+1} = & (M_{RR}^2 + M_{RA}M_{RA}^T)\mathbf{v}_R^k + (M_{RR}M_{RA} + M_{RA}M_{AA})\mathbf{v}_A^k + M_{RA}M_{AC}\mathbf{v}_C^k \\ \mathbf{v}_A^{k+1} = & (M_{RA}^T M_{RR} + M_{AA}M_{RA}^T)\mathbf{v}_R^k + \\ & (M_{RA}^T M_{RA} + M_{AA}M_{AA}^T + M_{AC}M_{AC}^T)\mathbf{v}_A^k + M_{AA}M_{AC}\mathbf{v}_C^k \\ \mathbf{v}_C^{k+1} = & M_{AC}^T M_{RA}^T \mathbf{v}_R^k + M_{AC}^T M_{AA}^T \mathbf{v}_A^k + M_{AC}^T M_{AC}\mathbf{v}_C^k \end{array}$$

Il apparaît maintenant dans cette formulation trois relations de récurrence permettant de calculer les trois vecteurs de scores $\mathbf{v}_R^{k+1} \in \mathbb{R}^n$, $\mathbf{v}_A^{k+1} \in \mathbb{R}^m$ et $\mathbf{v}_C^{k+1} \in \mathbb{R}^p$ associés aux chercheurs, articles et conférences respectivement. En faisant intervenir différentes fonctions des variables $\mathbf{v}_R^k$, $\mathbf{v}_A^k$ et $\mathbf{v}_C^k$, cette dernière formulation, tout à fait équivalente à la relation de récurrence exprimée par l' Équation (3.24), peut aussi être réécrite :

$$\begin{array}{rl} \mathbf{v}_R^{k+1} = & f_{RR}(\mathbf{v}_R^k) + f_{RA}(\mathbf{v}_A^k) + f_{AC}(\mathbf{v}_C^k) \\ \mathbf{v}_A^{k+1} = & f_{AR}(\mathbf{v}_R^k) + f_{AA}(\mathbf{v}_A^k) + f_{AC}(\mathbf{v}_C^k) \\ \mathbf{v}_C^{k+1} = & f_{CR}(\mathbf{v}_R^k) + f_{CA}(\mathbf{v}_A^k) + f_{CC}(\mathbf{v}_C^k) \end{array} \tag{3.26}$$

avec :

1. $f_{RR} = M_{RR}^2 + M_{RA}M_{RA}^T$
2. $f_{RA} = M_{RR}M_{RA} + M_{RA}M_{AA}$
3. $f_{RC} = M_{RA}M_{AC}$
4. $f_{AR} = M_{RA}^T M_{RR} + M_{AA}M_{RA}^T$
5. $f_{AA} = M_{RA}^T M_{RA} + M_{AA}M_{AA}^T + M_{AC}M_{AC}^T$
6. $f_{AC} = M_{AA}M_{AC}$
7. $f_{CR} = M_{AC}^T M_{RA}^T$
8. $f_{CA} = M_{AC}^T M_{AA}^T$
9. $f_{CC} = M_{AC}^T M_{AC}$

Ces neufs fonctions, que l'on associe au concept de *fonction d'influence*, définissent les différents impacts qu'ont les scores des familles d'entités les unes sur les autres. Dans ce contexte, nous montrons qu'il peut-être souhaitable de redéfinir ces fonctions d'influences et en particulier lorsque l'autorité ou la qualité d'une entité n'est pas assimilable à sa popularité. Dans ce but, considérons le vecteur de scores associé aux conférences, calculé, selon la formulation originale de l'Équation (3.26), à partir des vecteurs de scores des chercheurs, des articles et des conférences :

$$\mathbf{v}_C^{k+1} = f_{CR}(\mathbf{v}_R^k) + f_{CA}(\mathbf{v}_A^k) + f_{CC}(\mathbf{v}_C^k) \tag{3.27}$$

avec $f_{CR} = M_{AC}^T M_{RA}^T$, $f_{CA} = M_{AC}^T M_{AA}^T$, $f_{CC} = M_{AC}^T M_{AC}$. Cette dernière fonction d'influence, $f_{CC} = M_{AC}^T M_{AC}$, ajuste le score d'autorité d'une conférence i selon le nombre d_i d'articles publiés dans la conférence. En effet :

$$f_{CC}(\mathbf{v}_C^k) = M_{AC}^T M_{AC} \mathbf{v}_C^k = \text{diag}((d_i)_{i \leq p}) \mathbf{v}_C^k = (d_1 \mathbf{v}_C^k(1), ..., d_p \mathbf{v}_C^k(p))$$

avec d_i le nombre d'articles publiés dans la conférence c_i. Or, comme cela est empiriquement suggéré par la Figure 3.8 construite à partir des données expérimentales issues du Microsoft Academic Search, le nombre d'articles publiés dans une conférence n'est pas nécessairement un critère de qualité. Ainsi, la prise en compte d'une telle fonction d'influence, pourtant inhérente à l'application immédiate de l'algorithme HITS, biaise les résultats en favorisant les entités populaires.

La solution développée dans la Section 4 permet alors de limiter ce biais en corrigeant le poids des fonctions d'influence de façon à privilégier la propagation des scores depuis certaines sources d'autorité connues et à limiter la propagation des scores depuis les sources de moindre qualité.

6 Conclusion

Ce chapitre est consacré à la définition d'un modèle de qualité basé sur une représentation par les graphes des plateformes et notamment sur l'exploitation d'un graphe d'influence. Plus particulièrement, pour pallier le manque de généricité de certaines approches de la littérature, nous avons proposé une approche basée sur l'exploitation des graphes hétérogènes (de La Robertie et al., 2016). Le caractère générique de cette première contribution a été suggéré en modélisant les trois plateformes collaboratives différentes que sont Wikipédia, Stack Exchange et Microsoft Academic Search.

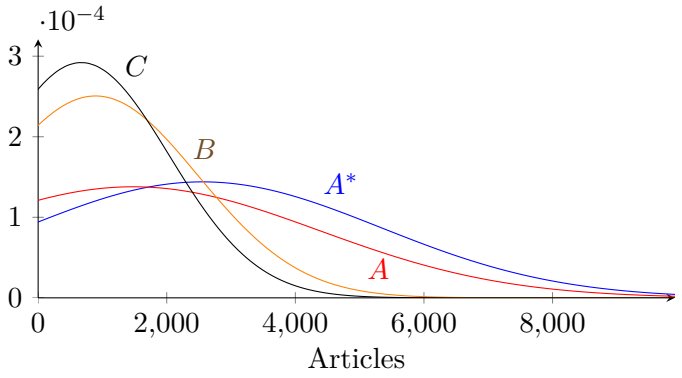

FIGURE 3.8 – Fonctions de densité associées aux nombres d'articles par catégorie de conférence construites à partir de 2 158 conférences issues des jeux de données Microsoft Academic Search et Core.edu.

Par ailleurs, un algorithme de propagation des scores a aussi été proposé (de La Robertie et al., 2017a), exploitant la notion introduite de fonction d'influence entre les familles d'entités. La discussion menée en Section 5 montre que certains calculs liés à l'application directe d'algorithmes de propagations de scores tels que HITS ne devraient pas être considérés pendant la propagation des scores. Pour limiter la propagation des scores depuis les noeuds les plus populaires, biaisant, par définition, les résultats, nous avons proposé une méthode de correction des paramètres des fonctions d'influence. L'ensemble des connaissances extérieurs représenté sous la forme d'une préférence utilisateurs, c'est à dire uniquement sur une échelle qualitative, permet la formulation d'un problème d'optimisation ne faisant intervenir aucun autre paramètre utilisateur que celui associé au taux d'apprentisage. La résolution du problème d'optimisation, réalisé par une méthode de descente de gradient, permet alors, via l'ajustement des paramètres d'influence, la correction des différents impacts entre les familles.

Chapitre 4

Intégration de signaux temporels

Introduction

Dans le chapitre précédant, une approche structurelle est proposée, permettant une représentation générique des plateformes collaboratives, ainsi que le calcul, par l'intermédiaire d'un graphe d'influence, de scores de qualité des utilisateurs et des productions associées. Néanmoins, les applications directes suggérées en Section 2 ne considèrent qu'un ensemble de facteurs statiques d'autorité ou de qualité. Pourtant, comme le suggère certains travaux de la litérature (Hu et al., 2007; Anderson et al., 2012) ainsi que les résultats de nos expérimentations (de La Robertie et al., 2015; de La Robertie et al., 2016), le modèle proposé gagnerait à exploiter certains indices de qualité de nature temporelle.

Dans ce chapitre, deux nouvelles instances des plateformes Stack Exchange [1] et Wikipédia [2] sont proposées, intégrant des indices temporels liés à l'activité des utilisateurs. Sur le site de Questions-Réponses Stack Exchange, c'est un graphe de réactivité des utilisateurs, construit à partir des délais d'apparition des réponses, qui est intégré comme graphe porteur de signaux d'autorité. Sur le site Wikipédia, c'est le processus de révision des articles qui est modélisé de façon à créditer davantage les portions de textes approuvées par l'ensemble des contributeurs et, par effet de chaîne, l'autorité de ces derniers.

L'organisation de ce chapitre est la suivante. La Section 1 motive empiriquement, sur un ensemble de données expérimentales réelles issu du site Stack Exchange, l'intérêt d'exploiter un graphe de réactivité puis propose une reformulation de l'instance MAS 1 en conséquence. La Section 2 propose une modélisation plus complète du site Wikipédia en

1. https ://stackexchange.com/
2. https ://fr.wikipedia.org

définissant le processus de révision à partir du formalisme introduit dans le chapitre précédent. Enfin, la Section 3 conclue le chapitre.

1 La réactivité des utilisateurs sur les sites de Q&R

Cette section propose une modélisation par les graphes de la réactivité des utilisateurs sur les plateformes de type Questions-Réponses ainsi que son exploitation par le modèle développé en Section 3. Dans un premier temps, l'intérêt d'exploiter la réactivité des utilisateurs comme signal d'autorité est empiriquement suggéré en Section 1.1. Dans un second temps, la représentation de cette réactivité utilisateur selon un graphe orienté est motivée en Section 1.2. Enfin, la nouvelle modélisation de la plateforme Stack Exchange, intégrant ce graphe de réactivité, est formalisée en Section 1.3.

1.1 Motivations

Nous postulons que les utilisateurs faisant autorité sur les sites de Questions-Réponses, c'est-à-dire les utilisateurs à l'origine des réponses de qualité, sont davantage réactifs que les autres. Pour quantifier cette notion de réactivité d'un utilisateur, le concept de délai d'apparition des réponses, introduit par la Définition 1.1, est utilisé.

Définition 1.1. (*Délai d'apparition d'une réponse*) Le délai d'apparition d'une réponse est le temps $dt > 0$ séparant la date de publication de la réponse de la date de publication de la question à laquelle la réponse est associée.

Soit la Figure 4.1, traçant l'évolution du délai d'apparition moyen des réponses publiées sur le site Stack Exchange en fonction de leur qualité [3]. Le graphique, réalisé sur la base de 2.4 millions de réponses, montre une forte tendance pour les réponses de qualité à être produites rapidement après la publication des questions. Ces réponses au faible délai d'apparition semblent être davantage appréciées par la communauté, créditant alors l'autorité de leur auteur. La prise en compte de ce signal de réactivité constitue la principale motivation de cette section.

3. Pour l'analyse, la qualité d'une réponse est assimilée au nombre de votes positifs reçu de la commaunté.

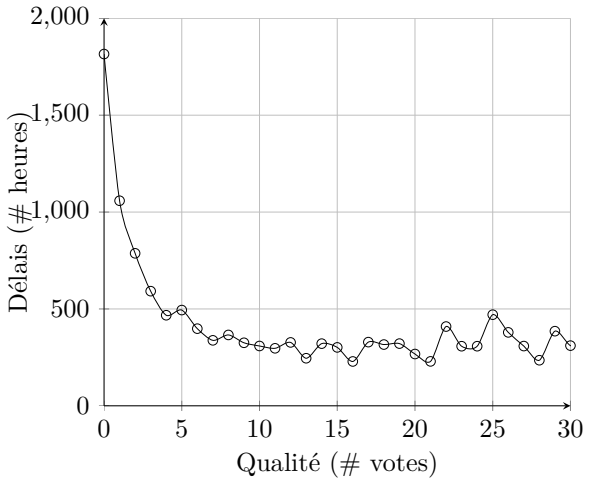

FIGURE 4.1 – Délais d'apparitions moyen des réponses en fonction de leur qualité calculés à partir de 99.5% des réponses publiées entre 2008 et 2014 sur la plateforme STACK EXCHANGE.

1.2 Représentation

La modélisation du signal d'autorité presenté en Section 1.1 est réalisée via la notion de *graphe de réactivité*.

Définition 1.2. (*Graphe de réactivité*) Un graphe de réactivité est un graphe orienté $G_R = (V, E)$ défini sur un ensemble V d'utilisateurs et une relation E de réactivité telle que $(i, j) \in E$ si et seulement si l'utilisateur i a répondu au moins une fois plus vite que j à une même question. La matrice d'adjacence A est telle que $A(i, j) \in \mathbb{R}^+$ est le nombre de questions auxquelles l'utilisateur i a répondu plus rapidement que l'utilisateur j.

Soit $G_R = (V, E)$ le graphe de réactivité construit à partir des données de la plateforme Stack Exchange. Soit $G_R^n = (V_n, E_n)$ le sous-graphe de G_R tels que E_n ne contient que les arcs dont le poids est supérieur à n et l'ensemble V_n ne contient que les nœuds dont le degré est non nul. Autrement dit, le graphe G_R^n est construit en supprimant de G_R tous les arcs dont le poids est strictement inférieur à n puis tous les sommets isolés.

Soit la Figure 4.1 traçant, en gris, la réputation moyenne des utilisateurs [4] associés aux nœuds restants dans les graphes G_R^n pour n variant

4. Les scores de réputation assignés par la platforme Stack Exchange sont utilisés.

de 0 à 30. Par ailleurs, en noir, est tracée la qualité moyenne des réponses apportées par l'ensemble des utilisateurs restant.

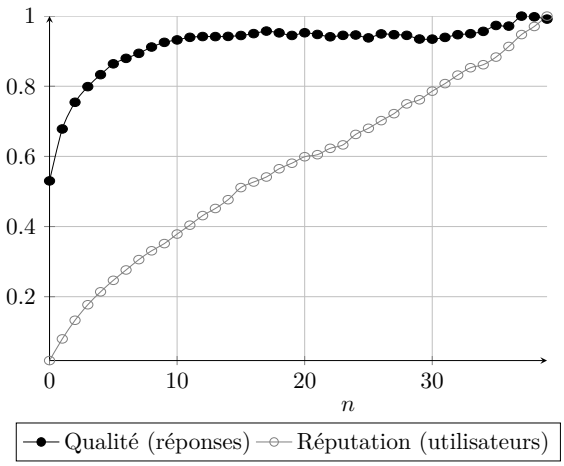

FIGURE 4.2 – Réputation moyenne des utilisateurs (en gris), ainsi que qualité moyenne des réponses produites (en noir), associés aux noeuds des graphes de réactivité G_R^n.

L'évolution globale de l'autorité des utilisateurs et de la qualité moyenne des réponses associées suggère à nouveau que les utilisateurs les plus réactifs, i.e., ceux dont la pondération des arcs sortant est la plus grande, sont à l'origine de réponses de qualité. L'exploitation de ce graphe de réactivité au sein de la modélisation STE 1 est décrite dans la section suivante.

1.3 Modélisation

Le graphe de réactivité $G_R = (V, E)$ est directement exploitable par le modèle présenté au Chapitre 3. Pour ce faire, la relation de réactivité E définie par la Définition 1.2 est considérée pour étendre le modèle STE 1. La nouvelle modélisation de la plateforme Stack Exchange est formellement proposée en Définition 2.6.

Définition 1.3. (STE 2) L'instance STE 2 est représentée par un graphe hétérogène $G = (H, E)$ d'ordre 4 défini sur V_1 l'ensemble des utilisateurs, V_2 l'ensemble des questions, V_3 l'ensemble des réponses et V_4 l'ensemble des commentaires. De plus, $T_E = 7$ avec E_{11} la relation de réactivité associée aux utilisateurs, E_{12}, E_{13} et E_{14} les relations d'auteur, E_{23}, E_{24}

et E_{34} les relations de liaison. Le modèle STE 2 est exprimé par :

$$\text{STE 2} \begin{cases} \mathbf{x}_1^{(t+1)} & = f_{\theta_{11}}(\mathbf{x}_1) + f_{\theta_{21}}(\mathbf{x}_2) + f_{\theta_{31}}(\mathbf{x}_3) + f_{\theta_{41}}(\mathbf{x}_4) \\ \mathbf{x}_2^{(t+1)} & = f_{\theta_{12}}(\mathbf{x}_1) + f_{\theta_{32}}(\mathbf{x}_3) \\ \mathbf{x}_3^{(t+1)} & = f_{\theta_{13}}(\mathbf{x}_1) \\ \mathbf{x}_4^{(t+1)} & = f_{\theta_{14}}(\mathbf{x}_1) \end{cases} \tag{4.1}$$

où les huit fonctions d'influence expriment l'influence :

- $f_{\theta_{11}}$ des utilisateurs ;
- $f_{\theta_{21}}$ des questions sur l'autorité d'un utilisateur ;
- $f_{\theta_{31}}$ des réponses sur l'autorité d'un utilisateur ;
- $f_{\theta_{41}}$ des commentaires sur l'autorité d'un utilisateur ;
- $f_{\theta_{12}}$ d'un utilisateur sur la qualité d'une question ;
- $f_{\theta_{32}}$ des réponses sur la qualité d'une questions ;
- $f_{\theta_{13}}$ de l'auteur sur la qualité d'un réponse ;
- $f_{\theta_{14}}$ de l'auteur sur la qualité des commentaires.

Une représentation graphique du modèle STE 2 est donnée en Figure 4.3.

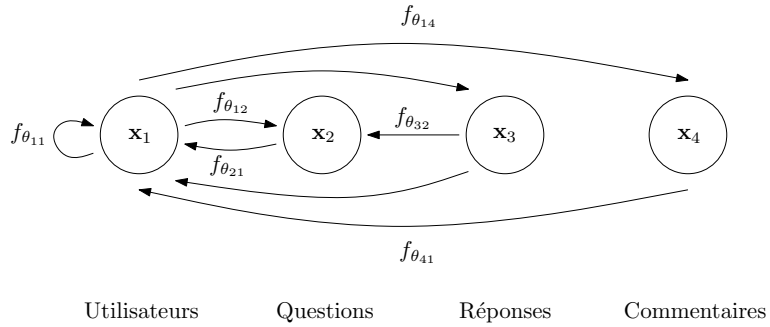

FIGURE 4.3 – Graphe d'influence du modèle STE 2.

On note la différence avec le modèle STE 1 pour lequel aucune relation entre utilisateur n'est considérée et aucune influence inter-utilisateurs.

2 Le processus de révision des Wikis collaboratifs

Le modèle WIKI 1 présenté dans la Section 2 du chapitre 3 se base sur une représentation très immédiate de la plateforme Wikipédia. Seuls deux ensembles d'entités sont considérés, les contributeurs d'une part et les articles d'autre part, ainsi qu'une relation d'éditions, associant à chaque auteur l'ensemble de ses contributions. De fait, les seuls

critères de qualité capturés au niveau des articles sont ceux associés au nombre de contributions réalisées par les utilisateurs. Pourtant, comme le suggère l'état de l'art réalisé au Chapitre 2, le processus de révision des articles permet d'extraire un important signal de qualité associé à l'ancienneté des mots des articles (Hu et al., 2007; Adler et al., 2008; de La Robertie et al., 2015; de La Robertie et al., 2017b).

