Intégration du web social dans les systèmes de recommandation

Coriane Nana Jipmo

université PARIS-SACLAY

CentraleSupélec

NNT : 2017SACLC082

Thèse de doctorat

Intégration du Web Social dans les Systèmes de Recommandation

Thèse de doctorat de l'Université Paris-Saclay
préparée à CentraleSupélec

École doctorale n°580 : sciences et technologies de l'information et de la communication (STIC)

Spécialité de doctorat : Informatique

Thèse présentée et soutenue à Orsay, le 19 décembre 2017, par

Coriane NANA JIPMO

Composition du Jury :

Chantal Reynaud Professeur Université Paris-Sud, LRI	Président
Jérôme Azé Professeur Université de Montpellier, LIRMM	Rapporteur
Nicolas Labroche Maître de Conférences Université François-Rabelais de Tours, LI	Rapporteur
Haïfa Zargayouna Maître de Conférences Université Paris 13, LIPN	Examinateur
Nacéra Bennacer Seghouani Professeur CentraleSupélec (MODHEL)	Directrice de thèse
Gianluca Quercini Maître de Conférences CentraleSupélec (MODHEL)	Co-Directeur de thèse
Uriel Berdugo Wepingo	Invité

École Doctorale Sciences et technologies de l'information et de la communication

Thèse de Doctorat de L' université Paris Saclay | CentraleSupelec

Spécialité: Informatique

Intégration du Web Social dans les Systèmes de Recommandation

Présentée par: Coriane Nana Jipmo

19 décembre 2017

Rapporteur :	Jérôme Azé	-	Université de Montpellier, LIRMM
Rapporteur :	Nicolas Labroche	-	Université François-Rabelais de Tours, LI
Examinateur :	Chantal Reynaud	-	Université Paris-Sud, LRI
Examinateur :	Haïfa Zargayouna	-	Université Paris 13, LIPN
Directrice de thèse :	Nacéra Seghouani Bennacer	-	CentraleSupelec, LRI
Encadrant :	Gianluca Quercini	-	CentraleSupelec, LRI
Invité :	Uriel Berdugo	-	Wepingo

Dédicaces

Je dédie cette thèse à mes parents, avec toute mon affection et tout mon respect. Vous ne m'avez pas seulement donné la vie : vous m'avez également permis de la vivre pleinement. Je ne vous en remercierai jamais assez.

Remerciements

Le résultat d'un travail est l'aboutissement de nombreux efforts et de beaucoup de persévérances. Chaque réussite cache derrière elle un ensemble de personnes qui se sont illustrées par leurs aides et conseils, que je souhaite remercier du fond du cœur.

Tout d'abord, je remercie chaleureusement monsieur Nicolas Labroche et monsieur Jérôme Azé d'avoir accepté de rapporter mon manuscrit. Les commentaires et observations ont été enrichissants. J'en suis très flattée d'avoir eu de tels rapporteurs.

Je tiens également à remercier madame Chantal Reynaud et madame Haïfa Zargayouna d'avoir accepté d'être dans mon jury de thèse.

Mille fois merci à ma directrice de thèse, Nacéra Seghouani Bennacer, et mon encadrant Gianluca Quercini de m'avoir toujours soutenue et accompagnée tout au long de cette thèse. Merci pour leur encadrement, orientations, discussions, conseils, motivations et investissement personnel qu'ils ont fournis pour mener au bout mon travail de thèse.

Mon travail a bénéficié de plusieurs collaborations et je tiens à remercier chacun de mes collaborateurs Xiyao Wang, Mohammad Ghufran, Dimitri Lasne, Yufan Zheng pour le travail accompli ensemble.

Je remercie grandement Francesca Bugiotti pour son soutien pendant les bons ainsi que les moments difficiles

Je remercie infiniment mes parents, mes frères et soeurs et toute ma famille pour leur soutien, confiance et encouragements.

Enfin, je ne pourrai jamais assez te remercier, Louis d'avoir été présent tout au long de ma thèse, de m'avoir toujours soutenue malgré la distance les deux premières années.

Table de matières

Chapitre 1

INTRODUCTION GÉNÉRALE

Contexte général

Le Web social croît de plus en plus et donne accès à une multitude de ressources très variées, accessibles à travers la toile. L'expression "Web social" fait référence à une vision d'internet considéré comme un espace de socialisation où les utilisateurs peuvent interagir, publier, produire et partager des ressources. Certaines de ces ressources proviennent entres autres, de sites de partage tels que del.icio.us, d'échange de messages comme Twitter, des réseaux sociaux à finalité professionnelle, comme LinkedIn, ou plus généralement à finalité sociale, comme Facebook et LiveJournal. De plus, un même individu peut être inscrit et actif sur différents réseaux sociaux ayant potentiellement des finalités différentes en publiant des informations diverses et variées.

Un *réseau social* peut être défini comme un ensemble d'individus reliés entre eux par des liens leur permettant de communiquer, de partager, et d'échanger des ressources telles que les photos, les vidéos, les données personnelles, les préférences et leurs activités. C'est un outil de communication qui ne cesse d'évoluer tout en mettant les utilisateurs au centre de leurs préoccupations, dans la perspective de les amener à produire de plus en plus de contenu sur la toile, d'où la création d'un espace personnel communément appelé profil utilisateur dédié à chaque utilisateur. Un *profil utilisateur* est l'ensemble des informations telles que le nom, la localité, les communautés, et les diverses activités mises en avant par un utilisateur dans un réseau social.

Cette croissance du Web social a suscité auprès des entreprises, des chercheurs et de bien d'autres acteurs un fort intérêt à exploiter ces sources d'informations très précieuses et d'une grande importance pour les applications cherchant à caractériser les utilisateurs afin de mieux comprendre leurs besoins et leurs intérêts personnels et professionnels. De nos jours de nombreuses applications permettent à leurs utilisateurs d'accéder à leurs différents profils afin de partager leurs avis sur les produits achetés ou sur des sujets de conversations. En l'occurrence, c'est le cas du site de vente en ligne Amazon[1] qui propose à ses utilisateurs un lien vers des réseaux sociaux comme Facebook, Twitter et Pinterest leur permettant ainsi d'y laisser des notes ou remarques sur les produits achetés ou recommandés. L'accès aux différents réseaux sociaux via ce type d'application pousse une fois de plus les utilisateurs à créer du contenu dans ces réseaux.

L'objectif de notre travail de thèse est d'exploiter le contenu édité par les utilisateurs dans différents réseaux sociaux afin de construire un profil élargi exploitable notamment dans les systèmes de recommandation. Ce travail a été réalisé dans le cadre d'une CIFRE (Convention Industrielle de Formation par la REcherche) en collaboration avec l'entreprise Wepingo.

[1] http://www.amazon.com/

1 Wepingo

Wepingo[2] est une entreprise qui possède un site internet qui assiste les utilisateurs pour effectuer l'achat de produits divers en répondant au mieux à leurs besoins. Le site gère des produits comme les appareils électroménagers, les ordinateurs, les téléviseurs et d'autres produits électroniques, comme les smartphones ou les GPS. Plus précisément, Wepingo agrège et compare les offres de produits provenant de différents sites Web marchands. Les informations sur les produits sont fusionnées et alignées de manière automatique afin de décrire chaque produit par un ensemble de propriétés bien définies. Par ailleurs, pour assister les utilisateurs dans leurs achats, le moteur d'affinité Wepingo propose un ensemble de questionnaires spécifiques pour chaque type de produit afin de guider au mieux l'utilisateur dans ses choix. Par exemple, dans le cas d'un smartphone, il est demandé à l'utilisateur de préciser certaines réponses par rapport aux performances de navigation, énergétiques et à ses préférences en termes de design. De plus, les utilisateurs peuvent se connecter à leurs comptes Facebook et/ou Twitter, ce qui donne à Wepingo les droits d'accès à leurs données personnelles conformément à la charte de confidentialité de Facebook et/ou Twitter. De plus, un utilisateur peut spécifier des liens vers ses profils sur d'autres réseaux sociaux, comme Twitter ou LinkedIn, ce qui permet d'accéder à des informations publiques complémentaires.

L'entreprise Wepingo vise dans le futur à élargir son spectre de produits, ainsi qu'à mettre en place un système qui exploite les informations recueillies sur les utilisateurs afin de fournir des recommandations sur divers produits personnalisés et adaptés à l'utilisateur.

Ce travail s'est déroulé dans le cadre d'une thèse CIFRE (Convention Industrielle de Formation par la REcherche), qui est un dispositif français subventionnant toute entreprise de droit français qui embauche un doctorant pour le placer au cœur d'une collaboration de recherche avec un laboratoire public.

2 Motivations et objectifs

L'exploitation des données du Web social constitue un énorme challenge qui est au cœur d'un nombre important de travaux de recherche. Ces travaux couvrent un large spectre allant de l'analyse des sujets de discussions et des blogs, des relations entre les individus, à l'identification des communautés.

Une *communauté* est un groupe d'utilisateurs partageant un ou plusieurs centres d'intérêt. Le problème de détection de communautés a été largement étudié dans la littérature [55, 54, 8, 14, 76, 56, 16]. Il consiste à identifier les groupes d'utilisateurs similaires par rapport à une métrique de similarité bien précise, par exemple, les utilisateurs partageant une même opinion ou un même genre littéraire. De manière générale, les travaux d'identification des communautés exploitent les graphes où les nœuds représentent les utilisateurs et les arêtes indiquent le type de lien existant entre eux. On distingue plusieurs types de liens, notamment les liens d'amitiés ou de *follower*, les liens familiaux (*sœur* ou *frère* sur Facebook), les liens

[2] http://www.affinytix.com/

de partage ou d'échange de messages. Cette représentation permet d'une part, d'exploiter les algorithmes liés à la théorie des graphes, et d'autre part, de visualiser efficacement les données. Les méthodes d'identification de communautés peuvent être groupées en quatre grandes catégories : celles basées sur le *partitionnement* [54], celles basées sur le *regroupement hiérarchique* [55, 16, 8], celles basées sur le *regroupement partiel* [82] et enfin celles basées sur le *regroupement spectral* [39]. Les approches basées sur le partitionnement fractionnent le graphe des utilisateurs en un nombre fixe de groupes de nœuds de telle sorte que le nombre de liens à l'intérieur (respectivement extérieur) des communautés soit maximal (respectivement minimal). Leur principal inconvénient vient du fait que le nombre de communautés souhaité à la fin du partitionnement doit être initialement fixé. Le regroupement hiérarchique, quant à lui, définit deux méthodes distinctes, à savoir *regroupement ascendant* et *regroupement descendant*. Pour les méthodes *ascendantes*, initialement chaque nœud du réseau constitue une communauté. Par la suite, les nœuds similaires sont successivement fusionnés jusqu'à l'obtention d'un nombre fixe de communautés. En revanche, pour les méthodes *descendantes*, au départ l'ensemble des nœuds du graphe constitue une seule communauté, qui est progressivement divisée en séparant les nœuds dissimilaires. La difficulté ici est de trouver une bonne fonction de similarité (dissimilarité) à appliquer aux nœuds. Quant aux approches basées sur le regroupement partiel, les nœuds du graphe sont répartis en k clusters fixes de manière à maximiser ou minimiser une fonction basée sur la distance entre les nœuds. Cette technique représente chaque nœud du graphe dans un espace à n dimensions, où n est le nombre d'attributs décrivant un nœud. Par contre, comme pour les approches basées sur le partionnement, son inconvénient vient également du fait que le nombre de communautés finale doit être initialement fixé. Et enfin, le partionnement spectral est une méthode de partitionnement en k groupes reposant sur le spectre d'une matrice de similarité.

Par ailleurs, les systèmes de recommandation ont besoin de comprendre leurs utilisateurs afin, d'une part, de personnaliser leurs services, et d'autre part, augmenter leur satisfaction personnelle face aux recommandations proposées. D'après l'encyclopédie WIKIPEDIA les *systèmes de recommandation* sont une forme spécifique de filtrage de l'information visant à présenter les éléments d'information ou items qui sont susceptibles d'intéresser un utilisateur. Les items peuvent être entre autres des films, des musiques, des livres, etc. Leur but est de guider les utilisateurs à découvrir des ressources susceptibles de correspondre au mieux à leurs attentes. De nombreuses approches pour la recommandation ont été proposées dans la littérature [34, 22, 37, 45, 61, 80, 92]. Ces approches peuvent être catégorisées comme suit : celles basées sur le contenu, celles basées sur le filtrage collaboratif, celles basées sur la connaissance, et celles hybrides. Plus précisément, les approches s'appuyant sur le contenu exploitent les items similaires à ceux déjà consommés par l'utilisateur cible pour affiner les recommandations. Cependant, le principal inconvénient de ces approches est qu'elles ont tendance à proposer aux utilisateurs uniquement les items semblables. Au contraire, le filtrage collaboratif exploite plutôt les items consommés par les utilisateurs similaires à l'utilisateur cible. Selon la technique employée, le raisonnement peut être de type *User to*

User (U2U), recherchant les utilisateurs similaires à la cible, ou *Item to Item* (I2I), recherchant les associations fréquentes d'items telles que "*Les personnes ayant acheté cet article ont également acheté ....*". Le principal inconvénient ici est lié aux nouveaux items pas encore consommés (I2I), ainsi qu'aux nouveaux utilisateurs n'ayant pas encore consommés de produits (U2U). Les méthodes basées sur la connaissance utilisent la description sémantique liée aux items et aux utilisateurs. Les approches hybrides quant à elles, combinent les approches sus-citées. En résumé, ces solutions proposées exploitent essentiellement les notes données par leurs utilisateurs sur les produits ou items consommés, ainsi que les interactions de ces utilisateurs avec le système telles que les *click stream*. Par contre, ils ne tiennent pas suffisamment compte des informations éditées par les utilisateurs dans leurs différents réseaux sociaux. De plus, nous observons que la performance des approches de recommandation décrites précédemment dépend fortement de la bonne volonté des utilisateurs à interagir avec le système, particulièrement en donnant des notes sur les items précédemment consommés. Or généralement certains utilisateurs tendent à ne noter que s'ils sont satisfaits et d'autres que s'ils ne le sont pas. Ce qui peut avoir un impact négatif sur le type de recommandation proposé, puisque le système est en possession de peu d'information et ne peut recommander de manière pertinente. Aussi, un système de recommandation devrait être capable de fournir des découvertes inattendues tout en restant pertinent et enrichissant pour l'utilisateur. La mise en œuvre d'une telle solution nécessite d'aller au delà des données liées directement aux items et tenir compte des différentes sources d'information concernant un utilisateur afin de construire un modèle mettant en avant ses activités favorites ainsi que ses intérêts.

L'objectif de ce travail de thèse est d'exploiter les données produites par les utilisateurs dans les différents réseaux sociaux en vue de mieux les caractériser. D'une part, nous nous sommes focalisés sur la mise en œuvre d'algorithmes automatiques et efficaces visant à découvrir les intérêts des utilisateurs à partir de leurs ressources textuelles disponibles sur le Web social. D'autre part, nous avons étudié et analysé la corrélation qui peut exister entre les intérêts des utilisateurs et leurs traits de personnalité, afin de les caractériser davantage.

Les challenges affrontés lors de le mise en œuvre de nos solutions sont les suivants :

- Au vu de sa dimension internationale, le Web est par nature multilingue et par conséquent les informations éditées par les utilisateurs sur le Web le sont également. Notamment, un utilisateur peut à un moment donné utiliser soit sa langue natale, soit sa langue adoptive, soit un mélange des deux langues pour s'exprimer dans ses profils. Comment concevoir une application qui tient compte de cette variation de la langue au niveau des données?

- Les expressions extraites des profils utilisateurs sont intrinsèquement ambiguës. Une expression peut avoir plusieurs interprétations ou significations possibles, déterminer sa bonne interprétation n'est pas une tâche triviale. Comment résoudre le problème

de désambiguïsation du langage naturel lors de l'analyse du contenu des profils utilisateurs?

- Le non respect des règles grammaticales, ainsi que l'utilisation des mots ne se trouvant pas dans les dictionnaires tels que *lol* (lots of laughs), *mdr* (mort de rire), nous empêchent d'utiliser les outils de traitement automatique de la langue naturelle qui intègrent plusieurs type d'analyses du texte, notamment l'analyse sémantique. Comment concevoir une méthode d'analyse sémantique des expressions utilisées dans les profils des utilisateurs?

- Un utilisateur peut avoir plusieurs profils, et ce dans plusieurs réseaux sociaux distincts sans toutefois les spécifier explicitement. Or l'identification des profils d'un même individu dans différents réseaux sociaux est d'une importance capitale, voir bénéfique, car elle permet d'avoir accès à une variété de ressources caractérisant l'utilisateur cible. Notamment, un utilisateur dans son profil Facebook peut mentionner sa vie sociale, par contre sur le réseau LinkedIn, faire part de sa vie professionnelle. Ainsi, agréger les informations d'une même personne, provenant de ses différents profils va nous permettre d'avoir une vue ou un profil complet sur cet utilisateur. Cependant, la difficulté ici est de savoir comment définir une approche qui permet d'identifier les différents profils d'un même utilisateur à partir de plusieurs réseaux sociaux?

- Nous observons généralement que l'évolution des évènements dans l'actualité amène certains individus à réagir de façon instantanée en mettant en avant, dans leurs multiples profils, leurs opinions par rapport à la thématique portée par ces évènements. Certains utilisateurs sont entraînés par les évènements du moment, par exemple les utilisateurs qui publient des commentaires portant sur les candidats lors des élections présidentielles. A cet effet, comment concevoir une application qui est en mesure de capturer dans le temps, les domaines d'intérêts dominants et réels d'un utilisateur?

3 Contributions de la thèse

Nos principales contributions pour traiter les différents challenges sus-cités sont les suivantes :

1. **Construction du profil d'intérêts d'un utilisateur.** Dans cette partie, notre objectif est de définir des approches automatiques, efficaces et non supervisées, qui exploitent les différentes sources d'informations du Web social afin de découvrir les intérêts d'un utilisateur. Nos approches proposées, qui utilisent essentiellement des ressources textuelles, sont robustes au bruit et tiennent compte de l'évolution des données dans le temps. De plus, compte tenu du fait que les informations éditées par les utilisateurs au sein de leurs réseaux sociaux sont fortement hétérogènes, la construction de leurs profils d'intérêts respectifs relève d'un processus non trivial. Nous

avons testé nos approches sur deux jeux de données réelles constitués des utilisateurs des réseaux sociaux LIVEJOURNAL et TWITTER.

2. **Analyse des traits de personnalité des utilisateurs.** Dans cette partie, notre but est d'exploiter les sources de données et les intérêts de chaque utilisateur, dans l'intention de faire une étude sur leurs différents traits de personnalité afin de les caractériser davantage. En d'autres termes, nous faisons ici une analyse sur la corrélation entre les intérêts d'un utilisateur et ses traits de personnalité. Notre analyse s'est faite sur un jeu de données réelles constitué principalement des informations extraites des profils des utilisateurs du réseau TWITTER.

3. **Réconciliation des profils utilisateurs.** Notre objectif dans cette section est d'exploiter les différents réseaux sociaux d'un même utilisateur afin de construire pour ce dernier un profil utilisateur élargi. Plus précisément, nous cherchons ici à découvrir à travers différents réseaux sociaux, les différents profils d'un même utilisateur afin d'exploiter les informations y provenant. L'idée est d'agréger les informations d'un même utilisateur provenant de ses différents espaces personnels afin de mieux comprendre ses différents besoins. Nous avons également testé notre approche sur un jeu de données réelles constitué de quatre réseaux sociaux LIVEJOURNAL, FLICKR, TWITTER et YOUTUBE.

4 Plan du manuscrit

Ce manuscrit est organisé en quatre chapitres.

Chapitre II: Découverte des intérêts des utilisateurs dans les réseaux sociaux.

Dans ce chapitre, nous présentons tout d'abord les travaux existants portant sur la caractérisation des individus dans les réseaux sociaux, suivis des concepts sous-jacents aux algorithmes que nous proposons. Ensuite, nous décrivons plus en détails nos deux approches de construction automatique des profils d'intérêts des utilisateurs à partir des réseaux sociaux, à savoir DELVE[3] et FRISK[4]; ainsi que les expérimentions et évaluations réalisées sur les données réelles. Et enfin, nous finissons par une conclusion.

Chapitre III: Analyse des traits de personnalité des individus.

L'objectif de ce chapitre est de faire une analyse de la corrélation entre les traits de personnalité et les intérêts des utilisateurs. Tout d'abord, nous présentons les travaux existants portant sur l'identification des traits de personnalité des individus. Ensuite, nous décrirons les concepts manipulés lors de notre analyse. Par la suite, nous présentons en détail notre méthodologie ASCERTAIN[5], suivie des expérimentations et des analyses faites sur un jeu de données réelles. Et enfin, nous achevons ce chapitre par une conclusion.

[3] DiscovEr LiVejournal intErests [4] Find twitteR InterestS via wiKipedia

[5] AnalySis Correlation pERsonality Traits And INterests

Chapitre IV: Réconciliation des profils utilisateurs dans les réseaux sociaux.

Un utilisateur peut avoir plusieurs profils dans différents réseaux sociaux. Dans ce chapitre nous présentons notre approche de réconciliation des profils utilisateurs à travers différents réseaux sociaux. Nous débutons par une présentation d'une part, des travaux de recherche liés à notre problématique, et d'autre part, des notions et notations utilisées pour formaliser notre problème. Ensuite nous détaillons notre approche de réconciliation proposée. En définitive, nous décrivons les expérimentations faites sur un jeu de données réelles, et nous terminons par une conclusion.

Chapitre V: Conclusions et perspectives.

Dans ce chapitre, nous faisons la synthèse de nos travaux de recherche et nous présentons nos différentes perspectives.

Chapitre 2

DÉCOUVERTE DES INTÉRÊTS DES UTILISATEURS DANS LES RÉSEAUX SOCIAUX

Introduction

La prolifération des réseaux sociaux au sein de notre société, ainsi que dans notre sphère quotidienne, prend de plus en plus une place grandissante. Selon la publication d'une étude de *l'observatoire des usages internet de Médiamétrie*[1], en août 2010, plus de 20 millions de Français se sont inscrits sur un site communautaire et plus de 8 millions ont consulté leur site quotidiennement. Ainsi, le nombre d'inscriptions sur ces sites a augmenté de 26% en un an. Cet engouement global envers les réseaux sociaux a un impact significatif sur notre quotidien, d'autant plus qu'ils incitent de plus en plus leurs utilisateurs à renseigner beaucoup plus d'informations sur la toile en rapport avec leurs vies privées, sociales ou professionnelles.

Dans ce chapitre, nous étudions le problème de découverte des intérêts des utilisateurs à partir des ressources textuelles extraites de leurs profils sur différents réseaux sociaux. Plus précisément, notre objectif est de construire pour chaque utilisateur un *profil d'intérêts* qui est l'ensemble constitué de ses différents intérêts ou domaines d'intérêts (politique, économie, sports) classés par ordre de pertinence. Le processus de découverte automatique des domaines d'intérêts est d'une importance capitale pour les applications cherchant à comprendre leurs utilisateurs comme dans les systèmes de recommandation.

Lors de la mise en œuvre de notre approche de construction automatique du profil d'intérêts d'un utilisateur, nous nous sommes confrontés aux challenges suivants:

1. Le Web est par nature multilingue au vue de sa dimension internationale, par conséquent les réseaux sociaux le sont également. De ce fait, les informations provenant des profils utilisateurs sont éditées par ces derniers, soit dans leur langue natale, soit dans leurs langues adoptives, soit dans un mélange de deux ou plusieurs langues. Or selon les estimations de Wikipedia, il existe entre 3 000 et 7 000 langues vivantes. Comment concevoir une approche qui tient compte de cette variation de la langue au niveau des données?

2. Chaque réseau social a son orientation personnelle, qui peut être par exemple, à finalité professionnelle (c'est le cas du réseau social LinkedIn) ou plus généralement à finalité sociale (comme Facebook et LiveJournal). En fonction de son orientation, on distingue différents types de ressources (photos, vidéos, messages, textes, etc.) partagées par les utilisateurs. Dans notre travail, nous nous sommes focalisés essentiellement sur des ressources textuelles. Ainsi, le problème posé se concentre principalement sur les différentes formes de description d'une ressource textuelle : soit un ensemble de liste de mots sans contexte, soit un texte avec un nombre de caractères fixe, soit un texte sans limite de nombre de caractères. Comment définir une approche qui tient compte de ces différentes formes que peut prendre une ressource textuelle?

3. Les expressions extraites des données issues d'un profil utilisateur sont intrinsèquement

[1] https://www.slideshare.net/onprenduncafe/mdiamtrie-observatoire-des-usages-internet

ambiguës. Autrement dit, une expression peut avoir plusieurs interprétations possibles. Par exemple, d'après la version française de Wikipedia l'expression "*Game*" signifie entre autres, soit un genre musical, soit un rappeur américain né en 1979, soit un film d'action indien de Abhinay Deo (2011). Choisir la bonne interprétation parmi les possibles n'est pas une tâche évidente. Par conséquent, comment résoudre le problème de désambiguïsation du langage naturel lors de l'analyse des ressources textuelles des profils utilisateurs?

4. Le non respect des règles grammaticales, l'utilisation de mots ne se trouvant pas dans les dictionnaires tels que *lol* (lots of laughs), *mdr* (mort de rire), ainsi que la nature multilingue des ressources exploitées nous empêche de bénéficier des outils de traitement automatique de la langue naturelle qui intègrent plusieurs programmes d'analyse textuelle, notamment l'analyse sémantique. De ce fait, comment concevoir une méthode d'analyse sémantique multilingue des expressions utilisées dans les profils utilisateurs en vue de comprendre leur sémantique et par la suite découvrir les domaines d'intérêts qui y sont évoqués?

Dans ce chapitre nous proposons une solution qui va au delà des challenges précédemment cités. Nos contributions sont les suivantes:

1. Nous exploitons les ressources textuelles des réseaux sociaux afin de concevoir une approche qui construit automatiquement le profil d'intérêts d'un utilisateur. Notre système s'appuie essentiellement sur l'encyclopédie Wikipedia afin de déterminer les possibles interprétations des différentes expressions issues du profil d'un utilisateur, afin de faire émerger leurs intérêts. Généralement, une personne peut avoir plusieurs intérêts, chacun ayant une importance particulière. D'où la nécessité de classer les intérêts découverts par ordre d'importance, ce qui permet de dissocier les activités favorites des utilisateurs de celles moins pratiquées. Par exemple, on peut préférer faire plus du sport que de suivre la politique, bien que nous sommes passionnés par ces deux domaines d'intérêts.

2. Nous définissons une approche multilingue qui découvre les intérêts des utilisateurs à partir des données écrites dans différentes langues. La méthode proposée ne fait aucune hypothèse sur la langue utilisée par les utilisateurs dans leurs différents profils, ce qui permet l'utilisation de notre approche dans un contexte multilingue.

3. Nous proposons une approche non supervisée, qui cherche à découvrir les intérêts d'un utilisateur et non à classer un utilisateur parmi un ensemble de catégories d'intérêts déjà prédéfinies. De plus, nous n'avons pas besoin au préalable de construire un modèle d'apprentissage, qui est très coûteux en temps car il nécessite une collection et une validation d'un jeu de données assez représentatif. Par ailleurs, compte tenu de la variation de la langue au sein des profils utilisateurs, les approches supervisées ont besoin d'un échantillon de données représentatives et validées par langue et par

catégorie cible, ce qui n'est pas une tâche facilement réalisable au vue des différentes langues existantes. Et même dans la mesure où on se dit qu'il y a un nombre de langues raisonnable, c'est toujours une nouvelle tâche fastidieuse qui se rajoute aux méthodes supervisées que nous évitons dans notre approche.

Le plan de ce chapitre se présente comme suit. Tout d'abord dans la section 1, nous présentons un aperçu des différents travaux existants portant sur la caractérisation des utilisateurs. Dans la section 2, nous décrivons les différents types de réseaux sociaux, ainsi que les concepts existants manipulés, notamment l'encyclopédie WIKIPEDIA, la mesure de similarité *Wikipédia Link Measure* (WLM) [49] et l'algorithme de *PageRank* [9]. Ensuite, dans les sections 3 et 4, nous présentons respectivement nos deux approches de découverte des intérêts des utilisateurs dans un réseau social: DELVE [35] et FRISK [**?**]. Chacune de ces approches exploitent deux types de ressources textuelles différentes. L'une utilise une liste d'expressions telles que *books*, *computers*, *cryptography*, *father ted*, *frasier*, et l'autre, des textes courts ayant un nombre de caractères très réduit tel que "*Jules going to the store without a budget*". L'intérêt de ces deux approches proposées est d'évaluer les ressources textuelles des profils utilisateurs suivant les configurations identifiées. Et enfin, nous terminons par une conclusion suivie de nos perspectives.

1 État de l'art

La caractérisation des utilisateurs à partir des ressources dont ils sont à l'origine sur le Web social, a fait et continue de faire l'objet de nombreux travaux de recherche. En fait, elle consiste à définir différents éléments d'information permettant de comprendre, ainsi que de connaître les besoins des utilisateurs tels que leurs intérêts. La plupart des approches dans la littérature exploitent le contenu textuel généré par ces utilisateurs sur le Web social pour inférer leurs intérêts. Dans le même ordre d'idée, nous nous sommes essentiellement focalisés sur les études portant sur l'exploitation des ressources textuelles. Les auteurs de ces études ont proposé plusieurs approches que nous avons réparti suivant trois grands aspects. Le premier infère les sujets d'intérêts des utilisateurs à partir de ce qu'ils, ou leur entourage, publient sur le Web [19, 47, 67, 74, 83, 84, 86, 87, 88, 89, 91]. Le second exploite les mots clés ou étiquettes utilisés par les utilisateurs pour organiser leurs ressources favorites (marque-page) [43, 77, 85]. En effet, le marque-page ou signet est la version électronique du marque-page papier, qui permet de garder en mémoire le numéro des pages désirées afin de pouvoir y accéder ultérieurement. Et enfin, le dernier aspect se base sur les célébrités que les utilisateurs aiment, ou suivent leurs actualités [6, 31] afin de les caractériser.

Sur TWITTER, certains utilisateurs mentionnent leurs intérêts dans leur biographie. A cet effet, Ding et al. [19] exploitent ces données biographiques par le biais de champs aléatoires conditionnels [42], une classe de modèles statistiques utilisés en reconnaissance des formes et plus généralement en apprentissage statistique. Malheureusement, très peu d'utilisateurs prennent le temps de remplir sérieusement leur biographie, d'où l'utilisation

des tweets dans la majorité des travaux. Vu et Perez [83] utilisent des expressions régulières pour extraire les mots clés contenus dans les tweets postés par les utilisateurs. D'après eux, ces mots clés, qui peuvent être constitués d'un ou de plusieurs termes, représentent au mieux les intérêts des utilisateurs. Après leur identification, les mots clés sont classés par fréquence d'apparition en utilisant les méthodes de pondération les plus populaires telles que le tf-idf ou TextRank [48]. Par contre, les techniques d'apprentissage supervisées [67, 74], quant à elles, créent des modèles à partir des données textuelles, qui sont généralement des valeurs d'attributs extraites des profils utilisateurs telles que le sexe, la localité ou le nombre de connexions d'un utilisateur. Plus clairement, Raghuram et al. [67] utilisent les techniques de classification traditionnelles pour déterminer les intérêts des utilisateurs de Twitter. Ils utilisent principalement trois grands attributs, à savoir : les informations propres à l'utilisateur (sexe, localité, et les profils des amis et followers), les caractéristiques des tweets des utilisateurs (tf-idf des termes), et les mesures statiques qui expriment la fréquence d'édition des tweets des utilisateurs telles que la moyenne, l'écart-type, le maximum ou le minimum. Quant à Spasojevic et al. [74], ils analysent le contenu des profils utilisateurs, en extrayant les expressions (classées par groupes) appartenant à un dictionnaire interne qui contient deux millions d'expressions et qui est généralement mis à jour, parfois manuellement ou en Freebase qui est un projet collaboratif libre de rassemblement et de connexion des connaissances du web, sous forme sémantique, soit les concepts Wikipédia ou soit les expressions les plus utilisées par les utilisateurs influents. A chaque groupe d'expressions est assigné, en s'inspirant d'une ontologie, un ensemble de sujets probables, parmi lesquels chaque sujet possède un score qui exprime son degré de liaison aux expressions du groupe cible. Chaque utilisateur est donc représenté par un vecteur de score de chaque sujet. Malencontreusement, les techniques d'apprentissage ont toujours besoin d'une intervention humaine pour la construction de la base de données d'apprentissage.

Il existe plusieurs autres approches [84, 88, 89] basées spécialement sur un principe nommé *l'allocation de dirichlet latente* [7] connu sous le sigle de LDA, très utilisé pour déterminer les intérêts des utilisateurs à partir d'un document textuel. En effet, LDA est un modèle probabiliste de sujets définis, dans lequel chaque sujet est représenté par une distribution de probabilité sur des termes. Il prend en entrée un ensemble de mots issus d'un document, qu'il voit comme un mélange de sujets, selon une certaine distribution de probabilité. Par exemple, 60% de « politique », 40% de « sport ». Chaque sujet est également considéré comme une distribution de probabilité par rapport aux mots qu'il contient. Ainsi, on peut dire dans la catégorie « politique » on a 3% de mots qui ont un lien avec les « élections », 2% avec la « campagne électorale », 5% avec le « Vote ». Pendant que Weng et al. [88] appliquent LDA tel que décrit à la collection des tweets postés par un utilisateur, Xu et al. [89] proposent une variation de LDA qui filtre les tweets bruités qui n'indiquent nécessairement pas les intérêts d'un utilisateur. Plus précisément, l'idée de Xu et al. est d'inclure dans le modèle LDA basique une information supplémentaire sur la prise en compte ou pas d'un tweet, notamment si c'est un tweet, une réponse à un tweet ou un retweet. Pour cela, ils ont fait un sondage, à l'issue duquel ils ont découvert que la plupart des retweets et

des tweets contenant des URLs sont beaucoup plus liés aux intérêts des utilisateurs, tandis que d'autres tweets (par exemple, ceux contenant les tags et les réponses) sont généralement des tweets conversationnels qui ne contiennent pas d'informations portant sur les intérêts des utilisateurs.

Les auteurs de [85] par contre ne cherchent pas à inférer les intérêts des utilisateurs, mais plutôt à déterminer si deux utilisateurs possèdent des intérêts similaires en fonction des labels qu'ils utilisent pour marquer leurs ressources comme les blogs. D'un autre côté, l'approche décrite par Bhattacharya et al. [6] infère les intérêts des utilisateurs à partir des célébrités qu'ils suivent. Plus clairement, chaque utilisateur célèbre (politicien, musicien ou journaliste) appartient habituellement à plusieurs listes créées par les utilisateurs réguliers. Chaque liste possède un nom et contient une description abrégée, à partir de laquelle ils extraient un ensemble de tags ou labels. Ainsi, on peut donc décrire chaque célébrité par une liste de tags, et par conséquent, représenter également les utilisateurs par un vecteur d'intérêts (intérêt = tag). A chaque intérêt est affecté un score, qui est simplement le nombre de célébrités qui porte cet intérêt et qui sont suivies par l'utilisateur cible. Les intérêts ayant un score d'au moins trois sont considérés comme ceux de l'utilisateur cible. La difficulté ici se présente lors du traitement d'une part des utilisateurs ne suivant aucune ou très peu de célébrités, et d'autre part, des listes mal éditées ou renseignées.