Dans cette section, nous étendons cette hypothèse, en supposant que la qualité d'une portion de texte croît d'une part avec son ancienneté et aussi avec le nombre d'éditeurs ayant *approuvé* la portion de texte. La définition proposée pour le calcul de la qualité d'un élément textuel à partir de l'historique de révision des articles permet alors la considération simultanée des trois signaux de qualité que sont l'ancienneté d'une production, le nombre de *réviseurs* ainsi que leur activité au sein de l'article. Cette section est organisée de la façon suivante.

La Section 2.1 introduit certaines définitions utiles pour la formalisation du modèle. La Section 2.2 définit la représentation de la plateforme Wikipédia et formalise le modèle WIKI 2. Le détail des calculs liés aux fonctions d'influence utilisées est donné en Section 2.3. Enfin un rapprochement avec l'état de l'art est établi et discuté Section 2.4.

2.1 Définitions

Définition 2.1. (*Révision*) Une révision est la modification d'un article par des opérations d'insertion, de modification ou de suppression de portions de texte.

Une révision est associée à un unique utilisateur. En revanche, un utilisateur est généralement associé à plusieurs révisions dans un même article et/ou dans plusieurs articles. Par ailleurs, une même révision peut impacter plusieurs endroits dans le texte de l'article. Dans tous les cas, l'auteur d'une révision devient l'auteur de l'ensemble des portions de textes qu'il ajoute à l'article.

Définition 2.2. (*Réviseur*) L'auteur d'une révision d'un article est appellé un *réviseur*. Un réviseur *approuve implicitement* toutes les portions de textes de l'article qu'il ne modifie pas. Ces portions de textes sont alors *approuvées* par lui.

Définition 2.3. (*Ancienneté d'un segment*) L'ancienneté $t_i \in \mathbb{N}$ d'une portion de texte est calculée avec le nombre de révisions qui suivent son ajout dans l'article.

Selon la Définition 2.3, une portion de texte est implicitement approuvée par un contributeur de l'article si ce dernier est l'auteur d'au

moins une révision du même article effectuée *après* l'insertion de la séquence.

Définition 2.4. (*Voisinage antérieur*) Une portion de texte d'ancienneté t_j appartient au voisinage antérieur d'une portion de texte d'ancienneté t_i si et seulement $t_j \geq t_i$.

Définition 2.5. (*Voisinage postérieur*) Une portion de texte d'ancienneté t_j appartient au voisinage postérieur d'une portion de texte d'ancienneté t_i si et seulement $t_j \leq t_i$.

Il convient de préciser que tous les segments de texte modélisés dans la suite sont présents dans la dernière version des articles. Aussi, les mots ou paragraphes ayant été supprimés pendant le processus de révision des articles ne sont pas considérés dans la représentation. La notion de voisinage antérieur (resp. postérieur) permet, étant donnée une portion de texte, d'identifier les portions de textes qui ont été ajoutées à l'article avant (reps. après) celle-ci et qui figurent dans la dernière version de l'article. Une illustration des notions de voisinages antérieur et postérieur est donnée dans la Figure 4.4.

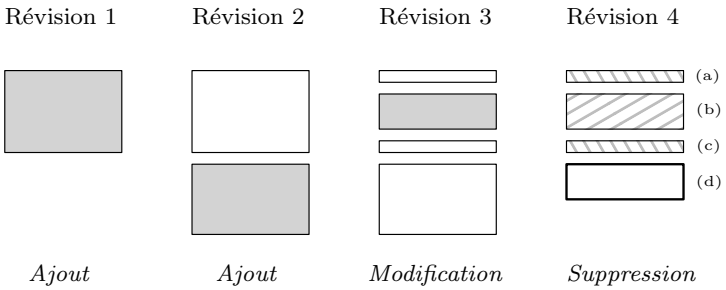

FIGURE 4.4 – Illustration des notions de voisinage antérieur et postérieur. Quatre révisions successives d'un article sont considérées. La version finale de l'article contient 4 portions de textes. L'ancienneté de la portion de texte (d), en gras sur la figure, à la révision 4, est 2 (ajoutée deux révisions auparavant). Les deux portions de textes (a) et (c) appartiennent au voisinage antérieur de (d). En effet, l'ancienneté de (a) et (c) vaut 3 (les deux poritions de texte ont été ajouté à l'article durant la première révision). La portion de texte (b) appartient au voisinage postérieur de (d). En effet, l'ancienneté de (b) vaut 1.

2.2 Modèle

La plateforme Wikipédia est représentée par un graphe hétérogène $G = (H, E)$ d'ordre 3 défini sur un ensemble des contributeurs V_1, un

ensemble de segments de textes V_2 et un ensemble d'articles V_3. Chaque contributeur est associé à l'ensemble de ses contributions par une relation E_{12}. De la même manière, chaque article est mis en relation avec l'ensemble des segments de texte qu'il contient par la relation E_{23}. Enfin, la relation E_{22} défini une relation de voisinage temporel à partir de la Définition 2.4 entre les portions de textes.

Définition 2.6. (WIKI 2) L'instance WIKI 2 est représentée par un graphe hétérogène $G = (H, E)$ d'ordre 3 défini sur V_1 l'ensemble des contributeurs, V_2 l'ensemble des segments de textes et V_3 l'ensemble des articles. De plus, $T_E = 4$ avec E_{11} une relation de co-édition, E_{12} une relation de liaison, E_{22} une relation de voisinage et E_{23} une relation d'appartenance. Le modèle WIKI 2 est exprimé par :

$$\text{WIKI 2} \begin{cases} \mathbf{x}_1 &= f_{\theta_{21}}(\mathbf{x}_2) \\ \mathbf{x}_2 &= f_{\theta_{12}}(\mathbf{x}_1) \\ \mathbf{x}_3 &= f_{\theta_{23}}(\mathbf{x}_2) \end{cases} \tag{4.2}$$

où les trois fonctions d'influence expriment l'influence :

- $f_{\theta_{12}}$ des réviseurs sur la qualité d'un segment ;
- $f_{\theta_{21}}$ des segments sur l'autorité d'un utilisateur ;
- $f_{\theta_{23}}$ des segments sur la qualité d'un article.

Une représentation du graphe d'influence est donnée en Figure 4.5.

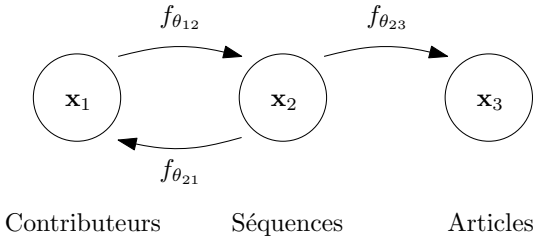

FIGURE 4.5 – Graphe d'influence du modèle WIKI 2.

2.3 Définition des influences

Les trois fonction d'influence $f_{\theta_{12}}$, $f_{\theta_{21}}$ et $f_{\theta_{23}}$, associées aux calculs d'autorité des contributeurs, de qualité des portions de textes et de qualité des articles sont instanciées dans cette section.

Qualité d'une portion de texte. La qualité d'une portion de texte est calculée à partir de l'autorité de l'ensemble des réviseurs du segment.

Plus précisemment, une combinaison $\phi : \mathbb{R}^2 \to \mathbb{R}$ de l'autorité de l'auteur de la séquence et de l'autorité des différents réviseurs de la séquence est proposée. Si z_i désigne l'auteur de la séquence $v_2^i \in V_2$, i.e., $(v_1^{z_i}, v_2^i) \in V_{12}$, alors la qualité de la séquence associée au nœud v_2^i est exprimée par l'Équation (4.3).

$$x_2^i = \sum_{(v_2^i, v_2^j) \in E_{22}} \phi(x_1^{z_i}, x_1^{z_j}) \tag{4.3}$$

Autorité d'un contributeur. L'autorité d'un utilisateur est calculée à partir de la qualité de l'ensemble des séquences révisées. Comme pour le calcul de qualité d'un segment, une combinaison de la qualité des séquences ajoutées et de la qualité des séquences révisées est proposée. Le calcul d'autorité d'un utilisateur associé au nœud v_1^i est formulé par l'Équation (4.4).

$$x_1^i = \sum_{(v_1^i, v_2^j) \in E_{12}} \sum_{(v_2^k, v_2^j) \in E_{22}} \phi(x_2^k, x_2^j) \tag{4.4}$$

Qualité d'un article. La qualité d'un article est calculée étant donnée la qualité de l'ensemble des séquences qu'il contient. Formellement,

$$x_3^j = \sum_{(v_2^i, v_3^j) \in E_{23}} x_2^i = \sum_{(v_2^i, v_3^j) \in E_{23}} \sum_{(v_2^i, v_2^j) \in E_{22}} \phi(x_1^{z_i}, x_1^{z_j}) \tag{4.5}$$

On note l'absence de fonction d'influence de la famille V_3 vers la famille V_2. Aussi, le modèle peut être résumé par l'Équation (4.6). La propagation des scores de qualité des articles pour le calcul de $\mathbf{x}_3$ pouvant être réalisée une seule fois à convergence.

$$\textsc{Wiki 2} \begin{cases} x_1^i &= \displaystyle\sum_{(v_1^i, v_2^j) \in E_{12}} \sum_{(v_2^k, v_2^j) \in E_{22}} \phi(x_2^k, x_2^j) \\ x_2^i &= \displaystyle\sum_{(v_2^i, v_2^j) \in E_{22}} \phi(x_1^{z_i}, x_1^{z_j}) \end{cases} \tag{4.6}$$

Dans les expérimentations, deux familles de fonctions ϕ sont étudiées, définissant les deux modèles suivants :

1. Wiki 2.1 lorsque $\phi(a, b) = a + b$. La fonction ϕ est la fonction somme. Dans ce cas, le modèle somme l'autorité de l'auteur d'une séquence avec celle du réviseur. La Section 2.4 montre que cette instance particulière généralise un travail de l'état de l'art ;

2. WIKI 2.2 lorsque $\phi(a,b) = (a \cdot b)^{\alpha(1-w_{ab})}$. La fonction ϕ réalise le produit de l'autorité de l'auteur d'une séquence avec l'autorité d'un réviseur en considérant la force de la relation de co-éditions entre les deux utilisateurs. Ici, w_{ab} est le nombre d'articles (normalisé) que les utilisateurs a et b ont co-édités. L'impact du paramètre utilisateur $\alpha \in [0,1]$ sera étudié en Section 4.

2.4 Discussion

Le modèle PEER REVIEW (Hu et al., 2007) est une solution de l'état de l'art exploitant le processus de révision des articles de façon à d'avantage créditer la qualité des mots les plus anciens avec l'autorité des différents réviseurs. Plus précisemment, la qualité d'un mot d'un article est calculée selon la somme de l'autorité des p distincts réviseurs du mot (incluant l'auteur lui même). Formellement, si $\{v_1^{z_i}\}_{i \leq p}$ est l'ensemble de ces p distincts réviseurs, la qualité d'un mot v_2^i est alors exprimée :

$$x_2^i = x_1^{z_1} + x_1^{z_2} + ... + x_1^{z_p} = \sum_{j=1}^{p} x_1^{z_j} \tag{4.7}$$

La proposition détaillée en Section 2.3, consistant à combiner l'autorité de l'auteur avec celle des réviseurs, est une généralisation de cette approche. En effet, si la fonction ϕ est la fonction somme, alors la qualité d'une portion de texte est exprimée par :

$$\begin{aligned} x_2^i &= \sum_{(v_2^i, v_2^j) \in E_{22}} \phi(x_1^{z_i}, x_1^{z_j}) \\ &= \sum_{(v_2^i, v_2^j) \in E_{22}} x_1^{z_i} + x_1^{z_j} \\ &= x_1^{z_i} + x_1^{z_{j_1}} + x_1^{z_i} + x_1^{z_{j_2}} + ... + x_1^{z_i} + x_1^{z_{j_s}} \\ &= s_i x_1^{z_i} + x_1^{z_{j_1}} + x_1^{z_{j_2}} + ... + x_1^{z_{j_s}} \end{aligned}$$

avec s_i le nombre de séquences qui ont été ajoutées à l'article après la portion de texte v_2^i, i.e., $s_i = |\{v_2^j : (v_2^i, v_2^j) \in E_{22}\}|$. En regroupant les scores d'autorité associés à un même contributeur, on obtient :

$$\begin{aligned} x_2^i &= \underbrace{s_i x_1^{z_i} + x_1^{z_{j_1}}}_{\alpha_1^i x_1^{z_1}} + \underbrace{x_1^{z_{j_2}} + x_1^{z_{j_3}}}_{\alpha_2^i x_1^{z_2}} + ... + \underbrace{x_1^{z_{j_s}}}_{\alpha_p^i x_1^{z_p}} \\ &= \alpha_1^i x_1^{z_1} + ... + \alpha_p^i x_1^{z_p} \\ &= \sum_{j=1}^{p} \alpha_j^i x_1^{z_j} \end{aligned}$$

avec $\alpha_j^i \in \mathbb{N}^+$ le nombre de séquences appartenant au voisinage postérieur de v_2^i et ajoutées par l'auteur $v_1^{z_j}$. Ainsi, en posant, $\alpha_j^i = 1, \forall j \leq p$, nous redéfinissons le modèle Peer Review (Hu et al., 2007). Dans notre approche, l'activité α_i^j des réviseurs est considérée pour pondérer leur autorité. Nous verrons dans la Section 6 que la prise en considération de l'activité des contributeurs permet d'améliorer significativement les résultats lors d'une tâche d'identification d'articles de qualité.

3 Conclusion

Nous avons proposé dans ce chapitre deux modèles intégrant des indices d'autorité et de qualité issus de données temporelles. En particulier, un graphe de réactivité construit à partir du flux des réponses des utilisateurs d'une plateforme de type Questions-Réponses a été construit et intégré dans la proposition, étendant le modèle proposé au chapitre précédant. Par ailleurs, l'historique des articles Wikipédia et plus particulièrement le processus de révision des articles a été modelisé via une relation de voisinage temporelle pour définir deux fonctions d'influence permettant la combinaison de l'autorité des différents réviseurs des éléments constituant les articles. Nous avons notamment démontré la généricité de l'approche en généralisant deux travaux de l'état de l'art.

Malgré la généricité de l'approche, la proposition, dans l'état actuel des choses, ne permet pas la considération du contenu des productions textuelles. Bien qu'il soit possible d'ajouter un ensemble d'entités modélisant par exemple le vocabulaire des productions textuelles et d'appliquer notre algorithme sur une telle représentation, les résultats ne rendraient pas compte d'une notion d'autorité ou de qualité par domaine d'expertise. Dans le chapitre suivant, nous proposons une nouvelle formulation permettant de représenter à la fois les domaines d'expertise des utilisateurs ainsi que les différents niveaux d'autorité associés calculés à partir d'indices de structures.

Chapitre 5

Intégration de données textuelles

Introduction

La solution proposée au Chapitre 3 est centrée sur la modélisation des notions de qualité et d'autorité sur les plateformes collaborative. L'approche exploite un graphe hétérogène d'influence et un principe de renforcement mutuel pour propager les signaux d'autorité de proche en proche sans tenir compte du contenu des productions. Si le Chapitre 4 propose une modélisation du contenu des articles, celle-ci reste abstraite car aucune information sémantique n'est réellement exploitée.

On propose dans ce chapitre de tenir compte de la sémantique apportée par l'ensemble des productions textuelles associées aux utilisateurs en vue d'effectuer un calcul d'autorité par domaine d'expertise. Pour ce faire, nous proposons une approche hybride, notée MAS-2, exploitant à la fois une représentation par les graphes des plateformes pour l'extraction des signaux d'autorité et le contenu des productions textuelles pour l'extraction des domaines d'expertise. La solution, formulée sous la forme d'un unique problème d'optimisation, permet un calcul simultané des notions d'autorité et d'expertise et donc la captation au niveau thématique des corrélations entre ces deux notions. Les profils d'autorité de domaines générés par la proposition permettent alors une caractérisation plus détaillée de la notion d'autorité et surtout la formulation d'une réponse au problème d'identification d'experts, c'est à dire d'utilisateurs faisant autorité dans un domaine d'expertise particulier. Aussi, deux problématiques sont abordées dans ce chapitre :

1. La définition et l'apprentissage d'une représentation conjointe des notions d'expertise et d'autorité des contributeurs étant donnés le contenu de l'ensemble de leurs productions et les différents signaux de qualité associés ;

2. L'exploitation des profils d'autorité de domaine calculés pour l'identification d'experts dans un domaine particulier exprimé sous la forme d'une requête textuelle ;

La Section 1 motive la proposition en illustrant l'intérêt d'une prise en compte de signaux d'autorité dans un calcul d'expertise. La Section 2 présente les représentations utilisées dans ce chapitre et introduit la notion d'*autorité de domaine*. Les Sections 3 et 4 formalisent le problème et détaillent sa résolution. La Section 5 exploite les solutions au prolème d'optimisation pour formuler un modèle d'ordonnancement et répondre à la tâche de recherche d'experts. Enfin, les Section 6 et 7 discutent la proposition et dressent les conclusions du chapitre.

1 Motivations

L'expertise utilisateur est généralement représentée dans la littérature par un vecteur unidimensionnel de $\mathbb{R}^p$, où p est le nombre de thématiques ou domaines d'expertises considérés. Chaque dimension du vecteur informe, par le poids réel qui lui est associé, du niveau d'expertise de l'utilisateur dans la thématique correspondante. Dans la pratique, les solutions de l'état de l'art exploitent le contenu textuel des productions associées à un utilisateur pour y extraire les différentes thématiques et quantifier, sur la seule base du vocabulaire utilisé, des différents niveaux d'expertise (Balog et al., 2012). Autrement dit, les profils d'expertise engendrés ne reflètent que le niveau d'utilisation des mots de vocabulaire associés aux thématiques. La notion d'expertise définie par la littérature et la représentation associée ne réflétent ni de notion d'autorité ni de notion qualité associés aux utilisateurs ou aux productions. Par conséquent, l'ordre induit sur les experts associés à une thématique particulière n'est pas fonction de la qualité de leurs productions.

Cette limitation est montrée dans l'Exemple 1.1 qui suggère une prise en compte de critères de qualité dans les modèles pour favoriser les utilisateurs à l'origine de productions de qualité.

Exemple 1.1. Soient $U = \{u_1, u_2\}$ un ensemble d'utilisateurs $D = \{d_1, d_2, d_3, d_4, d_5\}$ un ensemble de 5 documents définis sur un vocabulaire $V = \{w_1, w_2, w_3\}$. La qualité des documents est donnée par le vecteur de scores $\mathbf{x} = (0.35, 0.25, 0.15, 0.15, 0.1)$. De plus, on considère que l'utilisateur u_1 est l'auteur des documents d_1 et d_2, et que l'utilisateur u_2 est l'auteur des documents d_3, d_4 et d_5. Cette

information, ainsi que le modèle de langue associé aux documents sont définis par les matrices suivantes :

$$p(d_j|u_i) = \begin{array}{c} \\ d_1 \\ d_2 \\ d_3 \\ d_4 \\ d_5 \end{array} \begin{array}{c} \begin{array}{cc} u_1 & u_2 \end{array} \\ \begin{pmatrix} 1/2 & 0 \\ 1/2 & 0 \\ 0 & 1/3 \\ 0 & 1/3 \\ 0 & 1/3 \end{pmatrix} \end{array} \text{ et } p(w_i|d_j) = \begin{array}{c} \\ d_1 \\ d_2 \\ d_3 \\ d_4 \\ d_5 \end{array} \begin{array}{c} \begin{array}{ccc} w_1 & w_2 & w_3 \end{array} \\ \begin{pmatrix} 0.5 & 0.2 & 0.3 \\ 0.5 & 0.25 & 0.25 \\ 0.5 & 0.1 & 0.4 \\ 0.5 & 0 & 0.5 \\ 0.5 & 0.3 & 0.2 \end{pmatrix} \end{array} \tag{5.1}$$

Le modèle centré candidat (Balog et al., 2012) postule que les utilisateurs sont ordonnés, étant donnée une requête q, selon la probabilité

$$p(q|u_i) = \prod_{w \in q} p(w|u_i)$$

avec

$$p(w|u_i) = \sum_{d_j \in D} p(w|d_j)p(d_j|u_i)$$

Si la requête est $q = \{w_1\}$, les probabilités sont les suivantes :
- $p(w_1|u_1) = \sum_{d_j \in D} p(w_1|d_j)p(d_j|u_1) = 2(0.5 \times 1/2) = 0.5$
- $p(w_1|u_2) = \sum_{d_j \in D} p(w_1|d_j)p(d_j|u_2) = 3(0.5 \times 1/3) = 0.5$

Le modèle centré candidat ne tenant pas compte de la qualité des documents d_j considère que les utilisateurs u_1 et u_2 sont autant expert l'un que l'autre dans la thématique w_1. Or, les documents de l'utilisateur u_1 sont de meilleure qualité que ceux de l'utilisateur u_2. Il conviendrait alors d'attribuer un score plus grand à l'utilisateur u_1 qu'à l'utilisateur u_2 dans la thématique w_1, i.e., $p(q|u_1) > p(q|u_2)$.

2 Représentations

La représentation proposée dans ce chapitre se base sur un graphe hétérogène d'ordre 3 modélisant un ensemble de producteurs, un ensemble de documents et un ensemble de mots-clefs. Bien que d'autres descripteurs, comme les tags, spécifiques à certains types de documents pourraient être utilisés, ce dernier ensemble, par soucis de simplicité, est uniquement constitué du vocubulaire engendré par l'ensemble des documents. Pour relier les éléments du graphe, les deux relations suivantes sont considérées :
- une relation d'auteur entre les utilisateurs et leurs productions ;
- une relation de liaison entre les documents et les mots du vocabulaire

qu'ils contiennent.
Enfin, comme source de signaux de qualité, une relation de citation est considérée entre les documents. La Définition 2.1 formalise cette représentation, illustrée en Figure 5.1.

Définition 2.1. (MAS 2) L'instance MAS 2 est représentée par un graphe hétérogène $G = (H, E)$ d'ordre 3 définit sur V_1 l'ensemble des chercheurs, V_2 l'ensemble des articles et V_3 l'ensemble du vocabulaire. Par ailleurs, $T_E = 3$, avec $E = \{(V_1, V_2), (V_2, V_2), (V_2, V_3)\}$ où E_{12} est une relation d'auteur, E_{22} une relation de citation et E_{23} une relation de liaison.

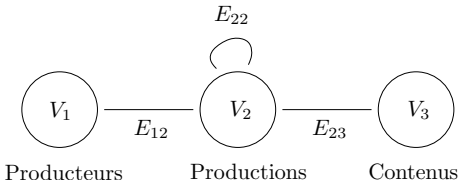

FIGURE 5.1 – Graphe des familles d'entité du modèle MAS 2.

Les modèles introduits dans les chapitres précédant propagent, via la notion de fonctions d'influence, les scores d'autorité et de qualité dans le graphe d'entités. Dans ce chapitre, c'est une *autorité de domaine* qui est calculée, notion formalisée en Définition 2.2, pour modéliser à la fois l'autorité et l'expertise d'une entité dans un même espace de représentation.

Définition 2.2. (*Autorité de domaine*) Le vecteur d'autorité de domaine de l'entité $v_k^j \in V_k$ est le vecteur $\mathbf{x}_k^j \in \mathbb{R}^p$ quantifiant à la fois l'autorité et l'expertise de v_k^j sur chacune des p thématiques. En particulier, $\mathbf{x}_k^j(i) \in \mathbb{R}$ est le score d'autorité de domaine de l'entité v_k^j pour la thématique i.

Dans notre représentation, $\mathbf{x}_1^j$ quantifie donc à la fois l'expertise et l'autorité de l'utilisateur j dans chacune des thématiques. Par exemple, $x_1^j(i)$ reflète à quel point l'utilisateur j est à l'origine de documents de qualité dans la thématique i. L'extraction des thématiques est discuté dans la section suivante. Il convient de préciser que $\mathbf{x}_k \in M_{n_i p}(\mathbb{R})$ pour $k \in \{1, 2, 3\}$.

3 Modèle

Le calcul de l'autorité de domaine des documents $\mathbf{x}_2$ est formulé par un problème d'optimisation combinant deux objectifs, l'un lié à l'extraction des thématiques, à partir de la relation E_{23}, l'autre à l'extraction des signaux d'autorité, à partir de la relation E_{22}. Formellement, pour un nombre $p \in \mathbb{N}$ fixé de thématiques, les vecteurs d'autorité de domaine sont solutions au problème d'optimisation suivant :

$$\begin{array}{ll} \min & \mathcal{L}_\lambda(\mathbf{x}_2, \mathbf{x}_3) \\ & \text{s.c. } \mathbf{x}_2 > 0, \mathbf{x}_3 > 0 \end{array} \tag{5.2}$$

où

$$\mathcal{L}_\lambda(\mathbf{x}_2, \mathbf{x}_3) = \lambda||A_{23} - \mathbf{x}_2 \cdot \mathbf{x}_3^T||_F^2 + (1-\lambda)||A_{22} \cdot \mathbf{x}_2 - \mathbf{x}_2||_F^2 \tag{5.3}$$

La valeur du paramètre utilisateur $\lambda \in [0, 1]$ permet de contrôler la sensibilité des deux critères. En particulier, pour $\lambda = 1$, le problème est réduit à un problème de factorisation en matrices non négatives. Dans ce contexte, seules les thématiques sont extraites à partir de la relation A_{23}. Lorsque $\lambda = 0$, l'expertise n'est plus considérée et seule la qualité des articles, via la relation de citation E_{22}, est calculée. La combinaison de ces critères est discutée en Section 6.

A convergence, l'autorité de domaine d'un utilisateur est alors calculée à partir de l'autorité de domaine des documents dont il est l'auteur. Formellement, si $\mathbf{x}_2^*$ est la solution au problème d'optimisation formulée par l'Équation (5.2), alors on calcule $\mathbf{x}_1 = A_{12}\mathbf{x}_2^*$.

4 Calculs

La fonction objectif $\mathcal{L}_\lambda$ étant convexe, la minimisation du problème d'optimisation formulé par l'Équation (5.2) peut être réalisée par une descente de gradient. Les dérivées partielles de la fonction $\mathcal{L}_\lambda$ par rapport aux variables $\mathbf{x}_2$ et $\mathbf{x}_3$ sont exprimées par l'Équation (5.4).

$$\begin{array}{l} \dfrac{\partial \mathcal{L}_\lambda}{\partial \mathbf{x}_2} = 2\left((1-\lambda)D(X_{22} \cdot \mathbf{x}_2) - \lambda(X_{23} - \mathbf{x}_2 \cdot \mathbf{x}_3^T)\mathbf{x}_3\right) \\ \dfrac{\partial \mathcal{L}_\lambda}{\partial \mathbf{x}_3} = -2\lambda(X_{23} - \mathbf{x}_2 \cdot \mathbf{x}_3^T)^T\mathbf{x}_2 \end{array} \tag{5.4}$$

avec $D = \text{diag}(X_{22}\mathbb{1} - 1)$. Il convient de préciser que les paramètres du modèle $\mathbf{x}_2$ et $\mathbf{x}_3$ ont $n_2 p$ et $n_3 p$ variables de décisions respectivement. L'ensemble des paramètres du modèle définit donc une métrique dans

$\mathbb{R}^{p(n_2 n_3)}$. La proposition, instanciée sur le jeu de données Microsoft Academic Search contenant 10^6 articles, $|V_3| = 500$ mots de vocabulaires et $p = 20$ thématiques, définit donc un modèle à 10^{11} variables de décision (100 milliards de variables). Dans la pratique, l'optimisation est réalisée avec l'algorithme L-BFGS (Liu and Nocedal, 1989) qui optimise le pas de gradient à chaque itération.

5 Modèle d'ordonnancement

Soit $\mathbf{x}_1(k) \in \mathbb{R}^{n_1}$ le vecteur colonne associé à la thématique $k \leq p$ extrait de la matrice $\mathbf{x}_1 \in M_{n_1 p}$ informant du niveau d'autorité de domaine des n_1 utilisateurs dans la thématique k. Soit σ, la permutation sur l'ensemble des utilisateurs induite par le trie en ordre décroissant de scores du vecteur $\mathbf{x}_1(k)$. Soit $q \subseteq V_3$ une requête de domaine exprimée par un ensemble de mots sur le vocabulaire V_3. L'ordonnancement des utilisateurs en réponse à la requête q est induit par le vecteur $\mathbf{x}_1(k)$ associé à la thématique $k*$ qui maximise la probabilité d'apparition de la requête. Formellement :

$$k* = \arg\max_k p(q|k) \tag{5.5}$$

avec

$$p(q|k) = \prod_{v_3^i \in q} p(v_3^i|k) = \prod_{v_3^i \in q} x_3^{ik}$$

Aussi, $\forall (v_1^i, v_1^j) \in V_1 \times V_1,\ \sigma(i) \leq \sigma(j) \iff x_1^{ik} \geq x_1^{jk}$.

Exemple 5.1. Soit $p = 3$ le nombre de thématiques et $\mathbf{x}_1 \in M_{n_1 p}$ la matrice des autorités de domaine suivante :

$$\mathbf{x}_1 = \begin{array}{c} \\ v_1^1 \\ v_1^2 \\ v_1^3 \end{array} \begin{array}{c} \begin{array}{ccc} k_1 & k_2 & k_3 \end{array} \\ \begin{pmatrix} 1.8 & 0.2 & 0 \\ 1 & 0.2 & 0.8 \\ 0 & 0.4 & 1.6 \end{pmatrix} \end{array}$$

Soit la matrice $\mathbf{x}_3 \in M_{n_3 k}$ encodant la probabilité $p(v_3^i|k)$:

$$\mathbf{x}_3^T = \begin{array}{c} \\ k_1 \\ k_2 \\ k_3 \end{array} \begin{array}{c} \begin{array}{ccccccc} v_3^1 & v_3^2 & v_3^3 & v_3^4 & v_3^5 & v_3^6 & v_3^7 \end{array} \\ \begin{pmatrix} 0.5 & 0.2 & 0.2 & 0.1 & 0 & 0 & 0 \\ 0.05 & 0.05 & 0 & 0 & 0.2 & 0.3 & 0.4 \\ 0.25 & 0 & 0.15 & 0.4 & 0.2 & 0 & 0 \end{pmatrix} \end{array}$$

Soit $q = (v_3^1, v_3^2)$ la requête. Le calcul de la probabilité $p(q|k)$ est détaillé de la façon suivante :

$$\begin{array}{c} \\ p(v_3^1|k) \\ p(v_3^2|k) \\ p(v_3^1|k)p(v_3^2|k) \end{array} \begin{array}{c} \begin{array}{ccc} k_1 & k_2 & k_3 \end{array} \\ \begin{pmatrix} 0.5 & 0.05 & 0.25 \\ 0.2 & 0.05 & 0 \\ 0.1 & 0.025 & 0 \end{pmatrix} \end{array}$$

Ainsi,

$$\arg\max_k \prod_{v_3^i \in q} p(v_3^i|k) = \arg\max_k p(v_3^1|k)p(v_3^2|k) = 1$$

avec $p(v_3^1|k)p(v_3^2|k) = 0.5 \times 0.2 = 0.1$. Les utilisateurs v_1^1, v_1^2 et v_1^3 sont donc triés par ordre décroissant de scores donnés par $\mathbf{x}_1(1)$, c'est à dire :

$$\mathbf{x}_1(1) = \begin{array}{c} \\ v_1^1 \\ v_1^2 \\ v_1^3 \end{array} \begin{array}{c} \begin{array}{c} k_1 \end{array} \\ \begin{pmatrix} 1.8 \\ 1 \\ 0 \end{pmatrix} \end{array}$$

La permutation renvoyée est alors $\sigma = (1, 2, 3)$ avec

$$x_1^{\sigma(1),1} \geq x_1^{\sigma(2),1} \geq x_1^{\sigma(3),1}$$

6 Discussion

La fonction objectif proposée regroupe deux objectifs. Le premier objectif, $f(A_{23}, \mathbf{x}_2, \mathbf{x}_3) = ||A_{23} - \mathbf{x}_2\mathbf{x}_3^T||_F^2$ formule un problème de factorisation de matrices dans le but de regrouper de façon homogène les documents sur la base des mots qu'ils contiennent. Autrement dit, cet objectif est responsable de l'extraction des thématiques latentes abordées dans les documents. Plus particulièrement, la minimisation de l'objectif f en gardant fixée la matrice A_{23} permet les deux calculs suivants :

1. La caractérisation des documents selon les thématiques extraites. La variable $\mathbf{x}_2 \in M_{n_2 p}(\mathbb{R})$ décrit alors le degré d'appartenance de chaque document à chaque thématique ;

2. La caractérisation des thématiques selon le vocabulaire. La variable $\mathbf{x}_3 \in M_{n_3 p}(\mathbb{R})$ décrit alors le degré d'appartenance de chaque mot du vocabulaire à chaque thématique.

Ces deux caractéristiques permettent de reconstruire les représentations initiales des documents. En particulier, le document i représenté par la ligne i de la matrice A_{23} est reconstruit par $\sum_{k=1}^{p} \mathbf{x}_2^i(k)\mathbf{x}_3(k)$.

Le second terme de la fonction objectif, $g(A_{22}, \mathbf{x}_2) = ||A_{22}\mathbf{x}_2 - \mathbf{x}_2||_F^2$, formule un problème de réduction de dimension qui, lorsque $p = 1$, conduit aux même solutions que l'approche proposée dans le Chapitre 3. En effet, supposons la qualité d'un article influencée par la qualité des articles le citant et posons $\mathbf{x}_2 = f_{\theta_{22}}(\mathbf{x}_2)$. Si l'influence est linéaire, c'est à dire si $f_{\theta_{22}} = A_{22}$, le système suivant doit être résolu :

$$\mathbf{x}_2 = A_{22}\mathbf{x}_2 \tag{5.6}$$

Une solution non nulle à l'Équation (5.6) est un point fixe de l'application engendrée par la matrice A_{22}, qui, par définition, vérifie

$$\mathbf{x}_2^* = A_{22}\mathbf{x}_2^* \iff \mathbf{x}_2^* - A_{22}\mathbf{x}_2^* = 0 \tag{5.7}$$

Ou, de façon équivalente

$$||\mathbf{x}_2^* - A_{22}\mathbf{x}_2^*||_F = 0 \tag{5.8}$$

Ainsi, le calcul des scores de qualité des documents initialement formulé par l'Équation (5.6) est traduit en un problème d'optimisation convexe. Il convient de préciser que la résultante d'une combinaison linéaire de deux fonctions convexes est convexe, permettant la résolution du problème par une descente de gradient.

7 Conclusion

Les modèles de l'état de l'art utilisés pour représenter l'expertise des utilisateurs se basent essentiellement sur le contenu de leurs productions textuelles. La notion même d'expertise semble avoir été définie uniquement à partir des mots d'un vocabulaire, sans considération de la qualité des productions textuelles exploitées. Pourtant, il parait important de considérer la qualité des productions dans un modèle d'expertise.

Ce chapitre répond à cette limitation en proposant une représentation unique exprimant à la fois la notion d'expertise telle que définie dans la litérature et celle d'autorité. Cette représentation, appelée autorité de domaine, est calculée en minimisant un problème d'optimisation formulé à partir d'un graphe hétérogène modélisant à la fois les producteurs, leurs produtions ainsi que le contenu de ces productions. Plus spécifiquement, la fonction objectif proposée combine deux objectifs de l'état de l'art,

l'un associé à un problème de factorisation de matrices pour l'extraction des thématiques et donc la caractérisation de l'expertise des utilisateurs, l'autre associé à un problème de réduction de dimension pour le calcul des scores de qualité des articles à partir de la relation de citations. Par ailleurs, les solutions au problème d'optimisation introduit sont réutilisées pour répondre à un problème de recherche d'experts et ainsi identifier les utilisateurs à l'origine de productions de qualité dans une thématique donnée.

Chapitre 6

Validation des propositions et expérimentations

Introduction

Ce chapitre est consacré à la validation de nos propositions. Nous proposons pour ce faire une analyse des résultats aux expérimentations conduites sur troix jeux de données réels associés à différentes tâches d'identification de contenus de qualité ou d'individus faisant autorité. Plus précisément, les modèles proposés dans ce manuscrit sont évalués sur les quatres tâches suivantes :

1. Une tâche d'identification des articles de qualité sur Wikipédia. En particulier, nous évaluons la capacité de la proposition à ordonner les articles par ordre de qualité décroissante. Nous verrons dans le Section 1 décrivant les données utilisées qu'une échelle ordinale mise en place par la communauté Wikipédia permet de comparer les prédictions des modèles à une verité terrain représentative de la qualité des articles en ligne ;
2. Une tâche d'identification des réponses de qualité sur le site de Questions-Réponses Stack Exchange. On se propose d'ordonner, pour chaque question publiée sur la plateforme, les réponses associées par ordre de qualité décroissante. L'évaluation et la comparaison avec les compétiteurs se feront sur la base des votes associés aux réponses donnés par la communauté des internautes ;
3. Une tâche d'identification de chercheurs faisant autorité sur la librairie digitale Microsoft Academic Search. La validation quantitative, via une liste d'experts fournie par le service en ligne AMiner, sera appuyée par une analyse qualitative des résultats liée à la qualité des publications des chercheurs retournés par les modèles ;

4. Enfin, une tâche d'identification d'experts, à nouveau réalisée sur les données de la plateforme Microsoft Academic Search. Les performances des modèles seront à nouveau évaluées en comparant les résultats obtenus avec la liste d'experts classée selon dix domaines d'expertise par le service en ligne AMiner.

L'organisation de ce chapitre est la suivante. Dans un premier temps, les trois jeux de données utilisés pour les expérimentations (Wikipédia, Stack Exchange et Microsoft Academic Search) sont présentés dans la Section 1. Ensuite, dans la Section 2, sont présentées les métriques d'évaluations pour comparer les propositions aux solutions de l'état de l'art. Dans la Section 3, sont présentés les résultats sur les quatres tâches introduites. La Section 4 présente des résultats supplémentaires obtenus sur les plateformes Wikipédia et Stack Exchange. Nous étudions notamment dans cette section l'impact du paramètre utilisateur α du modèle Wiki 2.2 ainsi que l'impact de la taille de l'ensemble $|\mathcal{P}|$ représentant l'ensemble des connaissances *a priori* sur la performance des modèles. Enfin, nos conclusions et critiques sont apportées dans la Section 5.

1 Jeux de données

Les trois jeux de données réelles suivants ont été utilisés pour évaluer les performances de nos modèles :

- Une collection d'environ 20 000 articles et 100 000 utilisateurs issus du site collaboratif Wikipédia [1] détaillée en Section 1.1 ;
- Un corpus comprenant l'intégralité des données publiées sur une période de cinq ans par le site de Questions-Réponses Stack Exchange [2] détaillé en Section 1.2 ;
- Un jeu de données contenant environs 1.5 millions d'articles et 130 000 chercheurs issus de la librairie digitale Microsoft Academic Search [3] détaillé en Section 1.3.

1.1 Wikipedia

Le jeu de données Wikipédia décrit dans cette section est utilisé pour évaluer les modèles Wiki 1, Wiki 2.1 et Wiki 2.2 sur une tâche d'identification d'articles de qualité. Les données brutes, les critères d'évaluation et les étapes de prétraitement sont détaillées ci-dessous.