Par ailleurs, comme nous certains approches cherchent à lier les mots ou expressions qui se trouvent dans les tweets aux articles Wikipédia [47, 91]. Notamment, Zarrinkalam et al. [91] qui ont créé une représentation sous forme de graphe des termes contenus dans les tweets postés par un utilisateur pendant un intervalle de temps. Les nœuds du graphe représentent les articles Wikipédia correspondant aux concepts associés aux termes des tweets et les arcs, pondérés par la valeur de similarité existant entre les nœuds. La principale différence avec notre approche est que les intérêts découverts ici sont représentés par une collection d'articles Wikipédia décrits par eux comme sémantiquement proches, une collection d'articles qui n'est pas toujours facile à interpréter. Par contre, nous proposons une méthode beaucoup plus précise qui découvre les domaines d'intérêts des utilisateurs, tels que *politique*, *économie* ou *sport*. D'un autre côté, Michelson et Macskassy [47] identifient les entités nommées existantes dans les tweets, ensuite les associent à leurs articles puis catégories Wikipédia respectifs, et enfin extraient parmi les catégories trouvées celles qui apparaissent les plus fréquemment. La reconnaissance des entités qui est une sous-tâche de l'activité d'extraction d'information dans les corpus documentaires, consiste à rechercher des objets textuels (c'est-à-dire un mot, ou un groupe de mots) pouvant être catégorisés dans des classes telles que noms de personnes, noms d'organisations ou d'entreprises, noms de lieux, quantités, distances, valeurs, dates, etc. Par exemple, aux mots anglais « Arsenal » et « Walcott » correspondent les articles Wikipédia « Arsenal F.C. » et « Theo Walcott » respectifs, et avec comme catégorie la plus fréquente « Football in England ». En réalité, leur méthodologie trouve les catégories qui sont les intérêts d'un utilisateur, situées à une distance de cinq et dans une direction bien précise, des nœuds articles représentant les entités identifiées. Le principal inconvénient ici est que les catégories les mieux classées comme par exemple « 2010 Fifa World Cup Players

» peuvent être trop spécifiques et ne mettent pas en avant l'intérêt de l'utilisateur, aussi bien que « sports » par exemple. De plus, les intérêts sont déterminés uniquement à partir des entités nommées extraites des tweets, ce qui est trop restrictif et entraîne une perte d'information nécessairement exploitable.

En résumé, à notre connaissance nous observons que parmi les challenges sus-citées, les points abordés dans la littérature tiennent compte que d'une part, du problème d'ambiguïté du langage naturel dans un contexte purement monolingue, et d'autre part, des différentes variations que peuvent prendre une ressource textuelle dans un profil utilisateur (signet, tags, document de texte). Cependant, le côté multilingue des ressources textuelles que nous abordons n'est pas pris en compte dans ces travaux sus-cités. Et par conséquent, nos approches proposées ne peuvent bénéficier des outils actuels de traitement automatique de la langue naturelle, qui intègrent plusieurs programmes d'analyse textuelle, notamment l'analyse sémantique. Autrement dit, les approches de désambiguïsation existantes ne peuvent être appliquées telles qu'elles sont décrites dans la littérature.

Dans la suite nous présentons les différents types de réseaux sociaux, ainsi que les concepts que nous avons manipulés dans nos approches.

2 Préliminaires

Comme mentionné précédemment, dans cette section nous présentons tout d'abord les différents types de réseaux sociaux existants. En effet, la typologie d'un réseau social nous renseigne sur le type de contenu que peuvent prendre les profils utilisateurs de ce réseau. Par la suite, nous poursuivons par une description successive des concepts sur lesquels nos approches s'appuient pour construire automatiquement le profil d'intérêts des utilisateurs. Nous débutons par définir l'encyclopédie Wikipedia, notamment ses différents types de pages : page article, page de désambiguisation, page de redirection et catégories, ainsi que les différents types de relation existants entre ces pages. Ensuite nous définissons la mesure de similarité *Wikipédia Link Measure* (WLM) [49], qui permet de quantifier à quel point deux pages Wikipedia sont similaires, c'est-à-dire décrivent une même notion. Enfin, nous présentons l'algorithme *PageRank* [9] qui permet de mesurer l'importance d'une page Web.

2.1 Typologie de réseaux sociaux

L'expression *réseau social* renvoie à un usage social d'internet notamment aux services de réseautage social. Un service de réseautage social est l'ensemble des moyens en ligne mis en œuvre pour relier via des liens hypertextes des profils utilisateurs. Le premier site Web de réseautage social fut *Classmates.com* créé en 1995. L'objectif de Classmates était d'aider ses inscrits à retrouver leurs amis d'école primaire, collège, lycée, université, anciens collègues et anciens combattants. En 2011, Classmates a changé de nom pour devenir *MemoryLane.com*. Leur objectif a été recentré sur la proposition de contenus multimédias à forte valeur nostalgique tels que les archives de magazines, films, chansons, etc.

Généralement, les sites de réseautage social sont orientés vers le web 2.0, c'est-à-dire qu'ils permettent à leurs utilisateurs d'être des participants actifs au sein du réseau, et non plus de simples visiteurs de pages statiques. Le profil d'un utilisateur au sein d'un réseau social peut être public, dans ce cas tous les individus reliés ou pas à l'utilisateur cible par un lien quelconque peuvent y accéder en lecture. Par contre, dans le cas d'un profil privé, le contenu n'est lisible que par les individus reliés à l'utilisateur cible par un lien bien précis. La plupart des réseaux sociaux intègrent communément deux types de liens : *lien d'amitié* et *lien de communauté*. Plus précisément, le lien d'amitié existe entre deux utilisateurs et indique que les utilisateurs cibles sont des amis, tandis que le lien de communauté existe entre un utilisateur (personne) et un organisme (communauté), et indique que l'utilisateur cible fait partie de la communauté rattachée.

Un réseau social fonctionne de manière récursive. On part d'une personne physique ou morale qui au préalable a déjà créé son profil utilisateur et qui envoie des messages invitant d'autres utilisateurs à se connecter à elle, c'est-à-dire à se lier à elle via un lien dont le type est défini dans l'invitation. Les nouveaux utilisateurs qui se sont alliés peuvent répéter le processus, accroissant le nombre de liens dans le réseau. De proche en proche le contenu et les liens au sein du réseau social se construisent. On distingue plusieurs types de réseaux sociaux, en fonction de leur orientation, ainsi que du type de données partagées par les utilisateurs dans leurs profils. Nous avons par exemple ceux organisés autour des relations d'affaires, c'est le cas de LinkedIn ou de Viadeo, dans lesquels les utilisateurs décrivent leurs parcours professionnels. De manière générale, nous catégorisons les réseaux sociaux en sept grands groupes, à savoir :

1. **Les réseaux sociaux dits « généralistes »** : ce sont des réseaux de personnes connectées par des relations d'amis, ou de fans pour discuter.

   Facebook : Chaque internaute a la possibilité de créer son profil dans lequel il existe une limite sur la taille du réseau d'amis (individus proches ou inconnus) qu'il peut construire. Ce réseau permet de partager : statut, photos, liens et vidéos. Il est également utilisé par certains établissements ou créateurs pour leur promotion grâce aux pages fans accessibles à tous.

   Twitter : outil de *microblogging* (service en ligne de textes courts) qui permet d'envoyer des « tweets » ou messages courts aux internautes (followers) qui suivent chaque compte. Autrement dit, les « followers » d'un compte sont ceux qui y sont abonnés. Par contre, les « followees » d'un utilisateur sont ses abonnements c'est-à-dire les comptes que ce dernier suit.

   MySpace : espace Web personnalisé dans lequel les utilisateurs ont la liberté de présenter des informations personnelles et de faire un blog. Ce réseau est notamment connu grâce aux nombreux chanteurs qui ont pris possession de cet espace, sa popularité a baissé ces dernières années.

   D'autres moins connus : *Beboomer*, pour les plus de 45 ans et *Cafemom*, pour les mamans.

2. **Les réseaux dits « de partage »** : ce sont des réseaux de personnes qui publient des contenus générés tels que audios, vidéos, photos, etc.

   YouTube & Dailymotion : ce sont des sites qui permettent de mettre en ligne et partager des vidéos, ainsi que des audios. On peut y trouver tous types de vidéos politiques, d'humour, de sport, de musique, de cinéma, d'art, etc. Ils proposent aux utilisateurs la possibilité de laisser un commentaire sous chaque vidéo.

   Flickr : c'est un site web de partage de photos et de vidéos gratuit, avec certaines fonctionnalités payantes. Il est aussi souvent utilisé par des photographes professionnels pour présenter leurs œuvres.

3. **Les réseaux sociaux dits « professionnels »** : ce sont des réseaux de personnes partageant leur vie professionnelle.

   LinkedIn : c'est un réseau social professionnel qui permet à ses utilisateurs de mettre en ligne leur Curriculum Vitae, afin qu'ils soient accessibles aux recruteurs.

   Viadeo : il permet d'établir des contacts professionnels, c'est-à-dire faire connaître leurs utilisateurs en publiant leurs Curriculum Vitae et en proposant également des offres d'emploi.

   Ziki : il a pour but d'aider les entreprises à trouver les meilleurs prestataires de services pour la réalisation de leurs projets.

   InterFrench : c'est un réseau social francophone mondial pour les projets à l'international.

4. **Les réseaux dits de « rencontre »** : ce sont des réseaux qui permettent de faire des rencontres amoureuses.

   Nous avons les réseaux sociaux tels que *Meetic*, *Match.com*, *Be2*, *Adoptunmec* ou encore *Lovinside*.

   Il existe également des réseaux sociaux dédiés à la communauté gay tels que *CiteGay*, *Cleargay* et *Ohlalaguys* qui lui est européen.

5. **Les réseaux dits de « services »** : ce sont des réseaux utilisés pour se rendre services entre utilisateurs.

   Ma-residence : c'est un lieu d'échange, de services et de discussions entre les voisins.

   Copains d'avant et *Trombi* : permettent de retrouver des anciens camarades de classe. Par contre, *RéseauxLycée* et *Etnoka* sont des réseaux pour lycéens et étudiants où il est possible de discuter, organiser des soirées ainsi que partager des cours.

   BeGlob : c'est un réseau dédié aux passionnés de voyages, qui permet d'échanger les bons plans, conseils, et expériences.

6. **Les réseaux dits de « Politique »**

   Coolpol : c'est le réseau social du parti socialiste de « toutes celles et de tous ceux qui veulent débattre et agir à gauche! » selon le site. C'est un lieu de discussion ou

les sympathisants du parti peuvent échanger. On y retrouve les évènements, débats, partage d'idées, des vidéos, etc.

Créateurs de possible : c'est le réseau social de l'UMP lancé en janvier 2010. Il propose des fonctionnalités similaires à Coolpol.

7. **Les réseaux dits de « géolocalisation »**:

   Foursquare et *Gowalla* : c'est un réseau qui donne la possibilité d'ajouter des amis lorsqu'on se rend dans un endroit précis, ainsi que de signaler sa présence.

   Yelp quant à lui son but est de connecter les gens aux meilleurs commerces de proximité.

2.2 Wikipedia

Wikipedia est un projet d'encyclopédie multilingue, universelle, créé par Jimmy Wales et Larry Sanger le 15 janvier 2001 sous le nom de domaine wikipedia.org. L'encyclopédie est en libre accès, en lecture ainsi qu'en écriture, en d'autres termes, n'importe qui peut y accéder et modifier la quasi-totalité du contenu de ses pages. Tous les rédacteurs des pages Wikipedia sont des bénévoles qui coordonnent leurs travaux au sein d'une communauté collaborative. Le but de Wikipedia est d'offrir un contenu librement réutilisable, objectif et vérifiable. D'après Wikipedia, en 2017, il a été classé comme le sixième site le plus fréquenté au monde et en février 2014, prêt de 500 millions de visiteurs dans le monde l'ont exploré chaque mois.

Il existe plusieurs versions ou éditions de Wikipedia, chacune d'elle dans une langue bien précise. En effet, l'ensemble constitué des pages écrites dans une même langue et des liens existants entre ces pages forment une version de Wikipedia dans cette langue. En 2016, on a compté environ 283 langues parmi lesquelles l'anglais est la principale langue utilisée, qui a toujours conservé cette importance, et avec un nombre de pages qui s'élève à plus de 5 millions.

Une page Wikipedia ou plus simplement *page* est semi-structurée, c'est-à-dire son contenu n'est pas organisé comme dans une base de données, mais comporte néanmoins des métadonnées qui le rend plus facile à traiter par rapport aux données brutes. Chaque page est identifiée par un *titre* qui donne une idée générale sur son contenu. Ce contenu est généré à partir d'un code appelé *wikicode* ou *wikitexte* qui est un langage de balisage léger qui définit la mise en forme de saisies du contenu d'une page Wikipedia . Il existe quatre types de pages Wikipedia : *article*, *désambiguïsation*, *redirection* ou *catégorie*.

Page article ou encore appelée simplement *article*. C'est une page dont le contenu décrit un concept ou un sujet bien spécifique tels que des évènements historiques, l'art, des personnes etc. Le titre d'une page de type article désigne le sujet traité dans son contenu. La figure 2.1 est l'entête de l'article ayant pour titre "Paris" dans la Wikipedia française. On constate que les éditeurs de la Wikipedia française ont été majoritairement d'accord sur le fait que le mot "*paris*" doit être associé à la capitale française. Globalement, si un mot

est associé à une page de type article, alors cette page est appelée la *page par défaut* de ce mot. Par exemple, l'article ayant pour titre "Paris" est la page par défaut du mot "*paris*". Cependant, le mot "*paris*" peut avoir d'autres interprétations, d'où la nécessité d'associer à sa page par défaut une page de désambiguïsation, qui contient toutes ses interprétations possibles.

Figure 2.1: Entête de l'article de la WIKIPEDIA française ayant pour titre Paris.

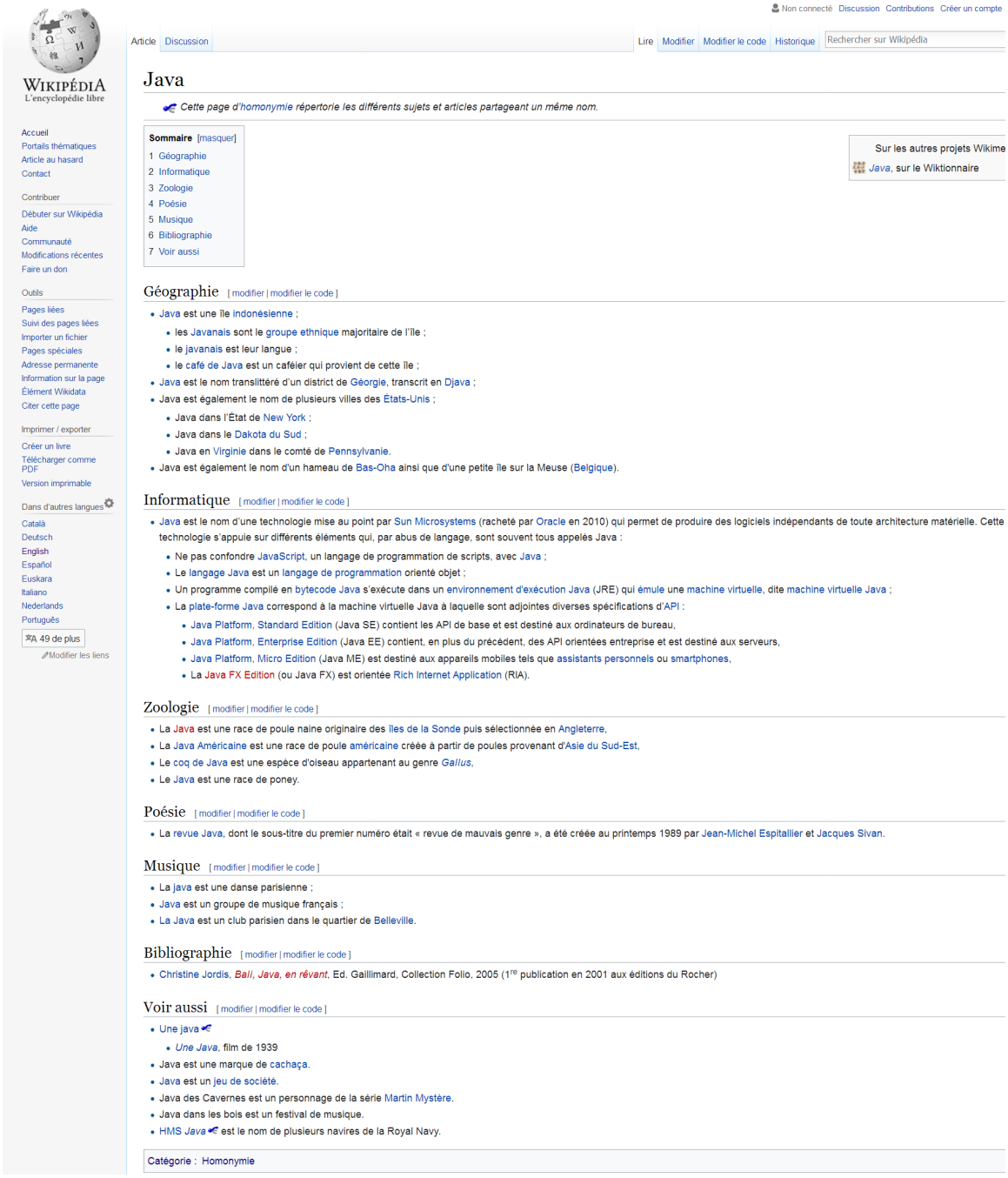
Page de désambiguïsation. C'est une page dont le titre est une expression ambiguë d'après WIKIPEDIA. De manière générale, une page est ambiguë lorsque les éditeurs de la WIKIPEDIA à laquelle elle appartient, n'ont pas été majoritairement d'accord sur la signification par défaut à associer au titre de cette page. C'est le cas de la page de désambiguïsation de la figure 2.2 associée à l'expression "*java*" qui est ambiguë car elle peut faire référence entre autres, au langage de programmation Java, à une île en Indonésie, ou encore à un village dans l'état de Virginie aux États-Unis. Le contenu d'une page de désambiguïsation est constitué de toutes les interprétations (homonymes) possibles, qui sont également des pages WIKIPEDIA que peuvent prendre son titre. Quelquefois, certains articles (ayant pour titre *title*) peuvent être associés à une page de désambiguïsation. Dans ce cas, le titre de la page de désambiguïsation associée est de la forme *title*(*homonymie*) dans le cas de la

WIKIPEDIA française, *title*(*disambiguation*) dans le cas de l'anglaise, etc. Par exemple, la page de désambiguïsation associée à l'article "Paris" sus-cité précédemment, est "Paris (homonymie)". Cependant, d'autres expressions comme "*cryptographie*" associée à la page par défaut *Cryptographie*, ne sont pas ambigües par nature et n'ont donc pas de page de désambiguïsation associée.

Page de redirection ou tout simplement *redirection*. Ce type de page n'a pas de contenu textuel, il fournit juste un lien vers une page article dont le titre est un alias. En d'autres termes, ce type de page nous redirige vers une page de type article. Le titre d'une page de redirection est un synonyme du titre de la page article vers laquelle elle nous redirige. Par exemple, la page de redirection de titre "*Paris (France)*" indiquée à la figure 2.3 nous redirige vers la page article ayant pour titre "*Paris*" de la figure 2.1.

Catégorie. Le titre de ce type de page fait référence à une thématique bien précise par exemple, *Sport*, *Politique*, *Jeu*, etc. De même, le titre d'une page catégorie est préfixé de l'expression "Catégorie:" pour la WIKIPEDIA française, "Category:" pour la WIKIPEDIA anglaise, etc. Chaque page de type article est classée selon sa description dans une ou plusieurs catégories parents, qui sont organisées de manière hiérarchique. En d'autres termes, les catégories sont classées dans d'autres pages catégories ayant une thématique plus large. De proche en proche, on accède à des catégories couvrant des domaines de plus en plus vastes, ce qui constitue donc une arborescence de catégories. Par exemple, l'article `https://fr.wikipedia.org/wiki/Football` de la WIKIPEDIA française appartient aux catégories ayant pour titre *Catégorie:Sport originaire d'Angleterre*, *Catégorie:Football* comme l'indique le schéma 2.4 . En d'autres termes les catégories *Catégorie:Football* et *Catégorie:Sport originaire d'Angleterre* sont les catégories directes de l'article *Football*. On observe que la catégorie *Catégorie:Football* possède trois catégories parents à savoir *Catégorie:Sport collectif*, *Catégorie:Sport olympique*, et *Catégorie:Sport de ballon*. De même, la catégorie *Catégorie:Sport originaire d'Angleterre* possède une catégorie parent *Sport en Angleterre*. Au fur et à mesure qu'on remonte dans l'arbre des catégories WIKIPEDIA on accède à des catégories de plus en plus génériques telles que *Sport* ou *Jeu de ballon*.

Type de liens. Les pages d'une même langue sont liées entre elles à travers des liens nommés *links*. Notamment, les articles ayant pour titre *Tour Eiffel* et *Louvre* de la WIKIPEDIA française sont liés par un lien de type *link*. Par contre les pages appartenant à deux versions différentes de WIKIPEDIA et traitant le même sujet sont reliées par des liens de type *crosslink*. Par exemple, l'article *Paris* de la WIKIPEDIA française est relié à l'article *Parigi* de la WIKIPEDIA italienne par un lien de type *crosslink*. Plus clairement, les liens de type *crosslink* relient deux pages dont le contenu respectif fait référence à un même concept, mais édité dans deux langues différentes. Par ailleurs, les liens de type *belongTo* existant entre un article et une catégorie, permettent d'identifier les catégories associées à un article. Par contre, les liens de type *childOf* existant entre deux catégories permettent de mettre en évidence le type de relation existant entre les catégories cibles, c'est-à-dire faire la distinction



WIKIPÉDIA
L'encyclopédie libre

Non connecté Discussion Contributions Créer un compte

Article Discussion | Lire Modifier Modifier le code Historique | Rechercher sur Wikipédia

Accueil
Portails thématiques
Article au hasard
Contact

Contribuer
Débuter sur Wikipédia
Aide
Communauté
Modifications récentes
Faire un don

Outils
Pages liées
Suivi des pages liées
Importer un fichier
Pages spéciales
Adresse permanente
Information sur la page
Élément Wikidata
Citer cette page

Imprimer / exporter
Créer un livre
Télécharger comme PDF
Version imprimable

Dans d'autres langues
Català
Deutsch
English
Español
Euskara
Italiano
Nederlands
Português
49 de plus
Modifier les liens

Java

Cette page d'homonymie répertorie les différents sujets et articles partageant un même nom.

Sommaire [masquer]
1 Géographie
2 Informatique
3 Zoologie
4 Poésie
5 Musique
6 Bibliographie
7 Voir aussi

Sur les autres projets Wikime
Java, sur le Wiktionnaire

Géographie [modifier | modifier le code]

- Java est une île indonésienne ;
  - les Javanais sont le groupe ethnique majoritaire de l'île ;
  - le javanais est leur langue ;
  - le café de Java est un caféier qui provient de cette île ;
- Java est le nom translitteré d'un district de Géorgie, transcrit en Djava ;
- Java est également le nom de plusieurs villes des États-Unis ;
  - Java dans l'État de New York ;
  - Java dans le Dakota du Sud ;
  - Java en Virginie dans le comté de Pennsylvanie.
- Java est également le nom d'un hameau de Bas-Oha ainsi que d'une petite île sur la Meuse (Belgique).

Informatique [modifier | modifier le code]

- Java est le nom d'une technologie mise au point par Sun Microsystems (racheté par Oracle en 2010) qui permet de produire des logiciels indépendants de toute architecture matérielle. Cette technologie s'appuie sur différents éléments qui, par abus de langage, sont souvent tous appelés Java :
  - Ne pas confondre JavaScript, un langage de programmation de scripts, avec Java ;
  - Le langage Java est un langage de programmation orienté objet ;
  - Un programme compilé en bytecode Java s'exécute dans un environnement d'exécution Java (JRE) qui émule une machine virtuelle, dite machine virtuelle Java ;
  - La plate-forme Java correspond à la machine virtuelle Java à laquelle sont adjointes diverses spécifications d'API :
    - Java Platform, Standard Edition (Java SE) contient les API de base et est destiné aux ordinateurs de bureau,
    - Java Platform, Enterprise Edition (Java EE) contient, en plus du précédent, des API orientées entreprise et est destiné aux serveurs,
    - Java Platform, Micro Edition (Java ME) est destiné aux appareils mobiles tels que assistants personnels ou smartphones,
    - La Java FX Edition (ou Java FX) est orientée Rich Internet Application (RIA).

Zoologie [modifier | modifier le code]

- La Java est une race de poule naine originaire des îles de la Sonde puis sélectionnée en Angleterre,
- La Java Américaine est une race de poule américaine créée à partir de poules provenant d'Asie du Sud-Est,
- Le coq de Java est une espèce d'oiseau appartenant au genre *Gallus*,
- Le Java est une race de poney.

Poésie [modifier | modifier le code]

- La revue Java, dont le sous-titre du premier numéro était « revue de mauvais genre », a été créée au printemps 1989 par Jean-Michel Espitallier et Jacques Sivan.

Musique [modifier | modifier le code]

- La java est une danse parisienne ;
- Java est un groupe de musique français ;
- La Java est un club parisien dans le quartier de Belleville.

Bibliographie [modifier | modifier le code]

- Christine Jordis, *Bali, Java, en rêvant*, Ed. Gaillimard, Collection Folio, 2005 (1re publication en 2001 aux éditions du Rocher)

Voir aussi [modifier | modifier le code]

- Une java
  - *Une Java*, film de 1939
- Java est une marque de cachaça.
- Java est un jeu de société.
- Java des Cavernes est un personnage de la série Martin Mystère.
- Java dans les bois est un festival de musique.
- HMS *Java* est le nom de plusieurs navires de la Royal Navy.

Catégorie : Homonymie

Figure 2.2: Page de desambiguisation de la WIKIPEDIA française ayant pour titre Java.

entre la catégorie la plus spécifique, de celle qui est la plus générique. Sur le schéma 2.4, il existe un lien de type *belongTo* entre l'article "Football" et les pages catégories "Football" et "Sport originaire d'Angleterre". De même il existe un lien de type *childOf* entre la catégorie

Figure 2.3: Page de redirection de la WIKIPEDIA française ayant pour titre Paris (France).

"Discipline sportive" et la catégorie "Sport". La catégorie "Sport" est la catégorie parent. C'est elle qui est la plus générique par rapport à la catégorie "Discipline sportive".

Graphe WIKIPEDIA

L'encyclopédie WIKIPEDIA (toutes langues confondues) peut être modélisée comme un graphe orienté $\mathcal{W}$, dont les nœuds sont des pages et les arêtes des liens orientés dont le type est l'un des types indiqués précédemment. Notons p^{α} un nœud ou une page dans la version de WIKIPEDIA dont la langue est α. Plus formellement, $\mathcal{W} =< V, E >$ avec V étant l'union des pages (nœuds) de chaque version de WIKIPEDIA c'est-à-dire $V = \cup V^{\alpha}$. Et E, l'ensemble des arcs tels que le lien $(p_1^{\alpha}, p_2^{\beta})$ relie les nœuds p_1^{α} et p_2^{β} appartenant respectivement aux versions de la WIKIPEDIA dans les langues α et β. Autrement dit, $p_1^{\alpha} \in V^{\alpha}$ et $p_2^{\beta} \in V^{\beta}$ avec $\alpha \neq \beta$ ou $\alpha = \beta$.

Pour formaliser les différents types de liens existants entre les nœuds de $\mathcal{W}$, on a défini un ensemble de prédicats. En effet, un *prédicat* est une fonction qui renvoie une valeurs, vrai ou faux, selon qu'une condition est vérifiée ou pas. Les prédicats définis s'appliquent à deux pages p_1^{α} et p_2^{β} d'une même version ($\alpha = \beta$) ou d'une version différente ($\alpha \neq \beta$) de WIKIPEDIA. Pour plus de lisibilité, on note p_1 et p_2 si $\alpha = \beta$. Le prédicat $link(p_1, p_2)$ avec p_1 et p_2 deux articles est vrai s'il existe un lien de type $link$ qui part du nœud p_1 vers le nœud p_2. Et inversement $link(p_2, p_1)$ est vrai s'il existe un lien de type $link$ qui part du nœud p_2 vers le nœud p_1. Considérons le graphe de la figure 2.5, dans lequel les nœuds en gris sont des articles, tandis que ceux en violet sont des catégories. Le prédicat $link(p_4, p_1)$ est vrai, or $link(p_1, p_4)$ est faux. Quant au prédicat $crosslink(p_1^{\alpha}, p_2^{\beta})$, il est vrai s'il existe un lien de type $crosslink$ entre les nœuds p_1^{α} et p_2^{β} avec $\alpha \neq \beta$. En général, $crosslink(p_1^{\alpha}, p_2^{\alpha}) = crosslink(p_2^{\beta}, p_1^{\beta})$ car la direction du lien n'a pas d'importance dans ce cas. Ainsi, d'après notre figure 2.5 $crosslink(p_3, p_4) = crosslink(p_4, p_3)$ est vrai. Par ailleurs, le prédicat $cat(p_1, p_2)$ est vrai si et seulement si p_1 est une page qui n'est ni une

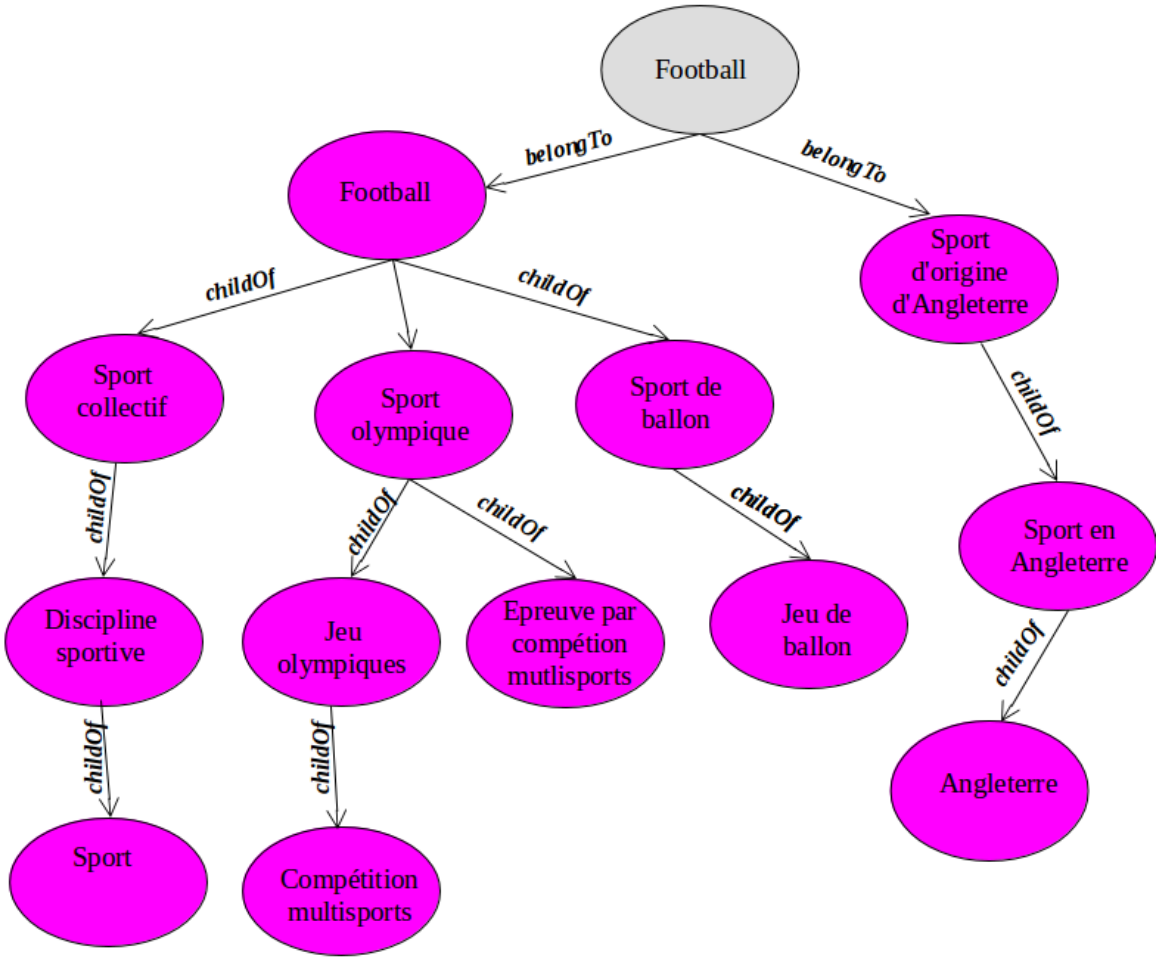


Figure 2.4: Catégories de la page article WIKIPEDIA Football.

catégorie, ni une redirection, p_2 une page de type catégorie et s'il existe un lien de type *belongTo* entre p_1 et p_2. Notamment, $cat(p_1, p_9)$ est vrai, or $cat(p_{15}, p_1)$ est faux. Le prédicat $childOf(p_1, p_2)$ est vrai si et seulement si p_1 et p_2 sont des pages de type catégorie et s'il existe un lien de type $childOf$ de p_1 vers p_2. Par exemple, $childOf(p_9, p_{11})$ est vrai, mais $childOf(p_{10}, p_9)$ est faux. Et enfin, le prédicat $ancestor^l(p_1, p_2)$ est vrai si et seulement si p_1 et p_2 sont des pages de type catégories, et la distance du chemin dirigé partant de p_1 à p_2 en passant uniquement par des liens de type $childOf$ vaut l.

En plus, chaque nœud du graphe $\mathcal{W}$ est caractérisé par un ensemble d'attributs ou de mesures qui le distingue des autres, définies comme suit:

- $in\text{-}neighbors(p_1)$: c'est l'ensemble des liens de type *links* entrants au nœud p_1. On le

définit comme suit:

$$in\text{-}neighbors(p_1) = \{p_2 \mid link(p_2, p_1)\ est\ vrai\} \tag{2.1}$$

D'après la figure 2.5, $in\text{-}neighbors(p_1) = \{p_4, p_7\}$, et $in\text{-}neighbors(p_7) = \{p_1, p_8, p_6\}$.

- *out-neighbors*(p_1): c'est l'ensemble des liens de type *link* sortant du nœud p_1. Plus formellement il est définit comme suit:

$$out\text{-}neighbors(p_1) = \{p_2 \mid link(p_1, p_2)\ est\ vrai\} \tag{2.2}$$

De même, sur notre figure 2.5, on a $out\text{-}neighbors(p_1) = \{p_2, p_7\}$.

- *categories(p_1)*: c'est l'ensemble des catégories associées au nœud p_1. En d'autres termes ce sont les catégories auxquelles la page de type article ou de desambiguisation p_1 appartient. Plus formellement, on le définit comme suit:

$$categories(p_1) = \{p_2 \mid cat(p_1, p_2)\ est\ vrai\} \tag{2.3}$$

On a $categories(p_1) = \{p_9, p_{15}\}$ et $categories(p_7) = \{p_{18}, p_{13}\}$.