1. https ://fr.wikipedia.org
2. https ://stackexchange.com/
3. http ://academic.research.microsoft.com/

1.1.1 Description des données

Plus de 20 000 articles Wikipédia, écrits en langues Anglaise, ont été utilisés pour valider nos propositions. Le choix de ces articles a été guidé par le projet Maine de Wikipédia [4], apportant une classification de ces articles le long d'une échelle de qualité ordinale. Un outil de crawling a donc été développé pour récupérer, dans un premier temps, depuis la page du projet Maine, les identifiants uniques associés à chacun de ces articles ainsi que les labels de qualité associés. Dans un second temps, via l'API Wikipédia, ont été récupérés les historiques complets de ces articles, c'est à dire l'ensemble ordonné des différentes versions des contenus des articles depuis leur première révision. Par ailleurs, l'ensemble des contributeurs associés à chacune des révisions des articles a aussi été enregistré. L'ensemble de ces données brutes, principalement textuelles, représente environs 130 Go.

1.1.2 Vérité terrain

Pour l'évaluation, l'échelle de cotations des articles établie par l'équipe éditoriale du projet Maine de Wikipédia [5] a été utilisée. Celle-ci, établie en fonction de divers critères de qualité, d'exhaustivité ou de présentation des articles est régulièrement mise à jour. A la date de rédaction de ce manuscrit, l'échelle de cotation comporte 7 niveaux de qualité croissants allant de la classe Stub, représentant les articles de pauvre qualité ou à des ébauches d'articles, à la classe FA, attribuée aux articles complets et profesionnels et donc de haute qualité. Les différents niveaux de qualité de l'échelle de cotation utilisée sont détaillés en Figure 6.1. Notons qu'un article est associé à une unique classe de qualité, du moins à un instant donné de son évolution. Dans notre cas, les labels de qualité informent de la qualité des articles à la date de crawling.

Plus formellement, à chaque article appartenant à notre jeu de données, est associé un label $y_i \in \mathcal{Y}$, où $\mathcal{Y} = \{S, C, B, GA, A, FA\}$ est l'ensemble ordonné représentant l'echelle de cotation Wikipédia. En particulier, la préférence utilisateur induite sur la qualité des articles est la suivante :

$$FA \succ A \succ GA \succ B \succ C \succ S$$

Pour l'automatisation du protocole d'évaluation, nous associons un nombre entier, de 0 à 5, à chaque classe de S à FA.

4. https ://en.wikipedia.org/wiki/Wikipedia :WikiProject_Maine/Assessment
5. https ://en.wikipedia.org/wiki/Template :Grading_scheme

FA	Featured article
A	Article is well written, reasonably complete and referenced; possible featured article candidate.
GA	This is a good article.
B	We have a decent article, but it needs further editing.
C	Improvement is needed with content provided
Start	Significant cleanup or expansion needed.
Stub	Article is either a very short, or rough.

FIGURE 6.1 – Échelle de qualité des articles Wikipédia extraite de la page https //en.wikipedia.org/wiki/Template :Grading_scheme.

1.1.3 Prétraitements

La relation entre l'ensemble des séquences de texte d'un article et l'ensemble des contributeurs est construite à partir des révisions successives d'un article. Puisqu'à chaque révision est associé un unique contributeur, toutes les modifications réalisées pendant une révision lui sont attribuées. L'évolution des modifications du contenu d'un article permet alors d'identifier l'auteur de chacune des portions de texte. Un outil de comparaison de contenu a donc été développé pour suivre l'évolution des portions de texte d'un article et mettre à jour la relation d'auteur au fil du processus de révision. Le graphe de collaboration est alors construit étant donnée cette relation d'auteur, autrement dit, à partir des productions présentes dans la dernière version des articles. Les statistiques des données résultant de cette étape de prétraitement sont résumées dans le Tableau 6.1. En particulier, environ 110 000 utilisateurs sont à l'origine de près de 3 millions de portions de textes. Le graphe de collaboration, établit donc entre 110 000 utilisateurs, comporte environ 5 millions d'arcs.

Classe	FA	A	GA	B	C	S
Articles	245	51	346	1 012	1 946	18 823
Auteurs par article*	61	37	37	46	41	7
Séquences*	1 114	963	809	695	439	40
Age d'une séquence*	275	166	154	141	126	12
Taille d'une séquence*	78	44	86	85	88	73
Collaborations (10^3)	546	62	406	1 505	2 279	369

TABLE 6.1 – Statistiques des données Wikipédia (*en moyenne par classe) après prétraitement.

1.2 Stack Exchange

Le jeu de données Stack Exchange présenté est utilisé pour évaluer les modèles STE 1 et STE 2 sur une tâche d'identification des réponses de qualité sur les sites de Questions-Réponses. Les données utilisées et les critères d'évaluation sont détaillés ci-dessous.

1.2.1 Description des données

L'ensemble des données, rendues publiques par le portail Stack Exchange, couvrant la période Octobre 2008 - Septembre 2014 [6] est utilisé pour évaluer nos propositions. En particulier, les données, au format XML, regroupent les activités des utilisateurs de plus de 109 sites appartenant à Stack Exchange. Au total, c'est plus d'1.5 millions de questions, 2.5 millions de réponses et 6.5 millions de commentaires qui ont été produits par environ un million d'utilisateurs. Le graphe de réactivité établi à partir de ces données compte près de 3 millions d'arcs.

1.2.2 Vérité terrain

La plupart des sites de Questions-Réponses permet à ses utilisateurs d'intervenir sur l'ordonnancement des réponses via un système de notations. Plus les utilisateurs votent pour une réponse, plus celle-ci remonte dans l'affichage des résultats. D'une certaine manière, les votes d'une réponse reçus par la communauté d'utilisateurs reflètent alors sa qualité et sa pertinence vis-à-vis de la question posée. Nous utilisons donc le nombre de votes qu'a reçus une réponse comme vérité terrain. Plus précisemment, nous comparons, pour chaque question, l'ordonnancement induit par les scores de qualité retournés par notre modèle à celui induit par les votes des réponses.

Dans le jeu de données utilisé, les votes des réponses sont des scores numériques et prennent leur valeur dans l'intervalle $[-65, 2\,182]$. L'histogramme de répartition des votes des réponses est donnée en Figure 6.2 et les principales valeurs de la fonction de répartition associée sont données dans le Tableau 6.2. On remarque que la moitié des réponses publiées sur la plateforme ont un score inférieur ou égal à 2 et que seulement 1% des réponses publiées ont un score supérieur à 24.

Pour automatiser l'évaluation, les scores des réponses sont discrétisés selon deux échelles de valeurs. Une première échelle binaire, notée b_s, associe à toutes les réponses dont le score est négatif la classe des réponses de mauvaise qualité ($y_i = 0$) et à toutes les réponses dont le score est

6. http ://blog.stackoverflow.com/2009/06/stack-overflow-creative-commons-data-dump/

Scores (X)	1	2	4	7	11	24	34	2182
$p(X \leq x)$	0.25	0.5	0.75	0.9	0.95	0.99	0.995	1

TABLE 6.2 – Fonction de répartition des votes des réponses de la plateforme Stack Exchange.

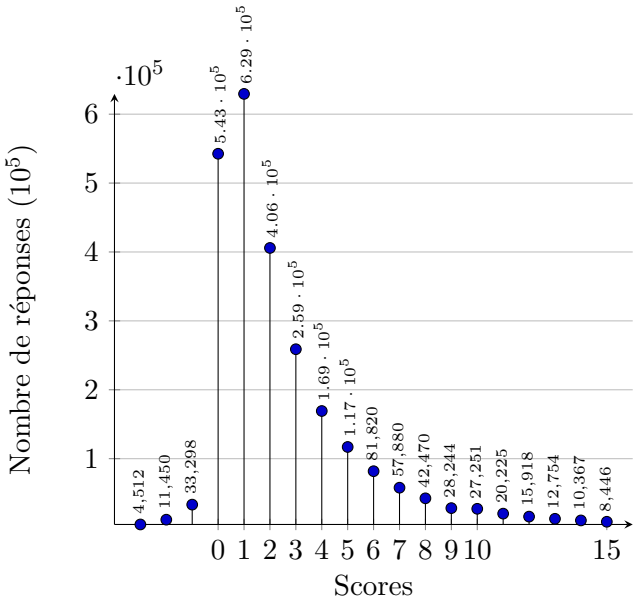

FIGURE 6.2 – Histogramme de répartition des scores des réponses issues de la plateforme de Question/Réponses STACK EXCHANGE. Plus de 98% des données sont représentées.

positif la classe des réponses de bonne qualité ($y_i = 1$). La deuxième échelle, plus expressive, notée r_s, comporte 5 niveaux de qualité relativement homogènes détaillés dans le Tableau 6.3.

En utilisant l'échelle binaire, la capacité des modèles à séparer les réponses de qualité des réponses de mauvaise qualité est essentiellement évaluée. En utilisant la seconde échelle ordinale, on évalue la capacité des modèles à ordonner les réponses par ordre de qualité décroissante.

1.3 Microsoft Academic Search

Le jeu de données Microsoft Academic Search [7] détaillé dans cette section est utilisé pour évaluer le modèle de qualité MAS 1 ainsi que le

7. http ://academic.research.microsoft.com/

Classe	A	B	C	D	E
Votes	$]-\infty,-1]$	$\{0\}$	$\{1\}$	$\{2,3\}$	$]3,\infty[$
Réponses	52 540	542 562	629 443	629 443	651 825
Label (y_i)	0	1	2	3	4

TABLE 6.3 – Discrétisation des votes sur la plateforme Stack Exchange.

modèle d'autorité de domaine AD sur les deux tâches d'identification de chercheurs faisant autorité et d'experts sur les librairies digitales.

1.3.1 Description des données

Tous les articles, auteurs et conférences de la communauté scientifique en informatique disponibles sur la plateforme Microsoft Academic Search sont utilisés pour construire les modèles. En particulier, ce sont environ 1.5 millions d'articles, 130 000 chercheurs et plus de 2 000 conférences qui sont utilisés pour construire les modèles. Il convient de préciser que seul le résumé des articles est récupéré, représentant alors 4 Go de données textuelles.

1.3.2 Vérité terrain

Les deux tâches sur le jeu de données Microsoft Academic Search sont évaluées sur la même vérité terrain construite à partir de la liste d'experts publiée par le service en ligne AMiner [8]. Parmi les 1 270 experts publiés, classés dans 10 thématiques d'expertise différentes, plus de 900 ont été retrouvés dans la base de données Microsoft Academic Search. La réconciliation entre les chercheurs présents dans la base Microsoft Academic Search et ceux listés par AMiner est réalisée en utilisant un ensemble d'expressions régulières comparant à la fois les noms, les prénoms et les initiales des chercheurs.

Formellement, à chaque chercheur du jeu de données, est assigné un vecteur de labels $\mathbf{y} \in \mathbb{B}^{10}$, informant si oui $\mathbf{y}(j) = 1$ ou non $\mathbf{y}(j) = 0$ le chercheur en question est expert dans la thématique associée à la dimension j. Il convient de préciser que certains chercheurs sont experts dans plusieurs thématiques. Le nombre d'experts par thématique est synthétisé dans le Tableau 6.5.

8. https ://aminer.org/

1.3.3 Connaissance *a priori*

Le site Core.edu [9] fournit une classification des principales conférences scientifique en informatique selon 4 niveaux de qualité. En particulier, ce sont plus de 2 000 conférences, listées par le site Core.edu et étiquetées avec les labels A*, A, B ou C, qui ont été retrouvées dans le jeu de données Microsoft Academic Search. Le label A* est reservé aux grandes conférences internationales connues pour leur exigences en terme d'attente de qualité des articles soumis. À nouveau, un ensemble d'expressions régulières a été utilisé pour réaliser le processus de matching entre les noms et les initiales des conférences des deux jeux de données Core.edu et Microsoft Academic Search.

Formellement, à chaque conférence est alors associé un label $y_i \in \mathcal{Y}$, où $\mathcal{Y} = \{A^*, A, B, C\}$ vérifie

$$A^* \succ A \succ B \succ C$$

Cette connaissance sur la qualité des conférences sera utilisée dans nos modèles comme connaissance extérieure ou *a priori*. En outre, elle permettra d'apporter une évaluation qualitative à propos de la qualité des articles associés aux chercheurs retournés par les modèles. La distribution des conférences par classe de qualité est résumée dans le Tableau 6.4.

Catégorie	**A***	**A**	**B**	**C**
Conférences (%)	5.88	19.37	31.9	42.82
Articles (10^3)	138.9	264.1	262.2	261.2
Label (y_i)	4	3	2	1

TABLE 6.4 – Répartition des articles par catégorie de conférences dans le jeu de données utilisé pour les évaluations.

1.3.4 Prétraitements

À partir des données brutes décrites précédemment, dix graphes d'expertise, utilisés pour les évaluations, ont été construits. Soit k un domaine d'expertise, le graphe G_k est construit en ajoutant successivement à l'ensemble des nœuds :

1. les experts dans la thématique k ;
2. les co-auteurs de ces experts ;

9. http ://www.core.edu.au/

3. les articles publiés par l'union des experts et de leur co-auteurs ;
4. les articles qui citent ou qui sont cités par l'ensemble des articles précédemment ajoutés ;
5. les auteurs de tous les articles ajoutés.

De plus, un graphe plus large, G_{exp}, est reconstruit, fusionnant les 10 graphes d'expertise ainsi créés, pour l'évaluation de la tâche de recherche d'experts. Les statistiques sur les 11 graphes construits sont résumées dans le Tableau 6.5.

Domaine d'expertise	**Experts**	**Chercheurs**	**Articles**
G_0 - Boosting	43	52 228	94 172
G_1 - Fouille de données	221	86 786	243 071
G_2 - Extraction d'informations	72	36 880	80 983
G_3 - Agents intelligents	28	36 323	60 246
G_4 - Apprentissage	52	37 277	69 025
G_5 - Traitement du langage	36	20 175	36 684
G_6 - Alignement d'ontologies	42	30 216	48 601
G_7 - Planification	13	22 809	32 710
G_8 - Web sémantique	274	81 039	244 855
G_9 - SVM	70	33 448	60 319
G_{exp} - Tous	851	131 303	1 427 317

TABLE 6.5 – Les 11 sous-graphes d'expertise construits à partir du jeu de données Microsoft Academic Search utilisés pour les expérimentations.

Par ailleurs, le contenu des articles est traité avec la librairie Python Natural Language Toolkit [10]. Les noms sont extraits des résumés et titres des articles pour constituer un premier vocabulaire. En particulier, ceux présents dans plus de 80% et moins de 10% des articles sont supprimés, jugés peu discriminant. Au total, environs 5 000 noms, radicalisés avec l'outil Lancaster Stemmer, constituent l'ensemble du vocabulaire utilisé pour la description des articles. La force de la relation entre les articles et les mots est estimée grâce à la métrique TF-IDF.

2 Métriques d'évaluation

Les performances des modèles sont évaluées avec les métriques usuelles issues du domaine de la Recherche d'Information. En particulier, nous

10. http ://www.nltk.org/

utilisons les métriques Précision, Rappel, AUC et NDCG pour évaluer la qualité des ordonnancements des modèles.

1. **Rappel**. Le rappel à k, noté R@k, traduit combien d'éléments recherchés sont effectivement renvoyés par le modèle. Pour un rang k donné, le rappel à k mesure, parmi les k premiers éléments renvoyés par le modèle, combien d'éléments sont pertinents au regard du nombre total d'éléments pertinents dans le jeu de données. Formellement, si N_p est le nombre d'élements pertinents total qu'il conviendrait de retrouver et N_k est le nombre d'éléments pertinents figurant parmi les k premiers éléments de l'ordonnancement, alors

$$R@k = \frac{N_k}{N_p} \tag{6.1}$$

   Lorsque $R@k$ vaut 1, tous les éléments pertinents sont renvoyés par le modèle. Il convient donc de maximiser le rappel.

2. **Précision**. La précision à k, notée P@k, mesure le bruit d'un modèle en calculant le ratio entre le nombre d'éléments pertinents dans l'ensemble renvoyé par le modèle et le nombre d'éléments renvoyés par le modèle. Formellement, pour un rang k donné, la précision est calculée par :

$$P@k = \frac{N_k}{k} \tag{6.2}$$

   Lorsque $P@k$ vaut 1, tous les éléments renvoyés par le modèle sont pertinents. Il convient donc de maximiser la précision.

3. **AUC**. Informellement, la métrique AUC estime la probabilité de placer un élément pertinent, choisi aléatoirement, devant un élément non pertinent. Elle correspond à l'air sous la corbe ROC, construite à partir des deux indicateurs classiques TVP (Rappel ou Sensibilité) et TFP (1-Spécificité). Si le modèle place tous les éléments pertinents devant les autres, AUC=1.

4. **NDCG**. Etant donné un rang k, le NDCG@k renvoie un réel entre 0 et 1 traduisant le rapport entre la qualité de l'ordonnancement retourné par le modèle et la qualité de l'ordonnancement optimal en ne considérant que les k premiers éléments. Formellement, si σ est la permutation retournée par le modèle à évaluer et $\sigma(i)$ est la position de l'élément i, alors :

$$NDCG(\sigma, k) = \frac{DCG(\sigma, k)}{DCG(\sigma^*, k)} \tag{6.3}$$

où la qualité d'une permutation σ au rang k est définie par

$$DCG(\sigma, k) = \sum_{i=1}^{k} \frac{2^{y_{\sigma(i)}} - 1}{\log(1+i)} \tag{6.4}$$

avec y_j le score associé à l'étiquette de l'élément j et σ^* la permutation optimale sur l'ensemble à trier. En particulier, la valeur de la métrique NDCG@k(σ^*, k) vaut 1 à tout rang k.

3 Validation des propositions

Les résultats aux expérimentations sont regroupés par jeu de données. Dans la Section 3.1, nous présentons les résultats aux expérimentations réalisées sur le jeux de données Microsoft Academic Search concernant les deux tâches d'identification de chercheurs faisant autorité et d'identification d'experts en réponse à une requête de domaine. Dans la Section 3.2, sont présentés les résultats associés au jeu de données Stack Exchange relatifs à la tâche d'identification de réponses de qualité sur les sites de Questions-Réponses. Enfin, les résultats aux expérimentations réalisées sur le jeu de données Wikipédia sur la tâche d'identification d'articles de qualité sont présentés dans la Section 3.3.

3.1 Microsoft Academic Search

Les modèles proposés sont évalués sur les deux tâches suivantes :

1. *Tâche d'identification de chercheurs faisant autorité.* Le modèle MAS-1 décrit en Section 2 du chapitre 3 est utilisé pour répondre à cette problématique. Deux variantes de la proposition sont évaluées. La première n'intégre pas de connaissance extérieure. L'algorithme 1, décrit en page 55, est donc utilisé. La seconde intégre les signaux *a priori* sur la qualité de certaines conférences. L'algorithme 2, décrit en page 58, est alors utilisé. Dans les deux cas, les scores retournés par les modèles sont utilisés pour trier par ordre décroissant de score d'autorité les chercheurs de la librairie. Les ordonnancements obtenus sont alors évalués avec les métriques présentées en Section 2 ;

2. *Tâche de recherche d'experts.* Nous répondons à cette seconde problématique en utilisant le modèle MAS-2 présenté dans la Section 3 du chapitre 5. En particulier, une fois le modèle appris sur le graphe d'expertise complet G_{exp}, le modèle d'ordonnancement présenté dans la Section 5 du chapitre 5 est utilisé pour ordonnner, en fonction des dix requêtes de domaines, les profils des chercheurs.

3.1.1 Compétiteurs

Sur la première tâche d'identification de chercheurs faisant autorité, nous comparons notre approche à trois applications de l'algorithme HITS (Kleinberg, 1999) sur trois représentations différentes de la librairie. L'objectif premier est de montrer l'intérêt de préciser les fonctions d'influence entre les familles d'entités plutôt que les relations entre elles. De plus, la proposition est comparée à l'ordonnancement obtenu en utilisant comme score d'autorité le score de h-index des chercheurs, calculé par la librairie Microsoft Academic Search. Pour résumer, les quatre compétiteurs suivants sont considérés :

- **Hi**. Les chercheurs sont triés par ordre décroissant de scores h-index. La répartition des valeurs de h-index dans le jeu de données utilisé est donnée en Figure 6.3 ;
- **Co**. Les chercheurs sont triés par ordre décroissant de scores calculés par l'algorithme HITS appliqué au graphe de collaboration des chercheurs. On rappelle que le graphe de collaboration des chercheurs est le graphe $G = (V, E)$ défini sur l'ensemble V des chercheurs et l'ensemble E définissant la relation de co-auteur. En particuler, $(i, j) \in E$ si les chercheurs associés aux nœuds i et j sont co-auteurs d'au moins un article. La matrice d'adjacence associée à la relation E est définie à partir du nombre d'articles que chaque paire de chercheurs a co-écrit ;
- **Cit**. L'algorithme HITS est appliqué au graphe de citation des articles. Le score d'un chercheur est alors calculé selon la somme des scores associés aux articles qu'il a écrits. On rappelle que le graphe de citation est le graphe $G = (V, E)$ défini sur l'ensemble V des articles et l'ensemble E des relations de citations. En particulier, $(i, j) \in E$ si l'article i cite l'article j. La matrice d'adjacence associée est binaire ;
- **HITS**. L'algorithme HITS est appliqué sur le graphe hétérogène tel que modélisé par la proposition MAS-1 décrite dans la Section 2 du chapitre 3. Les nœuds et les relations sont homogénéisées. Les chercheurs sont alors triés par ordre décroissant de scores obtenus.

De plus, toujours dans le cas de la première tâche, les deux variantes suivantes de la proposition décrite par la Définition 2.2 à la page 52 sont évaluées :

- **MAS 1-A**. Aucune préférence utilisateur n'est spécifiée en entrée de l'algorithme. Autrement dit, aucune correction des paramètres des fonctions d'influence n'est réalisée pendant la propogation et seule la phase de propagation des scores est effectuée. L'algorithme 1, décrit en page 55, est donc utilisé ;

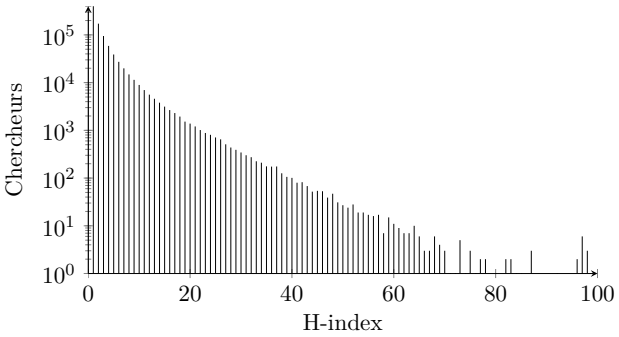

FIGURE 6.3 – Répartition des scores de h-index des chercheurs du jeu de données Microsoft Academic Search.

- **MAS 1-B**. Une préférence utilisateur est précisée en entrée de l'algorithme sur un sous ensemble des conférences. A chaque itération, les paramètres des fonctions d'influences sont alors corrigés de façon à minimiser la fonction de coût empirique exprimée par l'Equation (3.17). L'algorithme 2, décrit en page 58, est donc utilisé.

Sur la seconde tâche de recherche d'experts, nous comparons notre proposition aux approches de l'état de l'art suivantes :

- **NMF**. Lorsque $\lambda = 1$, la proposition résoud un problème de factorisation en matrices non négatives équivalent à l'algorithme des k-moyennes appliqué à la matrice A_{23}. Les articles sont regroupés en groupes homogènes en fonction du vocabulaire utilisé. Les chercheurs sont alors caractérisés par la somme des profils associés à leurs articles ;

- **Cos**. Le modèle classique de Recherche d'Information supppose que le score d'expertise d'un chercheur, étant donné une requête q, est la similarité de q avec le profil du chercheur. Pour aligner nos travaux avec les recherches antérieures, le profil d'un chercheur est construit en utilisant une concaténation des articles écrits par lui. Les requêtes et les profils des chercheurs sont représentés par des sacs de mots. Une pondération TF-IDF est utilisée ;

- **Cm** (Balog et al., 2012). Le modèle génératif centré candidat utilise un formalisme basé sur les modèles de langue. Etant donné une requête q, les chercheurs sont triés par ordre décroissant de probabilité $p(q|\theta_i) = \prod_{w \in q} p(w|\theta_i)$, avec θ_i le modèle de langue associé à l'utilisateur i. En particulier, $p(w|\theta_i) = \sum_{d \in D} p(w|d)p(d|u_i)$;

- **Cms** (Balog et al., 2012). Il s'agit d'une version lissée du modèle CM précédent. Les probabilités sont lissées par les fréquences d'apparition

des termes de la collection. Formellement, $\tilde{p}(w|d) = \alpha p(w|d) + (1 - \alpha)p(w|D)$, avec $\alpha = 0.5$.

3.1.2 Résultats

Identification de chercheurs faisant autorité

Analyse quantitative. Les résultats aux évaluations pour cette première tâche d'identification de chercheurs faisant autorité sur les librairies digitales évaluée avec les métriques P@k, R@k et NDCG@k sont présentés dans les Figure 6.4, Figure 6.5 et Figure 6.6 respectivement. Par ailleurs, la qualité des ordonnancements sur les conférences pour lesquelles nous disposons de connaissances *a priori* produits par les compétiteurs est aussi évaluée et détaillée dans le Tableau 6.6.

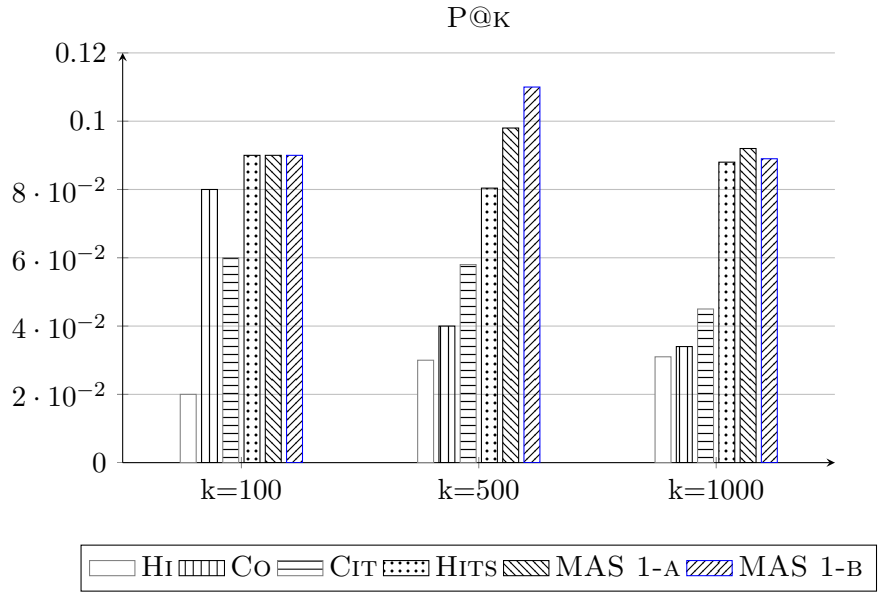

FIGURE 6.4 – Précision@k des modèles d'identification de chercheurs faisant autorité sur la librairie Microsoft Academic Search.

D'une façon générale, les résultats aux évaluations de nos propositions MAS 1-A et MAS 1-B sont légèrement meilleurs. En particulier, l'apport en connaissance *a priori* sur la qualité des conférences place le modèle MAS 1-B systématiquement devant les autres compétiteurs selon la métrique NDCG à différents rangs k d'évaluation. L'ordonnancement produit par la contribution sur l'ensemble des chercheurs semble donc s'approcher davantage de la vérité terrain. Par ailleurs, les résultats présentés dans le Tableau 6.6 confirment que l'ordonnancement obtenu

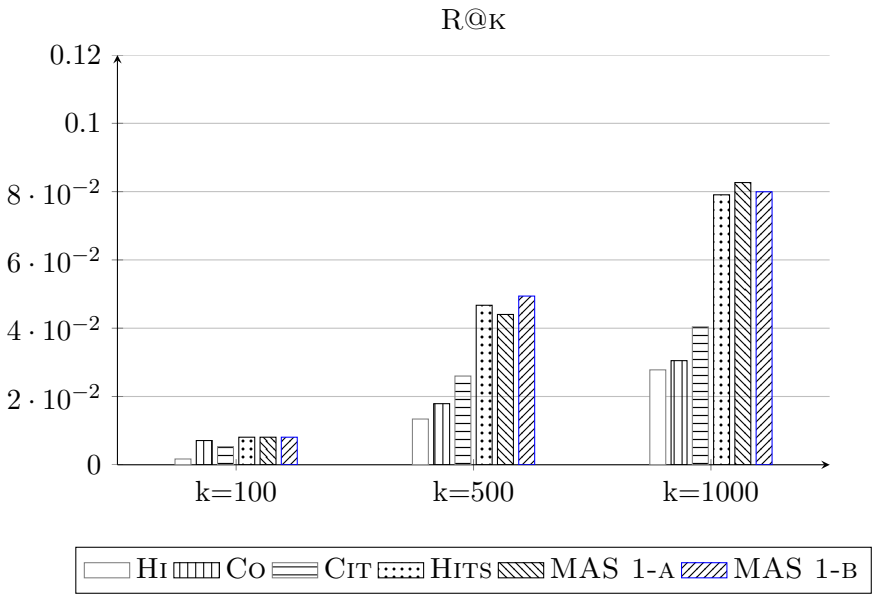

FIGURE 6.5 – Rappel@k des modèles d'identification de chercheurs faisant autorité sur la librairie Microsoft Academic Search.

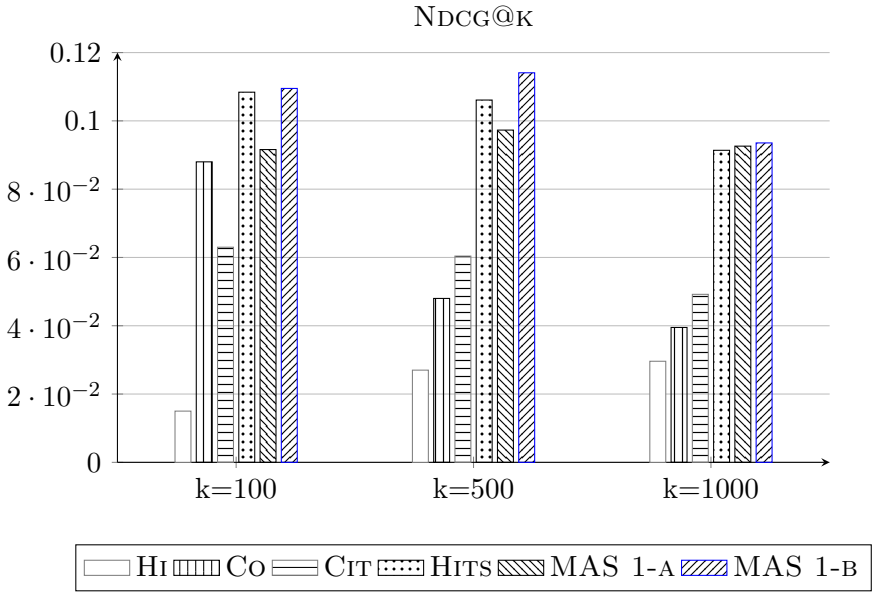

FIGURE 6.6 – NDCG@k des modèles d'identification de chercheurs faisant autorité sur la librairie Microsoft Academic Search.

sur l'ensemble des conférences labelisées est de meilleure qualité lorsque

Classe	A*	B	C	S
Hi	0.4919	0.6355	0.7570	0.8754
Co	0.4191	0.5578	0.7034	0.8422
Cit	0.5123	0.6362	0.7666	0.8813
Hits	0.6429	0.6974	0.7890	0.8899
MAS 1-a	0.5466	0.5730	0.7476	0.8628
MAS 1-b	**0.6979**	**0.7572**	**0.8882**	**0.9344**

Table 6.6 – Evaluation des ordonnancement sur les conférences du jeu de données Microsoft Academic Search avec la métrique Ndcg@k.

l'algorithme exploite l'ensemble des connaissances *a priori* $\mathcal{P}_3$. En effet, nous pouvons constater que le modèle Mas 1-b entraine un gain de plus 30% par rapport à Mas 1-a.

Il est intéressant de constater que la précision du modèle Mas 1-a est très compétitive face à l'algorithme HITS. En effet, pour $k = 500$ et $k = 1000$, la précision de la proposition, sans connaissance extérieure donc, dépasse les performances obtenues avec l'algorithme HITS. Ces premiers résultats suggèrent donc un premier intérêt à exploiter le graphe d'influence entre les familles d'entités plutôt qu'un graphe basé sur les relations.

Par ailleurs, rappelons que le nombre de multiplications réalisées par la proposition Mas 1-a est bien inférieur à celui effectué par l'algorithme HITS. En effet, le nombre de fonctions d'influence, d'après la Section 5 du Chapitre 3, passe, de 9, pour l'algorithme HITS, à 6, pour le modèle Mas 1-a. Surtout, chacune des 9 fonctions intervenant dans le calcul de HITS réalise au moins un produit entre matrices, alors que dans notre proposition, exactement un produit entre une matrice et un vecteur est réalisé, réduisant drastiquement le nombre d'opérations élémentaires réalisées.

Analyse qualitative. La vérité terrain issue du site AMiner utilisée pour l'évaluation automatique quantitative peut être discutée. Aussi, nous étudions la qualité des publications des chercheurs renvoyés par les différents modèles. Les différents ensembles constitués des top-5 chercheurs renvoyés par les modèles étudiés sont listés dans le Tableau 6.7. Le Tableau 6.8 recense le nombre de publications produit par l'ensemble des différents groupes de 5 chercheurs par qualité de conférence.

Tout d'abord, nous constatons que les 5 chercheurs retrouvés par nos deux contributions sont à l'origine d'un plus grand nombre de publications en général mais surtout à l'origine d'un plus grand nombre

	HI	CIT	CO	HITS	MAS 1-A	MAS 1-B
S. Shenker	•					
J-M. Dricot	•					
H. Garcia-Molina	•					
J. D. Ullman	•					
I. T. Foster	•					
M. R. Garey		•				
D. S. Johnson		•				
T. H. Cormen		•				
C. E. Leiserson		•				
R. L. Rivest		•				
D. Dubois			•			
H. Prade			•			
W. Wang			•	•	•	
P. S. Yu			•	•	•	
T. S. Huang			•	•	•	•
D. F. Towsley				•		•
J. J. Dongarra				•	•	•
E. Bertino					•	•
S. Chawla						•

TABLE 6.7 – Top-5 chercheurs retrouvés par les modèles étudiés.

Classe	**A***	**A**	**B**	**C**	**Total**
CO	377	407	294	340	1419
CI	290	328	112	86	816
HITS	381	483	309	248	1421
MAS 1-A	384	490	**353**	305	1532
MAS 1-B	**410**	**518**	347	309	**1584**

TABLE 6.8 – Nombre de publications produites par catégorie de conférence par les top-5 chercheurs retrouvés par les différents modèles évalués.

de publications de qualité, i.e., publiées dans des conférences de qualité. En effet, les chercheurs retrouvés par le modèle MAS 1-B ont produits 410 publications dans des conférences de rang A^* et 518 dans des conférences de rang A, soit une trentaine en moyenne de plus que tous les compétiteurs. Par ailleurs, on note que même la proposition MAS

1-A, n'exploitant pas de connaissance *a priori*, identifie un ensemble de cinq chercheurs produisant davantage d'articles de qualité que les compétiteurs CO, CI et HITS. D'après le Tableau 6.8, la différence avec le modèle HITS viendrait du chercheur D.F.Towsley, remplacé par le chercheur E.Bertino par notre proposition.

Identification d'experts

Analyse quantitative. Les performances des cinq compétiteurs évalués avec la métrique AUC sur les 10 requêtes de domaine sont regroupées dans le Tableau 6.9. On rappelle que les performances des modèles ont toutes été évaluées sur un vocabulaire de taille 5 000 mots. Les meilleurs résultats, par domaine, sont en gras dans le tableau.

Topic query	NMF	COS	CM	CMS	MAS 2
Boosting	0.829	0.703	0.703	0.703	**0.842**
Data Mining	0.671	0.664	0.635	**0.682**	0.681
Information Extraction	0.607	0.601	0.676	**0.696**	0.623
Intelligent Agents	0.628	0.766	0.676	**0.771**	0.717
Machine Learning	0.745	0.622	0.553	0.635	**0.781**
Language Processing	0.464	0.488	0.492	0.487	**0.567**
Ontology Alignments	0.386	0.492	0.499	0.492	**0.512**
Planning	0.837	0.607	0.617	0.617	**0.904**
Semantic Web	0.541	0.648	0.550	**0.651**	0.622
Support Vector Machine	0.723	0.712	**0.786**	0.779	0.743
Moyenne	0.643	0.646	0.619	0.651	**0.699**

TABLE 6.9 – Performances du modèle MAS 2 évaluée avec la métrique AUC sur les dix requêtes de domaines.

En moyenne, on observe que la proposition MAS 2 améliore les résultats de l'état de l'art, avec une valeur de la métrique AUC à 0.699 contre 0.562 pour les autres compétiteurs. En particulier, on observe un gain de 7% à utiliser la proposition par rapport au meilleur compétiteur CMS faisant état de l'art. Cela dit, on constate que sur les 10 requêtes de domaines testées, la proposition n'est meilleure que sur 5 d'entre elles. Les cinq requêtes Boosting, Machine Learning, Language Processing, Ontology Alignments et Planning produisent des meilleurs résultats avec notre proposition. En revanche, les requêtes Data Mining, Information Extraction, Intelligents Agents, Semantic Web et Support Vector Machine réalisent des meilleurs performances avec le modèle de l'état de

l'art Cms ou Cm.

Analyse qualitative. Nous étudions à nouveau la qualité des articles publiés par les top-5 chercheurs retournés par les modèles de recherche d'experts. Pour cela, la répartition des publications des différents groupes de 5 chercheurs est publiée dans le Tableau 6.10.

Model	**A***	**A**	**B**	**C**
Nmf	20.83	41.66	22.91	14.58
Cos	21.90	34.28	28.25	15.55
Cm/Cms	21.54	31.64	27.60	19.19
Mas 2	**40.85**	**35.10**	15.74	8.29

Table 6.10 – Répartition des publications (en %) des top-5 chercheurs retrouvés par les différents modèles en fonction du type de conférence.

Les résultats concordent avec les résultats quantitatifs. En effet, la proposition permet d'identifier un sous ensemble de cinq chercheurs dont plus 75% des publications sont associées à des conférences A^* ou A. En particulier, les chercheurs retournés par le modèle Mas 2 publient 40% de leurs articles dans des conférences A^* lorsque les chercheurs retournés par les compétiteurs en publient environ 20%. Enfin, on observe que le groupe de chercheurs retournés par notre proposition publie moins de 10% de leurs articles dans des conférences de rang C, lorsque cette valeur pour les autres groupes tourne autour de 15%. Cette analyse qualitative permet alors de confirmer, du moins de suggérer, le réel intérêt de la proposition et, en particulier, celui d'utiliser notre représentation d'expertise de domaine pour les profils utilisateurs.