2.3 Wikipedia Link-based Measure

***W**ikipedia **L**ink-based **M**easure* (**WLM**) [49] est une mesure de similarité appliquée aux pages WIKIPEDIA, qui utilise essentiellement le graphe WIKIPEDIA constitué uniquement des liens de type *link* et de tous les nœuds à l'exception des nœuds catégories. Plus précisément, WLM calcule la similarité entre deux articles d'une même version de WIKIPEDIA, qui est représentée par une valeur comprise entre 0 et 1. Plus la valeur est proche de 1, plus les pages sont sémantiquement similaires. Formellement, WLM est la moyenne de deux mesures de similarité inspirées respectivement de la similarité cosinus[2] et de la distance de Google normalisée[3].

En effet, la similarité cosinus permet de calculer la similarité entre deux vecteurs de n dimensions, en déterminant le cosinus de l'angle entre eux. Deux vecteurs sont dits opposés si leur valeur de similarité cosinus vaut -1. Ils sont dits indépendants ou orthogonaux si elle vaut 0, et similaires si elle vaut 1. Les valeurs intermédiaires permettent d'évaluer le degré de similarité des vecteurs associés. Dans le cas de WLM, on cherche à comparer deux articles WIKIPEDIA p_1 et p_2, de ce fait chaque article est représenté par un vecteur nommé $\overrightarrow{vecOut}$ dont la taille est fonction du nombre de liens de type *link* sortant de cet article, c'est-à-dire $|\overrightarrow{vecOut}(p_1)| = |out\text{-}neighbors(p_1)|$. De manière générale, le vecteur d'un nœud p_1 se définit comme suit:

[2] https://en.wikipedia.org/wiki/Cosine_similarity
[3] https://en.wikipedia.org/wiki/Normalized_Google_distance

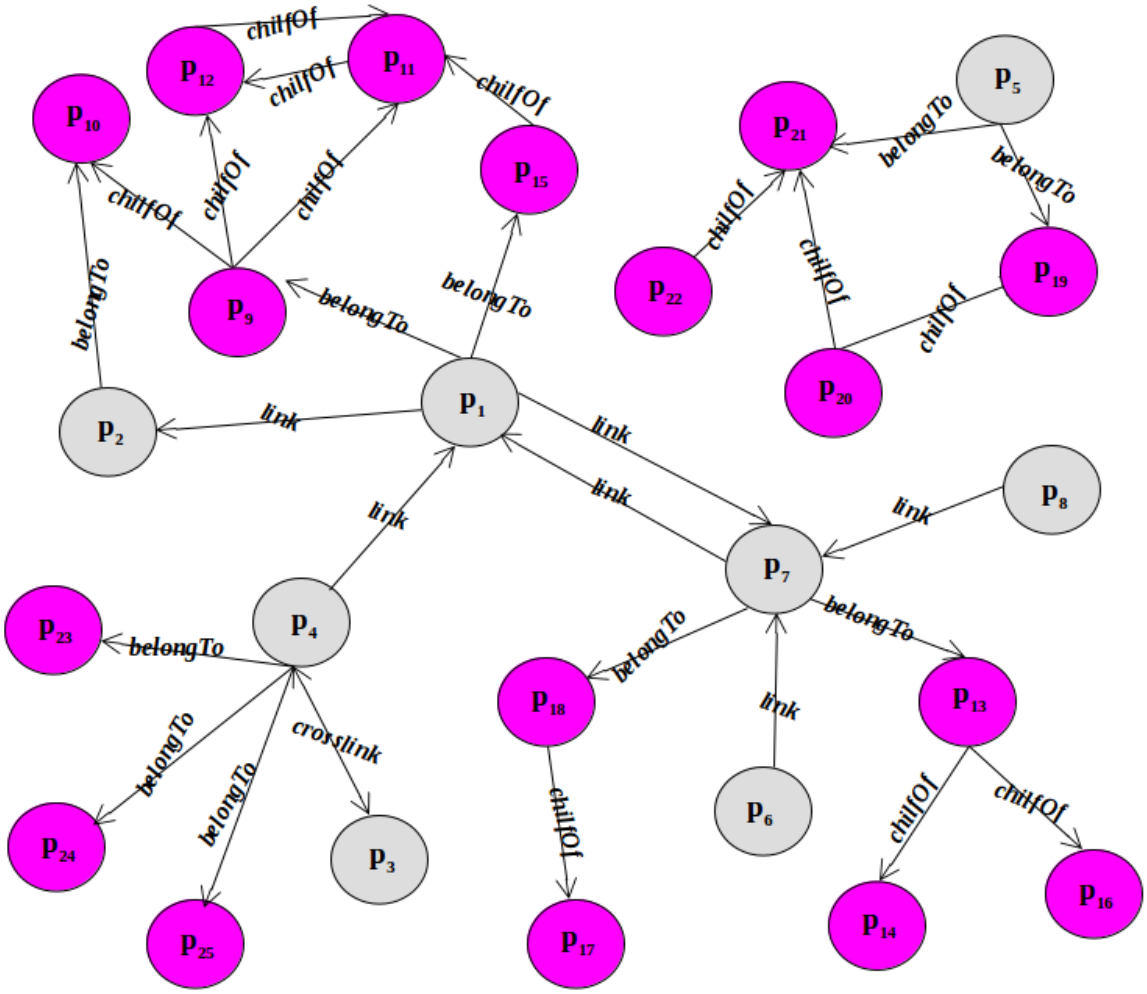


Figure 2.5: Exemple de graphe Wikipédia

$$\overrightarrow{vecOut}(p_1) = \bigcup_{p_i} \left(w_{p_i} = log \times \frac{|W|}{|in\text{-}neighbors(p_i)|} \right)^T \tag{2.4}$$

avec $p_i \in out\text{-}neighbors(p_1)$, log le logarithme et $|W|$ le nombre de nœuds qui ne sont pas de type catégorie de la version WIKIPEDIA à laquelle appartient le nœud p_1 et T la transposée d'une matrice.

La figure 2.6 représente une partie de la structure des articles *Camogie* et *Softball* de la WIKIPEDIA française. En fait *Camogie* est un sport collectif d'extérieur, exclusivement féminin, qui se joue avec une batte et une balle. Par contre, *Softball* est un sport collectif pratiqué par deux équipes de neuf à douze joueurs alternant entre l'attaque et la défense. Nous avons donc:

$$\overrightarrow{vecOut}(Camogie) = \Big(log \times \frac{|W|}{|in\text{-}neighbors(Hurling)|}, \quad \dots \quad , \quad , log \times \frac{|W|}{|in\text{-}neighbors(Kilkenny\ GAA)|}\Big)^T$$

$$\overrightarrow{vecOut}(Softball) = \Big(log \times \frac{|W|}{|in\text{-}neighbors(Universite\ Yale)|}, \quad \dots \quad , \quad , log \times \frac{|W|}{|in\text{-}neighbors(Gants\ de\ boxe)|}\Big)^T$$

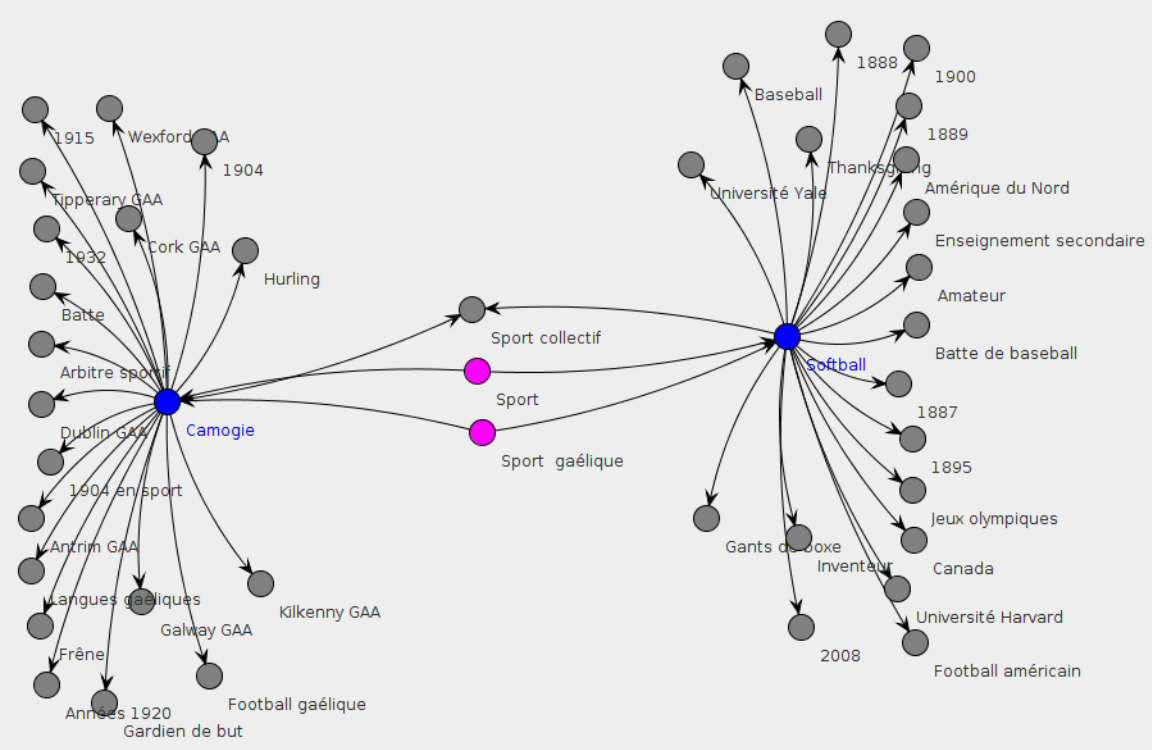


Figure 2.6: Extrait de la structure des articles *Camogie* et *Softball* de la WIKIPEDIA française .

L'idée de cette formule est d'attribuer un poids w_{p_i} à chaque lien de type *link* sortant du nœud p_1 c'est-à-dire les liens tels que $link(p_1, p_i)$ est vrai. Ce poids est fonction du nombre de liens entrants au nœud p_i. En d'autres termes, plus un article p_i a un nombre important de liens entrants, plus il est considéré comme une page "générique" qui ne caractérise pas nécessairement p_1 et par conséquent a un poids w_{p_i} faible. Par exemple, l'article *Sport collectif* commun aux deux articles *Camogie* et *Softball* possède 129 liens entrants de type *links*, on a donc $w_{Sport\ collectif} = 11.49$ en considérant que notre base de données WIKIPEDIA contient 12 671 709 articles. La similarité cosinus s'obtient en combinant le produit scalaire

et la norme des vecteurs *vecOut* des articles associés. Ainsi, après normalisation, on obtient $simCos(Camogie,\ Softball) = 0,94$.

Par ailleurs, la distance normalisée de google (Normalized Google Distance) [15] est une mesure de similarité sémantique qui, à l'origine, s'applique entre deux mots. Cette mesure est d'autant plus importante si les mots cibles apparaissent le plus souvent simultanément dans plusieurs pages Web. La distance normalisée de google notée *ngd*, vaut 0 si les mots associés sont considérés comme similaires. Par contre, si $ngd \geq 1$, alors ils sont complètement différents, ou tout simplement n'apparaissent pratiquement jamais ensemble dans une même page Web et leur *ngd* tend vers l'infini. Dans notre cas, nous appliquons la distance normalisée de google entre deux pages WIKIPEDIA, plus précisément deux articles d'une même version de WIKIPEDIA. A cet effet, deux pages WIKIPEDIA sont d'autant plus similaires si elles sont référencées simultanément par plusieurs autres pages.

La *ngd* de deux articles p_1 et p_2 se définit comme suit:

$$ngd(p_1, p_2) = \frac{log(max(|inN_{p_1}|, |inN_{p_2}|)) - log(|inN_{p_1} \cap inN_{p_2}|)}{log(|W|) - log(min(|inN_{p_1}|, |inN_{p_2}|))} \tag{2.5}$$

avec $inN = in\text{-}neighbors$.

Si nous reprenons note exemple précédent, nous avons $|in\text{-}neighbors(Camogie)| = 26$, $|in\text{-}neighbors(Softball)| = 430$, et $|inN_{Camogie} \cap inN_{Softball}| = |(Sport\ gaelique,\ Sport)| = 2$. Après normalisation on obtient $ngd(Camogie,\ Softball) = 0,59$.

La mesure de similarité WLM se définit dont ainsi:

$$wlm(p_1, p_2) = \frac{simCos(p_1, p_2) + ngd(p_1, p_2)}{2} \tag{2.6}$$

Ainsi $wlm(Camogie,\ Softball) = \frac{0,94+0,59}{2} = 0,765$.

2.4 PageRank

PageRank (PR) [9] est un algorithme inventé par Larry Page et Sergey Brin et utilisé par Google Search pour mesurer l'importance ou la popularité d'une page Web. La popularité d'une page Web est le nombre de pages Web qui lui font référence. Imaginons un utilisateur naviguant sur internet, en passant de pages à pages en suivant des liens hypertextes. Le principe de base de PageRank est d'attribuer à chaque page une valeur ou score proportionnel au nombre de fois que passerait par cette page un utilisateur naviguant aléatoirement sur internet, autrement dit, un utilisateur cliquant aléatoirement, sur un des liens apparaissant sur chaque page. Le déplacement de l'utilisateur est semblable à une marche aléatoire sur le graphe du Web. Sachant que l'utilisateur choisit chaque lien indépendamment des pages précédemment visitées, l'algorithme PageRank peut être vu comme un processus de Markov.

Plaçons-nous dans un contexte où nous disposons d'un graphe dirigé comme indiqué sur la figure 2.7. La taille de chaque nœud de ce graphe est fonction de sa popularité ou de son PageRank. Un nœud a un PageRank d'autant plus important qu'est grande la somme des PageRanks des nœuds qui pointent vers lui. C'est le cas du nœud rouge B de la figure 2.7, qui a un PageRank très important, car il possède plusieurs nœuds qui pointent sur lui, à savoir les nœuds D, C, E, F, et trois nœuds violets. Ces nœuds qui pointent sur le nœud B, viennent booster son PageRank. De plus, ce même nœud rouge possède un seul lien sortant vers le nœud orange C, par conséquent le PageRank du nœud orange est également considérable par rapport aux autres nœuds. Par contre, les nœuds violets ont un faible PageRank car ils n'ont pas de liens entrants, donc aucun nœud ne peut venir booster leur PageRank respectifs. Le PageRank tient aussi compte des boucles ou liens internes à un nœud, c'est-à-dire ceux qui pointent vers lui même.

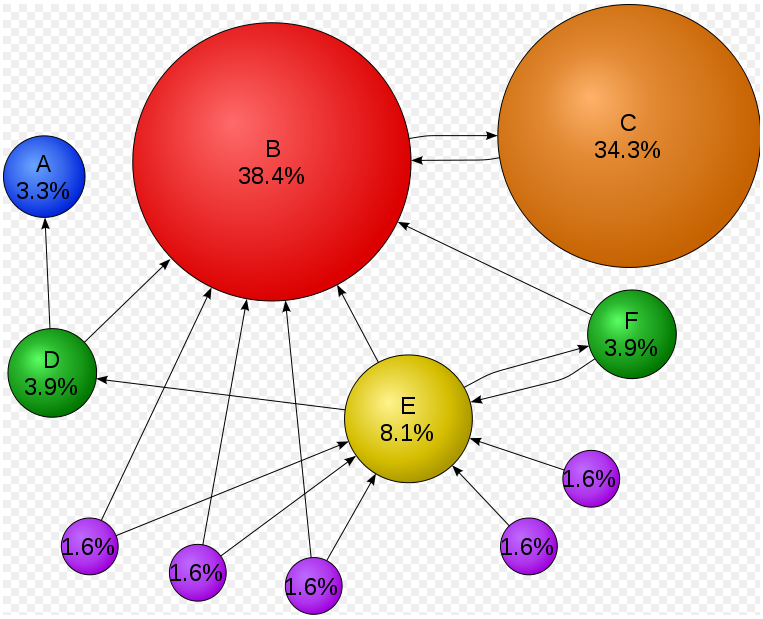


Figure 2.7: Graphe extrait du site WIKIPEDIA mettant en avant le PageRank de chaque nœud.

Soit $\mathcal{G} =< V, E >$ un graphe orienté pondéré, p_1 un nœud ayant pour poids initial w_1 (valeur numérique), et w_{12} une valeur numérique également désignant le poids du lien existant entre les nœuds p_1 et p_2. Pour calculer le PageRank d'un nœud du graphe, la somme des poids des liens ainsi que celle des nœuds de tout le graphe doit être égale à 1. En d'autres termes $\sum w_{ij} = 1$ et $\sum w_i = 1$. Le *PageRank* d'un nœud p_i noté $pr(p_i)$ est défini comme suit:

$$pr(p_i) = c \times \sum_{p_j \in in\text{-}neighbors(p_i)} \frac{w_{ij} \times pr(p_j)}{|out\text{-}neighbors(p_j)|} \tag{2.7}$$

où c est un facteur de normalisation constant < 1.

Dans ce qui suit nous présentons plus en détails nos approches proposées de construction du profil d'intérêts des utilisateurs dans un réseau social: DELVE [35] et FRISK [?]. En fait, la première approche DELVE est une étude préliminaire dont le but est de se familiariser avec les différents concepts sus-cités, ainsi que d'identifier les différents objets ou éléments de WIKIPEDIA que nous pouvons exploiter à notre profit afin de construire le profil d'intérêts des utilisateurs.

3 DELVE: DiscovEr LiVejournal intErests

L'intérêt d'un individu est l'élément pour lequel cette personne est passionnée ou se sent pleinement concernée, ou encore captive son attention. Dans cette section, nous décrivons les différentes étapes de notre approche de découverte des intérêts des utilisateurs dans un réseau social, nommée DELVE. En effet, cette approche utilise les informations extraites du profil d'un utilisateur pour construire son profil d'intérêts. Plus précisément, DELVE prend en entrée essentiellement les ressources textuelles renseignées par les utilisateurs sous forme de liste d'expressions dans leurs profils et fournit en sortie leurs intérêts classés par ordre de pertinence. Nous avons testé DELVE sur un jeu de données contenant les profils des utilisateurs du réseau social LIVEJOURNAL.

En effet, LIVEJOURNAL est un service de réseautage social où les utilisateurs peuvent créer, peupler et administrer un blog ou un journal. Il a été conçu en 1999 par le programmeur américain Brad Fitzpatrick afin de garder ses relations avec ses amis du lycée. Comme dans la plupart des réseaux sociaux, LIVEJOURNAL met à la disposition de ses utilisateurs un profil utilisateur qui contient une variété de données, telles que les informations de contact, une biographie, des images et des listes d'amis, d'intérêts, de communautés et même d'écoles auxquelles l'utilisateur a participé dans le passé ou fréquente actuellement. La figure 2.8 est un exemple de profil utilisateur du réseau social LIVEJOURNAL. De ce profil, DELVE se sert principalement de la liste des intérêts renseignées sous l'onglet INTÉRÊTS à savoir: *books*, *computers*, *cryptography*, *father ted*, *frasier*, *ian m.banks*, *mp3*, et *music*. Dans la suite, par soucis d'ambiguïté avec le mot "intérêt" du profil d'intérêts, nous utilisons le mot "expression intérêt" pour faire référence à l'un des intérêts de la liste des intérêts indiquée dans le profil d'un utilisateur, notamment *books*, *mp3*, *computers*, etc. Ces expressions intérêts renseignées par les utilisateurs à travers un vocabulaire libre sont généralement ambiguës, par conséquent lever cette ambiguïté est l'un des challenges de DELVE.

3.1 Approche proposée

Plaçons-nous dans un contexte où nous disposons des profils utilisateurs d'un réseau social dans lesquels sont renseignés de manière explicite une liste d'expressions intérêts. De même, considérons que la langue α utilisée par chaque utilisateur pour éditer son profil est connue.

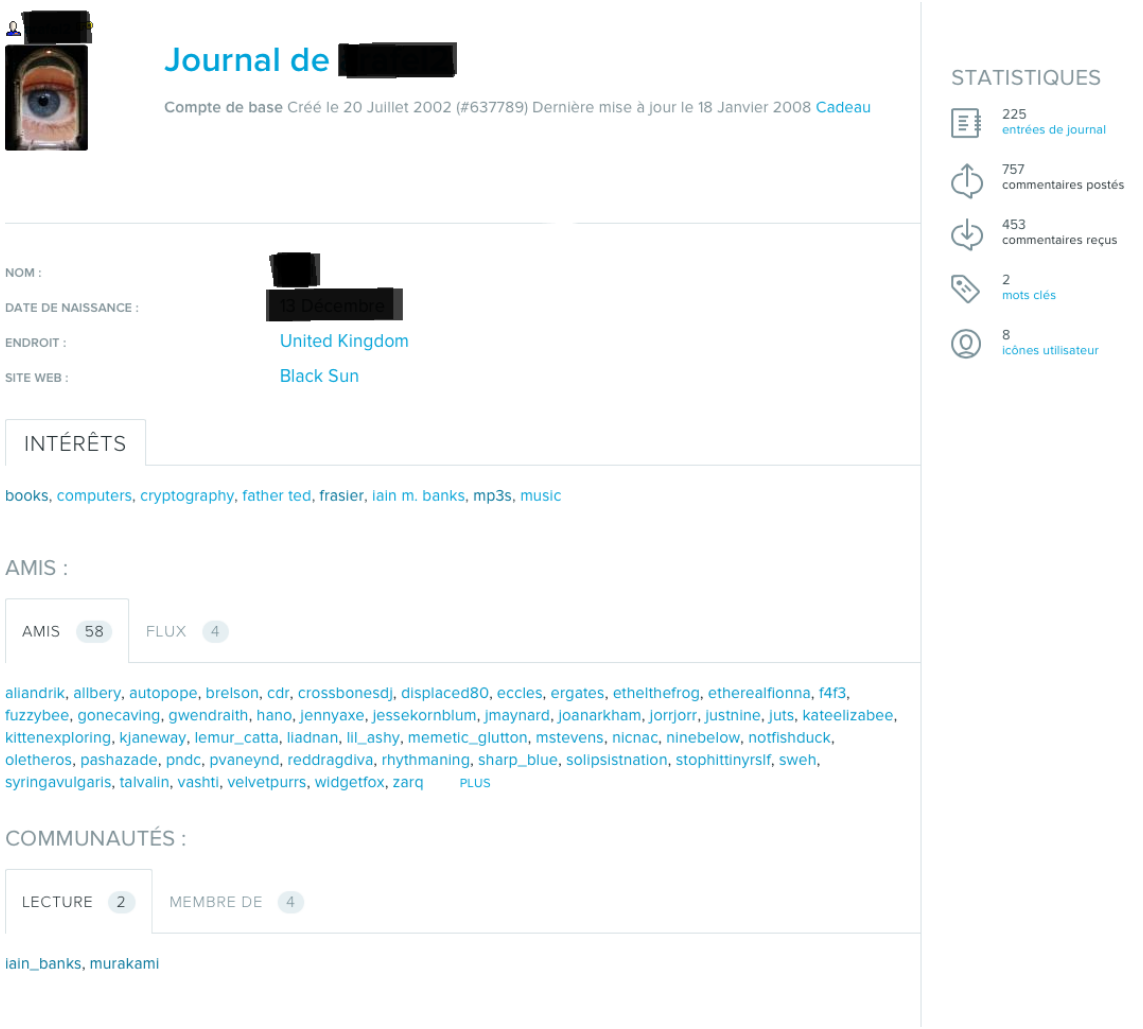


Figure 2.8: Profil d'un utilisateur du réseau social LIVEJOURNAL.

Le problème que nous posons ici se résume donc ainsi:

$$\textbf{Entrée} \quad \rightarrow \qquad \mathcal{EXP}^u,$$

$$\textbf{Sortie} \quad \rightarrow \quad (\mathcal{P}^u, \geqslant) = \{I_1, I_2, ..., I_m\}$$

où $\mathcal{EXP}^u = \{exp_1, exp_2, ..., exp_i\}$ est la liste des expressions intérêts renseignées par l'utilisateur u dans son profil et $\mathcal{P}^u$ son profil d'intérêts. Plus précisément, $(\mathcal{P}^u, \geqslant)$ est l'ensemble ordonné des domaines d'intérêts découverts I_m de l'utilisateur u. La relation d'ordre $\geqslant$ permet de présenter les intérêts par ordre de pertinence, de telle sorte que I_1 est plus important que I_2, qui à son tour est plus important que I_3, etc. L'algorithme 1 montre les différentes étapes principales de notre approche DELVE: la *détermination des interprétations* des expressions intérêts, et *la découverte des intérêts* d'un utilisateur.

Algorithm 1 : DELVE

1: **function** DELVE($\mathcal{EXP}^u$): $\mathcal{P}^u$
2: $\quad \mathcal{I}nter^u \leftarrow$ DETERMINEINTERPRETATIONS($\mathcal{EXP}^u$)
3: $\quad \mathcal{P}^u \leftarrow$ DISCOVERYINTERESTS($\mathcal{I}nter^u$)
4: **end function**

3.1.1 Détermination des interprétations des expressions intérêts

Considérons un profil utilisateur de LIVEJOURNAL, dans lequel nous avons les expressions intérêts suivantes: *beer, computers, ocaml, food, hansei, hiking, macintosh, java, perl, politics, san francisco, wine, writing.* Ces expressions intérêts exprimées dans un vocabulaire libre, sont intrinsèquement ambiguës, c'est-à-dire que certaines peuvent avoir plusieurs *interprétations* ou *significations* possibles. Par exemple, l'expression intérêt *java* mentionnée dans cette liste peut faire référence au langage de programmation Java, à une île en Indonésie, ou encore à un village dans l'état de Virginie aux États-Unis. Comment choisir la meilleure interprétation parmi ses différentes interprétations possibles?

Pour déterminer la meilleure signification d'une expression intérêt, nous avons tout d'abord utilisé l'encyclopédie WIKIPEDIA pour identifier toutes ses interprétations possibles. Ensuite, nous nous sommes servis de la mesure de similarité WLM défini à la section 2.3 pour rapprocher les interprétations similaires entre elles. L'idée est de dire que si un utilisateur est intéressé par le sport par exemple il va renseigner un ensemble d'expressions intérêts sémantiquement liées au sport. Et enfin, nous avons utilisé l'algorithme PageRank défini à la section 2.4 pour affecter un score d'importance à chacune des interprétations possibles. L'idée ici est de classer pour chaque expression intérêt ses interprétations par ordre d'importance afin de choisir comme meilleure interprétation, celle ayant le plus grand score.

Plus précisément, pour identifier les interprétations possibles d'une expression intérêt *exp* de $\mathcal{EXP}^u$, nous cherchons dans la version WIKIPEDIA dont la langue est α (langue utilisée par l'utilisateur u pour éditer son profil) toutes les pages ayant pour titre *exp*. Les pages obtenues peuvent être soient des articles, soient des pages de désambiguïsation ou soient des pages de redirection. Notamment, la table 2.1 présente les différents types de pages trouvées dans la WIKIPEDIA anglaise, à partir des quatre expressions intérêts suivantes: *cryptography*, *java*, *computers* et *read*. Pour plus de lisibilité, nous avons en ligne et en bleu les titres des pages associées à chaque expression intérêt, et en colonne le type de chaque page.

Nous remarquons que les expressions *cryptography* et *java* sont associées à leur page par défaut ayant pour titre *Cryptography* et *Java* respectivement. Par contre, l'expression intérêt *read* est reliée à la page de désambiguïsation ayant pour titre *Read*. En rappel une page de désambiguïsation ayant pour titre *title* est une page qui contient toutes les interprétations possibles que peut prendre le mot *title*. A cet effet, la page de désambiguïsation *Read* indiquée à la figure 2.9 contient toutes les interprétations possibles que peut prendre l'expression *read* à savoir: *Read (process), Read (magazine), Rural Educational and*

	Pages		
Expressions	**Article**	**désambiguïsation**	**Redirection**
cryptography	Cryptography	-	-
java	Java	Java (disambiguation)	-
computers	Computer	Computer (disambiguation)	Computers
read	-	Read	-

Table 2.1: Pages WIKIPEDIA obtenues pour chaque expression.

Development Foundation, Read (computer), etc.

Par ailleurs, l'expression intérêt *computers*, est associée à la page de redirection ayant pour titre *Computers* indiquée à la figure 2.10. Compte tenu de leur description, les pages de redirection sont remplacées par les pages articles vers lesquelles elles redirigent. C'est le cas de la page de redirection *Computers*, qui est remplacée par l'article *Computer*. En revanche, certains articles ont une page de désambiguïsation associée, c'est le cas des articles *Java* et *Computer* comme indiqué dans la table 2.1 à l'aide de la couleur violette. Dans ces conditions, les interprétations possibles d'une expression intérêt sont, en plus de sa page par défaut, les pages contenues dans sa page de désambiguïsation. Par exemple, les interprétations de l'expression *java* sont *Java* (sa page par défaut) et toutes les pages contenues dans sa page de désambiguïsation *java (disambiguation)*.

Cependant, d'autres expressions peuvent être associées à leur page par défaut, qui ne possède pas de page de désambiguïsation. C'est le cas des expressions non ambiguës, leur unique interprétation est leur page par défaut. En occurrence, l'expression *cryptography* qui a pour unique interprétation, sa page par défaut *Cryptography*.

De manière générale, lorsqu'on cherche les interprétations d'une expression intérêt *exp*, nous avons trois cas de figure qui se présentent:

1. **Cas 1**: si *exp* est associé à une page par défaut qui n'a pas de page de désambiguïsation, alors cette page par défaut est l'unique interprétation de *exp*. C'est le cas de l'expression intérêt *cryptography*.

2. **Cas 2**: si *exp* est associé à une page par défaut qui possède une page de désambiguïsation, alors la page par défaut et les pages contenues dans la page de désambiguïsation sont les différentes interprétations possibles de *exp*. C'est le cas de l'expression intérêt *java*.

3. **Cas 3**: si *exp* est plutôt associé uniquement à une page de désambiguïsation, alors ses interprétations sont les pages contenues dans cette page de désambiguïsation. C'est le cas de l'expression intérêt *read*.

Les différentes interprétations possibles de chaque expression intérêt *exp* déterminées, nous cherchons à choisir la meilleure interprétation pour chacune de ces expressions. Intu-

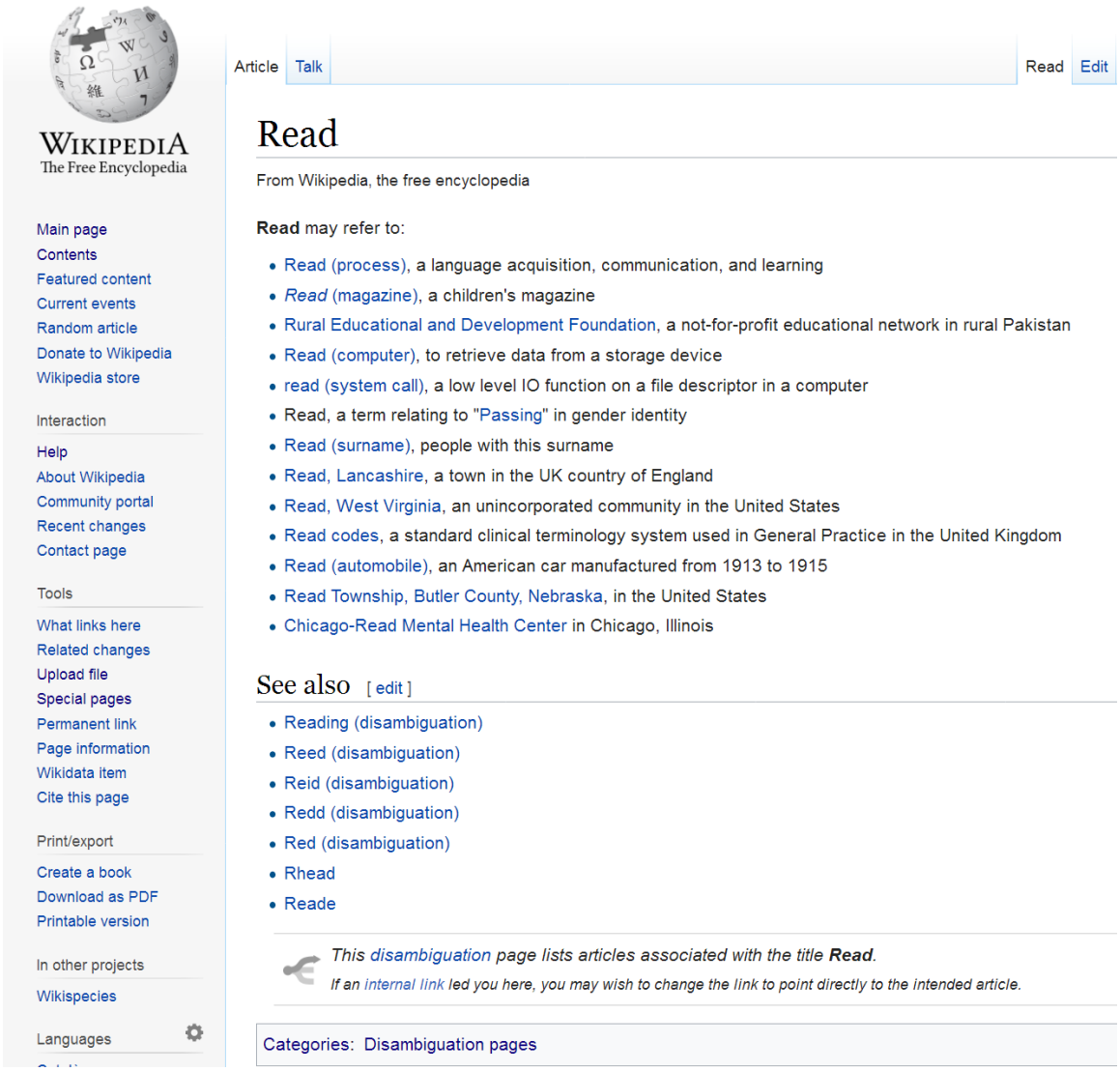


Figure 2.9: Page de désambiguïsation ayant pour titre *Read*

itivement, l'idée est de classer les interprétations d'une même expression de la plus probable vers la moins probable, de telle sorte que celle ayant la plus forte probabilité soit la meilleure possible. De ce fait nous avons fait une analyse des différentes interprétations possibles que peut prendre une expression. Dans le but de déterminer la pertinence des interprétations renvoyées par WIKIPEDIA, nous avons expérimenté 3 algorithmes différents.

1. Tout d'abord, nous avons défini un algorithme nommé DÉFAUT, qui considère uniquement si elle existe la page par défaut associé à une expression comme "meilleure"

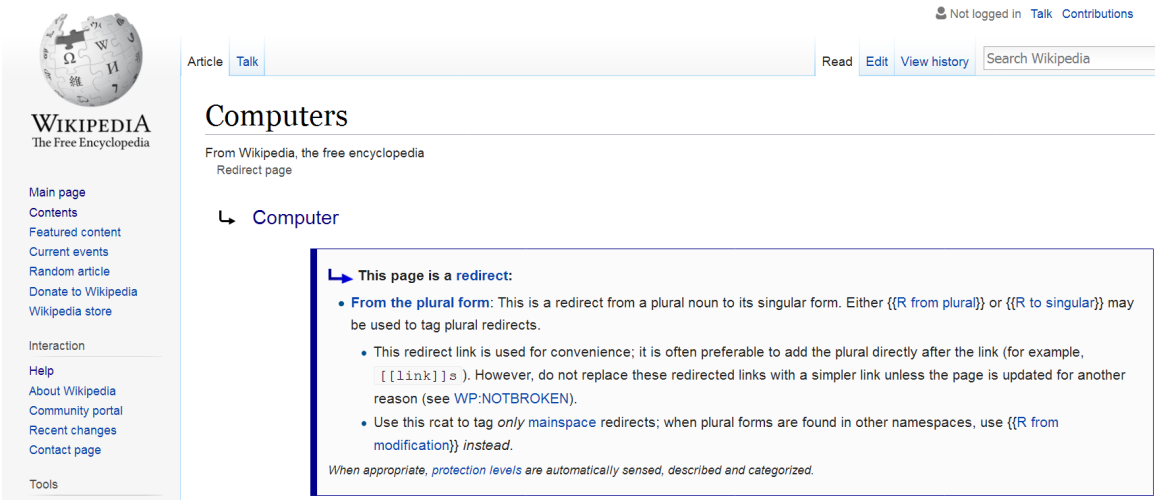


Figure 2.10: Page de redirection ayant pour titre *Computers*

interprétation.

2. Ensuite nous avons défini un deuxième algorithme, nommé DÉSAMBIG, qui exploite les pages de désambiguïsation pour chaque expression même si elle a une page par défaut. Les interprétations d'une expression sont sa page par défaut et celles listées dans sa page de désambiguïsation, si elles existent.

3. Et enfin nous avons le troisième algorithme, nommé HYBRIDE, qui mixe les deux premiers algorithmes. Autrement dit, si une expression a une page par défaut c'est elle qui est son interprétation, même si elle a une page de désambiguïsation. Par contre si une expression n'a pas de page par défaut mais plutôt une page de désambiguïsation alors ses différentes interprétations sont les pages mentionnées dans cette dernière.

Nous discutons dans la section évaluation des résultats de ces trois d'algorithmes suscités. Dans la suite nous considérerons l'un de ces algorithmes pour identifier les interprétations des expressions intérêts.

Les interprétations étant identifiées, nous nous inspirons de l'algorithme PageRank pour attribuer un score d'importance à chacune d'elles. En effet, PageRank est un algorithme de scoring qui s'applique sur des graphes orientés ou non, pondérés ou non. Nous représentons donc nos différentes interprétations sous forme de graphe non orienté pondéré, nommé *graphe des interprétations*.

Plus formellement, chaque utilisateur u est représenté par un graphe non orienté et pondéré, le graphe des interprétations noté $\mathcal{G}^u =< V, E >$, où chaque nœud $p_n \in V$ représente une interprétation d'une expression intérêt de $\mathcal{EXP}^u$, et chaque lien pondéré existant uniquement entre deux nœuds correspondant à deux interprétations de deux ex-

pressions intérêts distinctes. Plus clairement, un lien $(p_1, p_2) \in E$ est pondéré par la valeur de similarité $wlm(p_1, p_2)$, qui définit le degré de ressemblance sémantique existant entre ces nœuds. Cette représentation d'un utilisateur sous forme de graphe permet à chaque interprétation de "voter" pour les interprétations des autres expressions intérêts, et par la suite PageRank va se charger d'attribuer un score à chacune de ces interprétations en fonction du nombre de votes qu'elles ont reçues. Un vote est une valeur numérique qui varie en fonction du poids d'une part, de l'interprétation votant, et d'autre part, du lien existant entre le votant et le nœud qui reçoit le vote. Pour ne considérer que les votes importants nous avons décidé de lier uniquement les interprétations ayant un lien sémantique assez significatif, par conséquent nous avons choisi d'indiquer dans le graphe des interprétations exclusivement les liens entre les nœuds dont la valeur de similarité WLM est supérieure à un seuil fixé ε. La figure 2.11 est un extrait du graphe des interprétations du profil utilisateur de la figure 2.8. Les nœuds de même couleur représentent les interprétations d'une même expression intérêt, notamment, la couleur bleu représente les interprétations de l'expression intérêt *music*, le rouge ceux de *mp3*, le jaune ceux de *computer*, etc. Après l'exécution de l'algorithme de PageRank sur $\mathcal{G}^u$, l'interprétation ayant le plus grand score pour chaque expression intérêt est considérée comme sa meilleure interprétation. A ce niveau nous avons pour chaque expression intérêt, une unique interprétation et nous notons $\mathcal{I}nter^u$, l'ensemble de toutes les meilleures interprétations de l'utilisateur u.

L'algorithme 2 est un récapitulatif des différentes phases de l'étape de détermination des interprétations de notre approche DELVE. La fonction *getInterpretations* de la ligne 2, détermine les interprétations de chaque *exp* de $\mathcal{EXP}^u$ à l'aide de l'un des trois algorithmes DÉFAUT, DÉSAMBIG et HYBRIDE. Quant à la fonction *buildGraphInterpretations* de cette même ligne, elle prend en entrée la sortie de la fonction *getInterpretations*, c'est-à-dire $\mathcal{I}nter^u$ et retourne le graphe des interprétations de l'utilisateur cible. Ensuite, la fonction *executePageRank* de la ligne 3 exécute l'algorithme de PageRank sur le graphe des interprétations $\mathcal{G}^u$. Et enfin, la fonction *getMaxScoreInterpretation* de la ligne 4, affecte à chaque expression intérêt *exp*, l'interprétation ayant le plus grand score.

Algorithm 2 Détermination des interprétations

1: **function** DETERMINEINTERPRETATIONS($\mathcal{EXP}^u$): $\mathcal{I}nter^u$
2: $\quad \mathcal{G}^u \leftarrow buildGraphInterpretations(getInterpretations(\mathcal{EXP}^u))$
3: $\quad vertexScore \leftarrow executePageRank(\mathcal{G}^u)$
4: $\quad \mathcal{I}nter^u \leftarrow getMaxScoreInterpretation(vertexScore)$
5: **end function**

Une fois que nous avons déterminé les interprétations de chaque expression intérêts, nous passons à l'étape de découverte des intérêts.

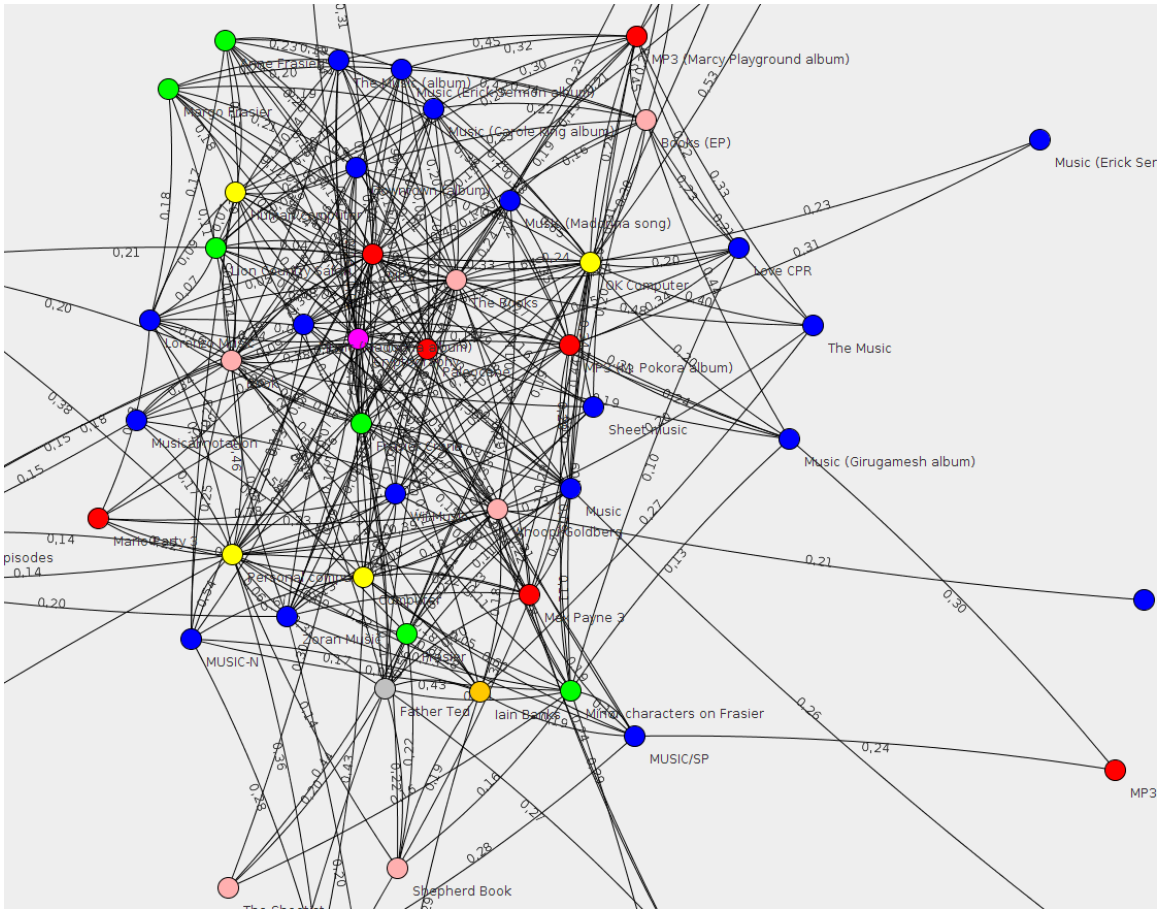


Figure 2.11: Extrait d'un exemple de graphe des interprétations

3.1.2 Découverte des intérêts d'un utilisateur

Le principe de découverte des domaines d'intérêts d'un utilisateur consiste tout d'abord à regrouper les interprétations de l'ensemble $\mathcal{I}nter^u$ par similarité ou par affinité, ensuite, affecter un ou plusieurs intérêts à chaque bloc d'interprétations similaires, et enfin, classer les intérêts découverts par ordre de pertinence.

Reprenons notre exemple précédent, dans lequel la liste des expressions intérêts est la suivante : $\mathcal{EXP}^u = \{$ *beer, food, wine, computers, ocaml, macintosh, java, perl*$\}$. De plus, considérons que l'ensemble $\mathcal{I}nter^u = \{$*Beer, Food, Wine, Computer, OCaml, Macintosh, Java (programming language), Perl*$\}$ sont les interprétations obtenues à l'issue de l'étape de détermination des interprétations sur $\mathcal{EXP}^u$. En fait, à la fin de l'étape de découverte des intérêts, nous cherchons à avoir en sortie:

- l'intérêt *Gastronomy* associé au bloc d'interprétations (*Beer, Food, Wine*) et
- l'intérêt *Programmation* ou *Informatique* pour le bloc d'interprétations (*Computer,*

OCaml, Macintosh, Java (programming language), Perl).

- Et au final on a $\mathcal{P}^u$= { Programmation, Informatique, Gastronomy}

A cet effet, nous utilisons un algorithme de clustering hiérarchique ascendant pour regrouper les interprétations par bloc similaire. Cet algorithme prend en entrée l'ensemble $\mathcal{I}nter^u$, ainsi que les valeurs de similarité existant entre ces différentes interprétations (calculées à partir de WLM), et retourne un diagramme qui illustre un arrangement de groupes d'interprétations générées par un regroupement hiérarchique. Ce diagramme est communément appelé *dendrogramme* et se représente comme l'indique la figure 2.12.

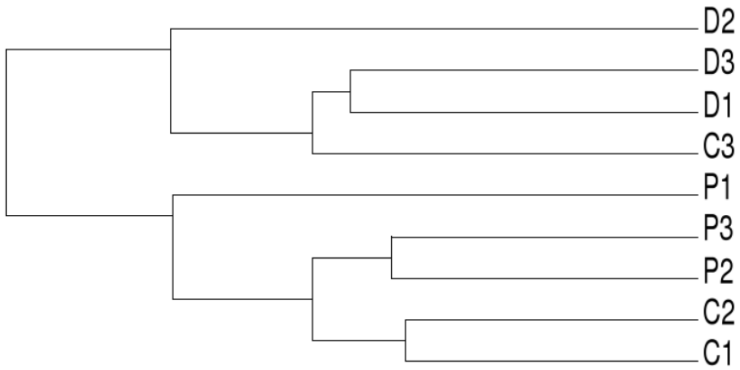


Figure 2.12: Forme d'un dendogramme.

Plus précisément, l'algorithme de clustering hiérarchique ascendant fonctionne comme suit:

- Initialement, chaque interprétation de $\mathcal{I}nter^u$ constitue un bloc d'interprétation.
- A chaque itération, les paires de blocs d'interprétations les plus similaires sont fusionnées et ne forment plus qu'un seul nouveau bloc.
- De proche en proche, les blocs similaires sont reliés jusqu'à l'obtention d'un seul bloc contenant toutes les interprétations initiales.

Sur la figure 2.12, à la première itération, on obtient les blocs suivants: (D2), (D3, D1), (C3), (P1), (P3, P2) et (C2, C1). Ensuite à la deuxième itération, nous avons les blocs: (D2), (D3, D1, C3), (P1) et (P3, P2, C2, C1). Et enfin, à la quatrième itération l'algorithme s'arrête puisque toutes les interprétations initiales sont contenues dans un même bloc.

De manière générale, pour obtenir les blocs d'interprétations similaires d'un utilisateur, nous avons décidé de couper le dendogramme construit au niveau d'une itération quelconque, qui est fonction du nombre de blocs d'interprétations existants déjà. Plus on coupe le dendogramme à une itération basse ou petite, plus on a de blocs d'interprétations

et plus les domaines d'intérêts découverts sont spécifiques. L'algorithme 3 représente un récapitulatif des différentes phases de découverte des intérêts d'un utilisateur. La fonction *getBlocInterpretations* de la ligne 3 implémente le principe du clustering hiérachique décrit précédemment pour regrouper les interprétations $\mathcal{I}nter^u$ par blocs similaires. Quant à la fonction *getCommonAncestor* de la ligne 5, elle détermine les intérêts de chaque bloc d'interprétations $\mathcal{B}$. En effet, on exploite le graphe WIKIPEDIA, notamment, essentiellement les pages et les liens de type catégories et $chilfOf$ respectivement, pour découvrir l'intérêt associé à un bloc $\mathcal{B}$. L'intérêt associé à un bloc est la catégorie WIKIPEDIA parent commune à toutes les interprétations contenues dans ce bloc. Plus clairement, pour chaque interprétation de $\mathcal{B}$, nous déterminons tout d'abord leurs catégories directes respectives à l'aide de la fonction *categories* de l'équation 2.3 décrite à la section 2.2. A partir des catégories obtenues, nous explorons les liens de type $chilfOf$ du graphe WIKIPEDIA à la recherche de la première catégorie parent commune aux catégories directes initiales. Par exemple, d'après la figure 2.4 de la section 2.2, "Football" est la catégorie parent commune aux catégories "Discipline sportive" et "Jeu de ballon". La fonction *rank* de la ligne 7 classe les intérêts de $\mathcal{P}^u$ par ordre d'importance décroissant. L'importance d'un intérêt est définie par le nombre d'interprétations contenues dans le bloc d'interprétations ayant permis sa découverte. Les intérêts de chaque bloc d'interprétations ainsi découverts et classés forment l'ensemble des domaines d'intérêts $\mathcal{P}^u$ de l'utilisateur u.

Algorithm 3 Découverte des intérêts d'un utilisateur

```
1: function DISCOVERYINTERESTS(Inter^u): (P^u, ⩾)
2:     P^u ← ∅
3:     BlocInter^u ← getBlocInterpretations(Inter^u)
4:     for all B ∈ BlocInter^u do
5:         P^u ← getCommonAncestor(B) ∪ P^u
6:     end for
7:     rank(P^u)
8: end function
```

3.2 Expérimentations et évaluations

Pour l'évaluation de DELVE, nous avons sélectionné au hasard 50 urls de profil utilisateurs de la base de données de [10]. Tout d'abord, nous avons mis à jour nos données, en rajoutant pour chaque utilisateur la liste de ses expressions intérêts, ensuite nous avons évalué nos trois algorithmes d'identification des interprétations d'une expression à savoir DÉFAUT, DÉSAMBIG et HYBRIDE. Et enfin, nous avons évalué DELVE sur nos 50 profils utilisateurs mis à jour.

3.2.1 Collecte des données

A l'origine, la base de données [10] contient des profils utilisateurs de quatre réseaux sociaux, à savoir LIVEJOURNAL, FLICKR, TWITTER et YOUTUBE[4]. Ces profils contiennent uniquement un seul attribut, le pseudonyme qui représente l'identifiant d'un utilisateur dans un réseau social. Après avoir sélectionné aléatoirement 50 profils utilisateurs du réseau LIVEJOURNAL, nous nous sommes appuyés sur l'API de LIVEJOURNAL et les identifiants de chacun des 50 utilisateurs sélectionnés pour extraire de leurs pages Web respectifs leurs différentes expressions intérêts. A l'issue du processus d'extraction, nous avons obtenu entre 7 et 15 expressions intérêts par profil utilisateurs, pour un total de 392 expressions intérêts, avec plus précisément 257 distinctes associées à une page WIKIPEDIA (article, page de désambiguïsation ou redirection) de la version anglaise, et 36 n'étant reliées à aucune page (non trouvées).

Par ailleurs, pour évaluer DELVE sur notre jeu de données nous avons voulu tout d'abord évaluer nos trois algorithmes d'identification des interprétations possibles d'une expression précédemment décrits à savoir DÉFAUT, DÉSAMBIG et HYBRIDE.

3.2.2 Résultats des algorithmes DÉFAUT, DÉSAMBIG et HYBRIDE

L'évaluation faite dans cette section consiste à utiliser l'un des trois algorithmes DÉFAUT, DÉSAMBIG et HYBRIDE pour identifier les interprétations possibles des expressions intérêts des utilisateurs de notre base de données. Ensuite construire le graphe d'interprétations et utiliser PageRank pour classer ces interprétations. Et enfin, comparer les meilleures interprétations obtenues pour chaque expression intérêt à celles que nous avons acquis comme vérité terre. En fait, l'idée est d'évaluer la pertinence de chaque algorithme à partir des mesures de performances de la littérature.

La comparaison des meilleures interprétations obtenues pour chaque interprétation nécessite l'existence d'une vérité terre ou base de vérité, qui contient les bonnes interprétations de chaque expression intérêts des profils utilisateurs de notre jeu de données. De ce fait, nous (deux personnes) avons fait une évaluation manuelle de nos 50 profils dans l'optique de construire notre base de vérité. En effet, l'évaluation consiste à inspecter visuellement chaque profil utilisateur à la recherche des différents aspects possibles nous permettant de distinguer parmi les différentes interprétations proposées par nos algorithmes, celles pouvant être les meilleures. Plus précisément, l'idée est de rechercher manuellement au sein du réseau social LIVEJOURNAL, les ressources décrivant chaque expression intérêt, afin d'y extraire des informations pouvant nous aider à identifier parmi les interprétations proposées par chacun des trois algorithmes, celle la plus probable. Les ressources associées à une expression intérêt peuvent être entre autres la page la décrivant ou l'ensemble de ses communautés. Cette évaluation, qui a été faite par deux évaluateurs (mon encadrant et moi) est loin d'être triviale pour ces derniers. Lors de la phase d'inspection des profils utilisateurs, pour chaque expression intérêt, nous avons décidé d'avoir quatre évaluations possibles de chacune de ses

[4] http://www.ursino.unirc.it/pkdd-12.html

interprétations proposées :

- -1 si l'évaluateur considère qu'elle est fausse,
- 0 si l'évaluateur ne sait pas,
- 1 si l'évaluateur considère qu'elle est correcte,
- 2 s'il considère qu'elle est corrélée à l'intérêt.

Plus généralement, une interprétation de $\mathcal{I}nter^u$ est correcte si à l'issue de son évaluation nous avons la valeur 1 ou 2. Compte tenu du fait que nous avons deux évaluateurs, par conséquent, pour prendre en compte leurs différentes évaluations, nous avons défini deux situations possibles:

1. **Situation 1** : une interprétation est correcte, si au moins un évaluateur l'a évaluée à correcte, noté min.
2. **Situation 2** : une interprétation est correcte, si tous les 2 évaluateurs l'ont évaluée à correcte, noté max.

Par ailleurs, pour mesurer la pertinence de chaque algorithme nous avons utilisé trois mesures de performances, la *précision*, le *rappel* et la *f-mesure* qui se décrivent comme suit.

Imaginons une application qui prend en entrée un document de texte et retourne pour chaque mot du document sa bonne interprétation en fonction de son contexte. La précision est le nombre d'interprétations pertinentes retrouvées rapporté au nombre d'interprétations totales proposées par l'application. En effet, quand un utilisateur interroge l'application, il souhaite que les interprétations proposées en réponse correspondent à celles attendues en fonction du contexte de chaque mot. Toutes les interprétations erronées constituent du bruit. La précision, qui s'oppose à ce bruit, est élevée si peu d'interprétations proposées sont erronées. Notons p_e, la précision d'un algorithme pour une évaluation précise c'est-à-dire dans le cas où une interprétation est correcte si elle est évaluée soit à 1, soit à 1 ou 2. Plus nettement nous avons :

$$p_e = \frac{\text{nbre interprétations correctes trouvées}}{\text{nbre interprétations trouvées}} \tag{2.8}$$

Le rappel, quant à lui, est défini par le nombre d'interprétations pertinentes retrouvées au regard du nombre d'interprétations pertinentes existantes réellement dans le document. Cela signifie que, lorsqu'on passe un document de texte à l'application, on souhaite voir apparaître toutes les meilleures interprétations de chaque mot. Autrement dit, on veut que le résultat trouvé soit à la limite semblable à ce que peut fournir manuellement un être humain sur ce même document. S'il existe une adéquation importante entre le résultat proposé par l'application et la réalité, alors le taux de rappel est élevé. À l'inverse si l'application retourne des interprétations inattendues, on parle de silence qui s'oppose au

rappel. Notons r_e, le rappel d'un algorithme pour une évaluation précise. Plus nettement nous avons:

$$r_e = \frac{\text{nbre interprétations correctes trouvées}}{\text{nbre interprétations correctes ou fausses ou non trouvées}} \tag{2.9}$$

La F-mesure ou moyenne harmonique est une mesure populaire qui combine la précision et le rappel. Notons f_e, la F-mesure d'un algorithme pour une évaluation précise. Plus nettement nous avons :

$$f_e = \frac{2p_e r_e}{p_e + r_e} \tag{2.10}$$

Les résultats de l'algorithme DÉFAUT sur nos 50 profils utilisateurs sont décrits dans le tableau 2.2. La première colonne du tableau indique la précision, le rappel et la F-mesure obtenus en considérant comme correcte une interprétation si et seulement si elle est notée 1 par au moins un évaluateur (min), ou par les deux évaluateurs (max). La deuxième colonne illustre le second cas, c'est-à-dire une interprétation est correcte si elle est évaluée à 1 ou 2.

	correcte =1			correcte =1 ou 2		
	p_e	r_e	f_e	p_e	r_e	f_e
$e = min$	85.39	77.35	81.17	85.67	77.06	81.14
$e = max$	79.49	72.01	75.57	80.61	73.02	76.63

Table 2.2: Résultats de l'algorithme DÉFAUT.

Bien que ces résultats soient encourageants, l'algorithme DÉFAUT a deux limites principales. Premièrement, il ne peut pas affecter une interprétation à une expression intérêt qui n'a pas de page par défaut. Deuxièmement, il affecte toujours la même interprétation à une même expression, et ne tient donc pas compte du contexte dans lequel l'expression est mentionnée. Par exemple, soit la liste d'expressions intérêts suivantes: *c, cats, computers, unix, video games*. La page par défaut renvoyée pour l'expression intérêt *c* correspond à l'interprétation "caractère de l'alphabet". L'interprétation de *c* comme langage de programmation référencée dans la page de désambiguïsation semble la plus probable car elle est liée sémantiquement aux autres expressions intérêts *computers* et *unix*. De même, prenons une autre liste d'expressions d'intérêts: *beer, computers, ocaml, food, hansei, hiking, macintosh, java, perl, politics, san francisco, wine, writing*. Au vu de cette liste, il est clair que la page par défaut de l'expression *java*, qui fait référence à une île en Indonésie, est une interprétation erronée. Toutefois, l'interprétation *Java (programming language)* référencée dans la page de désambiguïsation est correcte d'autant plus qu'elle est liée sémantiquement aux interprétations des intérêts *perl, ocaml, computers*. Pour exploiter ce lien sémantique entre les interprétations, nous avons utilisé dans notre approche la similarité structurelle (WLM) existant entre les paires d'interprétations.

Les résultats de l'algorithme DÉSAMBIG sur notre jeu de données sont décrits dans le tableau 2.3. Par rapport à l'algorithme DÉFAUT, la précision et le rappel baissent sensiblement. En effet, on constate que la plupart du temps l'interprétation correcte d'une expression intérêt correspond à sa page par défaut.

	correcte =1			correcte =1 ou 2		
	p_e	r_e	f_e	p_e	r_e	f_e
$e = min$	69.17	62.65	65.75	63.76	57.76	60.61
$e = max$	64.51	58.43	61.32	59.17	53.60	56.25

Table 2.3: Résultats de l'algorithme DÉSAMBIG

Par ailleurs, il faut noter que l'algorithme DÉSAMBIG détermine l'interprétation correcte d'une expression intérêt en se basant sur les interprétations possibles des autres expressions d'intérêts, et cela peut parfois générer des erreurs. À titre d'exemple, considérons l'expression intérêt *cat*, dont l'interprétation par défaut correspond à l'article WIKIPEDIA décrivant les chats. Si cette expression est indiquée dans le profil d'un utilisateur avec d'autres expressions tels que *unix* et *computer*, l'algorithme DÉSAMBIG a tendance à affecter un score élevé à l'interprétation *Cat (Unix)*, correspondant à la commande *cat* du système d'exploitation *Linux*. Une solution possible pour la résolution de ce problème est de combiner les deux algorithmes DÉFAUT et DÉSAMBIG, c'est-à-dire pour une expression intérêt on utilise DÉFAUT, si sa page par défaut existe, sinon on utilise DÉSAMBIG.

Les résultats de l'algorithme HYBRIDE sur les 50 utilisateurs de notre jeu de données sont décrits dans le tableau 2.4. On constate une amélioration considérable par rapport aux résultats des deux algorithmes DÉSAMBIG et DÉFAUT.

	correct =1			correct = 1 ou 2		
	p_e	r_e	f_e	p_e	r_e	f_e
$e = min$	89.04	80.66	84.64	90.16	81.67	85.71
$e = max$	82.58	74.80	78.50	84.26	76.33	80.10

Table 2.4: Résultats de l'algorithme HYBRIDE

L'algorithme HYBRIDE repose sur l'hypothèse, qu'une expression intérêt pour laquelle WIKIPEDIA propose une page par défaut n'est pas ambiguë. En effet, WIKIPEDIA est le produit de la contribution de millions d'individus sur le Web. De manière générale, si une certaine expression comme par exemple *cat*, est associée à une page par défaut "Cat", décrivant les chats, cela signifie que plusieurs individus ont été majoritairement d'accord sur cette interprétation. Par ailleurs, l'expression *cat* peut être associée à l'article *Cat (Unix)*, même si c'est un cas de figure jugé rare. Dans le même ordre d'idée, si une expression n'a pas de page par défaut, mais plutôt une page de désambiguïsation, cela signifie que les contributeurs de WIKIPEDIA n'ont pas trouvé un accord quant à la signification par défaut

de cette expression, dans ce cas, il faut donc appliquer un algorithme de désambiguïsation en utilisant le contexte, c'est-à-dire les autres expressions intérêts.

3.2.3 Résultats de DELVE

Au vu des résultats obtenus à l'issue de l'évaluation de nos trois algorithmes, nous avons choisi d'utiliser l'algorithme HYBRIDE pour déterminer les différentes interprétations possibles de chaque expression intérêt lors de l'évaluation de DELVE. De plus, nous avons fait plusieurs évaluations préliminaires visant à identifier la valeur du seuil ε reliant les nœuds similaires dans le graphe des interprétations et nous avons obtenu de meilleurs résultats en fixant ε à 0.4.

La figure 2.13 montre la sortie de notre approche DELVE sur deux utilisateurs, dont le premier est celui possédant le profil indiqué à la figure 2.8. Plus clairement, les résultats montrent qu'à la deuxième itération, pour l'utilisateur 1, nous avons découvert les blocs d'interprétations et les catégories suivants:

- $\mathcal{B}$**loc 1**: *{Father Ted, Frasier}* associé à la catégorie commune {Category:English-language television programming}.
- $\mathcal{B}$**loc 2**: *{Book, Music}* associé aux catégories {Category:Mind, Category:Creativity, Category:Mass media, Category:Popular culture, Category:Behavior}.
- $\mathcal{B}$**loc 3**: *{ MP3, Computer, Cryptography}* associé aux catégories {Category:Digital technology, Category:Information Age}.

Cependant le bloc constitué uniquement de l'expression intérêt *Iain Banks* est supprimé puisqu'il contient une seule expression intérêt, ce qui n'est pas représentatif. De plus, nous précisons que les catégories communes obtenues sont celles qui possèdent aux moins 4 catégories parents. L'idée est d'éliminer les catégories communes trop génériques d'une part, notamment *Category* : *Industries* obtenue pour le bloc 2. D'autre part, nous avons constaté que, plus une catégorie commune a de catégories parents plus elle représente au mieux les interprétations de son bloc. Le profil d'intérêts ainsi construit est donc:

$\mathcal{P}^{u_1}$ = { *Digital technology, Information Age, English-language television programming, Mind, Creativity, Mass media, Popular culture, Behavior*}.

Par ailleurs, à la première itération pour l'utilisateur 2, nous avons les blocs d'interprétations et les catégories suivants:

- $\mathcal{B}$**loc 1**: *{Fantasy, Speculative fiction }* associé à la catégorie commune {Category: Speculative fiction}.
- $\mathcal{B}$**loc 2**: *{Chocolate, Wine}* associé à la catégorie commune {Category:Foods}.

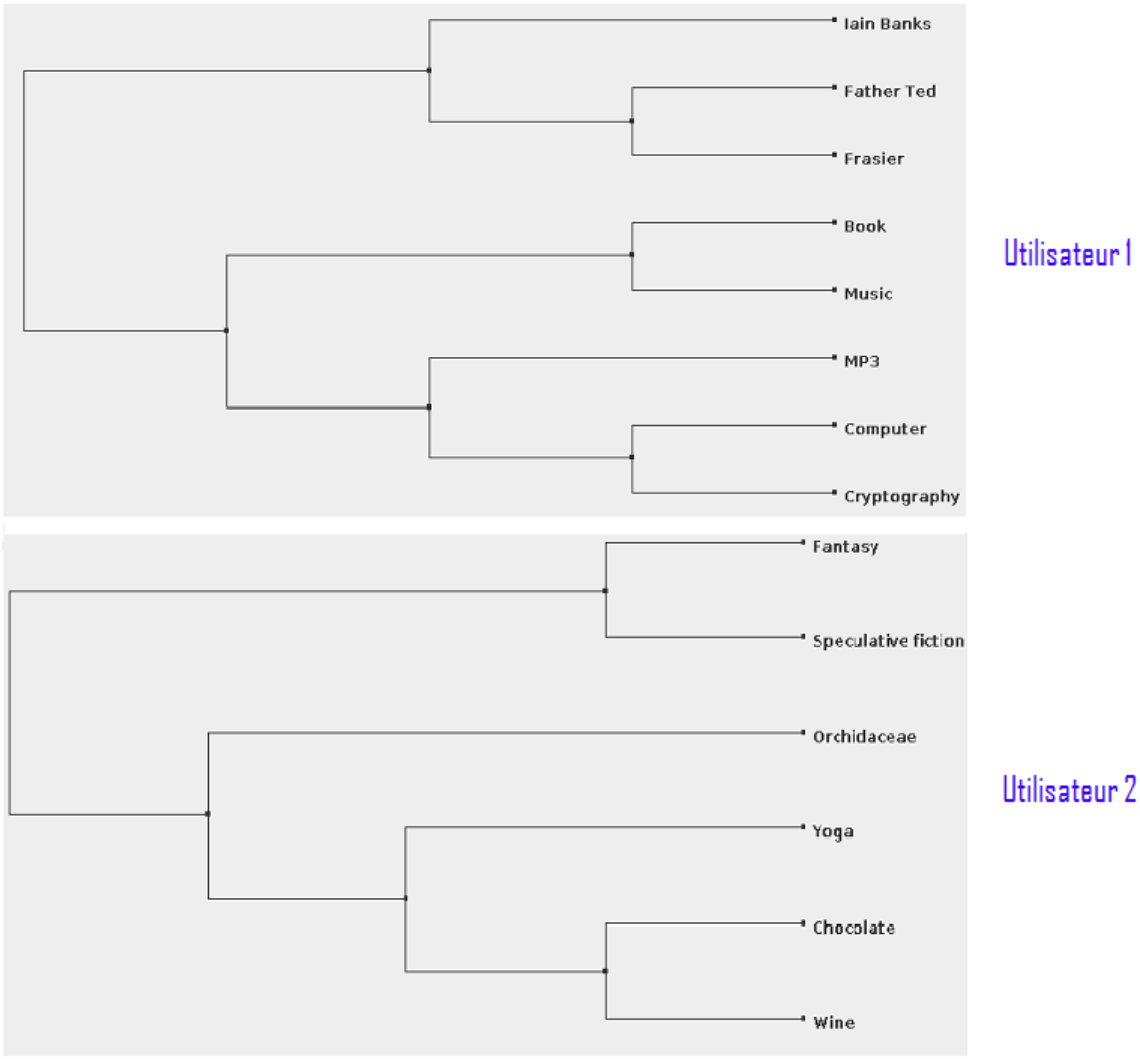


Figure 2.13: Sortie de DELVE sur deux profils utilisateurs

Pour la même raison que précédemment, les blocs d'interprétations constitués de *Orchidaceae* et *Yoga* sont supprimés. Le profil d'intérêts de l'utilisateur 2 est donc:

$\mathcal{P}^{u_2}$= { *Speculative fiction, Foods*}.

Globalement, nous constatons que les résultats obtenus par DELVE sont encourageants puisqu'il découvre des catégories qui ont un lien sémantique avec les expressions intérêts mentionnées par les utilisateurs dans leurs profils. A partir de ces résultats, on peut déduire que l'utilisateur 2 est intéressé par les genres littéraires présentant un ou plusieurs éléments surnaturels qui relèvent souvent du mythe. Cette information est d'une grande importance pour les applications cherchant à comprendre les besoins de leurs utilisateurs tels que les systèmes de recommandation afin de leurs faire des propositions de produits ou articles. Par exemple, pour le cas de l'utilisateur 2 on peut lui proposer entres autres des livres de fiction

ou des articles similaires à ce genre.