3.2 Stack Exchange

Cette section présente les résultats aux expérimentations conduites sur le jeu de données Stack Exchange sur la tâche d'identification des réponses de qualité. La tâche évaluée consiste donc à ordonner, pour chaque question publiée sur la plateforme, les réponses associées par ordre de qualité décroissante. On rappelle que les résultats sont moyennés sur l'ensemble des questions.

3.2.1 Compétiteurs

Nous comparons nos propositions aux deux méthodes de l'état de l'art suivantes :

- **HITS** (Jurczyk and Agichtein, 2007). Le premier compétiteur est une application directe de l'algorithme HITS sur un graphe simple ne modélisant que les utilisateurs de la plateforme. À chaque utilisateur de la plateforme est associé un nœud et il existe un arc orienté d'un utilisateur i vers un utilisateur j si l'utilisateur associé au nœud j a répondu au moins une fois à une question formulée par l'utilisateur i ;
- **NCR** (Zhang et al., 2014). Le second compétiteur est une méthode structurelle non supervisée définie explicitement pour une plateforme de type Question/Réponses modélisée par exactement 3 types d'entités. Les auteurs définissent un algorithme de propagation de scores le long des différentes relations entre utilisateurs, questions et réponses. Les détails de l'approche sont décrits dans la Section 2.2.3 du Chapitre 2.

Nous rappelons que les deux propositions évaluées dans cette section sont :

- **Ste 1**. Ce premier modèle, décrit dans la Section 2 du chapitre 3, étend la proposition NCR (Zhang et al., 2014) en intégrant un quatrième ensemble d'entités (les commentaires) ;
- **Ste 2**. Notre second modèle, décrit dans le chapitre 4, étend aussi la proposition NCR en considérant d'une part un quatrième ensemble d'entités (les commentaires) ainsi qu'un graphe de réactivité entre les utilisateurs.

3.2.2 Résultats

Dans le Tableau 6.11 sont présentés les résultats aux expérimentations évalués selon la métrique Précision. L'évaluation de l'ordonnancement des réponses avec la métrique NDCG sur les deux échelles introduites figure dans le Tableau 6.12.

L'intérêt de la proposition est immédiat. Pour les deux métriques de classification et d'ordonnancement, la proposition améliore les résultats des modèles de l'état de l'art HITS et NCR. En l'occurence, comparé au meilleur compétiteur (NCR), nous observons des améliorations proches de 0.5% en précision. Par exemple, pour $k = 4$, les résultats passent de 0.7989 à 0.8055, ce qui étant donnée la taille du jeu de données, correspond à environ 7 000 réponses de qualité supplémentaires correctement identifiées par notre proposition. Par ailleurs, on constate que le modèle Ste 2 est toujours légèrement meilleur que Ste 1, suggérant un intérêt à utiliser le graphe de réactivité pour identifier les utilisateurs faisant autorité et donc les réponses de qualité.

En général, ces résultats montrent que considérer de nouvelles relations entre les différentes familles d'entités de la plateforme permet d'améliorer l'identification des utilisateurs faisant autorité et donc les contenus de qualité. En particulier, la relation de réactivité, injectée dans le système d'influence apporte des signaux d'autorité supplémentaires que notre proposition parvient à propager dans le reste du graphe pour identifier davantage de réponses de qualité.

Modèle	k=**2**	k=**3**	k=**4**	k=**5**	k=**10**	k=**20**
HITS	0.8141	0.8038	0.7914	0.7785	0.4992	0.2663
NCR	0.8252	0.8128	0.7989	0.7831	0.5000	0.2665
STE 1	0.8276	0.8149	0.7992	0.7871	0.5004	0.2666
STE 2	**0.8283**	**0.8185**	**0.8055**	**0.7877**	**0.5010**	**0.2666**

TABLE 6.11 – Evaluation de la précision des modèles d'ordonnancement de réponses sur Stack Exchange.

Modèle		k=**2**	k=**3**	k=**4**	k=**5**	k=**10**	k=**20**
HITS	b_s	0.8838	0.8889	0.9013	0.9247	0.9501	0.9539
	r_s	0.6727	0.7126	0.7564	0.8029	0.8521	0.8598
NCR	b_s	0.8922	0.8964	0.9081	0.9300	0.9537	0.9572
	r_s	0.6933	0.7307	0.7726	0.8126	0.8621	0.8691
STE 1	b_s	0.8938	0.8992	0.9115	0.9327	0.955	0.9582
	r_s	0.6949	0.7359	0.7785	0.8208	0.8641	0.8705
STE 2	b_s	**0.8947**	**0.8996**	**0.9119**	**0.9331**	**0.9552**	**0.9585**
	r_s	**0.6967**	**0.7447**	**0.7795**	**0.82016**	**0.8648**	**0.8711**

TABLE 6.12 – Evaluation de l'ordonancement des modèles sur Stack Exchange avec la métrique NDCG.

3.3 Wikipedia

Cette section présente et discute les résultats aux évaluations des modèles WIKI 1 et WIKI 2 sur la tâche d'identification d'articles de qualité sur Wikipédia. Une fois les modèles appliqués sur les différentes représentations introduites, les articles sont triés par ordre décroissant de score, puis, les métriques sont ensuite utilisées pour comparer l'ordonnancement obtenu avec l'ordonnancement optimal.

3.3.1 Compétiteurs

Nous comparons nos propositions aux trois approches suivantes de l'état de l'art :

- **Naive** (Blumenstock, 2008). Les articles sont triés par ordre décroissant de taille. Dans notre cas, nous trions les articles selon le nombre de séquences qu'ils contiennent ;
- **Basic** (Hu et al., 2007). La représentation immédiate de la plateforme Wikipédia est utilisée. Un graphe biparti est défini entre l'ensemble des contributeurs et l'ensemble des articles. Une seule relation d'auteur est considérée et la matrice d'adjence associée est telle que $A(i, j)$ est le nombre de mots que l'utilisateur i a ajouté dans l'article j ;
- **Peer** (Hu et al., 2007). Le processus de révision par les pairs est exploité mais l'autorité de l'auteur d'une séquence est calculé comme une somme non pondérée des autorité des réviseurs qui ont approuvés la séquence (voire Section 2.4) ;
- **Prob** (Hu et al., 2007). Ce modèle étend la version précédente en modifiant la façon dont l'autorité d'un réviseur influe sur l'autorité d'un auteur. Le calcul de l'autorité d'un auteur devient une fonction décroissante de l'autorité des réviseurs avec la distance qui sépare la séquence de l'auteur et celle du réviseur. L'hypothèse qu'un réviseur a davantage de chance d'éditer une séquence à proximité des siennes est exploitée. Formellement, la fonction de lissage $f(d) = \frac{1}{\max(0, d-\beta)+1}$ (avec d la distance, en nombre de mots, entre les séquences de l'auteur et du réviseur) est utilisée pour pondérer l'influence de l'autorité d'un réviseur sur l'autorité d'un auteur. Pour les expérimentations, la valeur du paramètre β, contrôlant la distance entre deux séquences au delà de laquelle l'influence d'un réviseur n'est plus considérée, i.e. $f(d) = 0$, a été mise à $\beta = 1000$.

Dans l'objectif d'illustrer la généricité de notre approche ainsi que l'intérêt de considérer le graphe de collaboration et le processus de révision par les pairs dans un processus de renforcement mutuel, les trois instances suivantes de notre modèle sont considérées :

- **Wiki 1**. Il s'agit du modèle proposé dans la Section 2 du Chapitre 3. Une représentation immédiate de la plateforme Wikipédia et deux ensembles d'entités sont considérés (articles et contributeurs). De plus, est utilisée le graphe de collaboration entre les utilisateurs ;
- **Wiki 2.1**. Il s'agit du définit en Section 2.2 du Chapitre 4 et utilisant comme fonction d'influence $\phi(x_2^k, x_2^j) = x_2^k + x_2^j$. Ainsi, la qualité d'une

séquence prend comme valeur la somme de l'autorité de l'auteur et de l'autorité du réviseur ;

- **Wiki 2.2**. Il s'agit à nouveau modèle définit en Section 2.2 du Chapitre 4 avec la fonction d'influence $\phi(x_2^k, x_2^j) = (x_2^k \cdot x_2^j)^{\alpha(1-w_{ij})}$.

3.3.2 Résultats

Les résultats aux expérimentations sur le jeu de données Wikipédia avec les métriques NDCG et Précision sont présentés dans le Tableau 6.13 et Tableau 6.14 respectivement. Les résultats pour le modèle Wiki 2.2 ont été obtenus avec la valeur du paramètre $\alpha = 1$. L'impact de ce paramètre sur les performances du modèle sera discuté en Section 4.

Modèle	**FA**	**A**	**GA**	**B**	**C**	**S**
Naive	0.3164	0.3563	0.5694	0.7165	0.8089	0.8268
Basic	0.3951	0.4021	0.4256	0.5551	0.5984	0.7997
Peer	0.6730	0.6847	0.7906	0.8065	0.8095	0.8099
Prob	0.2840	0.3066	0.4055	0.5768	0.7206	0.7973
Wiki 1	0.7439	0.7575	**0.8154**	**0.8119**	0.8316	**0.9380**
Wiki 2.1	0.7492	0.7735	0.7916	0.8117	**0.8949**	0.9312
Wiki 2.2	**0.8057**	**0.8245**	0.8036	0.7742	0.8363	0.9342

Table 6.13 – Résultats aux évaluations avec la métrique NDCG des différentes solutions appliquées au jeu de données Wikipedia.

Modèle	**FA**	**A**	**GA**	**B**	**C**	**S**
Naive	0.2530	**0.0392**	0.1156	**0.3221**	**0.5966**	0.9868
Basic	0.2741	0	0.0029	0.2975	0.3190	0.9108
Peer	0.3498	0.0196	0.1028	0.2866	0.5514	**0.9894**
Prob	0.2122	0	0.0405	0.2816	0.3921	0.9678
Wiki 1	0.6490	0	**0.1792**	0.2955	0.3027	0.9316
Wiki 2.1	0.6735	0.0196	0.1069	0.3043	0.4913	0.9789
Wiki 2.2	**0.7551**	0.0196	0.0809	0.3152	0.5313	0.9449

Table 6.14 – Résultats aux évaluations avec la métrique Precision des différentes solutions appliquées au jeu de données Wikipedia.

Les meilleurs résultats pour la tâche d'identification d'articles de qualité sont obtenus avec le modèle Wiki 2.1. En effet, l'ordonnancement sur les 245 premiers articles est évalué, avec la métrique NDCG, à

0.8057, lorsque le meilleur compétiteur de l'état de l'art retourne un ordonnancement évalué à 0.6730. Ce gain de près de 20% montre l'intérêt de notre proposition lorsqu'il s'agit d'identifier les articles de qualité. En particulier, cette observation montre l'importance des indices d'autorité issus du graphe de collaboration ainsi que ceux construits à partir du voisinage temporel d'une séquence.

Cela dit, on observe que Wiki 2.1 ne réalise pas les meilleures performances pour toutes les catégories d'articles. Si le modèle Wiki 2.1 renvoie un ordonnancement plus proche de l'optimale que les compétiteurs jusqu'à la catégorie A, c'est le modèle Wiki 2.1, intégrant uniquement le graphe de collaborations, qui obtient les meilleurs résultats sur une majorité des catérogies d'articles restantes. Cette observation confirme l'importance seule du graphe de collaborations et l'efficacité de la proposition.

4 Expérimentations

On s'intéresse dans cette section à l'étude des différents paramètres des modèles proposés sur l'impact des performances ansi qu'à l'intégration des scores produits par différentes propositions dans des modèles de prédictions. En l'occurence, nous présentons et discutons les résultats associés aux quatres expérimentations suivantes :

1. Dans la Section 4.1, nous étudions l'impact de la taille de l'ensemble $\mathcal{P}$ des connaissances *a priori* sur les performances de l'Algorithme 2 détaillé à la page 58 ;
2. Dans la Section 4.2, nous étudions l'impact du paramètre utilisateur λ, pondérant les deux fonctions objectif, sur les performances du modèle Mas 2 proposé dans le chapitre 5 ;
3. Dans la Section 4.3, nous étudions l'impact du paramètre α, pondérant la relation de co-édition sur Wikipédia, sur les performances du modèle Wiki 2.2 du chapitre 4 ;
4. Enfin, dans la Section 4.4, nous étudions l'intérêt de considérer les scores produits par les modèles Wiki 2.1 et Wiki 2.2 dans un scenario supervisé.

4.1 Ensemble de connaissances *a priori*

Nous étudions dans cette section l'impact de la taille de l'ensemble $\mathcal{P}$ représentant l'ensemble des connaissances *a priori* sur les performances de l'Algorithme 2. La propagation des scores est étudiée sur un graphe simple $G = (V, E)$ mettant en relation les auteurs des questions avec les

auteurs des réponses sur la plateforme Stack Exchange (Jurczyk and Agichtein, 2007). En l'occurence, le nombre d'utilisateurs est $|V| = 959\,093$ et le nombre de relations $|E| = 2\,191\,185$. Nous considérons une tâche d'identification des utilisateurs faisant autorité sur la plateforme, c'est à dire, les 1% d'utilisateurs avec le plus grand score de réputation (Movshovitz-Attias et al., 2013). Les évaluations sont réalisées avec la métrique NDCG.

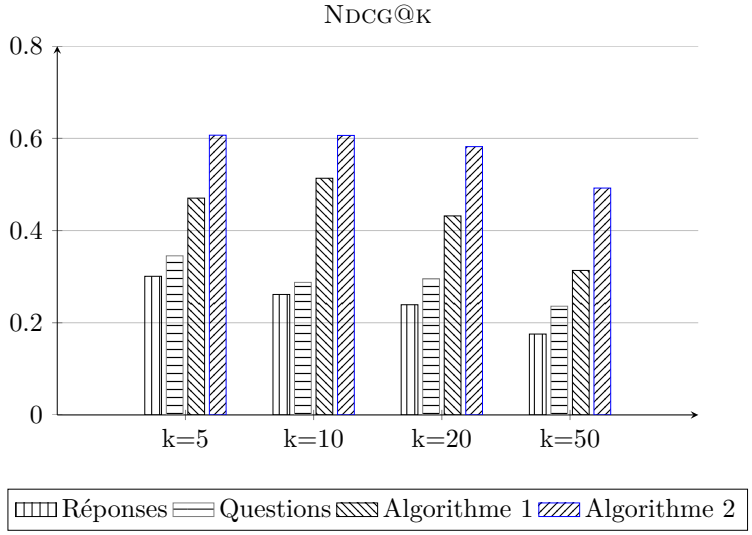

FIGURE 6.7 – Evaluation de l'ordonnancement des utilisateurs de la plateforme Stack Exchange avec la métrique NDCG pour les quatres modèles considérés.

Dans la Figure 6.8, nous traçons l'évolution des performances de l'Algorithme 2 avec la taille de l'ensemble P selon la métrique NDCG. Nous avons, pour chaque valeur de $|P|$ de 1 à 20, réalisé une dizaine d'évaluations différentes de l'algorithme pour différents ensembles P générés de façon aléatoire à partir de paires d'utilisateurs dont l'un appartient au top 1% de la plateforme. La courbe rouge résume ces évaluations en traçant la performance moyenne en fonction du cardinal de P. Bien que la qualité de l'ordonnancement produit par l'Algorithme 2 soit relativement variable lorsque l'on considère les 5 premiers utilisateurs (NDCG@5), la tendance est globalement à la hausse et suit l'accroissement de la taille de l'ensemble de connaissances *a priori* défini par $\mathcal{P}$. En particulier, la valeur de la métrique NDCG au rang 5 double entre $|P| = 1$ et $|P| = 20$. Pour des rangs supérieurs à 5, la variance des résultats diminue et la tendance reste dans l'ensemble positive.

Dans la Figure 6.7, nous soulignons l'efficacité de l'Algorithme 2 en

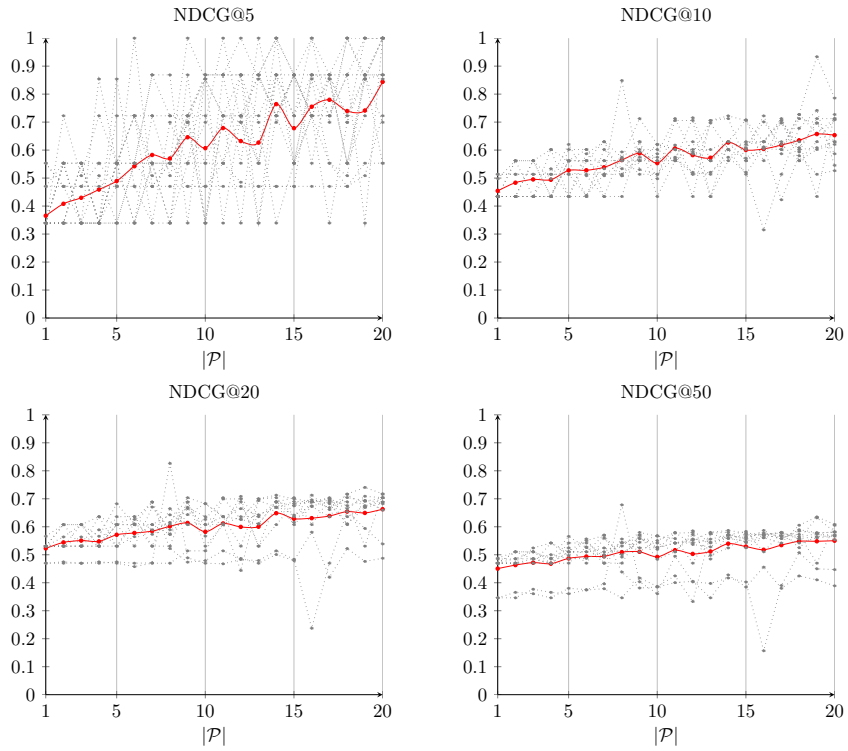

FIGURE 6.8 – Évolution des performances du modèle de correction selon la métrique NDCG@k pour $k \in \{5, 10, 20, 50\}$ en fonction de la taille de l'ensemble $\mathcal{P}$.

comparant, pour $|P| = 10$, la moyenne des différents résultats face à l'Algorithme 1. Sont aussi indiqués sur la figure la qualité des ordonnancements qui consistent à trier les utilisateurs selon le nombre de questions ou de réponses qu'ils ont publiées sur la plateforme. Le gain de 20% apporté par l'exploitation de l'ensemble P calculé avec la métrique NDCG@k pour quatre valeurs de k confirme l'intérêt de la proposition.

4.2 Coefficient pour l'expertise de domaine

Dans cette section, nous étudions l'impact du paramètre λ du modèle MAS 2 sur ses performances dans une tâche d'identification des chercheurs faisant autorité sur la plateforme digitale Microsoft Academic Search. Pour ce faire, différentes optimisations sont conduites pour chaque valeur de λ entre 0 et 1 par pas de 0.1 sur chacun des 11 graphes d'expertises introduits en Section 1.3.

Rappelons que lorsque $\lambda = 0$, la proposition MAS 2 est une va-

riante de l'algorithme Pagerank qui calcule alors un score de qualité pour chaque article à partir des signaux issus de la relation de citation. Dans ce cas, le modèle, noté PR, calcule alors un score pour chaque chercheur égal à la somme des scores de qualité des articles écrits par lui. Au contraire, lorsque $\lambda = 1$, la proposition n'est autre qu'un problème de factorisation de matrices en matrices non négatives sur la matrice A_{23}. Nous notons NMF cette solution. Les résultats, évalués avec la métrique AUC, sont détaillés dans le Tableau 6.15.