D'un autre côté, nous avons pris dans notre jeu de données deux autres utilisateurs, et leurs résultats sont indiqués dans la figure 2.14. On observe que plus un utilisateur renseigne dans son profil des expressions intérêts similaires, plus on en découvre des blocs d'interprétations et par conséquent ses domaines d'intérêts spécifiques. Le profil d'intérêts de l'utilisateur 3 est $\mathcal{P}^{u_3}$= { *2010s fashion, Entertainment, Visual arts*}. Il est obtenu à partir de trois blocs d'interprétations issus de la première itération à savoir :

- $\mathcal{B}$**loc 1**: *{Painting, Drawing}* associé à la catégorie {Category:Visual arts}.
- $\mathcal{B}$**loc 2**: *{Laugther, Music}* associé à la catégorie {Category:Entertainment}.
- $\mathcal{B}$**loc 3**: *{ Tattoo, Body piercing}* associé à la catégorie {Category:2010s fashion}.

Par contre pour l'utilisateur 4 on a le profil d'intérêts suivant $\mathcal{P}^{u_4}$= { *Mass media, Communication, Design, Hobbies, Natural resources, Personal life, Environment*}.

Nous venons de décrire notre approche DELVE de construction automatique de profil d'intérêts des utilisateurs dans un réseau social, en particulier dans LIVEJOURNAL, qui exploite essentiellement les ressources textuelles renseignées dans les profils utilisateurs sous forme de liste d'expressions. Dans la section suivante, nous décrirons une deuxième approche automatique et multilingue de construction de profil d'intérêts également des utilisateurs dans un réseau social, FRISK dans laquelle le type de ressources textuelles exploitées est différent de celui de DELVE.

4 FRISK : Find twitteR InterestS via wiKipedia

Comme indiqué précédemment, FRISK est une approche de découverte des intérêts des utilisateurs dans un réseau social au même titre que DELVE. Cependant, il prend en entrée des ressources textuelles décrites dans les profils utilisateurs sous forme de texte court. De plus, nous avons évalué FRISK sur un jeu de données contenant les profils des utilisateurs du réseau social TWITTER.

En effet, TWITTER est un réseau social de microblogage, créé en 2006 et géré par l'entreprise Twitter Inc. D'après l'encyclopédie WIKIPEDIA, il est rapidement devenu populaire jusqu'à rassembler plus de 500 millions d'utilisateurs dans le monde depuis février 2012. En fait, il permet à ses utilisateurs d'envoyer gratuitement sur internet ou par messagerie instantanée de brefs messages, appelés *tweets* dont la taille est limitée à 140 caractères. Pour respecter cette limite de caractères, certains utilisateurs ont tendance à violer les règles grammaticales ou orthographiques, notamment en abrégeant les mots. La figure 2.15 représente un tweet d'un utilisateur, dans lequel on peut déduire qu'il parle du joueur français de football "Paul Pogba". Nous observons la présence de plusieurs mots abrégés, notamment *contr* pour dire "contre", *Mé* pour dire "mais", *enigm* pour dire "énigme", etc. Le plus souvent, ces abréviations sont à l'origine de nombreuses pertes d'informations lors de l'analyse des

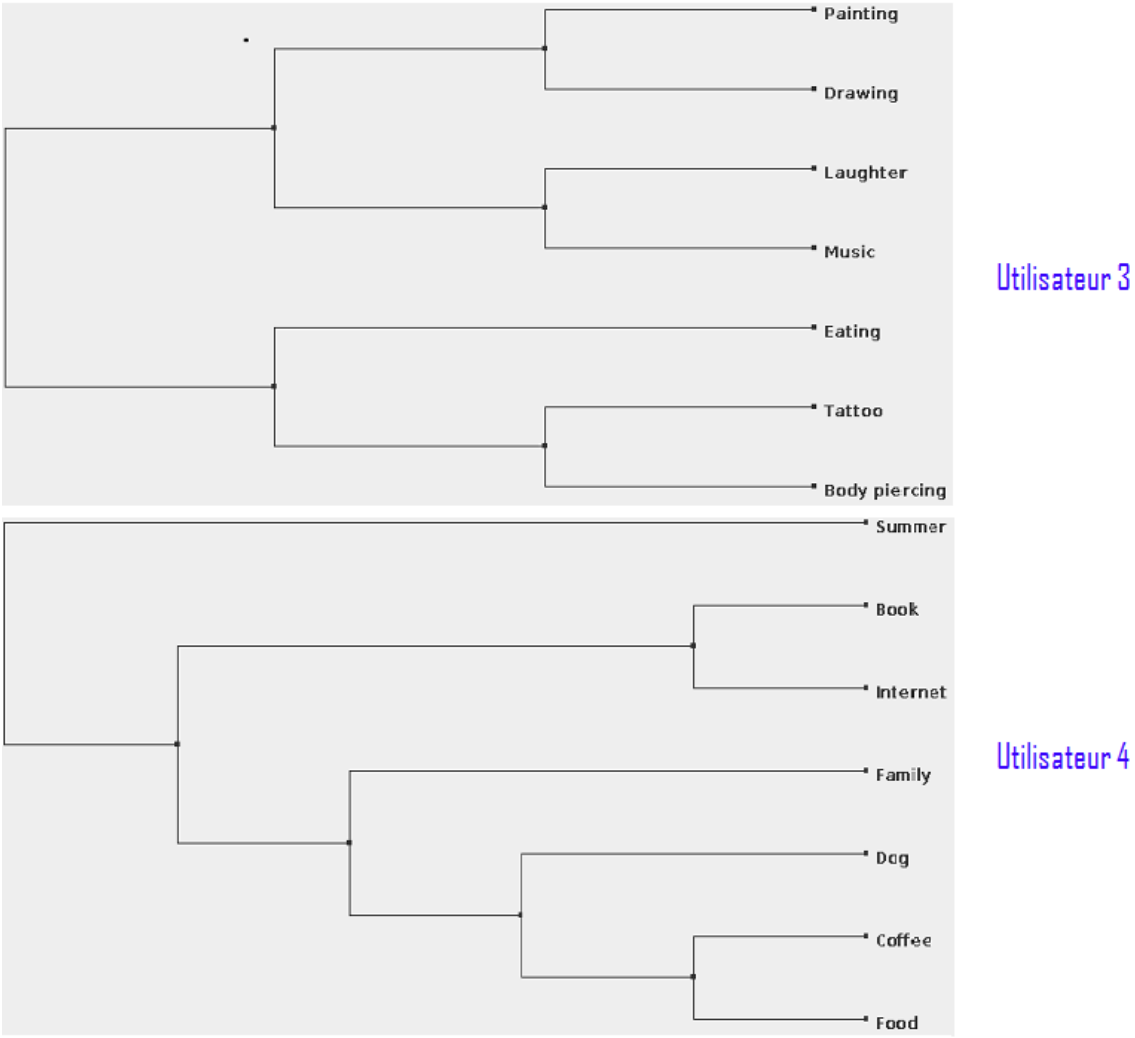


Figure 2.14: Sortie de DELVE sur deux autres profils utilisateurs.

tweets, car ces mots abrégés ne sont pas reconnus par un dictionnaire lexical courant et par conséquent ne peuvent être analysés.

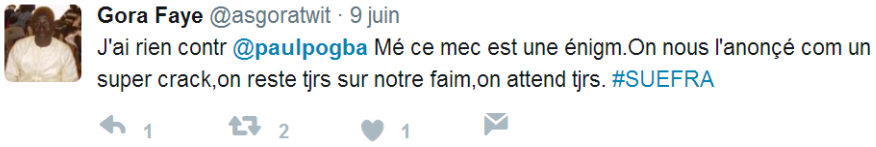


Figure 2.15: Exemple d'un tweet d'un utilisateur

Comme dans la majorité des réseaux sociaux, le profil d'un utilisateur dans TWITTER

contient les informations telles que son nom, sa localité, son site Web, sa date de naissance, ses tweets, ses contacts, ses photos, etc. La figure 2.16 est un exemple de profil, qui appartient à l'utilisateur "Barack Obama". Plusieurs études ont montré que les tweets postés par les utilisateurs de TWITTER sont des bons indicateurs pour déterminer les intérêts de ses utilisateurs. Ceci s'explique par le fait que les utilisateurs ont tendance le plus souvent à s'exprimer à travers leurs tweets sur des sujets mettant en avant ce qu'ils aiment ou n'aiment pas. Par exemple, les utilisateurs qui s'intéressent principalement à la *politique* publient régulièrement des tweets en rapport avec les *élections*, la *législation* et/ou les *événements politiques*. Par conséquent, l'analyse des tweets des utilisateurs en vue de découvrir leur domaines d'intérêts peut s'avérer prometteur.

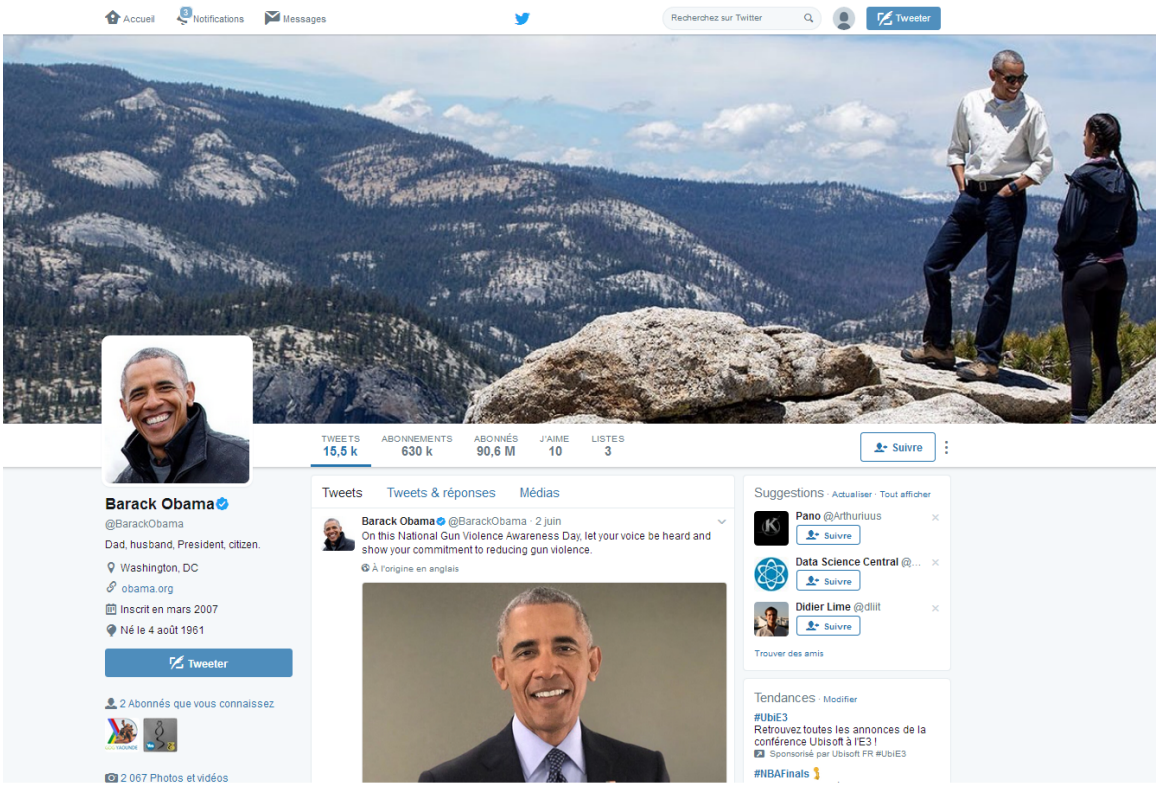


Figure 2.16: Profil TWITTER de l'utilisateur Barack Obama

Cependant, les tweets sont édités selon plusieurs langues, car TWITTER est utilisé par des millions de personnes dans différents pays. Cette variation de la langue au sein des tweets complique davantage la tâche d'analyse des tweets des utilisateurs puisqu'on ne peut appliquer les outils de traitement automatique de la langue naturelle car ils sont essentiellement monolingue. De plus, le problème de désambiguïsation de la langue naturelle évoqué lors de la description de l'approche DELVE est également un challenge a relever ici. Nous pro-

posons une approche automatique, non supervisée et multilingue, qui fait face aux difficultés sus-citées.

Dans ce qui suit, nous détaillons les différentes étapes de notre approche FRISK.

4.1 Approche proposée

Pour découvrir les domaines d'intérêts des utilisateurs, FRISK utilise également l'encyclopédie WIKIPEDIA au même titre que DELVE. Par contre, elle prend en entrée un sous ensemble des tweets publics et récents d'un utilisateur u, que nous notons $\mathcal{T}^u$. Le problème posé ici se résume donc ainsi :

$$\textbf{Entrée} \rightarrow \mathcal{T}^u$$

$$\textbf{Sortie} \rightarrow (\mathcal{P}^u, \geqslant) = \{I_1, I_2, ..., I_m\}$$

où $(\mathcal{P}^u, \geqslant)$ est l'ensemble ordonné des intérêts découverts. Plus précisément, $\mathcal{P}^u$ est constitué de l'ensemble des intérêts I_m de l'utilisateur u classés par ordre de pertinence. L'algorithme 4 présente les différentes étapes de FRISK à savoir :*l'analyse des tweets* et *l'exploration des domaines d'intérêts.*

Algorithm 4 FRISK

```
1: function FRISK(T^u) :P^u
2:     Articles^u ← getArticles(T^u)
3:     P^u ← getInterests(Articles^u)
4: end function
```

Plus précisément, la figure 2.17 est une représentation de l'architecture de FRISK, qui met en avant les différentes phases de chacune de ces deux étapes. Ces différentes étapes, ainsi que leurs phases se présentent comme suit :

1. **Analyse des tweets.** Elle consiste à analyser les tweets de chaque utilisateur et d'en ressortir pour chacun d'eux l'ensemble des articles WIKIPEDIA qui peuvent les caractériser. Elle s'effectue en trois sous-étapes comme le montre la figure 2.17.

   (a) *Prétraitement.* Son but est de nettoyer l'ensemble des tweets $\mathcal{T}^u$ d'un utilisateur, en supprimant les stopwords, qui sont des expressions utilisées fréquemment par tous les utilisateurs, telles que les jours de la semaine, les chiffres, les pronoms personnels.

   (b) *Bag Of Words.* Lors de cette phase, le document obtenu à l'issue de la phrase de prétraitement est transformé en une collection de mots appelées Bag Of Words et notée $\mathcal{BOW}^u$.

   (c) *Bag Of wikipedia Articles.* Quant à cette phase, elle transforme l'ensemble $\mathcal{BOW}^u$ obtenu précédemment en une collection d'articles WIKIPEDIA appelés Bag Of

wikipedia Articles, notée $\mathcal{BOA}^u$ qui représente au mieux les domaines d'intérêts de l'utilisateur cible.

2. **Exploration des domaines d'intérêts.** Elle utilise le graphe WIKIPEDIA, plus précisément les pages de type catégorie, pour découvrir les domaines d'intérêts de chaque utilisateur. Elle se déroule en deux sous étapes :

   (a) *Découverte des intérêts.* Tout au long de cette phase, nous cherchons à découvrir les intérêts possibles liés aux différents articles WIKIPEDIA de l'ensemble $\mathcal{BOA}^u$.

   (b) *Classification des intérêts.* Les intérêts découverts sont classés par ordre d'importance en vue de capturer uniquement ceux qui sont les plus pertinents.

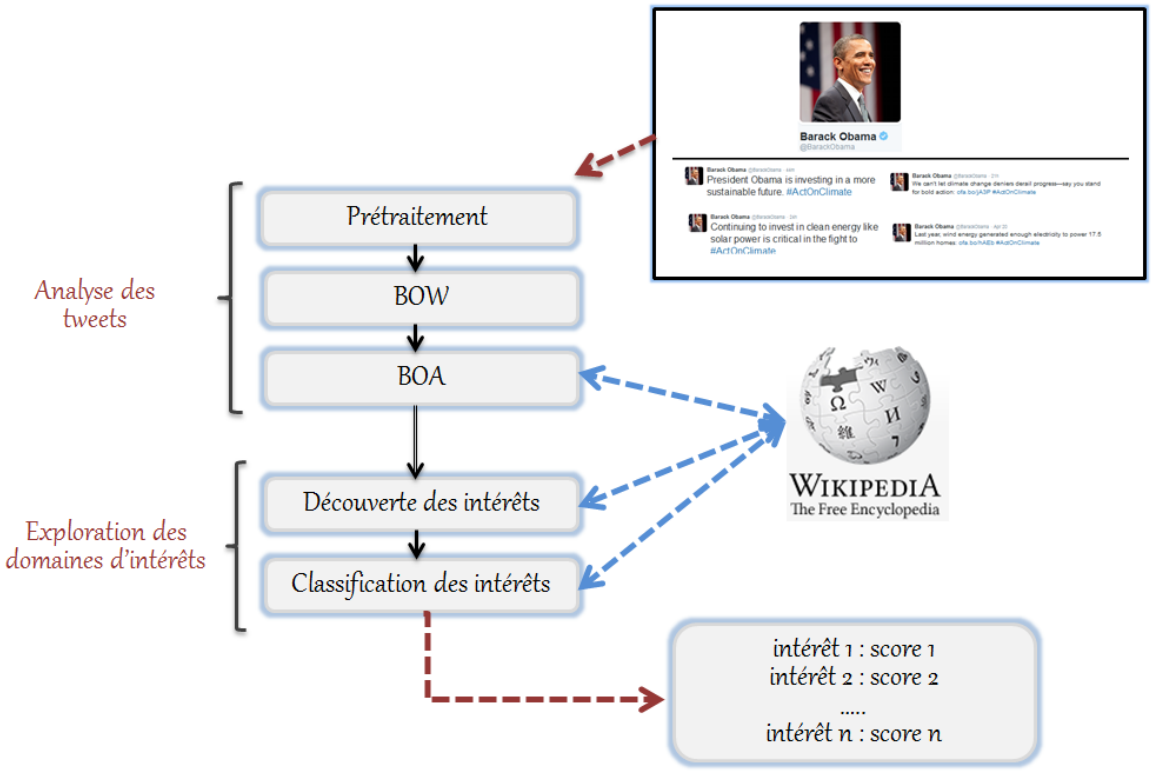


Figure 2.17: L'architecture de FRISK

4.1.1 Analyse des tweets

L'étape d'analyse de tweets consiste à déterminer à partir de l'ensemble de tweets d'un utilisateur, les articles WIKIPEDIA qui représentent au mieux ses intérêts. La description détaillée des différentes phrases de cette étape d'analyse est la suivante :

Prétraitement. Les tweets $\mathcal{T}^u$ d'un utilisateur sont prétraités dans le but de supprimer les stopwords, qui sont entre autres :

- les mots fréquents tels que *the*, *will*, *be*, *at*, *of*, etc.,
- les nombres,
- les caractères spéciaux par exemple /, \, #, etc., et
- les URLs.

En effet, les stopwords ne permettent ni de caractériser un utilisateur, ni de le distinguer des autres utilisateurs puisqu'ils sont utilisés dans les tweets par la quasi totalité des utilisateurs.

Bag Of Words. Le réseau social TWITTER possède un ensemble de marqueurs utilisés par leurs utilisateurs dans leurs tweets, notamment les mentions et les hashtags. En fait, une *mention* est une référence vers un autre utilisateur TWITTER, comme par exemple *@alain*. Il est généralement utilisé dans un tweet pour s'adresser à un utilisateur bien précis. Par ailleurs, les *hashtags* sont des mots-clés précédés du caractère " # ". Ils sont généralement composés de plusieurs mots sans espaces, notamment, "#androidgames" ou "#creditchat". Compte tenu de leur définition, nous avons décidé de supprimer les mentions de l'ensemble des tweets $\mathcal{T}^u$ des utilisateurs. Quant aux hashtags, nous leurs ôtons le caractère "#" avant de les conserver, car ils peuvent être associés à des articles WIKIPEDIA. De plus, les hashtags sont des mots porteurs d'informations importantes, puisqu'ils sont utilisés généralement pour spécifier le sujet d'un tweet. En l'occurrence, avec les hashtags "#androidgames" et "#creditchat" nous imaginons que les tweets auxquels ils appartiennent font allusion probablement au "jeu vidéo" et à la "finance" respectivement. Comme indiqué précédemment, nous les transformons en "androidgames" et "creditchat" respectivement avant de les analyser plus tard. Malheureusement, la plupart des hashtags ne sont généralement pas associés à des articles WIKIPEDIA, principalement à cause de leur syntaxe. Plus clairement, nous avons constaté que la correspondance hashtag ⇒ page WIKIPEDIA associée, n'arrive pas fréquemment, dûe au manque d'espace entre les expressions qui constituent les hashtags (android games et credit chat).

En résumé, la sortie de cette phase est une collection de mots nommée $\mathcal{BOW}^u$, provenant des tweets issus de la phase de *prétraitement*, auquelle on a supprimé les mentions et omit le caractère "#" sur les hashtags. La deuxième et la troisième colonne du tableau 2.5, montrent le résultat des deux premières phases de l'étape d'analyse sur un tweet précis.

Bag Of wikipedia Articles. Le principe ici consiste à associer à chaque mot de $\mathcal{BOW}^u$, sa page WIKIPEDIA si elle existe. Le langage naturel étant ambigu, alors la transformation de la collection de mots $\mathcal{BOW}^u$ en collection d'articles WIKIPEDIA $\mathcal{BOA}^u$ devient ici pour nous un challenge a relever. Par exemple, le mot en anglais *gender* a plusieurs significations ou interprétations possibles, décrites à travers différents articles WIKIPEDIA. Il peut s'agir

Tweet	**Prétraitement**	**Représentation** $\mathcal{BOW}$	**Représentation** $\mathcal{BOA}$
@Jules going to the store without a budget #creditchat	@Jules store budget #creditchat	store; budget; creditchat	Retail; Budget

Table 2.5: Exécution de l'étape d'analyse sur un tweet.

entre autres, soit d'une distinction entre l'homme et la femme, soit du genre grammatical. De plus, comme FRISK est multilingue, c'est-à-dire ne fait aucune hypothèse concernant la langue employée dans les tweets par les utilisateurs, ce challenge devient encore plus important.

Pour associer un article à un mot w de l'ensemble $\mathcal{BOW}^u$, FRISK cherche dans toute la WIKIPEDIA (toutes les langues confondues), les pages ayant pour titre w. Les pages obtenues peuvent être, soit des articles, soit des pages de désambiguïsation ou soit des pages de redirection. Les pages de redirection sont directement remplacées par leurs articles vers lesquelles elles redirigent. Compte tenu du fait que nous sommes dans un contexte multilingue, par conséquent un même mot peut être associé à plusieurs pages appartenant à différentes versions de WIKIPEDIA. C'est le cas du mot *store* qui est associé à la page de désambiguïsation `https://en.wikipedia.org/wiki/Store` de la WIKIPEDIA anglaise, et à la page `https://it.wikipedia.org/w/index.php?title=Store&redirect=no` de la WIKIPEDIA italienne, qui est une redirection vers l'article *Štore*, qui à son tour est une petite ville en Slovénie orientale. Dans ce type de configuration où un même mot est associé à plusieurs pages WIKIPEDIA, nous exploitons les liens de type *crosslink* du graphe WIKIPEDIA, pour ramener toutes les pages obtenues en leurs pages correspondantes dans la WIKIPEDIA anglaise. Cependant, les pages n'ayant pas de correspondantes sont supprimées. Nous nous sommes intéressés particulièrement à l'anglais car elle est la principale langue utilisée dans WIKIPEDIA et a toujours conservé cette importance. De ce fait, nous transformons l'article italienne *Štore* en l'article anglais *Štore*. Le mot *store* est maintenant associé à deux pages distinctes de la WIKIPEDIA anglaise : l'article *Štore* et la page de désambiguïsation *Store*. A ce niveau, nous cherchons à désambiguïser toutes les pages ambiguës, c'est-à-dire à déterminer pour chaque page de désambiguïsation ses différentes interprétations possibles.

Pour résoudre le problème de désambiguïsation, nous nous sommes basés sur une observation faite lors de l'évaluation de la méthode DELVE [36], décrite à la section 3. Cette observation montre que la page par défaut d'un mot est le plus souvent l'interprétation qui est la plus probable pour ce mot. En l'occurrence, soit le mot *election*, d'après les éditeurs de WIKIPEDIA, ce mot fait référence probablement plus au choix par le vote d'un électeur (page par défaut), plutôt qu'au film hong-kongais réalisé par Johnnie To en 2005, qui est l'une de ses interprétations possibles. Plus précisément, FRISK s'inspire de l'algorithme HYBRIDE présenté lors de la description de l'approche DELVE pour déterminer les différentes

interprétations possibles d'un mot. Autrement dit, si un mot w a une page par défaut, nous considérons cette dernière comme meilleure interprétation que peut prendre w, même si sa page par défaut est rattachée à une page de désambiguïsation. Cependant, si un mot w est associé plutôt à une page de désambiguïsation, ses différentes interprétations possibles sont les articles mentionnées dans sa page de désambiguïsation.

Reprenons notre exemple avec le mot *store*, auquel est associé l'article *Štore* et la page de désambiguïsation *Store*. L'article *Štore* reste inchangé. Par contre, la page de désambiguïsation *Store* est remplacée par les articles mentionnés dans son contenu. Les interprétations du mot *store* à savoir l'article *Štore* et tous les articles contenus dans la page de désambiguïsation *Store* sont classées par score de popularité décroissant, et l'article ayant le plus grand score est considéré comme la meilleure interprétation du mot *store*. En effet, la *popularité* d'une page p est son nombre de liens entrants dans le graphe WIKIPEDIA. Elle se définit comme suit :

$$popularity(p) = \mid in\text{-}neighbors(p) \mid \tag{2.11}$$

avec $in\text{-}neighbors(p)$ l'ensemble des liens entrants au noeud p définit à l'équation 2.1 à la section 2.2.

Le tableau 2.6 indique les différentes interprétations du mot *store* classées par ordre de popularité décroissant. Nous observons que l'article *Štore* est classé en sixième position. L'interprétation choisie par FRISK pour le mot *store* est celle ayant le plus grand score de popularité, c'est-à-dire l'article "Retail".

Interprétations	**Popularité**
Retail	2847
Warehouse	968
App store	68
Store and forward	67
Aircraft ordnance	59
Štore	35
Store-within-a-store	24
JML Direct TV	19
The Store	8

Table 2.6: Interprétations du mot *store* classées par ordre de popularité décroissant.

L'algorithme 5 est un résumé de l'étape d'analyse de tweets décrit précédemment. En rappel, il prend en entrée la collection de tweets $\mathcal{T}^u$ d'un utilisateur u, et retourne une collection d'articles WIKIPEDIA $\mathcal{BOA}^u$ qui représentent au mieux les mots contenus dans $\mathcal{T}^u$. La phase de *prétaitrement* est faite à la ligne 3 à l'aide de la fonction *preprocessing*,

qui retourne un document $\mathcal{D}oc^u$ dans lequel on a supprimé tous les stopwords. Par ailleurs, la fonction *getBOW* de la ligne 4 permet dans un premier temps de supprimer les mentions et le caractère "#" des hashtags dans le document $\mathcal{D}oc^u$. Et dans un deuxième temps, d'extraire de ce même document tous les mots qui le constituent, qu'on stocke dans la variable $\mathcal{BOW}^u$. De la ligne 5 à la ligne 10 on construit l'ensemble $\mathcal{BOA}^u$, à partir de quatre fonctions. La fonction *getAllPages* prend en entrée un mot w et recherche, dans toutes les versions de WIKIPEDIA, toutes les pages ayant pour titre w (ligne 6). Quant à la fonction *getEnglishPages*, elle utilise les liens de type *crosslink* du graphe WIKIPEDIA pour remplacer toutes les pages obtenues (à l'exception des pages anglaises) en leur page correspondante dans la WIKIPEDIA anglaise si elles existent (ligne 7). Celles n'ayant pas de page correspondante sont supprimées. La fonction *getInterpretations* remplace toutes les pages de désambiguïsation par les articles indiqués dans leurs contenus respectifs (ligne 8). Et enfin, la fonction *getMostPopular* qui retourne la page la plus populaire de l'ensemble $\mathcal{I}nter_w$ (ligne 9).

Algorithm 5 Analyse des tweets

1: **function** ANALYSISTWEETS($\mathcal{T}^u$) :$\mathcal{BOA}^u$
2: $\quad\mathcal{BOA}^u \leftarrow \emptyset$
3: $\quad\mathcal{D}oc^u \leftarrow preprocessing(\mathcal{T}^u)$
4: $\quad\mathcal{BOW}^u \leftarrow getBOW(\mathcal{D}oc^u)$
5: $\quad$**for each** $w \in \mathcal{BOW}^u$ **do**
6: $\quad\quad\mathcal{P}ages_w \leftarrow getAllPages(w)$
7: $\quad\quad\mathcal{P}agesEn_w \leftarrow getEnglishPages(\mathcal{P}ages_w)$
8: $\quad\quad\mathcal{I}nter_w \leftarrow getInterpretations(\mathcal{P}agesEn_w)$
9: $\quad\quad\mathcal{BOA}^u \leftarrow getMostPopular(\mathcal{I}nter_w) \cup \mathcal{BOA}^u$
10: $\quad$**end for**
11: **end function**

L'ensemble $\mathcal{BOA}^u$ étant construit, passons à la découverte des domaines d'intérêts des utilisateurs.

4.1.2 Exploration des domaines d'intérêts

Pour découvrir les domaines d'intérêts des utilisateurs, FRISK utilise les liens de type *belongTo* et *chilOf* du graphe WIKIPEDIA pour accéder aux catégories directes et parents des articles de l'ensemble $\mathcal{BOA}^u$. Par ailleurs, nous nous appuyons également sur une base de connaissances d'intérêts, nommée *ListInterests* que nous avons construite manuellement. En effet, *ListInterests* contient 545 domaines d'intérêts existants dans la vie réelle tels que : *Art, Jeu, Finance, Littérature, Politique, Sport*. Les différentes phases de cette étape se présentent comme suit:

Découverte des intérêts. Le processus de détermination des intérêts d'un utilisateur est décrit dans l'algorithme 6. Il prend en entrée l'ensemble des articles $\mathcal{BOA}^u$ qui représente un utilisateur, ainsi que notre base de connaissances d'intérêts $ListInterests$, et retourne la liste des intérêts $\mathcal{P}^u$

de l'utilisateur cible. De manière générale, les intérêts d'un utilisateur $\mathcal{P}^u$ sont l'union des intérêts associés à chaque article de $\mathcal{BOA}^u$.

Algorithm 6 Découverte des intérêts d'un utilisateur

```
1: function DISCOVERYINTERESTS(BOA^u, ListInterests): P^u
2:     P^u ← ∅
3:     for each p ∈ BOA^u do
4:         BOC_p ← categories(p) ∪ getAncestors(categories(p), l)
5:         BOI_p ← getInterests(BOC_p, ListInterests)
6:         P^u ← BOI_p ∪ P^u
7:     end for
8: end function
```

Plus précisément, d'après l'algorithme 6, pour chaque article p de $\mathcal{BOA}^u$ (ligne 3), on va tout d'abord chercher dans le graphe WIKIPEDIA toutes les catégories qui lui sont associées, qu'on stocke dans la variable $\mathcal{BOC}_p$, pour Bag Of wikipedia Categories (ligne 4). Pour le faire, nous utilisons d'une part, la fonction *categories* de l'équation 2.3 de la section 2.2, qui exploite les liens de type *belongTo* du graphe WIKIPEDIA pour déterminer les catégories directes d'un article. D'autre part, on va également utiliser la fonction *getAncestors*, qui renvoie les catégories parents à une distance l de chaque catégorie de $categories(p)$. Autrement dit, $getAncestors(categories(p), l) = \{p_2 \mid$ le prédicat $ancestor^l(p_1, p_2)$ est vrai avec $p_1 \in categories(p)\}$. Quant à la fonction *getInterests* de la ligne 5, elle extrait de l'ensemble $\mathcal{BOC}_p$ les intérêts qui existent dans notre base de connaissances d'intérêts *ListInterests*. L'utilité d'exploiter une base de connaissances est de pouvoir identifier les intérêts de la vie réelle, qui sont par la suite classés en utilisant une formule que nous présentons plus tard, qui tient compte de l'occurrence de chaque intérêt. Par exemple, prenons l'article *Cricket* de la WIKIPEDIA anglaise, le tableau 2.7 montre le résultat de la fonction *getInterests* avec $l = 0$, c'est-à-dire $getAncestors(categories(p), 0) = \emptyset$. Dans ce cas, on a $\mathcal{BOC}_p$= $categories(Cricket)$ = {*Category:Cricket terminology, Category:Cricket equipment, Category:Cricket laws and regulations, Category:Cricket*}. Plus clairement, la fonction *getInterests* parcourt les termes de chaque catégorie de $\mathcal{BOC}_p$, et vérifie s'ils existent dans la base de connaissances d'intérêts *ListInterests*, si oui, on les ajoute à l'ensemble $\mathcal{BOI}_p$. D'après la seconde ligne du tableau 2.7, des termes *cricket* et *terminology* issus de la catégorie *Category:Cricket terminology*, on a découvert l'intérêt *Cricket*. A la fin de l'exécution de la fonction *getInterests* sur l'ensemble des catégories directes de l'article *Cricket*, on obtient $\mathcal{BOI}_p$= { (*Cricket*, 0); (*Equipment*, 0); (*Law*, 0) }. En fait, les éléments de $\mathcal{BOI}_p$ sont sous la forme (I, d), où I est l'intérêt découvert et d la distance entre la catégorie ayant permis la découverte l'intérêt I et les catégories directes de p. Dans notre exemple, tous les intérêts sont découverts à partir des catégories directes de l'article *Cricket* d'où $d = 0$.

De même, pour $l = 1$, on a les resultats indiqués dans le tableau 2.8, c'est-à-dire $\mathcal{BOC}_p$ = $categories(Cricket) \cup getAncestors(categories(Cricket), 1)$. Après calcul, on a $\mathcal{BOI}_p$= {(*Cricket*, 0); (*Equipment*, 0); (*Law*, 0); (*Sports*, 1); (*Ball*, 1); (*Games*, 1)}. Nous notons que les intérêts découverts lorsque $l = 0$ sont ignorés à $l = 1$. Plus généralement, les intérêts découverts à une itération l sont ajoutés à $\mathcal{BOI}_p$ si et seulement si, ils n'ont pas été découverts à une itération inférieure à l. Nous constatons l'apparition de la catégorie générique *Sports*. Par contre, nous notons l'introduction de quelques intérêts erronés telles que *Law* et *Equipment*. Pour éliminer les intérêts erronés, nous fixons, pour chaque intérêt découvert, un seuil d'occurrence *min*. En d'autres

categories(Cricket)	**Termes**	**Intérêts**
Category:Cricket terminology	cricket	Cricket
	terminology	×
Category:Cricket equipment	cricket	Cricket
	equipment	Equipment
Category:Cricket laws and regulations	summer	×
	cricket	Cricket
	laws	Law
	and	×
	regulations	×
Category:Cricket	cricket	Cricket

Table 2.7: Résultat de la fonction *getInterests* pour $l = 0$ pour l'article *Cricket*.

termes, les intérêts dont leurs occurrences est inférieure au seuil fixé sont éliminés. Le but est de retenir pour chaque article uniquement ses intérêts pertinents. D'après la table 2.8, les occurrences de chaque intérêt sont : *Cricket*=4; *Equipment*=1; *Law*=1; *Sports*=4; *Ball*=1; *Games*= 1. Pour $min = 2$, les intérêts *Equipment*, *Law*, *Ball*, et *Games* sont éliminés car leur nombre d'occurrences respectives est strictement inférieure à $min = 2$. On a donc $\mathcal{BOI}_p$= {(*Cricket*, 0); (*Sports*, 1)} et les intérêts découverts à partir de la page *Cricket* sont {*Cricket*, *Sports*}. Ainsi, en incrémentant la valeur de l, on découvre les intérêts beaucoup plus génériques d'un article. Dans la section 4.2, nous discutons de la valeur de la variable l, qui ne doit pas être trop grande pour éviter d'atteindre des catégories parents qui contiennent des intérêts trop génériques. Dans le cas de notre article *Cricket*, avec $l = 3$, on a des intérêts beaucoup plus généraux tels que *Leisure* et *Hobbies*.

Par ailleurs, nous observons clairement l'importance de notre base de connaissances d'intérêts, puisqu'elle nous a permis d'associer à l'article *Cricket* les intérêts {*Cricket*, *Sports*} qui sont beaucoup plus expressif que les titres des catégories {*Cricket terminology, Cricket equipment, Cricket laws and regulations, Cricket*}. Le profil d'intérêts $\mathcal{P}^u$ d'un utilisateur est donc l'union des intérêts de chaque article de l'ensemble $\mathcal{BOA}^u$. Dans ce qui suit, nous présentons leur classification.

Classification des intérêts. Pour classer les intérêts découverts I de $\mathcal{P}^u$ par ordre de pertinence, nous affectons à chacun un score d'importance noté $score_I$ et calculé comme suit:

$$score_I = \sum_{p} \sum_{(I,l)\in\mathcal{BOI}_p} \frac{1}{l + 0.5} \tag{2.12}$$

En fait, plus un intérêt est proche en terme de distance (nombre de liens) de l'article ayant permis sa découverte, plus il a un score élevé. Les intérêts découverts sont donc classés par score d'importance décroissant. Les top-k intérêts dans le classement constituent le profil d'intérêts de l'utilisateur cible.

Après plusieurs exécutions dans le temps de l'algorithme FRISK sur différents tweets d'un même utilisateur, des intérêts découverts à chaque exécution, on peut distinguer les intérêts propres à un utilisateur (ceux qui durent dans le temps) par rapport aux intérêts à court terme, dûs à la présence des évènements populaires telles que les élections, les drames ou les fêtes nationales/internationales.

categories(Cricket)	*getAncestors* **avec** $l = 1$	**Intérêts**
Category:Cricket terminology	Category:Cricket	Cricket
	Category:Sports terminology	Sports
Category:Cricket equipment	Category:Cricket	Cricket
	Category:Sports equipment	Sports
		Equipment
Category:Cricket laws and regulations	Category:Cricket	Cricket
	Category:Sports rules and regulations by sport	Sports
Category:Cricket		Cricket
	Category :Team sports	Sports
	Category :Ball and bat games	Ball
		Games

Table 2.8: Résultat de la fonction *getInterests* pour $l = 1$ à partir de l'article *Cricket*.

Par exemple, lors des élections présidentielles en France, la plupart des utilisateurs français ainsi que dans le monde s'expriment ou donnent leur avis par rapport aux programmes des candidats à la présidentielle à travers leurs tweets. Dans leurs différents avis, ces utilisateurs ont tendance à utiliser les mots qui ont un lien fort avec, par exemple, la politique ou la finance. Les utilisateurs emportés par le phénomène électoral, vont cesser de s'exprimer sur la politique après le passage des élections. A cet effet, lors de la redécouverte des intérêts de ce type d'utilisateur à partir de leur nouveaux tweets, on ne va plus voir apparaître l'intérêt politique découvert à partir de ses tweets précédents à cause de l'évènement populaire et ponctuel, élection présidentielle.

4.2 Expérimentations et évaluations

Pour l'évaluation de FRISK, nous avons tout d'abord extrait du réseau social TWITTER un jeu de données nommé MULTIDS constitué de 1 347 profils utilisateurs dans lesquels les tweets sont écrits dans l'une des quatre langues sélectionnées —*Anglais*, *Français*, *Italien* ou *Espagnol*. Ensuite, nous avons exécuté FRISK sur notre jeu de données construit MULTIDS. Et enfin, nous avons comparé nos résultats à ceux obtenus à partir des méthodes d'apprentissage les plus connues telles que Naives Bayes, Machine à vecteur de support, Allocation de dirichlet latente et Random forest.

4.2.1 Collecte des données

Le processus de collecte des données multilingue MULTIDS s'est fait en deux étapes : la phase de *sélection des profils utilisateurs*, pendant laquelle nous avons sélectionné un sous ensemble de profils utilisateurs du réseau TWITTER et la phase de *collecte des tweets utilisateurs* au cours de laquelle nous avons collecté les tweets récents et publics des utilisateurs sélectionnés.

Sélection des profils utilisateurs. Pour identifier les profils utilisateurs de notre base de données MULTIDS, nous avons utilisé le moteur de recherche Twitter, qui prend en entrée un mot

clé et fournit en sortie une liste de profils utilisateurs qui s'intéressent principalement à un domaine d'intérêt représenté par le mot clé passé en entrée. Par exemple, si le mot clé de recherche est "politique" il nous renvoie les profils des utilisateurs qui s'intéressent à la politique tels que Barack Obama, Nicolas Sarkozy, François Fillon, etc. De ce fait, nous avons choisi au hasard cinq intérêts —*Politique*, *Économie*, *Jeux vidéo*, *Gastronomie* et *Sports* que nous stockons dans un ensemble $\mathcal{C}$. Chaque intérêt de l'ensemble $\mathcal{C}$ est utilisé comme mot clé de recherche en entrée du moteur de recherche Twitter afin d'obtenir les utilisateurs susceptibles d'appartenir à chaque catégorie d'intérêts. Le nombre de profils obtenus pour chaque catégorie n'étant pas représentatif, nous avons cherché une autre technique de sélection de nouveaux profils. En fait, cette nouvelle technique consiste à chercher des expressions similaires aux intérêts de l'ensemble $\mathcal{C}$ et par la suite utiliser ces expressions dans le moteur de recherche Twitter pour en sélectionner davantage de profils utilisateurs. Pour le faire, nous nous sommes inspirés de la hiérarchie de catégories indiquée sur Google AdWords[5] afin de déterminer les expressions similaires à chacun des intérêts de $\mathcal{C}$. Les expressions choisies pour chaque intérêt se présentent comme suit :

- **Politique** :gouvernement, campagne électorale, élection.
- **Économie** :économie, finance, banque.
- **Jeux vidéo** :jeu , jeux de société.
- **Gastronomie** : nourriture, boisson, restaurant.
- **Sports** :basketball, baseball, bowling, football.

Les profils utilisateurs obtenus à partir d'un mot clé sont classés dans la catégorie associée au mot clé cible. Par exemple, les profils utilisateurs obtenus à partir du mot clé "finance" sont classés dans la catégorie intérêt "Économie". Pour rendre nos données multilingue nous avons successivement traduit les expressions employées pour chaque catégorie intérêt dans l'une des quatre langues suscitées. Et par la suite, nous avons utilisé à tour de rôle ces expressions comme mot clé en entrée du moteur de recherche Twitter, afin de sélectionner de nouveaux profils utilisateurs. Le tableau 2.9 présente le nombre de profils utilisateurs obtenus par intérêts et par langue.

Intérêts	**Anglais**	**Français**	**Italien**	**Espagnol**	**Total**
Politique	92	56	69	57	**274**
Économie	93	58	57	68	**276**
Jeux vidéo	85	58	47	58	**248**
Gastronomie	89	53	73	50	**265**
Sports	88	66	67	63	**284**
Grand total	447	291	313	296	**1 347**

Table 2.9: Distribution du nombre d'utilisateurs par intérêt et par langue dans MultiDS.

[5] developers.google.com/adwords/api/docs/appendix/productsservices

Collecte des tweets utilisateurs. Pour collecter les tweets des utilisateurs sélectionnés précédemment, nous avons utilisé plutôt l'API de recherche Twitter. Plus clairement, nous avons extrait par le biais de cette API, pour chaque utilisateur, au plus 5 000 tweets récents et publics disponibles. Pour plus de fiabilité, nous avons manuellement lu les premiers tweets de chaque utilisateur dans le but de contrôler si les intérêts affectés aux utilisateurs par le moteur de recherche Twitter sont conformes à nos observations. Les utilisateurs classés dans des catégories d'intérêts différentes de celle obtenue après observation sont supprimés de MULTIDS. Nous notons que le tableau 2.9 présente le nombre d'utilisateurs obtenus après suppression des utilisateurs mal classés. Cette vérification de la crédibilité des intérêts associés aux utilisateurs est un long processus manuel indispensable pour l'évaluation de notre approche FRISK sur MULTIDS. À la fin du processus de collecte d'informations, MULTIDS contient plus de 3 millions de tweets utilisateurs, dont la répartition est indiquée dans le tableau 2.10. Dans ce tableau, la colonne "#tweets" représente le nombre total de tweets par intérêt pour toutes langues confondues. Nous observons un nombre de tweets par utilisateur plus élevé dans la catégorie Sports et faible dans celle de Gastronomie. En fait, à la base nous avons essayé de collecter au plus 5 000 tweets publics et récents par utilisateur. Cependant, ce tableau montre que pour certains utilisateurs on n'a pas pu avoir exactement 5 000 tweets, c'est notamment le cas des utilisateurs de la catégorie Gastronomie. Quant au tableau 2.11, il indique le nombre moyen de tweets "avg(#tweets)", de mots "avg(#mots)" et d'articles "avg(#articles)" obtenus par utilisateur d'une catégorie donnée lorsqu'on considère $\mathcal{T}^u \leq 500$. Le constat fait précédemment se confirme une fois de plus sur ce nouveau tableau, puisque même en considérant au maximum 500 tweets par utilisateur, nous notons un nombre moyen de tweets par utilisateur faible toujours dans la catégorie Gastronomie. Et par conséquent, le nombre moyen de mots total, et d'articles (nombre de mots associé à une page WIKIPEDIA) obtenus à partir de ces tweets sont également faibles. La colonne "% disc (mots)" représente le pourcentage de mots n'étant pas associés à un article WIKIPEDIA. Nous remarquons qu'il est sensiblement le même dans toutes les catégories.

Intérêts	**#tweets**	**avg(#tweets)**
Politique	739 499	2 698,9
Économie	596 331	2 160,6
Jeu vidéo	625 108	2 520,6
Gastronomie	492 197	1 857,3
Sports	822 597	2 896,5
Grand total	**3 275 732**	**12 133,9**

Table 2.10: Répartition des 5 000 tweets utilisateurs de MULTIDS par intérêt.

Après construction de notre base de données MULTIDS, nous passons, dans ce qui suit, à la présentation des résultats de FRISK.

Intérêts	avg(#tweets)	avg(#mots)	avg(#articles)	%disc(mots)
Politique	465,6	3 734,3	2 319,2	38
Économie	426,8	3 311,2	2 084,8	37
Jeu vidéo	443,3	3 040,8	1 956,1	36
Gastronomie	414,1	2 918,7	1 905,4	35
Sports	462,7	3 392,1	2 185,1	36

Table 2.11: Statistiques en considérant au plus 500 tweets utilisateurs par intérêt

4.2.2 Résultats de Frisk

À la suite de l'exécution de Frisk sur MultiDS, nous avons mesuré son efficacité à partir des trois mesures de performance utilisées également lors de l'évaluation des algorithmes de l'approche Delve, à savoir la précision, le rappel et la F-mesure.

Notons P_i, la précision de Frisk pour un intérêt $i \in \mathcal{C}$. Plus nettement nous avons :

$$P_i = \frac{|TP_i|}{|TP_i| + |FP_i|} \tag{2.13}$$

où TP_i encore appelé vrais positifs est l'ensemble des utilisateurs dont l'intérêt réel et prédit est i. En d'autres termes, TP_i est le nombre d'utilisateurs dont Frisk a attribué comme intérêt dominant i, et qui sont associés à l'intérêt i dans MultiDS. Cependant, FP_i ou faux positifs est l'ensemble des utilisateurs dont l'intérêt prédit et réel sont respectivement i et j avec $i \neq j$. Autrement dit, FP_i est le nombre d'utilisateurs qui sont associés à l'intérêt j dans MultiDS, alors que Frisk a prédit pour ces derniers comme intérêt dominant i.

Par ailleurs, notons R_i, le rappel de Frisk pour un intérêt $i \in \mathcal{C}$. Plus nettement nous avons :

$$R_i = \frac{|TP_i|}{|TP_i| + |FN_i|} \tag{2.14}$$

où FN_i encore appelé faux négatifs est l'ensemble d'utilisateurs dont l'intérêt prévu et réel est respectivement j et i avec $j \neq i$. En d'autres termes FN_i est le nombre d'utilisateurs qui sont associés à l'intérêt i dans MultiDS, alors que Frisk à prédit pour ces derniers comme intérêt dominant j.

Et enfin, nous notons F_i, la F-mesure de Frisk pour un intérêt $i \in \mathcal{C}$. Elle est définit comme suit:

$$F_i = \frac{2 \times P_i \times R_i}{P_i + R_i} \tag{2.15}$$

De manière générale la précision mesure le nombre de prédictions correctes pour un intérêt i précis, tandis que, le rappel mesure le nombre d'utilisateurs dont l'intérêt réel i est correctement prédit. Par contre, la F-mesure est un compromis entre la précision et le rappel. Les mesures de performance (précision, rappel et la F-mesure) globales de Frisk c'est-à-dire pour tous les intérêts $i \in \mathcal{C}$ sont définies comme suit:

$$P_{\text{FRISK}} = \frac{\sum P_i}{|\mathcal{C}|} \; ; \quad R_{\text{FRISK}} = \frac{\sum R_i}{|\mathcal{C}|} \; ; \quad F_{\text{FRISK}} = \frac{\sum F_i}{|\mathcal{C}|} \tag{2.16}$$

Notons que, les valeurs de la précision P_i, du rappel R_i et de la F-mesure F_i de chaque intérêt i , et par conséquent P_{FRISK}, R_{FRISK}, F_{FRISK}, varient en fonction du nombre de tweets que considère FRISK pour chaque utilisateur. Autrement dit, la variation de la taille de $\mathcal{T}^u$ (nombre de tweets de l'utilisateur u) a une répercussion sur les résultats de FRISK.

Nous avons fait plusieurs évaluations préliminaires de FRISK sur MULTIDS visant à identifier la meilleure valeur de la variable l. En effet, la variable l mentionnée à la section 4.1.2 lors de l'étape de découverte des intérêts des utilisateurs permet d'atteindre les catégories parents à une distance de l des articles de l'ensemble $\mathcal{BOA}^u$. Nous avons obtenu de meilleurs résultats avec $l = 2$. La figure 2.18 présente les résultats de FRISK sur MULTIDS avec une variation de la taille de $\mathcal{T}^u$ d'un pas de 50 pour chaque utilisateur. Nous nous apercevons que, plus le nombre de tweets par utilisateur est élevé plus l'exactitude globale de FRISK est aussi élevée. De même, nous constatons que l'utilisation de plus de 250 tweets par utilisateur n'a plus un impact considérable sur la précision et le rappel de FRISK puisqu'elle se stabilise au-delà de cette valeur. Cependant, nous observons des résultats déjà assez prometteurs (comme le montre la figure 2.18, précision de $0,82\%$ et rappel de $0,81\%$)avec seulement 50 tweets par utilisateur (les 50 tweets plus récents).

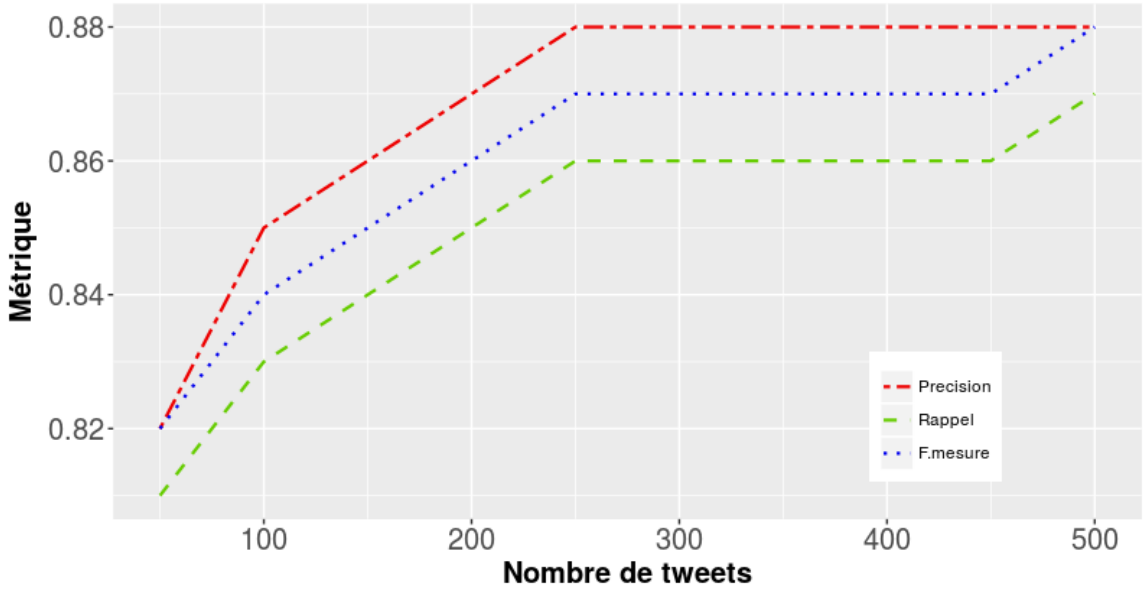


Figure 2.18: FRISK sur MULTIDS avec $50 \leq \mathcal{T}^u \leq 500$ sur un pas de 50.

Par ailleurs, la figure 2.19 montre également les résultats de FRISK par contre avec une variation de la taille de $\mathcal{T}^u$ d'un pas de 15 pour chaque utilisateur. Nous remarquons qu'à partir de 25 tweets FRISK atteint déjà une précision de 0.80%, ce qui est très remarquable. Ainsi, on peut dire que FRISK est en mesure de découvrir les intérêts des utilisateurs moins actifs, c'est-à-dire ceux qui postent moins de tweets dans leurs profils. De même, le fait que FRISK n'a pas besoin d'un nombre important de tweets pour atteindre une précision acceptable est particulièrement intéressante, car le temps de calcul de FRISK augmente en fonction du nombre de tweets considérés par utilisateur

c'est-à-dire en fonction de la taille de $\mathcal{T}^u$.

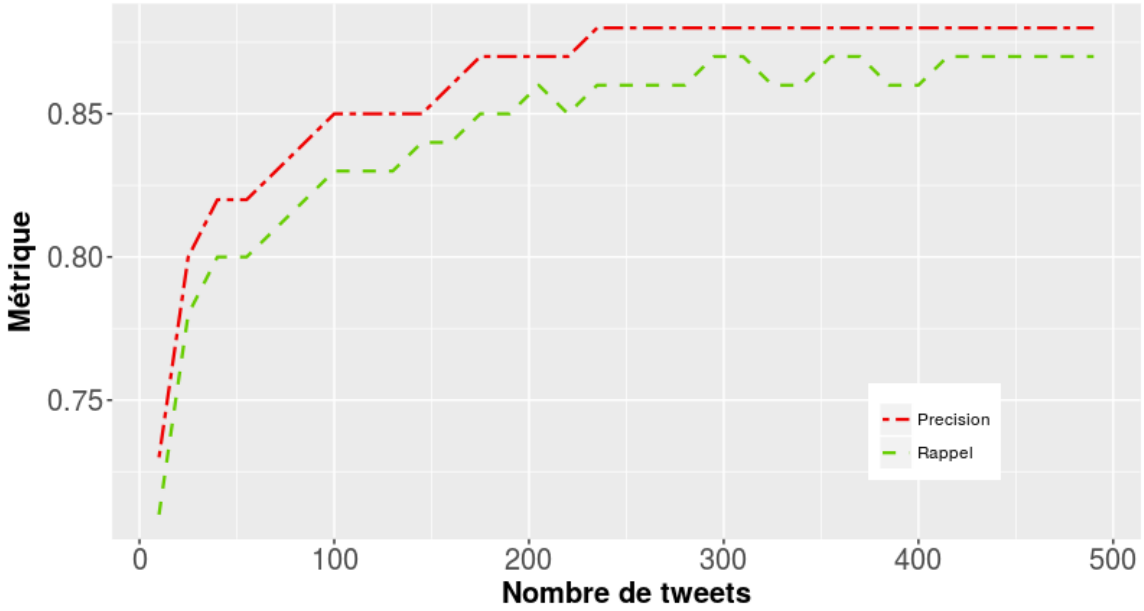


Figure 2.19: FRISK sur MULTIDS avec $10 \leq \mathcal{T}^u \leq 500$ avec un pas de 15.

De même, il existe des tableaux permettant d'analyser plus en détails les résultats d'une approche telle que la matrice de confusion. En effet, la matrice de confusion est un tableau de valeurs mettant en avant la qualité d'un système de classification. Chaque colonne de la matrice représente le nombre d'occurrences d'une classe estimée, tandis que chaque ligne représente le nombre d'occurrences d'une classe réelle (ou de référence). Le tableau 2.12 présente la matrice de confusion de FRISK dans le cas où $|\mathcal{T}^u| = 350$ tweets.

	Intérêt prédit par FRISK					
	Politique	Économie	Jeu vidéo	Gastronomie	Sports	Indéterminé
Politique	245	22	0	0	7	0
Économie	59	216	0	0	1	0
Jeu vidéo	9	19	195	1	24	0
Gastronomie	6	12	1	235	11	0
Sports	3	2	3	0	275	1

Table 2.12: Matrice de confusion de FRISK sur MULTIDS avec $|\mathcal{T}^u| = 350$ tweets

Nous constatons que FRISK se trompe le plus souvent lorsqu'il s'agit de traiter les utilisateurs appartenant à l'intérêt "Politique" en les mettant dans l'intérêt "Économie". Nous avons constaté

que cette erreur peut être due au fait que, contrairement aux autres intérêts, les catégories "Politique" et "Économie" sont deux intérêts en quelque sorte liés. Après tout, la politique d'un pays a une influence sur son économie et l'inverse. De même, FRISK se trompe également lorsqu'il s'agit de classer les utilisateurs de la catégorie "Sport", qu'il met dans "Jeu vidéo". En fait dans WIKIPEDIA, "Jeu" est une catégorie parent à "Sports" et comme dans notre approche nous parcourons le graphe WIKIPEDIA de façon ascendante (de la catégorie fille vers la catégorie parent), alors il y a un partage d'un sous-ensemble d'intérêts communs entre ces deux catégories : ce qui explique la confusion faite par FRISK. Par ailleurs, FRISK se comporte d'une manière inattendue en classant certains utilisateurs de la catégorie "Économie" dans "Jeu vidéo" ou "Gastronomie". Après exploration de notre base de données MULTIDS, nous avons constaté la présence des tweets de la catégorie "Économie" s'exprimant sur des sujets en relation avec la gastronomie ou les jeux parce qu'ils racontent des évènements que les auteurs de ces tweets ont sponsorisés.

4.2.3 Comparaison de FRISK avec les méthodes existantes

Pour la comparaison de FRISK avec les méthodes de classification existantes dans la littérature, nous avons sélectionné tout d'abord trois classifieurs de texte les plus répandus, à savoir les machines à vecteurs de support ou SVM, la classification naïve bayésienne ou tout simplement Naive Bayes, et les forêts d'arbres décisionnels ou Random Forests. Nous avons aussi considéré l'approche LDA dans notre comparaison. Ensuite, nous avons divisé notre jeu de données MULTIDS en deux parties : une pour l'apprentissage (à partir de laquelle les classifieurs se basent pour construire leurs modèles respectifs) et l'autre pour la prédiction. Et enfin, nous avons comparé les résultats obtenus à ceux de FRISK.

Méthodes de classification. Les SVM sont un ensemble de techniques d'apprentissage supervisées destinées à résoudre des problèmes de discrimination et de régression. En d'autres termes, elles permettent soit de décider à quelle classe appartient un échantillon, soit de prédire la valeur numérique d'une variable. Par ailleurs, Naive Bayes est un type de classification Bayésienne probabiliste simple basée sur le théorème de Bayes. Quant aux forêts d'arbres décisionnels, elles font partie des techniques d'apprentissage automatique, qui effectuent un apprentissage sur de multiples arbres de décision entraînés sur des sous-ensembles de données légèrement différentes. Toutes ces méthodes sont basées sur une technique d'apprentissage automatique où l'on cherche à produire automatiquement des règles à partir d'une base de données d'apprentissage contenant des «exemples».

Construction des données d'apprentissage et de prédiction. Pour construire nos bases de données d'apprentissage notée MULTIDS_TRAIN et de prédiction notée MULTIDS_TEST, nous avons fractionné aléatoirement MULTIDS en deux ensembles de profils utilisateurs tout en veillant à ce que dans chacun de ces ensembles les utilisateurs soient uniformément répartis (sur les cinq catégories), afin d'éviter le déséquilibre de classes dans les quatre langues. A la fin du fractionnement, MULTIDS_TRAIN contient au total 753 utilisateurs et MULTIDS_TEST 594 utilisateurs répartis comme le montre le tableau 2.13.

FRISK Vs. Méthodes de classification. Le tableau 2.14 montre la comparaison de FRISK avec les méthodes d'apprentissage sus-citées. Nous constatons que FRISK dépasse largement ces trois classificateurs de texte dans toutes les métriques, et ce malgré l'utilisation d'un grand nombre de tweets par utilisateur pour construire le modèle d'apprentissage. De même, nous observons que la capacité de ces trois classificateurs de texte à prédire les intérêts des utilisateurs dans un contexte

		Langues				Total
		Anglais	*Français*	*Italien*	*Espagnol*	
Politique	Tr	46	31	39	32	148
	Te	46	25	30	25	126
Économie	Tr	50	31	35	37	153
	Te	43	27	22	31	123
Jeu vidéo	Tr	42	36	32	36	146
	Te	43	22	15	22	102
Gastronomie	Tr	44	32	39	33	148
	Te	45	21	34	17	117
Sports	Tr	43	35	41	39	158
	Te	45	31	26	24	126
Grand total		447	291	313	296	**1,347**

Table 2.13: Répartition des données dans MultiDS_Train et MultiDS_Test.

multilingue est très limitée (SVM étant le meilleur des trois).

Intérêt	Frisk			SVM			Naive Bayes			R. Forest		
	P	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
Politique	0.81	0.88	**0.84**	0.68	0.41	0.51	0.57	0.41	0.47	0.45	0.22	0.29
Economie	0.80	0.83	**0.82**	0.56	0.45	0.50	0.56	0.50	0.52	0.37	0.46	0.41
Jeu vidéo	0.98	0.81	**0.89**	0.86	0.76	0.81	0.57	0.78	0.65	0.75	0.70	0.72
Gastronomie	0.99	0.91	**0.95**	0.53	0.91	0.67	0.51	0.60	0.55	0.44	0.70	0.54
Sports	0.89	0.97	**0.93**	0.41	0.42	0.42	0.42	0.32	0.36	0.43	0.32	0.36
Average	0.89	0.88	**0.89**	0.60	0.58	0.57	0.52	0.51	0.50	0.48	0.47	0.45

Table 2.14: Frisk vs. classifieurs de texte sur MultiDS_Test ($\mathcal{T}^u \leq 500$).

Plus globalement, la figure 2.20, montre la métrique "F-mesure" de Frisk et de nos trois classifieurs sur un nombre de tweets compris entre 50 et 500 avec un pas de 50. L'observation faite précédemment sur les limites des classifieurs se confirme, puisque Frisk les dépasse nettement quelque soit le nombre de tweets considéré. De plus, les méthodes de classification n'arrivent pas à dépasser 60% quelque soit le type de configuration, ce qui confirme effectivement qu'elles ne sont pas adaptées à un contexte multilingue.

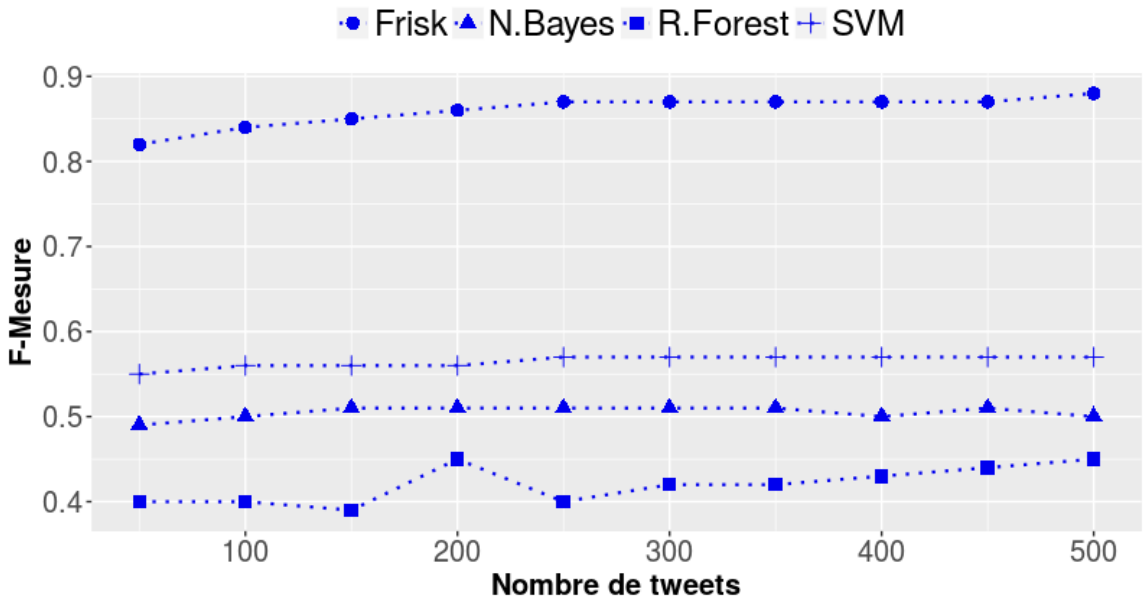


Figure 2.20: Comparaison sur MULTIDS_TEST avec $50 \leq \mathcal{T}^u \leq 500$ sur un pas de 50.

FRISK VS. LDA. LDA est un modèle génératif probabiliste permettant d'expliquer des ensembles d'observations, par le moyen de groupes non observés, eux-mêmes définis par des similarités de données. En d'autres termes, c'est un modèle probabiliste de sujets définis, dans lequel chaque sujet est représenté par un document contenant une collection de termes. Il prend en entrée un ensemble de mots issus d'un document, qu'il voit comme un mélange de sujets, selon une certaine distribution de probabilité. Nous avons comparé FRISK à l'approche proposée par Weng et ses collègues [88], qui appliquent LDA pour découvrir les intérêts des utilisateurs de Twitter. Pour le faire nous avons tout d'abord fait un prétraitement sur nos données. En fait, la phase de prétraitement consiste tout d'abord à mettre tous les mots des tweets utilisateurs en minuscule et par la suite supprimer les stopwords, afin d'éviter d'inclure des mots généralement utilisés par la quasi totalité des utilisateurs. Ensuite, nous avons généré le modèle LDA à partir des tweets obtenus à l'issue de la phase de prétraitement. Le modèle de sujet construit se compose de cinq groupes de mots comme l'indique la table 2.15. Chaque colonne représente un intérêt et n'est pas étiquetée, par conséquent nous sommes obligés de deviner à quelle catégorie d'intérêt correspond chaque groupe en se basant sur les mots qui s'y trouvent.

De la table 2.15, nous pouvons voir que certaines catégories intérêts sont facilement reconnaissables en visualisant les mots qui les décrivent, c'est le cas des colonnes "Intérêt 2" et "Intérêt 5" qui correspondent respectivement à la "Politique" (d'après les mots tels que "minister", "donald", "president" et "clinton") et au "Jeux vidéo" ("videogrames", "xbox", et "nintendo"). Par contre, les autres catégories intérêts ne semblent pas être faciles à identifier sans ambiguïté. Notamment la colonne, "Intérêt 3" qui comprend des mots en italien, dont certains représentent l'intérêt "Sports" (" calcio " qui signifie football), et d'autres indiquent l'intérêt "Politique" ("politica " qui signifie politique et «renzi», qui est le nom de famille de l'ancien Premier ministre italien). Cependant, la colonne "Intérêt 1" est une collection de mots en français, parmi lesquels certains indiquent les catégories intérêts "Jeux vidéo" ou "Sports" (" jeu " et " jeux ") d'une part, et d'autre part, celle de

Intérêt 1	Intérêt 2	Intérêt 3	Intérêt 4	Intérêt 5
jeu, bien	day, finance	politica, calcio	juego, nuevo	videogames, food
faire, j'ai	win, watch	solo	gracias	xbox, trailer
merci	support	ecco, dopo	partido	nintendo
monde, faut	team, week	prima, renzi	ver, ser	foodie
après, contre	release, world	cosa, italia	día, españa	foodporn, wars
entreprises	work, donald	ora, anni	juegos, gran	super, play
soir, bonjour	check	roma, grazie	mundo, mejor	day, sale
jour, jeux	big, campaign	sempre, lavoro	ahora, gobierno	lego, week
français	minister	sapevatelo	economia	gameplay
nouveau	president	tramite	millones	playstation
ans, demain	clinton, top	ancora	nueva, semana	world, gamedev

Table 2.15: Modèle de sujets LDA appris avec MultiDS_Train.

"Économie" (" entreprises "). Enfin, la colonne " Intérêt 4 " est un groupe de mots en espagnol qui ne correspondent visiblement pas à l'un des cinq intérêts que nous avons sélectionnés. A cet effet, nous constatons donc que LDA a du mal à identifier nos cinq catégories d'intérêts, ce qui entraîne une prédiction pas assez conforme à nos attentes.

Lors de l'évaluation de LDA, nous avons testé différents paramètres, visant à modifier le nombre et le type de sujets identifiés afin d'obtenir en sortie nos cinq catégories d'intérêts. En particulier, compte tenu du fait que nous avons cinq catégories d'intérêts et quatre langues, nous fixons le nombre de sujets à 20, dans l'espoir d'obtenir quatre groupes de mots (un pour chaque langue et pour chacun de nos cinq intérêts prédéfinis). Malgré cette configuration nous n'avons toujours pas pu obtenir les résultats attendus. Nous avons soumis au modèle de sujets appris (celui de la figure 2.15) les tweets des utilisateurs de l'ensemble MultiDS_Test et par la suite, nous avons calculé la précision, le rappel et la F-measure pour les deux catégories d'intérêts "Jeux vidéo" et "Politique" identifiées sans ambiguïté par LDA lors de la phase d'apprentissage. En définitive, pour la catégorie intérêt "Politique", nous obtenons une précision de $0,29$, un rappel de $0,37$ et une f-mesure de $0,33$. Par contre, pour la catégorie intérêt "Jeux vidéo", nous avons eu une précision de $0,55$, un rappel de $0,34$ et une f-mesure de $0,42$. Globalement, dans les deux cas, on se rend bien compte que les valeurs de ses mesures de performance sont bien inférieures à celles obtenues avec Frisk.