λ	**PR**	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	**NMF**
G_0	0.683	0.682	**0.693**	0.681	0.683	0.683	0.678	0.672	0.671	0.663	0.621
G_1	0.666	0.671	0.664	0.666	0.660	**0.672**	0.665	0.671	0.669	0.668	0.659
G_2	0.644	0.643	0.642	**0.653**	0.653	0.641	0.642	0.636	0.639	0.631	0.558
G_3	0.671	**0.684**	0.676	0.675	0.672	0.669	0.681	0.682	0.674	0.662	0.565
G_4	0.674	**0.680**	0.667	0.673	0.675	0.672	0.671	0.677	0.670	0.667	0.582
G_5	0.635	0.636	**0.644**	0.634	0.641	0.629	0.643	0.644	0.637	0.628	0.548
G_6	0.642	0.641	**0.651**	0.639	0.634	0.643	0.648	0.638	0.631	0.622	0.586
G_7	0.688	0.688	**0.694**	0.692	0.691	0.688	0.690	0.688	0.672	0.664	0.612
G_8	**0.667**	0.648	0.655	0.662	0.656	0.647	0.658	0.641	0.654	0.649	0.593
G_9	0.671	**0.674**	0.673	0.672	0.670	0.663	0.665	0.667	0.663	0.666	0.559
G	0.647	**0.661**	0.649	0.653	0.659	0.658	0.656	0.649	0.649	0.651	0.553

TABLE 6.15 – Evaluation du modèle MAS 2 en fonction de la valeur du paramètre λ avec la métrique AUC.

Il est intéressant de constater que, contrairement à l'intuition, la solution PR n'obtient en général pas les meilleurs résultats. En l'occurence, une combinaison linéaire des deux objectifs liés aux deux calculs d'autorité et d'expertise semble améliorer à la fois le modèle PR et le modèle NMF. On constate plus particulièrement que les meilleurs résultats sont obtenus pour des faibles valeurs de λ. Ces résultats suggèrent que la prise en compte du contenu des documents permet d'améliorer la caractérisation de la notion d'autorité. Les deux dimensions d'autorité et d'expertise devant alors être considérées en même temps.

4.3 Paramètre de co-éditions

Dans cette section, l'impact du paramètre utilisateur α de la seconde fonction d'infuence ϕ du modèle WIKI 2.2 est étudié. Pour cela, nous faisons varier la valeur du paramètre α entre 0 et 1 par pas de 0.1 et traçons les performances du modèle WIKI 2.2 évaluées avec les métriques Précision et NDCG pour chacune des six catégories d'articles. Cette évolution est représentée dans la Figure 6.9.

On constate que les résultats sont dépendants de la valeur du paramètre α. Les performances tombent, pour toutes les catégories d'ar-

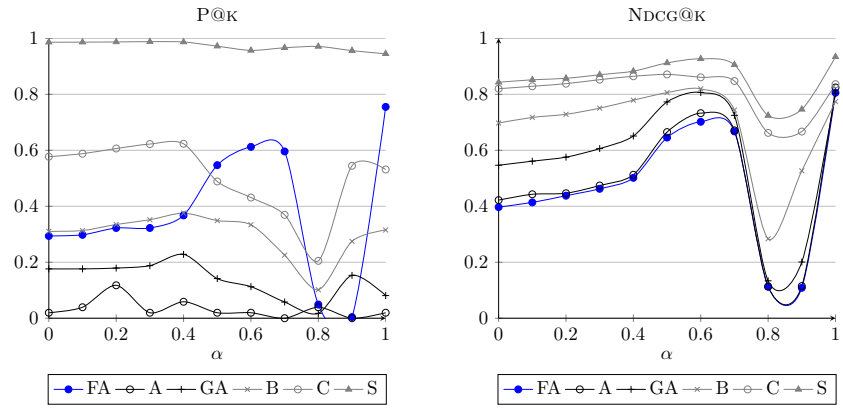

FIGURE 6.9 – Évolution des métriques Précision (à gauche) et NDCG (à droite) pour le modèle WIKI 2.2 en fonction du paramètre α pour les 6 classes d'articles.

ticles, lorsque $\lambda = 0.8$ ou $\lambda = 0.9$. S'il est difficile d'interpréter ces résultats, ils montrent une dépendance directe entre la pondération des relations de co-éditions dans le graphe de collaborations et les performances du modèle.

4.4 Evaluation des signaux de structure

Dans cette section, nous étudions combien les indices calculés par les propositions WIKI 2.1 et WIKI 2.2 permettent d'améliorer, dans un contexte supervisé, les modèles de prédiction classiques sur une tâche d'identification d'articles de qualité sur Wikipédia. Pour ce faire, nous considérons deux familles de fonctions prédictives différentes (basées sur les arbres de décisions et les forêts aléatoires) optimisées selon différentes configurations d'espaces de caractéristiques.

En l'occurence, nous considérons que les articles Wikipédia peuvent être décrits par une combinaison des trois espaces de caractéristiques associés (a) aux indices de contenu, (b) aux indices temporels et (c) aux indices de structures. Le Tableau 6.16 liste ces familles de caractéristiques utilisées dans nos expérimentations.

Aussi, nous comparons les performances des trois modèles prédictifs considérés entrainés sur les 7 espaces de représentations possibles en considérant (a) les indices de contenus uniquement, (b) les indices temporels uniquement, (c) les indices de structures uniquement, (d) les indices de contenu et les indices temporels, (e) les indices de contenu et ceux de structures, (f) les indices temporels et ceux de structures, enfin (g) les indices de contenus, les indices temporels et les indices de

Famille	**Indices**
Contenu	Nombre de caractères
	Nombre de mots
	Nombre de mots distincts
	Nombre de portions de textes
	Nombre de contributeurs distincts
Temporel	Nombre de révisions
	Temps entre deux révisions (moyenne)
	Temps entre deux révisions (variance)
	Durée de vie des séquences (moyenne)
	Durée de vie des séquences (variance)
Structure	Score de qualité calculé par WIKI 2.1
	Scores de qualité calculés par WIKI 2.2

TABLE 6.16 – Familles de caractéristiques des articles Wikipédia.

structures à la fois.

Plusieurs validations croisées basées sur des échantillonages stratifiés sont réalisées sur les $N \simeq 20\,000$ articles du corpus Wikipédia. Pour chaque valeur de $c \in \{2, 3, ..., 40\}$, un ensemble d'entrainement de taille $\frac{1}{c}$ est utilisé pour optimiser les paramètres des modèles considérés. L'évaluation est alors réalisée sur l'ensemble restant de taille $1 - \frac{1}{c}$. Par définition, pour chaque valeur de c, le processus est répété c fois de telle façon que chaque ensemble de taille $\frac{1}{c}$ soit utilisé exactement une fois pour l'apprentissage. Pour comparer les performances en classification des modèles selon les différents espaces de caractéristiques considérés, autant de modèles que de combinaisons d'espaces de caractéristiques possibles sont entrainés et évalués selon ce même protocole.

Les performances évaluées avec les métrique NDCG et Précision des deux modèles sur la tâche d'ordonnancement des articles de la catégorie FA en fonction des familles de caractéristiques considérées sont tracées dans la La Figure 6.10.

On constate d'abord que les meilleures performances sont toujours obtenues lorsque la famille d'indices de structure est prise en compte, soulignant l'expressivité des indices produits par les modlèles WIKI 2.1 et WIKI 2.2. De plus, on remarque que les performances des modèles entrainés uniquement avec les indice de structure (en bleu sur la figure) sont tout à fait comparables aux performances des modèles lorsque les autres familles d'indices sont considérés. Autrement dit, la seule information apportée par notre modèle permet une aussi bonne caractérisation des articles de qualité que les deux autres familles d'indices temporels

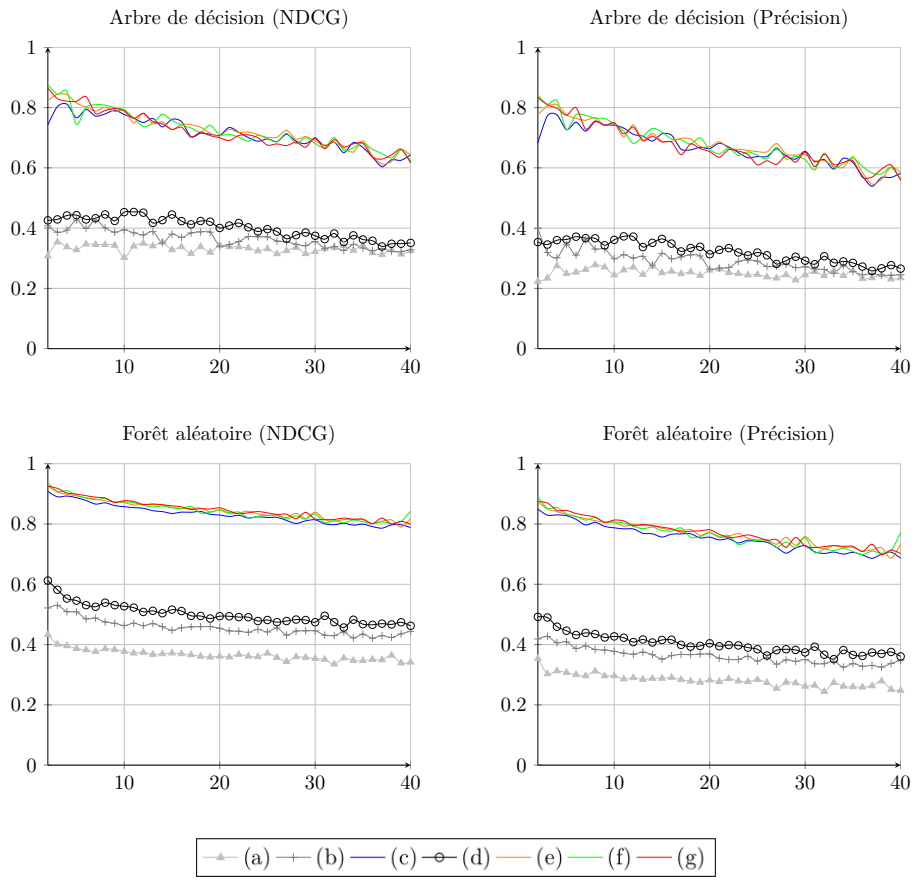

FIGURE 6.10 – Evolution des métriques Précision et NDCG pour les deux modèles considérés sur les 7 configurations de caractéristiques en fonction du nombre c de validation croisées.

ou de contenu réunies.

5 Conclusion

Dans ce chapitre, nous avons présenté les principaux résultats des expérimentations réalisées dans le cadre de cette thèse. Trois jeux de données réels, issus de trois plateformes collaboratives différentes, ont été utilisés. En particulier, ces expérimentations ont été conduites sur un jeu de données extrait à partir du site collaboratif Wikipédia, contenant environs 20 000 articles ; sur un corpus comprenant l'ensemble des données publiques d'un site de Questions-Réponses de plusieurs millions de questions et réponses ; sur un corpus de plus de 1.5 millions d'articles

scientifiques issus de la librairie digitale Microsoft Academic Search. Nos proposisions ont été évaluées et comparées à différentes approches de la litérature sur les quatres problématiques suivante :

1. *Identification des articles de qualité sur Wikipédia.* Les approches WIKI 1, WIKI 2.1 et WIKI 2.2 ont été proposées pour modéliser les données de la plateforme Wikipédia. Dans ce contexte, nous avons comparé les performances de nos propositions sur une même tâche d'identification d'articles de qualité. Les résultats ont montré une nette amélioration des modèles de l'état de l'art, notamment grâce à la prise en compte d'un graphe de collaboration entre les utilisateurs et à la modélisation du processus de révision des articles ;

2. *Identification des réponses de qualité sur Stack Exchange.* Les modèles STE 1 et STE 2 ont été proposés pour représenter les données de la plateforme Stack Exchange. Dans ce cadre, les performances de nos propositions ont été comparées sur une tâche d'identification des réponses de qualité. A nouveau, les résultats ont montré une amélioration des performances des modèles de l'état de l'art, notamment grâce à la prise en considération, dans la représentation de la plateforme, d'une nouvelle famille d'entités, ainsi qu'à l'exploitation d'un graphe de réactivité construit à partir des flux de réponses des utilisateurs ;

3. *Identification des chercheurs faisant autorité.* Les approches MAS 1.1 et MAS 1.2 ont été proposées pour représenter la librairie digitale Microsoft Academic Search. Les performances des propositions ont été comparées sur une tâche d'identification des chercheurs faisant autorité sur la plateforme. Les analyses quantitative et qualitative ont montré d'une part une amélioration des méthodes de l'état de l'art et surtout un réel intérêt à exploiter un ensemble de connaissances *a priori* sur la qualité de certaines conférences ;

4. *Tâche de recherche d'experts.* La proposition MAS 2 a été proposée pour donner une autre représentation de la plateforme Microsot Academic Search et répondre à une tâche de recherche d'experts. Nous avons proposé une représentation conjointe des notions d'autorité et d'expertise afin de prendre en considération, via un problème d'optimisation bi-objectifs, d'éventuelles correlations entre ces deux notions. Les résultats ont montré certaines améliorations comparées aux approches de la littérature. En particulier, il semblerait que les indices de contenus puissent améliorer l'identification d'utilisateurs faisant autorité, et, inversement, que les indices de qualité puisse améliorer la caractérisation du profil d'expertise des chercheurs.

Ces différentes propositions ont par ailleurs confirmé la généricité de notre proposition par les graphes hétérogènes, rendant particulièrement aisé la prise en considération de nouvelles familles d'entités ou de relations entre les éléments, et, par conséquent, facilement adaptable à l'évolution des plateformes.

Chapitre 7

Conclusion générale

1 Synthèse des contributions

Dans cette thèse, nous nous sommes intéressés à l'étude des approches structurelles permettant une modélisation et un calcul des notions de qualité, d'autorité et d'expertise sur les plateformes collaborative. Dans ce cadre, un ensemble de solutions et d'algorithmes ont été proposés améliorant les méthodes actuelles sur plusieurs tâches d'identification de contenus de qualité ou d'utilisateurs faisant autorité sur les plateformes collaboratives et librairies digitales. Trois contributions majeures ont été apportées.

D'abord, un modèle générique basé sur une représentation par les graphes hétérogènes des plateformes a été proposé. Cette première contribution a notamment permis la généralisation du principe clef de renforcement mutuel, largement exploité dans la littérature mais uniquement sur des représentations par les graphes ne considérant qu'un nombre restreint de familles d'entités. La généricité de notre approche a été démontrée en considérant différents problèmes d'identification de contenus de qualité sur trois plateformes collaboratives différentes et en généralisant divers travaux de la littérature.

Par ailleurs, une version paramétrique de notre approche a été proposée, et plus spécifiquement, une approche transductive, permettant l'intégration d'ensembles de connaisances *a priori* présentées de façon qualitative à l'algorithme. La pertinence de la proposition a été montrée sur une tâche d'identification de chercheurs faisant autorité sur la librairie Microsoft Academic Search. Les résultats obtenus ont suggérés qu'utiliser un ensemble de connaissances *a priori* sur la qualité de certaines conférences scientifiques améliore le ciblage de chercheurs à l'origine d'articles de qualité.

Finalement, une nouvelle représentation de la notion d'autorité de domaine a été proposée dans ce manuscrit, permettant l'unification des représentations actuelles de la littérature liées aux notions d'autorité et d'expertise. Exploitant un modèle de réconciliation simple entre une requête de domaine et l'ensemble des profils d'autorité de domaines calculs, l'approche a montré certaines améliorations comparées aux modèles de l'état de l'art sur une tâche de recherche d'experts sur la librairie Microsoft Academic Search.

2 Perspectives de recherche

Par définition, le modèle proposé dans le Chapitre 3 est restreint à un ensemble de relations binaires. Bien qu'il soit possible d'instancier notre modèle sur un nombre arbitraire de familles d'entités et de relations binaires, il n'est pas possible, dans l'état actuel des choses, de considérer plusieurs types de relations entre une même paire de familles d'entités (Interdonato et al., 2016; Interdonato et al., 2017). Or, il pourrait être, par exemple, intéressant de prendre en compte à la fois une relation de co-réponse et une relation de réactivité sur une plateforme de type Questions-Réponses. Lever cette limitation paraît donc une amélioration importante de la proposition.

Dans l'approche paramétrique actuelle, le gradient de la fonction de coût est exprimé et calculé en ne considérant que les voisins directs des noeuds, c'est à dire les paramètres sur les arcs incidents. Autrement dit, à chaque itération, la correction des paramètre d'influence ne se propage que localement. Or, il paraît indispensable d'exprimer le gradient de la fonction de coût proposée à des ordres de profondeurs supérieurs de façon à propager plus largement la correction des paramètres dans le graphe et corriger davantage de paramètres à chaque itération. D'ailleurs, il serait intéressant d'étudier un éventuel rapprochement de la proposition (algébrique) avec les approches probabilistes de l'état de l'art (réseaux bayesiens). De plus, il doit être possible de fusionner les deux étapes de propagation et de correction des paramètres d'influences en ne formulant qu'une unique fonction de coût et un problème d'optimisation sous-contrainte.

Les résultats aux expérimentations ayant prouvé l'efficacité de la proposition, il paraît maintenant intéressant de proposer une version parallélisable de la proposition ; les fonctions d'influence étant, d'une itération à l'autre, indépendantes les unes des autres.

De nombreuses autres perspectives de recherches pourraient améliorer la proposition actuelle. Par exemple, des solutions probabilistes devraient être utilisées pour résoudre d'une façon moins stricte le problème de

désambiguisation des noms des chercheurs dans les corpus. Il paraît important d'étudier différents types de fonctions d'influence, notamment au niveau des paires de familles d'entités. Il paraît intéressant d'étudier une extention des notions d'autorité ou de qualité actuelles en considérant des scores signés. Une telle représentation pourrait sans doute permettre la distinction entre les nouveaux entrants d'une plateforme et les détracteurs enregistrés.