Étant donné que les classifieurs de texte (SVM, Naives Baiyes, Random forest et LDA) fonctionnent généralement mieux dans un contexte monolingue, une alternative possible est de construire quatre modèles d'apprentissage monolingue (une pour chacune de nos quatre langues sélectionnées) et l'utiliser pour prédire les intérêts des utilisateurs en fonction de la langue dominante utilisée pour éditer leurs profils. Cette solution n'est clairement pas objective, puisque d'une part, la langue de l'utilisateur doit être connue à l'avance, et par conséquent on ne peut traiter les utilisateurs qui utilisent plusieurs langues simultanément dans leurs profils. D'autre part, nous n'ignorons sans doute pas que la construction d'une base de données d'apprentissage nécessite un temps considérable et de la patience pour sa réalisation. A cet effet, nous avons paramétré Frisk pour une version monolingue.

Frisk monolingue et comparaison. Dans la version monolingue de Frisk que nous nommons FriskMono, les différentes étapes de Frisk restent inchangées, sauf la phase de construction

de l'ensemble $\mathcal{BOA}^u$. Plus précisément, au lieu de chercher dans toutes les éditions de WIKIPEDIA les articles ayant pour titre $w \in \mathcal{BOW}^u$, on va plutôt faire la recherche dans l'unique version de WIKIPEDIA dont la langue est celle utilisée par l'utilisateur cible pour éditer son profil. Tout d'abord, nous présentons les résultats de FRISKMONO sur MULTIDS à travers la figure 2.21. Nous constatons une stabilité par rapport aux résultats de FRISK dans un contexte multilingue.

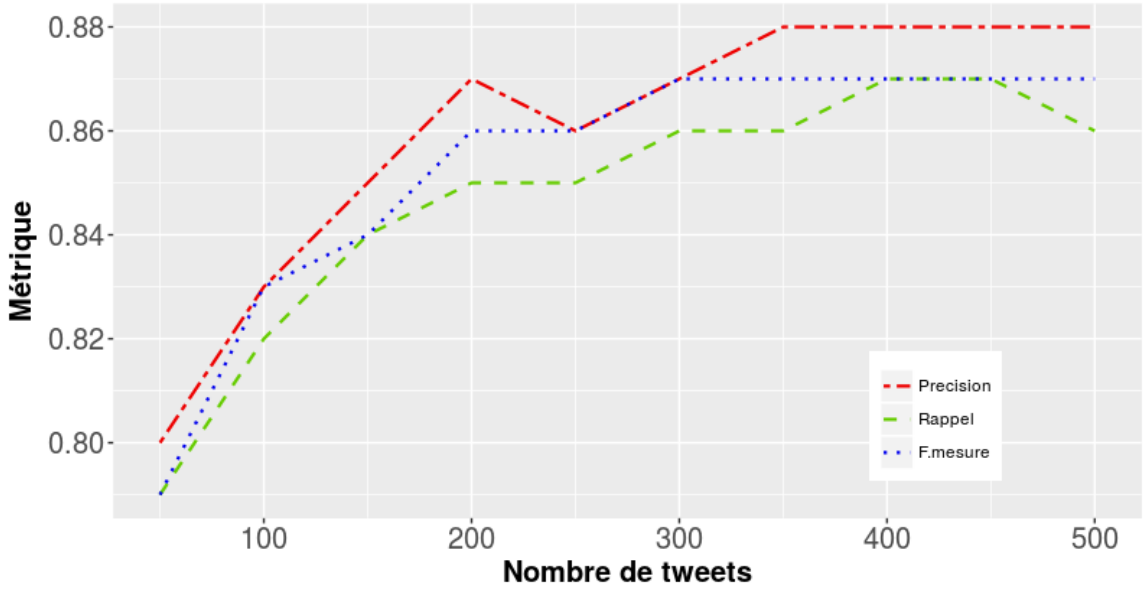
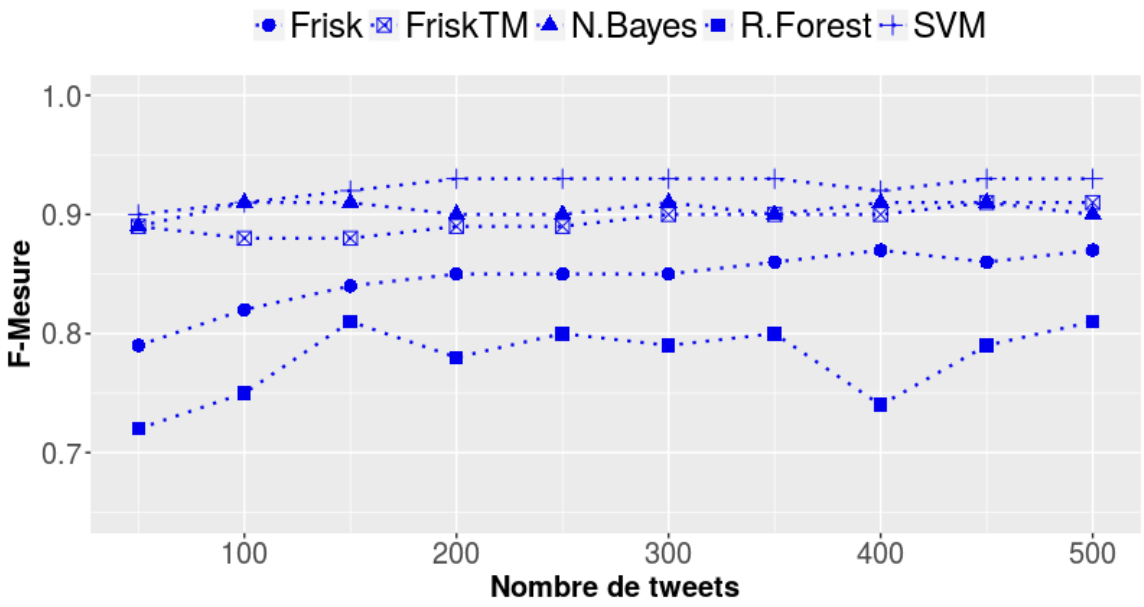


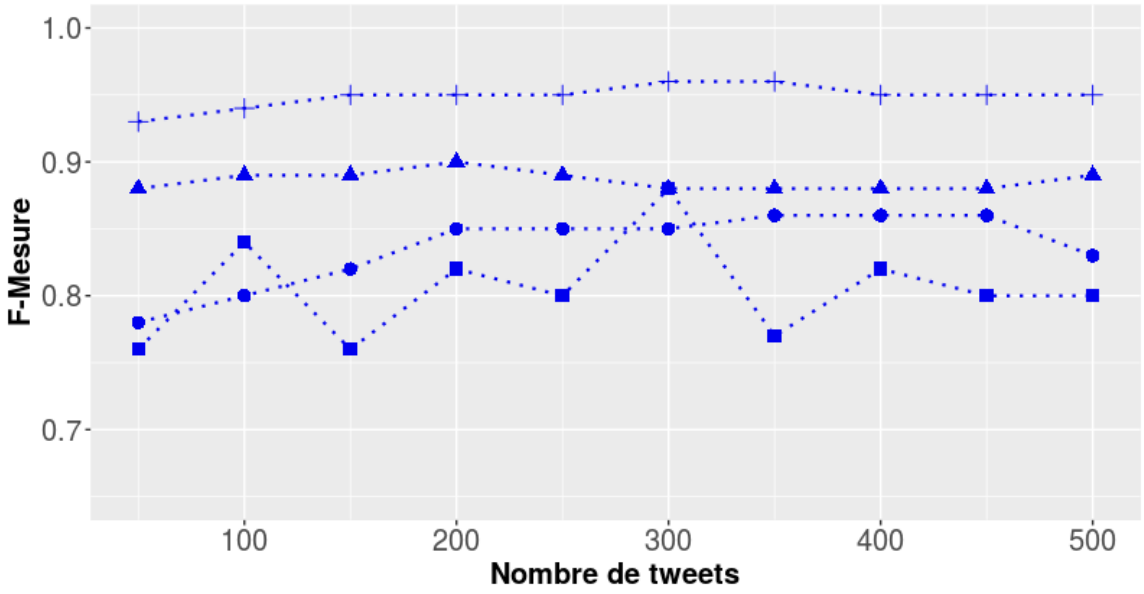
Figure 2.21: FRISKMONO sur MULTIDS avec $50 \leq \mathcal{T}^u \leq 500$ sur un pas de 50.

Pour la comparaison, nous avons séparé MULTIDS en quatre ensembles de données MONOEN, MONOFR, MONOIT et MONOES, où chaque ensemble représente les profils utilisateurs dans lesquels les tweets sont édités pour la plupart en utilisant respectivement les langues anglais, français, italien et espagnol. Ensuite nous avons divisé chacun de ces quatre ensembles de données en deux parties : les données d'apprentissage et de prédiction comme indiqué dans la figure 2.13. De plus, nous avons inclus dans l'évaluation une autre variation de FRISK que nous appelons FRISKTM, dans laquelle l'ensemble $\mathcal{BOA}^u$ de chaque utilisateur est construit en utilisant l'outil TagMe [20]. En effet, TagMe est un outil d'annotation des textes courts basé essentiellement sur WIKIPEDIA. Plus précisément, TagMe associe à chaque mot d'un tweet au plus un article WIKIPEDIA et en cas d'ambiguïté, il utilise le contexte pour sélectionner l'une des possibles interprétations du mot cible. Les figures 2.22 et 2.23 montrent la métrique F-mesure obtenue dans un contexte monolingue pour chacune de ces cinq approches. Nous constatons que SVM (respectivement Random Forest) est la meilleure (respectivement faible) dans tous les cas. Quant à FRISK, dans certains cas elle est relativement au dessus de Random forest et dans d'autres elle le croise. En fait, cela montre que la précision de FRISK ne dépend pas fortement de la langue utilisée par les utilisateurs, alors que c'est le cas contraire pour les classificateurs de texte.

Par ailleurs, on constate que la précision de FRISKTM dans la figure 2.22a est comparable à celle de SVM et Naive Bayes par rapport à FRISK. Ce qui est probablement dû à l'effet de TagMe qui tient compte du contexte d'un tweet lors de la détermination des interprétations des mots. Nous notons que TagMe traite aussi le cas de la langue italienne. La figure 2.23a montre les résultats



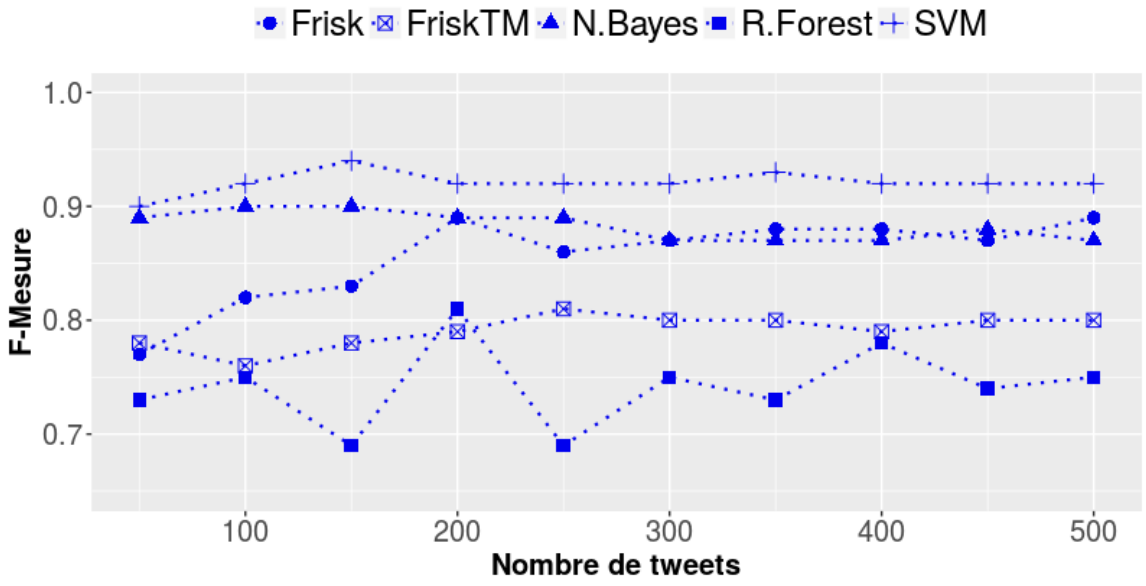
(a) Anglais



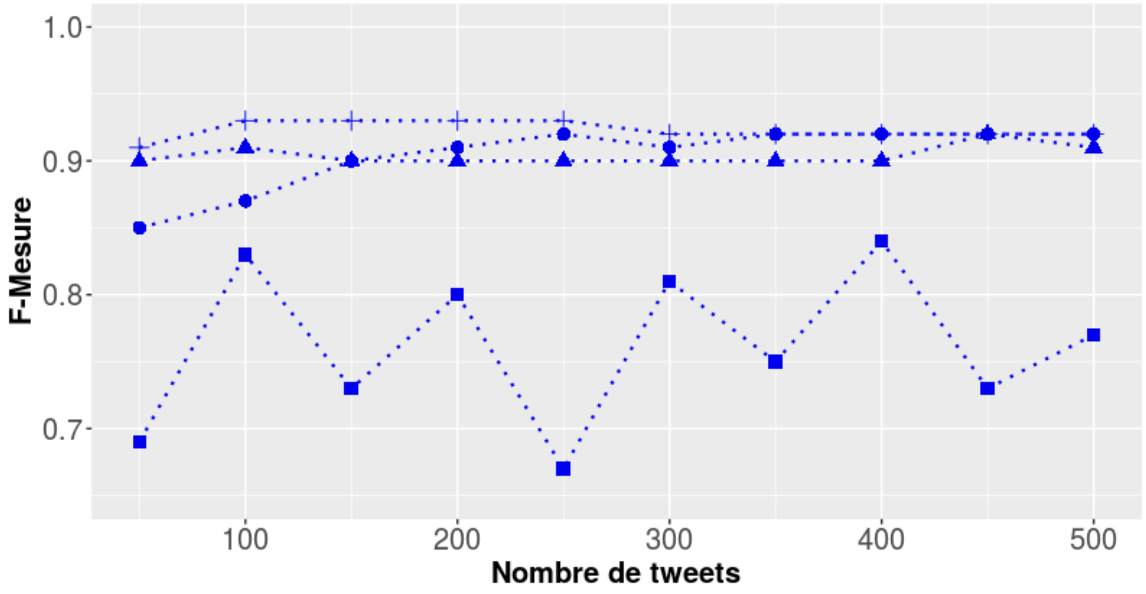
(b) Français

Figure 2.22: FRISK, FRISKTM et classifieurs sur MULTIDS_TEST pour les langues anglais et français

de TagMe sur le jeu de données italien. Nous constatons une baisse considérable des performances de TagMe jusqu'en dessous de FRISK, ce qui montre son instabilité en fonction de la variation de



(a) Italien



(b) Espagnol

Figure 2.23: FRISK, FRISKTM et classifieurs sur MULTIDS_TEST pour les langues Italien et Espagnol

la langue employée par les utilisateurs. Cependant, le seul inconvénient est que TagMe n'est pas vraiment multilingue car il doit savoir à l'avance la langue employée dans les tweets afin de calculer l'ensemble $\mathcal{BOA}^u$.

Conclusion

Nous avons consacré ce chapitre à la description des algorithmes de construction du profil d'intérêts des utilisateurs dans les réseaux sociaux. Tout d'abord nous avons commencé par une étude préliminaire à l'issue de laquelle nous avons proposée l'approche DELVE.

En résumé, DELVE prend en entrée une liste d'expressions issue d'un profil utilisateur qu'elle analyse afin de déduire les intérêts de l'utilisateur. L'approche se déroule principalement en deux étapes : la détermination des interprétations de chaque expression et la découverte des intérêts des utilisateurs. La détermination des interprétations exploite le graphe WIKIPEDIA, la mesure de similarité WLM et l'algorithme PageRank, pour tout d'abord déterminer les interprétations possibles de chaque expression, ensuite, affecter un score d'importance à toutes les interprétations, et enfin choisir parmi les interprétations éventuelles de chaque expression, celle ayant le score le plus élevé. Quant à l'étape de découverte des intérêts des utilisateurs, elle utilise également la mesure de similarité WLM, les articles et les catégories de WIKIPEDIA pour regrouper les interprétations déterminées par blocs similaires, et par la suite affecter une catégorie à chaque bloc d'interprétations similaires en fonction de leur pertinence.

Lors de la conception de DELVE, nous avons présenté une étude sur le problème de la désambiguïsation des expressions intérêts renseignées dans les réseaux sociaux. La difficulté principale réside dans le fait que la liste des expressions intérêts renseignées dans un profil utilisateur est fournie dans un contexte très limité. Nous avons expérimenté trois procédures, dont l'objectif est d'associer à une expression intérêt exprimée en langage naturel sous un vocabulaire libre, la meilleure interprétation possible. Tout d'abord, l'algorithme DÉFAUT choisit la page par défaut proposée par WIKIPEDIA comme interprétation d'une expression intérêt. Ensuite, l'algorithme DÉSAMBIG qui considère pour une expression intérêt, toutes les interprétations proposées par WIKIPEDIA et calcule un score pour chaque interprétation sur la base de leur proximité sémantique dans le graphe des interprétations. Et enfin, l'algorithme HYBRIDE qui applique DÉSAMBIG si et seulement s'il n'existe pas une page par défaut pour une certaine expression intérêt. Les résultats obtenus sont très encourageants, et nous comptons poursuivre nos expérimentations sur des données plus nombreuses.

Par ailleurs, l'approche FRISK prend en entrée des tweets utilisateurs et se déroule en deux étapes : l'analyse des tweets et l'exploration des domaines d'intérêts des utilisateurs. La première étape consiste à déterminer les articles WIKIPEDIA qui représentent au mieux les mots employés dans les tweets d'un utilisateur. La deuxième étape se sert du graphe WIKIPEDIA et d'une base de connaissances d'intérêts construite manuellement pour découvrir les intérêts des utilisateurs.

Contrairement aux approches non supervisées, FRISK qui est une approche multilingue et non supervisée ne caractérise pas les intérêts des utilisateurs comme des sacs de mots, mais avec des mots précis, tels que "Politique", "Économie", "Gastronomie". Ce type de résultat proposé par FRISK est beaucoup plus proche de ceux des approches supervisées, bien qu'elle n'a pas besoin de faire de l'apprentissage. De plus, son évaluation montre un comportement prometteur même lorsqu'elle est en possession de peu de tweets, contrairement aux méthodes supervisées.

Nous avons évalué chacune de nos approches sur des jeux de données réelles provenant des réseaux sociaux LIVEJOURNAL et TWITTER, et les résultats que nous avons obtenus sont très encourageants. Nous avons comparé FRISK à trois méthodes de classification supervisées (SVM, Naives Baiyes, Random Forest). Nous constatons que FRISK dépasse largement toutes ces méthodes même si nous les entraînons sur un grand nombre de tweets. De plus, on s'aperçoit que la capacité de ces méthodes à prédire les intérêts des utilisateurs dans un contexte multilingue est très limitée. En outre, SVM qui est le meilleur de tous n'arrive pas à dépasser 60% de précision quelque soit le type de configuration.

Étant donné que les classifieurs de texte fonctionnent généralement bien dans un contexte monolingue, nous avons voulu expérimenter ce point de vue. A cet effet, nous avons paramétré notre approche FRISK pour qu'elle puisse s'exécuter dans un contexte monolingue (FRISKMONO) et par la suite, nous avons comparé ses résultats à ceux des classifieurs SVM, Naives Baiyes, Random Forest et LDA. Lors de la comparaison nous avons également introduit une variante de FRISKMONO appelée FRISKTM, qui utilise la méthode de désambiguïsation des mots TagMe pour déterminer les interprétations (articles WIKIPEDIA) de chaque mot utilisé dans les tweets. Les résultats montrent que les méthodes de classification à l'exception de Random Forest sont meilleures que FRISK et FRISKTM. Néanmoins, FRISK est stable quelque soit la langue utilisée. Au vue des résultats de FRISKTM sur la langue anglaise, on peut dire que la phase de construction de la collection d'articles $\mathcal{BOA}^u$ a une répercussion sur la performance de FRISK. A cet effet, nous avons exploré un autre outil d'annotation multilingue nommé Babelfy [50] pour désambiguïser les mots des tweets des utilisateurs. Plus précisément, Babelfy utilise Wikipedia et WordNet pour annoter les mots et par conséquent on ne peut le comparer à FRISK, car les mots annotés en utilisant WordNet ne peuvent pas être représentés dans notre ensemble $\mathcal{BOA}^u$, ce qui peut biaiser nos résultats.

Au final, FRISK est une approche multilingue qui présente des résultats très encourageants, mais qui peut encore être améliorée.

Chapitre 3

ANALYSE DES TRAITS DE PERSONNALITÉ DES INDIVIDUS

Introduction

D'après le syndicat professionnel français FEVAD[1](Fédération du E-commerce et de la Vente à Distance) constitué de plus de 500 entreprises ayant une activité de vente à distance, 2016 a été une année marquante pour le e-commerce français. Cela se traduit par une augmentation de 9% entre 2015 et 2016 du nombre de transactions sur les sites d'e-commerce. Ainsi, compte tenu de cette augmentation, nous pouvons dire que la population française s'intéresse de plus en plus aux achats en ligne. Cet intéressement grandissant a entraîné la prolifération des sites d'e-commerce, qui par la suite a accentué une énorme concurrence dans ce marché. De plus, cette concurrence oblige les auteurs de ces sites à se démarquer des uns et des autres afin non seulement d'attirer plus de clients, mais aussi de les fidéliser davantage. C'est ainsi que les systèmes de recommandation se sont développés et sont désormais des outils incontournables du e-commerce.

Pour aller au delà des systèmes de recommandation traditionnels (collaboratifs et basés sur le contenu) la communauté recherche s'intéresse de plus en plus à mieux comprendre les utilisateurs en cernant leur personnalité [71, 70, 69, 12, 60].

La personnalité est l'ensemble des comportements, des attitudes et des aspects émotionnels qui caractérisent une personne. D'après WIKIPEDIA, on distingue deux approches différentes à la de la personnalité à savoir : les *théories des types* [38] et les *théories des traits* [4]. La théorie des types fait référence à une classification psychologique qualitative des différents types d'individus, alors que la théorie des traits fait allusion plutôt à une classification de manière quantitative des individus. Par exemple, si on prend deux caractères psychologiques en particulier, l'introversion et l'extraversion, pendant que la théorie des types cherche à classer chaque individu dans l'une de ces deux dimensions, la théorie des traits, quant à elle, essaie plutôt d'évaluer chaque individu suivant chacune de ces dimensions séparément, en chiffrant le degré selon lequel les individus sont introvertis d'une part, et extravertis, d'autre part. Le fait d'avoir à sa disposition les différents états psychologiques (personnalité) d'un individu est une information importante et complémentaire pour les applications cherchant à caractériser leurs utilisateurs telles que les systèmes de recommandation.

Certains travaux [71, 69] se sont focalisés sur la relation qui peut exister entre la musique et la personnalité. Dans [71] les auteurs ont exploré comment les préférences musicales sont liées au cinq traits de personnalité du modèle Big5 [26]. Ce modèle définit cinq dimensions psychologiques appelé OCEAN : —*Ouverture*, *Conscienciosité*, *Extraversion*, *Agréabilité* et *Neuroticisme*. Ces travaux ont montré que les musiques de type "réflexive, complexe, intense et rebelle" sont liées à l'ouverture à une nouvelle expérience, et que le type de musique "conventionnelle" est positivement corrélé avec les traits *extraversion*, *agreeableness*, et *conscientiousness*. Ils ont observé que les facteurs d'extraversion et d'ouverture sont les seuls qui expliquent la variance dans les préférences musicales. Dans [70, 12] les auteurs ont étendu les domaines de préférences des utilisateurs aux livres, magazines, films et émissions télévisées. Les auteurs dans [60] ont particulièrement étudié la relation entre les traits de personnalité et les émotions induites dans les films dans différents contextes sociaux.

Dans ces travaux, la découverte de la personnalité des utilisateurs nécessite de répondre à des questionnaires ou des tests de personnalité. Généralement, ces tests de personnalité reposent sur des réponses qui sont au préalable catégorisées : *d'accord*, *neutre* ou *pas d'accord* et parfois accompagnées du degré d'accord ou de désaccord : *fortement*, *moyennement* ou *faiblement d'accord* ou *pas d'accord*. Ce type de test requiert que les utilisateurs se livrent et révèlent leurs véritables informations, ce qui n'est pas toujours le cas.

Comme évoqué précédemment, notre travail de thèse vise à exploiter les ressources de plus

[1] https://www.fevad.com/

en plus nombreuses que les utilisateurs publient et partagent dans les réseaux sociaux. Exploiter les informations sur les activités favorites, les préférences et les intérêts des utilisateurs constitue une véritable mine d'or pour cerner leur personnalité. Dans le chapitre 2, nous avons proposé une approche automatique et multilingue FRISK de découverte des intérêts d'un utilisateur à partir de ses tweets dans le but de découvrir ses intérêts. Dans ce chapitre, nous présentons une méthodologie appelée ASCERTAIN[2] dont le but est de déterminer la corrélation entre les traits de personnalité et les intérêts découverts par FRISK. Nous nous intéressons essentiellement à la personnalité dans le sens de la théorie des traits puisqu'elle est assez représentée dans la littérature à travers les cinq traits centraux du modèle Big5. Nous exploitons l'outil Receptiviti qui est une extension de LIWC (Linguistic Inquiry and Word Count) afin d'extraire les traits de personnalité (appelés dimensions) à partir des tweets des utilisateurs dont les intérêts sont déjà déterminés par FRISK. Le but est de construire un modèle de prédiction capable de déterminer les traits de personnalité qui caractérisent le plus un intérêt.

Plus précisément, la méthodologie ASCERTAIN que nous proposons se décline en plusieurs étapes :

- Collecte des données qui consiste d'une part, à extraire les tweets des utilisateurs pour chaque intérêt cible, et d'autre part, à découvrir leurs traits de personnalité à l'aide de Receptiviti.
- Exploration des données qui consiste à appliquer une analyse en composantes principales sur les données collectées selon les dimensions Receptiviti. Le but de cette étape est d'identifier les traits de personnalité définis dans Receptiviti liés entre eux.
- Sélection des variables explicatives et construction du modèle de prédiction (régression logistique) afin d'identifier les dimensions Receptiviti les plus discriminantes par rapport aux intérêts cibles.
- Évaluation et interprétation des résultats.

De nombreuses analyses ont été effectuées pour chacune des étapes sur différentes collections construites à partir de 647 690 tweets appartenant à 1 446 utilisateurs du réseau social TWITTER catégorisés selon six intérêts : *Politique*, *Économie*, *Jeux vidéo*, *Gastronomie*, *Sports* et *Tourisme*. Nous avons également comparé les résultats de plusieurs modèles de prédiction.

Le plan de ce chapitre se présente comme suit. Tout d'abord, nous présentons dans la section 1, les travaux existants portant sur la détermination de la personnalité des individus. Puis, dans la section 2, nous décrivons les différents concepts existants manipulés, en particulier le modèle des Big Five, les outils LIWC et Receptiviti. Dans la section 3, nous exposons respectivement notre méthodologie ASCERTAIN et nos expérimentations faites qui consistent à appliquer la méthode décrite sur un jeu de données réelles. Et enfin, nous terminons par une conclusion.

1 État de l'art

Le problème de détermination de la personnalité des individus a fait et continue de faire l'objet de plusieurs travaux de recherche. Le challenge ici est d'analyser les ressources produites par un individu sur le Web social afin de découvrir ses traits de personnalités. A cet effet, deux directions d'études ont vu le jour. La première [24, 73, 40, 62, 2] est essentiellement basée sur les cinq traits de personnalité du modèle Big5, des caractéristiques LIWC [79] ainsi que certaines propriétés du Web social. LIWC est une application dont le but est de caractériser un document texte selon 80 valeurs

[2] AnalySis Correlation pERsonality Traits And INterests

d'attributs bien précis (*Family*, *Friends*, *Money*, etc). L'idée générale de cette première direction est d'étudier la corrélation existante entre les traits de personnalité du modèle Big5 et un groupe de variables constitué des attributs LIWC et/ou certaines propriétés du Web social. Par exemple, les utilisateurs des réseaux sociaux ayant beaucoup d'amis ont tendance à être extravertis.

Les auteurs de [24] cherchent à prédire la personnalité des utilisateurs à partir des ressources disponibles sur leur profil TWITTER. A cet effet, ils modélisent un utilisateur par un ensemble de valeurs d'attributs obtenues à partir de l'analyse de ses tweets. Plus clairement, les attributs mentionnés sont les 80 attributs LIWC, 14 attributs d'un dictionnaire nommé MRC qui contient 150 000 mots classés par catégories psychologiques et calcule le pourcentage de mots du document de l'utilisateur appartenant aux différents catégories, 9 attributs TWITTER (followers / following, les hastags, les retweets, etc.) et enfin les scores issus de l'analyse de sentiments du document de l'utilisateur qui est une moyenne du nombre de mots employés qui ont une connotation positive ou négative. L'analyse des sentiments s'inspire d'une base de données nommée "General Inquirer dataset" qui contient un ensemble de mots classés positivement ou négativement.

Par ailleurs, les auteurs de [73] proposent également deux approches de détermination de la personnalité des utilisateurs de FACEBOOK : l'une appelée approche ouverte et l'autre approche fermée. La première approche utilise un vocabulaire fermé, c'est-à-dire un dictionnaire de catégories pour extraire du document de tweets d'un utilisateur le pourcentage de mots appartenant aux différentes catégories définies. A ces pourcentages, ils ajoutent deux autres attributs, l'âge et le sexe de l'utilisateur avant d'exploiter la méthode des moindres carrés pour déterminer les corrélations existantes entre les différents attributs sélectionnés et les traits de personnalité du modèle Big5. Quant à l'approche ouverte, elle utilise également la méthode des moindres carrés mais, par contre, sur différents attributs. Les attributs considérés ici sont les mots, les phrases et les sujets contenus dans le document de l'utilisateur. Plus clairement, les sujets sont découverts par le biais du modèle probabiliste LDA (allocation de dirichlet latente) [7].

Les auteurs de [40], quant à eux, analysent les blogs des utilisateurs en vue d'extraire leurs traits de personnalité. Pour le faire, ils utilisent l'analyse en composantes factorielles pour réduire les valeurs des dimensions LIWC issues du document d'un utilisateur en cinq composantes principales. Par la suite, ils sélectionnent 12 dimensions psychologiques (nerveux, sociable, travailleur, etc.) à partir desquelles ils regroupent les utilisateurs. Et enfin, ils caractérisent chaque communauté d'utilisateurs à travers les dimensions sélectionnées.

D'un autre côté, les auteurs de [62] proposent une méthode collaborative pour découvrir la personnalité des utilisateurs. Étant données les préférences d'un utilisateur, les auteurs cherchent à calculer la probabilité qu'il ait les mêmes préférences que d'autres utilisateurs connus. Et par la suite, ils infèrent la personnalité de l'utilisateur cible à partir de celles de ses semblables. En plus d'inférer la personnalité, ils font également de la recommandation. Plus précisément, un utilisateur est représenté par un vecteur contenant de valeurs booléennes", qui représentent respectivement les items préférés ou ignorés (ou pas connus) par l'utilisateur. Le vecteur de score d'un utilisateur cible est comparé à ceux de plusieurs autres utilisateurs. A la fin, ceux dont leurs vecteurs sont sémantiquement proches de celui de l'utilisateur cible sont considérés comme des individus ayant la même personnalité et envie que l'utilisateur cible.

Les auteurs de [2] étudient les corrélations existantes entre les attributs extraits du profil FACEBOOK des utilisateurs et leurs personnalités. Les attributs FACEBOOK considérés sont, entre autres, le nombre d'amis, de groupes d'inscrits, le nombre de mises à jour de photos de profils, etc. Dans un premier temps, ils demandent aux utilisateurs de remplir un questionnaire de découverte de personnalité des individus. Le contenu textuel ainsi extrait des profils des utilisateurs est utilisé

pour calculer les valeurs LIWC. Les utilisateurs sont ensuite classés selon 10 groupes sur la base de chaque attribut. Chaque groupe est représenté par la moyenne des valeurs des différents attributs et comparé aux valeurs obtenues lors du questionnaire.

La seconde direction de travaux [58, 57, 68] qui a vu le jour tout récemment (plus précisément fin 2016) consiste à faire une étude sur la corrélation existant entre les dimensions psychologiques Receptiviti et les intérêts des individus. Receptiviti est une version commerciale de LIWC focalisée plutôt sur un ensemble de 59 dimensions décrivant les différents états psychologiques que peut avoir une personne à savoir *Artistic*, *Intellectual*, *Adventurous*, etc. Nous notons que Receptiviti inclut également les cinq traits du modèle Big5.

Les auteurs de [58] cherchent à savoir si la personnalité des entrepreneurs et des gestionnaires présente des différences systématiques. Par exemple, ils s'attendent à ce que les entrepreneurs soient plus audacieux voire aventureux dans leur personnalité que les gestionnaires. Ils ont utilisé une régression logistique simple pour prédire la probabilité qu'un utilisateur soit entrepreneur ou gestionnaire. Les variables exploitées sont un sous-ensemble des dimensions psychologiques Receptiviti et des attributs obtenus de Twitter tels que l'âge, le sexe, le nombre de followers et de tweets d'un utilisateur. L'évaluation de leur méthode est faite sur un jeu de données constitué de 106 individus (57 entrepreneurs et 49 gestionnaires) issus des données publiques portant sur des hautes personnalités aux États-Unis.

Par ailleurs, les auteurs de [57] présentent une analyse empirique des caractéristiques de personnalité d'un leader politique contemporain qui se considère plus comme un homme d'affaires qu'un politicien : Donald J. Trump. Ils utilisent le même jeu de données publiques des personnalités aux États-Unis pour comparer les traits de personnalité de Trump à ceux qui sont entrepreneurs et gestionnaires. Ils sélectionnent certaines dimensions psychologiques Receptiviti et calculent le score des utilisateurs entrepreneurs, gestionnaires ainsi que ceux de Trump et comparent ensuite les valeurs des entrepreneurs et gestionnaires à ceux de Trump. Dans ces travaux, les intérêts sont très spécifiques et liés à la notion d'entrepreneuriat.

De même, les auteurs de [68] utilisent les tweets des utilisateurs pour déterminer leurs traits de personnalité. A cet effet, ils font une analyse de la corrélation entre les tweets des utilisateurs (représentés par les dimensions LIWC et Receptiviti) et quatre types de personnalités (*mind*, *energie*, *nature* et *tactics*) qu'ils ont eux même définis. L'évaluation a été faite sur 450 profils Twitter.

Dans notre travail, nous proposons une étude qui exploite les 59 dimensions de Receptiviti et pas un sous-ensemble comme dans les méthodes précédentes. Le but est de déterminer la corrélation avec des intérêts qui sont plus généraux. De plus, pour notre évaluation, nous avons exploité un jeu de données nettement plus grand que ceux des études sus-citées construit à partir de 1 446 utilisateurs.