Liste des tableaux

2.1 Classification des approches structurelles de la littérature appliquées à une tâche d'identification de contenus de qualité sur les plateformes collaboratives et librairies digitales. 38

3.1 Résumé des notations utilisées dans ce manuscrit. 47

6.1 Statistiques des données Wikipédia (*en moyenne par classe) après prétraitement. 94
6.2 Fonction de répartition des votes des réponses de la plateforme Stack Exchange. 96
6.3 Discrétisation des votes sur la plateforme Stack Exchange. 97
6.4 Répartition des articles par catégorie de conférences dans le jeu de données utilisé pour les évaluations. 98
6.5 Les 11 sous-graphes d'expertise construits à partir du jeu de données Microsoft Academic Search utilisés pour les expérimentations. 99
6.6 Evaluation des ordonnancement sur les conférences du jeu de données Microsoft Academic Search avec la métrique NDCG@K. 106
6.7 Top-5 chercheurs retrouvés par les modèles étudiés. 107
6.8 Nombre de publications produites par catégorie de conférence par les top-5 chercheurs retrouvés par les différents modèles évalués. 107
6.9 Performances du modèle MAS 2 évaluée avec la métrique AUC sur les dix requêtes de domaines. 108
6.10 Répartition des publications (en %) des top-5 chercheurs retrouvés par les différents modèles en fonction du type de conférence. 109
6.11 Evaluation de la précision des modèles d'ordonnancement de réponses sur Stack Exchange. 111

6.12 Evaluation de l'ordonancement des modèles sur Stack Exchange avec la métrique NDCG. 111
6.13 Résultats aux évaluations avec la métrique NDCG des différentes solutions appliquées au jeu de données Wikipedia. 113
6.14 Résultats aux évaluations avec la métrique Precision des différentes solutions appliquées au jeu de données Wikipedia. 113
6.15 Evaluation du modèle MAS 2 en fonction de la valeur du paramètre λ avec la métrique AUC. 117
6.16 Familles de caractéristiques des articles Wikipédia. 119

Table des figures

2.1 Distributions du nombre de mots pour deux classes d'articles Wikipédia extraites du travail de (Blumenstock, 2008). 21

2.2 La notion d'autorité structurelle de l'algorithme PageRank. La taille d'un sommet du graphe reflète son autorité, fonction de l'autorité des sommets voisins. 28

2.3 Graphe biparti sur Wikipédia. 31

2.4 Construction d'un graphe biparti. A gauche, une représentation immédiate des données issues de la plateforme Yahoo ! Answers. A droite, une nouvelle représentation construite à partir de la représentation immédiate. L'auteur d'une réponse est associée à l'auteur de la question. 32

2.5 Graphe hétérogène du modèle NCR (Zhang et al., 2014). . 36

3.1 Graphe hétérogène associé à l'Exemple 1.1. 48

3.2 Influence d'une famille V_i sur V_j. 49

3.3 Principe de renforcement mutuel entre V_i et V_j. 51

3.4 Graphe d'influence du modèle WIKI 1. 52

3.5 Graphe d'influence du modèle MAS 1. 53

3.6 Graphe d'influence du modèle STE 1. 54

3.7 Graphe associé à l'Exemple 4.1. Les symboles $\{+, -\}$ associés aux arcs indiquent comment l'étape de correction met à jour les valeurs des paramètres du modèle pour que la contrainte $v_1^3 \succ v_1^4$ soit respectée. 59

3.8 Fonctions de densité associées aux nombres d'articles par catégorie de conférence construites à partir de 2 158 conférences issues des jeux de données Microsoft Academic Search et Core.edu. 68

4.1 Délais d'apparitions moyen des réponses en fonction de leur qualité calculés à partir de 99.5% des réponses publiées entre 2008 et 2014 sur la plateforme Stack Exchange. 71
4.2 Réputation moyenne des utilisateurs (en gris), ainsi que qualité moyenne des réponses produites (en noir), associés aux noeuds des graphes de réactivité G_R^n. 72
4.3 Graphe d'influence du modèle Ste 2. 73
4.4 Illustration des notions de voisinage antérieur et postérieur. Quatre révisions successives d'un article sont considérées. La version finale de l'article contient 4 portions de textes. L'ancienneté de la portion de texte (d), en gras sur la figure, à la révision 4, est 2 (ajoutée deux révisions auparavant). Les deux portions de textes (a) et (c) appartiennent au voisinage antérieur de (d). En effet, l'ancienneté de (a) et (c) vaut 3 (les deux poritions de texte ont été ajouté à l'article durant la première révision). La portion de texte (b) appartient au voisinage postérieur de (d). En effet, l'ancienneté de (b) vaut 1. 75
4.5 Graphe d'influence du modèle Wiki 2. 76

5.1 Graphe des familles d'entité du modèle Mas 2. 84

6.1 Échelle de qualité des articles Wikipédia extraite de la page https //en.wikipedia.org/wiki/Template :Grading_scheme. 94
6.2 Histogramme de répartition des scores des réponses issues de la plateforme de Question/Réponses Stack Exchange. Plus de 98% des données sont représentées. . . . 96
6.3 Répartition des scores de h-index des chercheurs du jeu de données Microsoft Academic Search. 103
6.4 Précision@k des modèles d'identification de chercheurs faisant autorité sur la librairie Microsoft Academic Search. 104
6.5 Rappel@k des modèles d'identification de chercheurs faisant autorité sur la librairie Microsoft Academic Search. . 105
6.6 NDCG@k des modèles d'identification de chercheurs faisant autorité sur la librairie Microsoft Academic Search. . 105
6.7 Evaluation de l'ordonnancement des utilisateurs de la plateforme Stack Exchange avec la métrique NDCG pour les quatres modèles considérés. 115
6.8 Évolution des performances du modèle de correction selon la métrique NDCG@k pour $k \in \{5, 10, 20, 50\}$ en fonction de la taille de l'ensemble $\mathcal{P}$. 116

6.9 Évolution des métriques Précision (à gauche) et NDCG (à droite) pour le modèle WIKI 2.2 en fonction du paramètre α pour les 6 classes d'articles. 118
6.10 Evolution des métriques Précision et NDCG pour les deux modèles considérés sur les 7 configurations de caractéristiques en fonction du nombre c de validation croisées. 120

Bibliographie

Adler, B. T., Chatterjee, K., de Alfaro, L., Faella, M., Pye, I., and Raman, V. (2008). Assigning trust to wikipedia content. In *Proceedings of the 4th International Symposium on Wikis*, WikiSym '08, pages 26 :1–26 :12, New York, NY, USA. ACM.

Adler, B. T. and de Alfaro, L. (2007). A content-driven reputation system for the wikipedia. In *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, pages 261–270, New York, NY, USA. ACM.

Agichtein, E., Castillo, C., Donato, D., Gionis, A., and Mishne, G. (2008). Finding high-quality content in social media. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, WSDM '08, pages 183–194, New York, NY, USA. ACM.

Anderson, A., Huttenlocher, D., Kleinberg, J., and Leskovec, J. (2012). Discovering value from community activity on focused question answering sites : A case study of stack overflow. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '12, pages 850–858, New York, NY, USA. ACM.

Anokhin, N., Lanagan, J., and Velcin, J. (2012). Social citation : Finding roles in social networks. an analysis of tv-series web forums. *Second International Workshop on Mining Communities and People Recommenders (COMMPER).*

Balog, K., Azzopardi, L., and de Rijke, M. (2006). Formal models for expert finding in enterprise corpora. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '06, pages 43–50, New York, NY, USA. ACM.

Balog, K., Fang, Y., de Rijke, M., Serdyukov, P., and Si, L. (2012). Expertise retrieval. *Found. Trends Inf. Retr.*, 6 :127–256.

Bellomi, F. and Bonato, R. (2005). Network analysis for wikipedia. In

Proceedings of Wikimania 2005, The First International Wikimedia Conference.

Betancourt, G. G., Segnine, A., Trabuco, C., Rezgui, A., and Jullien, N. (2016). Mining team characteristics to predict wikipedia article quality. In *Proceedings of the 12th International Symposium on Open Collaboration*, OpenSym '16, pages 15 :1–15 :9, New York, NY, USA. ACM.

Bian, J., Liu, Y., Zhou, D., Agichtein, E., and Zha, H. (2009). Learning to recognize reliable users and content in social media with coupled mutual reinforcement. In *Proceedings of the 18th International Conference on World Wide Web*, WWW '09, pages 51–60, New York, NY, USA. ACM.

Biancani, S. (2014). Measuring the quality of edits to wikipedia. In *Proceedings of The International Symposium on Open Collaboration*, OpenSym '14, pages 33 :1–33 :3, New York, NY, USA. ACM.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3 :993–1022.

Blumenstock, J. E. (2008). Size matters : Word count as a measure of quality on wikipedia. In *Proceedings of the 17th International Conference on World Wide Web*, WWW '08, pages 1095–1096. ACM.

Boucher, V. (2012). Structural Homophily. working paper or preprint.

Campbell, C. S., Maglio, P. P., Cozzi, A., and Dom, B. (2003). Expertise identification using email communications. In *Proceedings of the Twelfth International Conference on Information and Knowledge Management*, CIKM '03, pages 528–531, New York, NY, USA. ACM.

Chang, S. and Pal, A. (2013). Routing questions for collaborative answering in community question answering. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '13, pages 494–501, New York, NY, USA. ACM.

Chen, P., Xie, H., Maslov, S., and Redner, S. (2007). Finding scientific gems with google's pagerank algorithm. *Journal of Informetrics*, 1(1) :8 – 15.

Craswell, N., Hawking, A., Vercoustre, M., and Wilkins, P. (2001). Panoptic expert : Searching for experts not just for documents. In *usweb Poster Proceedings*.

Currarini, S., Matheson, J., and Vega-Redondo, F. (2016). A simple model of homophily in social networks. *European Economic Review*, 90 :18 – 39. Social identity and discrimination.

Dalip, D. H., Gonçalves, M. A., Cristo, M., and Calado, P. (2013). Exploiting user feedback to learn to rank answers in q&#38 ;a forums : A case study with stack overflow. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '13, pages 543–552, New York, NY, USA. ACM.

Dang, Q. V. and Ignat, C.-L. (2016). Quality assessment of wikipedia articles without feature engineering. In *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries*, JCDL '16, pages 27–30, New York, NY, USA. ACM.

De la Calzada, G. and Dekhtyar, A. (2010). On measuring the quality of wikipedia articles. In *Proceedings of the 4th Workshop on Information Credibility*, WICOW '10, pages 11–18, New York, NY, USA. ACM.

de La Robertie, B., Pitarch, Y., Takasu, A., and Teste, O. (2017a). Identifying Authoritative Researchers in Digital Libraries using External a Priori Knowledge. In *ACM Symposium on Applied Computing (SAC)*. ACM.

de La Robertie, B., Pitarch, Y., and Teste, O. (2015). Measuring article quality in wikipedia using the collaboration network. In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*, ASONAM '15, pages 464–471, New York, NY, USA. ACM.

de La Robertie, B., Pitarch, Y., and Teste, O. (2016). *Quality Prediction in Collaborative Platforms : A Generic Approach by Heterogeneous Graphs*, pages 19–26. Springer International Publishing, Cham.

de La Robertie, B., Pitarch, Y., and Teste, O. (2017b). *Structure-Based Features for Predicting the Quality of Articles in Wikipedia*, pages 115–140. Springer International Publishing, Cham.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by latent semantic analysis. *JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE*, 41(6) :391–407.

Deng, H., King, I., and Lyu, M. R. (2008). Formal models for expert finding on dblp bibliography data. In *Proceedings of the 2008 Eighth IEEE International Conference on Data Mining*, ICDM '08, pages 163–172, Washington, DC, USA. IEEE Computer Society.

Egghe, L., Rousseau, R., and Van Hooydonk, G. (2000). Methods for accrediting publications to authors or countries : Consequences for evaluation studies. *J. Am. Soc. Inf. Sci.*, 51(2) :145–157.

Emigh, W. and Herring, S. C. (2005). Collaborative authoring on the web : A genre analysis of online encyclopedias. In *Proceedings of the Proceedings of the 38th Annual Hawaii International Conference on System Sciences (HICSS'05) - Track 4 - Volume 04*, HICSS '05, pages 99.1–, Washington, DC, USA. IEEE Computer Society.

Farkas, I., Derényi, I., Jeong, H., Néda, Z., Oltvai, Z., Ravasz, E., Schubert, A., Barabási, A.-L., and Vicsek, T. (2002). Networks in life : scaling properties and eigenvalue spectra. *Physica A : Statistical Mechanics and its Applications*, 314(1) :25–34.

Flekova, L., Ferschke, O., and Gurevych, I. (2014). What makes a good biography ? : Multidimensional quality analysis based on wikipedia article feedback data. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14, pages 855–866, New York, NY, USA. ACM.

Flesch, R. (1948). *A New Readability Yardstick*.

Forestier, M., Stavrianou, A., Velcin, J., and Zighed, D. A. (2012). Roles in social networks : Methodologies and research issues. *Web Intelli. and Agent Sys.*, 10(1) :117–133.

Forestier, M., Velcin, J., and Zighed, D. (2011). Extracting social networks to understand interaction. In *2011 International Conference on Advances in Social Networks Analysis and Mining*, pages 213–219.

Gollapalli, S. D., Mitra, P., and Giles, C. L. (2011). Ranking authors in digital libraries. In *Proceedings of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries*, JCDL '11, pages 251–254, New York, NY, USA. ACM.

Golub, G. H. and Van Loan, C. F. (2012). *Matrix computations*, volume 3. JHU Press.

Gunning, R. (1968). *The technique of clear writing*. McGraw-Hill.

Güttler, F. N., Ienco, D., Poncelet, P., and Teisseire, M. (2016). Combining transductive and active learning to improve object-based classification of remote sensing images. *Remote Sensing Letters*, 7(4) :358–367.

Halfaker, A., Kittur, A., Kraut, R., and Riedl, J. (2009). A jury of your peers : Quality, experience and ownership in wikipedia. In *Proceedings of the 5th International Symposium on Wikis and Open*

Collaboration, WikiSym '09, pages 15 :1–15 :10, New York, NY, USA. ACM.

Hasan Dalip, D., André Gonçalves, M., Cristo, M., and Calado, P. (2009). Automatic quality assessment of content created collaboratively by web communities : A case study of wikipedia. In *Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL '09, pages 295–304, New York, NY, USA. ACM.

Hu, M., Lim, E.-P., Sun, A., Lauw, H. W., and Vuong, B.-Q. (2007). Measuring article quality in wikipedia : Models and evaluation. In *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*, CIKM '07, pages 243–252, New York, NY, USA. ACM.

Interdonato, R., Tagarelli, A., Ienco, D., Sallaberry, A., and Poncelet, P. (2016). Local community detection in multilayer networks. In *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 1382–1383.

Interdonato, R., Tagarelli, A., Ienco, D., Sallaberry, A., and Poncelet, P. (2017). Node-centric community detection in multilayer networks with layer-coverage diversification bias. *CoRR*, abs/1704.03441.

Javanmardi, S. and Lopes, C. (2010). Statistical measure of quality in wikipedia. In *Proceedings of the First Workshop on Social Media Analytics*, SOMA '10, pages 132–138, New York, NY, USA. ACM.

Jurczyk, P. and Agichtein, E. (2007). Discovering authorities in question answer communities by using link analysis. In *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*, CIKM '07, pages 919–922, New York, NY, USA. ACM.

Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. *J. ACM*, 46(5) :604–632.

Kossinets, G. and Watts, D. J. (2009). Origins of homophily in an evolving social network. *American Journal of Sociology*, 115 :405–450.

Lathauwer, L. D., Moor, B. D., and Vandewalle, J. (2000). A multilinear singular value decomposition. *SIAM J. Matrix Anal. Appl.*, 21(4) :1253–1278.

Li, B., Jin, T., Lyu, M. R., King, I., and Mak, B. (2012). Analyzing and predicting question quality in community question answering services. In *Proceedings of the 21st International Conference on World Wide Web*, WWW '12 Companion, pages 775–782, New York, NY, USA. ACM.

Li, X., Tang, J., Wang, T., Luo, Z., and de Rijke, M. (2015). Automatically assessing wikipedia article quality by exploiting article-editor networks. In *ECIR 2015 : 37th European Conference on Information Retrieval*. Springer, Springer.

Lih, A. (2004). Wikipedia as participatory journalism : reliable sources ? metrics for evaluating collaborative media as a news resource. In *In Proceedings of the 5th International Symposium on Online Journalism*, pages 16–17.

Liu, D. C. and Nocedal, J. (1989). On the limited memory bfgs method for large scale optimization. *Math. Program.*, 45(3) :503–528.

Liu, X., Bollen, J., Nelson, M. L., and de Sompel, H. V. (2005). Co-authorship networks in the digital library research community. *Information Processing and Management*, 41(6) :1462 – 1480. Special Issue on Infometrics.

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

Mimno, D. and McCallum, A. (2007). Mining a digital library for influential authors. In *Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL '07, pages 105–106, New York, NY, USA. ACM.

Movshovitz-Attias, D., Movshovitz-Attias, Y., Steenkiste, P., and Faloutsos, C. (2013). Analysis of the reputation system and user contributions on a question answering website : Stackoverflow. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '13, pages 886–893, New York, NY, USA. ACM.

Newman, M. E. (2004). *Who Is the Best Connected Scientist ?A Study of Scientific Coauthorship Networks*, pages 337–370. Springer Berlin Heidelberg, Berlin, Heidelberg.

Page, L., Brin, S., Motwani, R., and Winograd, T. (1999). The pagerank citation ranking : Bringing order to the web. Technical Report 1999-66, Stanford InfoLab.

Priedhorsky, R., Chen, J., Lam, S. T. K., Panciera, K., Terveen, L., and Riedl, J. (2007). Creating, destroying, and restoring value in wikipedia. In *Proceedings of the 2007 International ACM Conference on Supporting Group Work*, GROUP '07, pages 259–268, New York, NY, USA. ACM.

Serdyukov, P., Rode, H., and Hiemstra, D. (2008). Modeling multi-step relevance propagation for expert finding. In *Proceedings of the*

17th ACM Conference on Information and Knowledge Management, CIKM '08, pages 1133–1142, New York, NY, USA. ACM.

Serres, A. (2013). Dans le labyrinthe. *Bulletin des bibliothèques de France (BBF)*, pages 109–110.

Shah, C. and Pomerantz, J. (2010). Evaluating and predicting answer quality in community qa. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '10, pages 411–418, New York, NY, USA. ACM.

Smith, E., Senter, R., and (U.S.), A. F. A. M. R. L. (1967). *Automated Readability Index*. AMRL-TR-66-220. Aerospace Medical Research Laboratories.

Suryanto, M. A., Lim, E. P., Sun, A., and Chiang, R. H. L. (2009). Quality-aware collaborative question answering : Methods and evaluation. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, WSDM '09, pages 142–151, New York, NY, USA. ACM.

Suzuki, Y. (2015). Quality assessment of wikipedia articles using ¡i¿h¡/i¿-index. *Journal of Information Processing*, 23(1) :22–30.

Suzuki, Y. and Yoshikawa, M. (2013). Assessing quality score of wikipedia article using mutual evaluation of editors and texts. In *Proceedings of the 22Nd ACM International Conference on Conference on Information & Knowledge Management*, CIKM '13, pages 1727–1732, New York, NY, USA. ACM.

Tang, J., Zhang, J., Jin, R., Yang, Z., Cai, K., Zhang, L., and Su, Z. (2011). Topic level expertise search over heterogeneous networks. *Mach. Learn.*, 82(2) :211–237.

Twidale, B. S. M. B. (2005). Assessing information quality of a community-based encyclopedia. In *In Proceedings of the International Conference on Information Quality*, pages 442–454.

Vydiswaran, V. V., Zhai, C., and Roth, D. (2011). Content-driven trust propagation framework. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, pages 974–982, New York, NY, USA. ACM.

Wang, G., Gill, K., Mohanlal, M., Zheng, H., and Zhao, B. Y. (2013). Wisdom in the social crowd : An analysis of quora. In *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13, pages 1341–1352, New York, NY, USA. ACM.

Wang, X.-J., Tu, X., Feng, D., and Zhang, L. (2009). Ranking community answers by modeling question-answer relationships via ana-

logical reasoning. In *Proceedings of the 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, pages 179–186, New York, NY, USA. ACM.

Wilkinson, D. M. and Huberman, B. A. (2007). Cooperation and quality in wikipedia. In *Proceedings of the 2007 International Symposium on Wikis*, WikiSym '07, pages 157–164, New York, NY, USA. ACM.

Xia, F., Liu, T.-Y., Wang, J., Zhang, W., and Li, H. (2008). Listwise approach to learning to rank : Theory and algorithm. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, pages 1192–1199, New York, NY, USA. ACM.

Yeniterzi, R. and Calla, J. (2014). Constructing effective and efficient topic-specific authority networks for expert finding in social media. In *Proceedings of the ACM SIGIR First International Workshop on Social Media Retrieval and Analysis*, SoMeRA '14, pages 45–50, New York, NY, USA. ACM.

Yeniterzi, R. and Calla, J. (2015). Moving from static do dynamic modeling of expertise for question routing in cqa sites. In *Proceedings of the Ninth International Conference on Web and Social Media*, ICSWM'15, pages 702–705. AAAIPress.

Yeniterzi, R. and Callan, J. (2014). Analyzing bias in cqa-based expert finding test sets. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '14, pages 967–970, New York, NY, USA. ACM.

Zhang, J., Ackerman, M. S., and Adamic, L. (2007). Expertise networks in online communities : Structure and algorithms. In *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, pages 221–230, New York, NY, USA. ACM.

Zhang, J., Kong, X., Luo, R. J., Chang, Y., and Yu, P. S. (2014). Ncr : A scalable network-based approach to co-ranking in question-and-answer sites. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, CIKM '14, pages 709–718, New York, NY, USA. ACM.