2 Préliminaires

Les résultats obtenus dans les travaux [29, 53, 28, 32] ont montré que la personnalité des individus est corrélée aux mots qu'ils utilisent. LIWC et Receptiviti sont ainsi développées afin de fournir une application permettant d'étudier les facteurs structurels et/ou psychologiques présents dans les discours des individus. Dans cette section nous présentons tout d'abord le modèle des Big5 [26], qui est un repère pour la description et l'étude théorique de cinq dimensions psychologiques des individus. Par la suite, nous poursuivons par une description des concepts sur lesquels nous nous sommes appuyés pour décrire notre méthodologie Ascertain. Nous décrivons les outils utilisés

pour analyser les ressources textuelles à savoir : Linguistic Inquiry and Word Count (LIWC) et Receptiviti.

2.1 Modèle des Big5

C'est un modèle qui tire son origine de la recherche en psychologie et qui aboutit à l'idée selon laquelle les traits de personnalité les plus communs peuvent être répertoriés principalement sous cinq dimensions. Le modèle Big5, ou tout simplement Big5, représente en psychologie cinq traits centraux de personnalité proposés par Goldberg et développés par Costa et McCrae [26]. Généralement, dans la vie, les individus réagissent différemment face à une même situation en raison notamment de leur personnalité. En effet, chaque individu possède naturellement une combinaison de caractéristiques émotionnelles, d'attitudes et de comportements, qui influencent ses réactions au quotidien. A cet effet, dans la psychologie contemporaine, plusieurs équipes indépendantes de chercheurs ont découvert, à partir des études empiriques [81, 18, 27] fondées sur des données, qu'il existe cinq grands facteurs de personnalité. C'est ainsi qu'ont émergé après observation les cinq dimensions assez larges du modèle Big5. Plus encore, les cinq dimensions du modèle Big5 sont considérées comme des traits sous-jacents qui composent la personnalité globale d'un individu. Ces dimensions sont représentées à partir d'un modèle qu'on appelle communément le modèle *OCEAN* défini comme suit :

1. **(O)** *Ouverture* : mesure le degré auquel une personne est ouverte à de nouvelles idées et de nouvelles expériences. Par exemple, l'appréciation de l'art, de l'émotion, de l'aventure, des idées peu communes, la curiosité et l'imagination. Traditionnellement, ce sont les personnes qui aiment apprendre de nouvelles choses et ont souvent plusieurs centres d'intérêts.

2. **(C)** *Conscienciosité* : mesure le degré auquel une personne est fiable. Notamment, l'autodiscipline, le respect des obligations, l'organisation et l'orientation vers des buts. Généralement, ce sont des personnes qui ont un haut degré de conscience professionnelle, par conséquent sont fiables et ponctuelles.

3. **(E)** *Extraversion* : mesure le degré auquel une personne a l'amabilité à interagir avec d'autres personnes ou à s'engager dans des activités. Les expressions qui caractérisent les individus de cette dimension sont entre autres : énergie, émotions positives, tendance à chercher la stimulation et la compagnie des autres, fonceur, dynamique, assertif (capacité à s'exprimer et à défendre ses droits sans empiéter sur ceux des autres).

4. **(A)** *Agréabilité* : mesure le degré auquel une personne est prédisposée à satisfaire les autres. Autrement dit, ce sont ceux qui ont tendance à être compatissants et coopératifs plutôt que soupçonneux ou antagoniques envers les autres, telles que les personnes gentilles, affectueuses, sympathiques, amicales, coopérantes et douées de compassion.

5. **(N)** *Neuroticisme* ou *névrosisme* : mesure le degré auquel une personne exprime de fortes émotions négatives. En d'autres termes, cette dimension désigne chez une personne une tendance à éprouver facilement des émotions désagréables comme la colère, la vulnérabilité, l'inquiétude ou la dépression. Ce sont des personnes qui sont souvent confrontées à une instabilité émotionnelle, ainsi qu'une humeur changeante.

Globalement, le modèle Big5 ne classe pas les individus selon cinq dimensions psychologiques, mais il les évaluent plutôt indépendamment plusieurs (plus précisément cinq) fois différemment sur chacune de ces dimensions. Par exemple, l'ouverture à l'expérience d'une personne ne présume en

rien sa stabilité émotionnelle ou sa nervosité. De même, le degré d'extraversion d'un individu n'est lié ni à sa gentillesse ni à son caractère consciencieux. Ainsi, d'après le modèle OCEAN on connaît une personne si et seulement si on est capable de la jauger sur chacun de ces cinq aspects ou dimensions.

Par ailleurs, certaines applications, en particulier Receptiviti, que nous présentons plus loin, permettent à partir d'un document de texte de calculer un score ou plus précisément une valeur numérique associée respectivement à un ensemble d'attributs prédéfinis parmi lesquels on retrouve les dimensions psychologiques du modèle Big5. En effet, chaque score indique le degré selon lequel le document de texte en entrée de ces applications exprime une dimension ou un attribut bien précis.

2.2 Linguistic Inquiry and Word Count

LIWC[3] est une application permettant d'étudier les différents facteurs émotionnels, cognitifs et structurels présents dans un document de texte. Ces différents facteurs sont regroupés suivant cinq grandes catégories qui par la suite sont décrites par le biais de plusieurs dimensions. Au total LIWC compte 92 dimensions dans sa dernière version. La table 3.1 détaille les différentes catégories LIWC et leurs dimensions respectives. On constate que les catégories ne sont pas équilibrées puisque certaines d'entre elles contiennent plus de dimensions que d'autres.

Catégories	Attributs ou dimensions
Summary Variables	Analytical Thinking, Clout, Authentic, Emotional Tone
Language Metrics	Words per sentence, Words>6 letters, Dictionary words, Function Words, Total pronouns, Personal pronouns, 1st pers singular, 1st pers plural, 2nd person, 3rd pers singular, 3rd pers plural, Impersonal pronouns, Articles, Prepositions, Auxiliary verbs, Common adverbs, Conjunctions, Negations
Grammar Other	Regular verbs, Adjectives, Comparatives, Interrogatives, Numbers, Quantifiers, Affect Words, Positive emotion, Negative emotion, Anxiety, Anger, Sadness, Social Words, Family, Friends, Female referents, Male referents, Cognitive Processes, Insight, Cause, Discrepancies, Tentativeness, Certainty, Differentiation, Perpetual Processes, Seeing, Hearing, Feeling, Biological Processes, Body, Health/illness, Sexuality, Ingesting, Core Drives and Needs, Affiliation, Achievement, Power, Reward focus, Risk/prevention focus
Time Orientation	Past focus, Present focus, Future focus, Relativity, Motion, Space, Time
Personal Concerns	Work, Leisure, Home, Money, Religion, Death, Informal Speech, Swear words, Netspeak, Assent, Nonfluencies, Fillers, All Punctuation, Periods, Commas, Colons, Semicolons, Question marks, Exclamation marks, Dashes, Quotation marks, Apostrophes, Parentheses (pairs), Other punctuation

Table 3.1: Catégories LIWC et variables associées.

[3] https://liwc.wpengine.com/

Globalement, l'application LIWC prend en entrée un document de texte, ensuite utilise ses dictionnaires pour compter le pourcentage de mots qui appartient à chaque dimension de ses différentes catégories prédéfinies. La table 3.2 est un extrait de la sortie LIWC plus précisément les dimensions de la catégorie "Time Orientation" à partir d'un sous ensemble de tweets d'un utilisateur du réseau social TWITTER. On observe un score plus élevé pour la dimension "Relativity" par rapport aux autres dimensions.

Dimensions	Scores
Past focus	0.02
Present focus	0.09
Future focus	0.01
Relativity	0.14
Motion	0.02
Space	0.06
Time	0.06

Table 3.2: Dimension de la catégorie "Time Orientation" sur document de texte

Au début de la création du programme LIWC dans les années 90, ses fondateurs ont décidé de travailler sur des théories dominantes portant sur la psychologie, les affaires, la médecine et ceux du bon sens. Au fil des années, ils ont modifié les théories de base, car ils se sont rendus compte que certaines d'entre elles sont étroitement liées à la langue plus que d'autres. A cet effet, ils se sont concentrés essentiellement sur des catégories portant sur des états sociaux et psychologiques, qu'ils ont élargi à chaque nouvelle version.

La première version de LIWC a été développée dans le cadre d'une étude exploratoire de la langue et de la divulgation. Cette version est essentiellement basée sur un ensemble de dictionnaires constitués de plusieurs mots regroupés par catégories/dimensions et qui sont construits par ses créateurs. Cependant, les versions suivantes, notamment la deuxième LIWC2001, la troisième LIWC2007 et la plus récente LIWC2015 ont successivement mis à jour l'application originale, d'une part, en étendant principalement ses dictionnaires, et d'autre part, en développant un logiciel beaucoup plus moderne par l'ajout des options et des catégories liées à la personnalité. Par conséquent, l'outil LIWC2015 est devenu plus précis grâce à une amélioration des calculs des scores de chaque dimension des catégories. De même, il est devenu plus facile à utiliser tout en fournissant une gamme large de catégories sociales et psychologiques par rapport aux versions antérieures.

Concrètement, l'application LIWC est constituée de deux modules basés principalement sur un ensemble de dictionnaires intégrés : le module d'*analyse de texte* et le module de *traitement de texte*. Le module d'analyse de texte identifie et classe les mots du document de texte fournis en entrée dans leurs dimensions respectives. Par exemple, le mot *pleuré* peut appartenir à cinq dimensions distinctes à savoir : la *tristesse*, l'*émotion négative*, l'*affection globale*, le *verbe* et le *passé*. Par contre, le module de traitement de texte compte le nombre de mots appartenant à chaque dimension. Autrement dit, il calcule le pourcentage de mots total qui correspond à chacune des dimensions prédéfinies.

En effet, pour chaque dimension, les auteurs de LIWC ont crée un dictionnaire composé de mots corrélés c'est-à-dire caractérisant la dimension. Leur idée vient du fait que si on veut par

exemple mesurer le degré avec lequel un texte révèle des intérêts liés au *pouvoir*, au *statut* ou à la *domination*, alors LIWC considère que quelqu'un qui s'intéresse au pouvoir est plus susceptible d'évaluer les autres en termes de son statut relatif. Autrement dit, une personne attachée au pouvoir est plus susceptible d'utiliser des mots tels que *le patron, le sous-officier, le président, le directeur, forts, les pauvres* par rapport à quelqu'un qui ne se soucie pas du pouvoir et du statut. Ainsi, la partie la plus difficile ici pour eux est la construction pour chaque dimension d'un dictionnaire qui contient les mots sémantiquement proches à la définition des dimensions. A cet effet, ils se sont penchés sur une construction basée exclusivement sur un jugement humain. Tout d'abord, ils ont commencé par les mots appartenant aux dictionnaires standards et des thésaurus tels que "Roget", à partir desquels chaque membre de leur équipe a rajouté d'autres mots en s'appuyant sur leur connaissance. Par la suite, un nouveau groupe de juges a évalué les mots identifiés et a donné leurs avis par rapport à leurs appartenance aux différentes catégories cibles. Et enfin, tous les mots ayant reçu de tous les juges des avis favorables restent dans leur catégorie d'origine, par contre les autres sont supprimés.

Comme tous les autres outils d'analyse de texte, LIWC est relativement bruité. Parfois, il fait des erreurs lors de l'identification et le comptage de mots. Prenons par exemple le mot en anglais *mad*, il peut être classé dans les dictionnaires *Anger*, *Negative emotion* et dans d'autres dictionnaires qui indiquent l'émotion. Or nous savons que habituellement, le mot "mad" reflète de la colère, mais il peut aussi exprimer de la joie extrême (*he's mad for her*) et l'instabilité mentale (*mad as a hatter*). Pour palier à ce type de problème, LIWC compte sur la longueur des textes en entrée car plus le texte est long plus le bruit devient négligeable. Aussi, un texte de 10 000 mots donne des résultats beaucoup plus fiables qu'un de 100 mots. Par contre, la sortie LIWC obtenue à partir d'un texte de moins de 50 mots doit être analysé avec un esprit critique.

2.3 Receptiviti

Receptiviti[4] est une application semblable à LIWC qui, par contre, met l'accent sur les catégories ayant un lien avec la personnalité. De plus, elle a été lancée en collaboration avec LIWC2015. En réalité, Receptiviti est le côté commercial de LIWC, pour le rendre plus accessible aux communautés de développement de logiciels et de science de données, notamment aux entreprises qui souhaitent utiliser la personnalité dans leurs services. L'API Receptiviti prend en entrée un document de texte et renvoie en sortie trois types distincts d'analyse : Sortie LIWC, Receptiviti et LSM.

1. **Sortie LIWC.** Ce sont les attributs ou dimensions LIWC restructurées dans de nouvelles catégories comme le montre le tableau 3.3. Certaines dimensions telles que *Cognitive Processes*, *Insight*, *Certainty* et bien d'autres sont passées de la catégorie "Grammar Other" dans l'application LIWC à la catégorie "Psychological Processes" dans Receptiviti. Cependant, d'autres dimensions comme *Dashes*, *Quotation marks*, *Apostrophes* et bien d'autres également ont été supprimées de la sortie LIWC dans Receptiviti. Pour plus de détails sur toutes les catégories et dimensions LIWC, voir en annexe la section 1.
2. **Sortie Receptiviti.** Elle représente six catégories dans lesquelles sont listées plusieurs dimensions comme le montre la figure 3.4. Nous remarquons que la catégorie "Big 5 Insights" contient les dimensions psychologiques du modèle Big5, qui sont caractérisées plus finement à l'aide d'autres dimensions sous-jacentes indiquées entre parenthèses. Notamment, les dimensions sous-jacentes de "Openness" sont : *Artistic, Intellectual, Liberal, Imaginative, Emotionally Aware, Adventurous.* Celles de "Conscientiousness" sont: *Self-assured, Disciplined,*

[4] https://www.receptiviti.ai/

Catégories	Attributs ou dimensions
Summary Language Variables	Analytical thinking, Clout, Authentic, Emotional tone (positive/negative), Words/sentence, Words > 6 letters, Dictionary words
Linguistic Dimensions	Personal pronouns, 1st pers singular, 1st pers plural, 2nd person, 3rd pers singular, 3rd pers plural, Impersonal pronouns, Articles, Prepositions, Auxiliary verbs, Common Adverbs, Conjunctions, Negations
Time Orientation	Past focus, Present focus, Future focus, Relativity, Motion, Space, Time
Other Grammar	Common verbs, Common adjectives, Comparisons, Interrogatives, Numbers, Quantifiers
Psychological Processes	Affective processes, Positive emotion, Negative emotion, Anxiety, Anger, Sadness, Social processes, Family, Friends, Female references, Male references, Cognitive processes, Insight, Causation, Discrepancy, Tentative, Certainty, Differentiation, Perceptual processes, See, Hear, Feel, Biological processes, Body, Health, Sexual, Ingestion, Drives, Affiliation, Achievement, Power, Reward, Risk, Time orientations, Past focus, Present focus, Future focus, Relativity, Motion, Space, Time
Personal concerns	Work, Leisure, Home, Money, Religion, Death, Informal language, Swear words, Netspeak, Assent, Nonfluencies, Fillers

Table 3.3: Sortie LIWC dans Receptiviti.

Ambitious, Dutiful, Cautious, Organized. Ainsi de suite jusqu'à "Neuroticism" dont les dimensions sous-jacentes sont : *Impulsive, Stressed, Anxious, Aggressive, Melancholy, Self-conscious.* Pour plus de détails sur toutes les catégories et dimensions Receptiviti voir en annexe la section 2.

3. **Sortie LSM.** Cette partie permet de comparer des paires d'individus selon chaque dimension LIWC et Receptiviti. En fait, son but est de déterminer le degré avec lequel les gens affichent des taux d'utilisation de mots similaires lorsqu'ils communiquent. Ce taux permet d'identifier les points communs et divergents entre deux individus et peuvent être utilisés pour comprendre dans quelle mesure ces personnes sont synchronisées.

Intuitivement, comme dans le cas de l'application LIWC, plus on analyse de mots plus les différents résultats obtenus (sortie Receptiviti) sont dignes de confiance. A cet effet, Receptiviti recommande un minimum de 300 mots pour avoir une sortie crédible. Tandis que pour les scores LIWC générés par l'API Receptiviti, il est préférable d'avoir au minimum 50 mots.

Dans la section suivante, nous présentons en détail notre méthodologie Ascertain.

Catégories	Attributs ou dimensions
Cognitive/Thinking Style Insights	Thinking Style, Persuasive, Reward Bias
Big 5 Insights	***Openness***, (Artistic, Intellectual, Liberal, Imaginative, Emotionally Aware, Adventurous), ***Conscientiousness***, (Self-assured, Disciplined, Ambitious, Dutiful, Cautious, Organized), ***Extraversion***, (Sociable, Friendly, Assertive, Energetic, Cheerful, Active), ***Agreeableness***, (Generous, Trusting, Cooperative, Empathetic, Genuine, Humble), ***Neuroticism***, (Impulsive, Stressed, Anxious, Aggressive, Melancholy, Self-conscious)
Social Style Insights	Social Skills, Insecure, Cold, Family Orientation
Emotional Style Insights	Adjustment, Happiness, Depression
Working Style Insights	Independent, Power Driven, Type-A, Workhorse
Interests and Orientations	Friendship Focused, Body Focus, Health Oriented, Sexual Focus, Food Focus, Leisure Oriented, Money Oriented, Religion Oriented, Work Oriented, Netspeak

Table 3.4: Aperçu des dimensions Receptiviti.

3 Ascertain: AnalySis Correlation pERsonality Traits And INterests

Dans cette section nous essayons par le biais de notre méthodologie Ascertain de répondre à l'hypothèse selon laquelle il existe une relation entre les intérêts et les traits de personnalité d'un individu. L'objectif général de notre travail est de construire pour chaque utilisateur un profil élargi exploitable notamment par les systèmes de recommandation. L'étude que nous proposons dans ce chapitre va nous permettre d'enrichir davantage la caractérisation des utilisateurs.

La figure 3.1 décrit brièvement notre méthodologie Ascertain dont le but est de déterminer s'il existe une corrélation entre les dimensions Receptiviti extraites à partir des tweets des utilisateurs et leurs intérêts. Des profils utilisateurs on extrait leurs tweets, qu'on représente sous forme de dimensions Receptiviti. Par ailleurs, à l'aide de l'approche proposée dans le chapitre précédent Frisk, on découvre les intérêts des utilisateurs. Ainsi, à partir de ces deux informations construites (dimensions Receptiviti et intérêts de chaque utilisateur) nous faisons notre analyse.

Nous exploitons la régression logistique, une méthode statistique très utilisée pour ce type de problème et qui permet d'analyser les relations existantes entre une variable ou qualité et une ou plusieurs autres variables (explicatives).

Dans ce qui suit, nous présentons la régression logistique, ensuite nous détaillons la méthodologie proposée Ascertain, ainsi que les expérimentations et les analyses faites sur un jeu de données réelles.

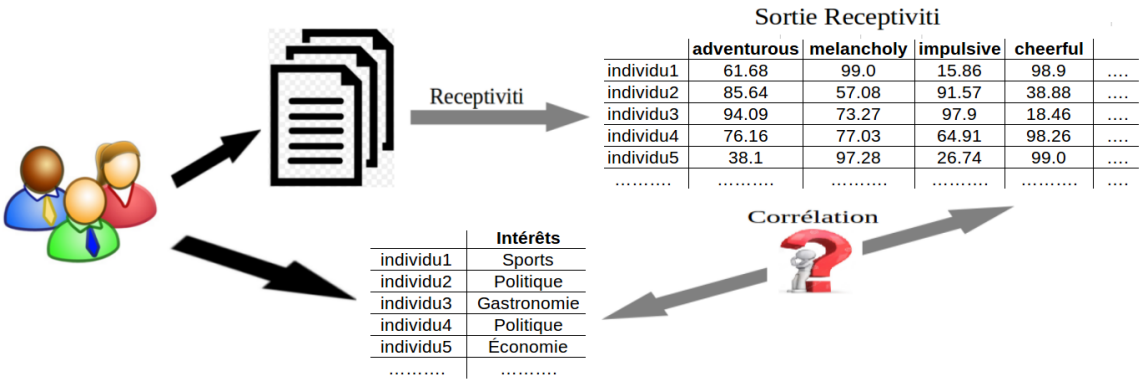


Figure 3.1: Schéma d'ASCERTAIN

3.1 Régression logistique

La régression logistique est un modèle de régression multivariables c'est-à-dire impliquant deux ou plusieurs variables, couramment utilisé dans la médecine, plus précisément en épidémiologie lors de l'étude des facteurs influant sur la santé et les maladies des populations. Elle s'utilise pour déterminer la corrélation existante entre une ou plusieurs variables explicatives quantitatives ou qualitatives notées X_i et une variable à expliquer qualitative notée Y. Pour cela, elle construit un modèle qui représente au mieux les variables impliquées dans l'analyse et qui peut être utilisé pour prédire les valeurs de Y à partir des valeurs des X_i.

Une variable est dite *quantitative* si les valeurs qu'elle peut prendre sont des nombres ou valeurs numériques tels que 10, 5, 99, etc. Par contre, une variable est dite *qualitative*, lorsque sa valeur ne représente pas une quantité, notamment, les jours de la semaine, présence ou absence d'une anomalie, etc. Dans notre cas d'étude, les variables explicatives X_i sont quantitatives (valeurs des dimensions Receptiviti) et la variable à expliquer qualitative (intérêt des utilisateurs). Aussi, le but d'ASCERTAIN est de répondre aux types de questions telles que: est-ce que ceux qui s'intéressent à la politique ou à l'économie sont fiables et nerveux. En fait, ces deux caractéristiques (fiables et nerveux) correspondent respectivement aux dimensions psychologiques "Conscientiousness" et "Neuroticism" du modèle Big5. D'où l'utilisation d'une régression logistique car elle convient à ce type d'analyse.

Cependant, il existe plusieurs types de régression logistique en fonction de la taille des variables explicatives X_i et de la modalité de la variable à expliquer Y. La modalité d'une variable est l'ensemble des valeurs distinctes que peut prendre cette variable. Le tableau 3.5 est un récapitulatif des différentes formes de régression logistique qu'on peut avoir en fonction du type et du nombre des variables.

Plus formellement, la régression logistique consiste à modéliser la probabilité que la variable Y se réalise étant donné X_i. Dans le cas de la régression logistique polytomique simple ou multiple elle se définit comme suit:

$$ln(\frac{P(Y=k/X)}{P(Y=K/X)}) = \beta_{0,k} + \beta_{1,k} \times X_1 + \ldots + \beta_{i,k} \times X_i + \varepsilon \tag{3.1}$$

	Modalité de $Y = 2$	Modalité de $Y > 2$
$\|X_i\| = 1$	simple ou binaire	polytomique simple ou multinomiale
$\|X_i\| \geq 2$	multiple	polytomique multiple

Table 3.5: Différentes formes de régression logistique.
$|X_i|$ représente le nombre de variables explicatives.

avec $k = 1, \ldots, K-1$ et K le nombre de valeurs que peut prendre Y, des variables explicatives X_i rattachées à leur coefficient $\beta_{i,k}$ et un terme ε représentant le bruit.

Dans le cas d'une régression logistique simple ou multiple, l'équation 3.1 se réduit à une fonction variant de façon monotone entre 0 et 1 et se définit comme suit :

$$ln(\frac{p}{1-p}) = logit(p) = \beta_0 + \beta_1 \times X_1 + \ldots + \beta_i \times X_i + \varepsilon \tag{3.2}$$

où le logit de la probabilité p est la probabilité de réalisation de la variable à expliquer Y qui est exprimé en fonction d'un intercept (ou ordonnée à l'origine) β_0, des variables explicatives X_i rattachées à leurs coefficients β_i et un terme ε représentant le bruit.

L'interprétation des résultats d'un modèle de régression logistique doit être précédée d'une vérification qui consiste à voir si les résultats obtenus sont statistiquement significatifs. Autrement dit, il faut vérifier si le modèle obtenu à l'issue de la régression logistique est fiable, c'est-à-dire correspond à une représentation des données initiales qui minimise les pertes d'informations. Cette vérification varie en fonction du type de problème, soit on utilise un test statistique, soit on calcule les mesures de performances du modèle telles que la précision et le rappel (dans le cas d'une prédiction), soit on calcul la vraisemblance du modèle obtenu. Pour rappel, un test statistique ou test d'hypothèse est une démarche consistant à rejeter ou pas une hypothèse statistique en fonction d'un jeu de données. En effet, quelque soit le type de vérification employé, l'objectif est d'évaluer l'écart existant entre le modèle et les données initiales. Si l'écart est petit, alors le modèle est crédible car il représente bien les données passées en paramètres.

Comme notre objectif est de faire une analyse de la corrélation entre les traits de personnalité des individus et leurs intérêts, nous proposons de construire un modèle de prédiction que nous utilisons par la suite pour prédire les intérêts des utilisateurs. Le modèle est considéré comme fiable s'il est capable de bien prédire les intérêts des utilisateurs. Ainsi, pour valider notre modèle on va calculer ses mesures de performances (précision et rappel).

Dans la section suivante, nous présentons les différentes étapes de notre étude.

3.2 Méthodologie

Plaçons-nous dans un contexte où nous disposons d'une part, de plusieurs documents chacun édité par un individu distinct, et d'autre part, les intérêts dominants de chacun de ces individus. Dans un premier temps, on va modéliser chaque document comme un vecteur de score, où chaque valeur représente une dimension Receptiviti. Nous nous sommes essentiellement focalisés sur les dimensions Receptiviti car elles contiennent en plus des cinq traits de personnalité du modèle Big5, d'autres dimensions psychologiques principalement larges et explicites. Concrètement, la sortie de Receptiviti présente 59 dimensions, et par conséquent, notre vecteur possède également 59 valeurs. Par la suite,

nous utilisons les vecteurs de score des utilisateurs et leurs intérêts dominants pour faire notre analyse.

De manière générale, ASCERTAIN se base essentiellement sur la régression logistique dans laquelle les variables explicatives X_i sont les dimensions psychologiques Receptiviti et la variable à expliquer Y sont les intérêts des utilisateurs. Nous optons pour une démarche principalement incrémentale dont les étapes successives se présentent comme suit :

- Collecte des données,
- Exploration des données,
- Sélection des variables explicatives,
- Construction du modèle de prédiction,
- Évaluation et interprétation des résultats du modèle de prédiction.

3.2.1 Collecte des données

L'objectif de cette étape est de construire la base de données SELFDS qui va servir plus tard à la construction de notre modèle de prédiction. De façon détaillée, le processus de collecte des données SELFDS s'est fait en trois étapes : la phase de *sélection des utilisateurs*, pendant laquelle nous sélectionnons un sous-ensemble de profils utilisateurs du réseau social TWITTER. Ensuite, la phase de *collecte des tweets des utilisateurs* au cours de laquelle nous collectons les tweets récents et publics des utilisateurs sélectionnés. Et enfin la phase de *calcul des dimensions psychologiques Receptiviti* pendant laquelle nous générons à partir des tweets de chaque utilisateur les différentes valeurs des 59 dimensions psychologiques Receptiviti.

Sélection des utilisateurs. Pour identifier les profils utilisateurs de notre base de données SELFDS, nous avons utilisé le moteur de recherche Twitter, comme expliqué dans la section 4.2 du chapitre 2. Nous nous sommes concentrés essentiellement sur les profils utilisateurs édités en anglais et dont l'intérêt dominant appartient à l'un des six intérêts choisis suivant : *Politique*, *Économie*, *Jeu vidéo*, *Gastronomie*, *Sports* et *Tourisme*. Nous avons choisi la langue anglaise car c'est elle qui est la mieux représentée parmi les dictionnaires de Receptiviti. En fait, nous nous sommes inspirés de la démarche décrite au chapitre précédent lors de la construction du jeu de données MULTIDS. Plus clairement, nous utilisons les expressions représentant nos intérêts, ainsi que leurs synonymes respectifs comme mots clés dans le moteur de recherche TWITTER pour découvrir les utilisateurs ayant comme intérêt dominant l'expression passée en paramètre. Les mots clés utilisés pour faire la recherche de profils utilisateurs sont identiques à ceux mentionnés dans la section 4.2 du chapitre 2. Cependant, pour la nouvelle catégorie "Tourisme" nous nous sommes toujours inspirés de la hiérarchie de catégories indiquée sur Google AdWords[5] pour déterminer les expressions sémantiquement liées au mot "tourisme". Ainsi, les mots clés utilisés dans le moteur de recherche TWITTER pour identifier les utilisateurs qui s'intéressent au "Tourisme" sont : *tourism*, *travel*, et *museum*. Le tableau 3.6 présente la distribution des profils utilisateurs par intérêt de notre base de données SELFDS. Nous notons que nous avons fait une vérification manuelle des premiers tweets de chaque utilisateur en vue de confirmer que l'intérêt dominant proposé par le moteur de recherche TWITTER est conforme à la réalité. À l'issue de cette phase de vérification, les profils utilisateurs dont les intérêts dominants déduits à partir du moteur de recherche TWITTER et qui ne correspondent

[5] developers.google.com/adwords/api/docs/appendix/productsservices

pas à ceux que nous avons trouvés manuellement sont supprimés de notre base de données. Nous précisons que le tableau 3.6 présente la distribution des utilisateurs après suppression des profils mals classés.

	Politique	**Économie**	**Jeu vidéo**	**Gastronomie**	**Sports**	**Tourisme**	
Profils	250	202	235	219	286	254	**1 446**

Table 3.6: Distribution du nombre d'utilisateurs par intérêt dans SelfDS.

Collecte des tweets utilisateurs. Pour collecter les tweets des utilisateurs sélectionnés précédemment, nous avons utilisé l'API de recherche Twitter, comme expliqué dans la section 4.2 du chapitre 2. Plus précisément, nous avons extrait au plus 500 tweets publics et récents par utilisateur sélectionné. À la fin du processus de collecte de données, SelfDS contient environ six cent mille tweets, dont le nombre de tweets (#tweets), le nombre moyen de tweets (avg(#tweets)) et le nombre moyen de mots (avg(#mots)) par utilisateur et par intérêt sont indiqués dans le tableau 3.7. Nous observons que les utilisateurs de la catégorie "Sports" de notre jeu de données SelfDS publient beaucoup plus de contenu (tweets) que les utilisateurs des autres catégories. Par contre, nous constatons que les tweets des utilisateurs de la catégorie "Jeu vidéo" sont moins consistants, c'est-à-dire contiennent moins de mots.

Intérêts	**#tweets**	**avg(#tweets)**	**avg(#mots)**
Politique	111 382	445,5	3 203
Économie	83 023	411	2 898,2
Jeu vidéo	105 595	449,3	2 682,5
Gastronomie	95 974	438,2	2 798,5
Sports	136 094	475,9	3 020,5
Tourisme	115 622	455,2	3 000,5
Grand total	**647 690**	**2 675,21**	**17 603,1**

Table 3.7: Statistiques sur les 500 tweets utilisateurs de SelfDS.

Calcul des dimensions psychologiques Receptiviti. Cette phase consiste à calculer à partir du document de tweets de chaque utilisateur sélectionné un vecteur de score dont les valeurs correspondent aux dimensions psychologiques Receptiviti. Pour le faire, nous avons utilisé l'API Receptiviti qui est une application écrite en Java et qui fonctionne uniquement en présence d'une clé d'une durée de validité d'un mois. Cette clé est générée par les auteurs de l'API lors de la création d'un espace personnel sur leur site. En fait, nous avons paramétré leur code Java de sorte qu'il prend en entrée un ensemble de documents de tweets représentant chacun un utilisateur et nous renvoie

en sortie leurs vecteurs de scores respectifs.

En définitive, SelfDS contient 1 446 vecteurs de scores qui représentent l'ensemble des utilisateurs sélectionnés et 60 attributs qui représentent les 59 dimensions psychologiques de Receptiviti à laquelle on ajoute un attribut qui définit les intérêts des utilisateurs.

3.2.2 Exploration des données

Le but de cette étape consiste d'une part, à se familiariser avec les données, en s'assurant qu'elles ne contiennent pas de valeurs rares ou manquantes, et d'autre part, de mesurer les liaisons existantes entre les variables explicatives. Autrement dit, cette étape consiste principalement à déterminer les variables explicatives discriminantes ainsi que celles qui sont corrélées entre elles. Compte tenu du fait que nos variables explicatives X_i sont quantitatives, nous avons choisi d'utiliser l'analyse en composantes principales.

L'analyse en composantes principales est une technique d'analyse des données qui, à partir de i variables numériques analysées ou explicatives, permet de construire m ($\leq i$) autres variables, appelées composantes principales ou facteurs, qui sont une combinaison linéaire des variables explicatives analysées. Ces facteurs présentent d'intéressantes caractéristiques telles que:

- Les composantes principales sont ordonnées selon l'information qu'elles restituent, la première étant celle qui restitue le plus d'information.
- À partir de la première composante, on ajoute successivement de nouvelles composantes (celles qui restituent également le plus d'informations parmi celles restantes) jusqu'à ce que l'information initiale soit nettement représentée avec une perte d'information minimale.
- Les composantes principales sont des vecteurs indépendants, c'est-à-dire des variables non corrélées linéairement entre elles.
- On a une inégalité stricte $m < i$ s'il existe des relations linéaires entre les variables explicatives X_i.

Comme mentionné précédemment, la première opération consiste à fiabiliser notre base de données, en évitant d'une part, d'y mettre des données incorrectes, et d'autre part, d'avoir dans les vecteurs de scores des dimensions ayant des valeurs manquantes. Une donnée incorrecte fait référence à la présence d'une valeur erronée dans le vecteur de scores d'un utilisateur. Pour rendre fiables nos données, nous avons généré notre base de données de manière automatique à l'aide de l'API Receptiviti.

La deuxième est une opération de réduction du nombre de dimensions initiales, qui est de 59 variables explicatives dans notre cas. L'idée de cette réduction est d'éliminer certaines variables "redondantes" tout en minimisant les pertes d'informations de telle sorte qu'on puisse représenter nos données sur un repère à deux ou trois dimensions. En fait, ce type de repère est pratique pour la visualisation des données. La réduction du nombre de variables consiste : soit à ignorer certaines variables trop corrélées entre elles d'une part, et d'autre part, absolument non pertinentes par rapport à l'objectif à atteindre, soit à rassembler plusieurs variables en une seule ou soit à transformer les variables au moyen d'une analyse factorielle. En effet, l'analyse factorielle est une technique statistique utilisée pour décrire un ensemble de variables explicatives au moyen de variables latentes. L'opération de réduction est faite ici à travers une analyse en composantes principales sur notre jeu de données.

Dans l'optique d'identifier la contribution de chaque variable explicative X_i de SelfDS, nous avons fait un regroupement de ces X_i suivant trois aspects ou représentations à savoir:

1. **Représentation 1 :** nommée Big5D, c'est l'ensemble des utilisateurs ou plus précisément, des vecteurs de scores décrits à partir de cinq variables explicatives X_i qui correspondent aux cinq traits de personnalité du modèle Big5 c'est-à-dire Ouverture, Conscienciosité, Extraversion, Agréabilité et Neuroticisme.
2. **Représentation 2 :** nommée Big5SD, c'est également l'ensemble des vecteurs de scores décrits par 35 variables explicatives X_i. Plus spécifiquement, les variables explicatives de Big5SD sont en plus des cinq traits centraux du modèle Big5, leurs 30 variables sous-jacentes respectives indiquées dans le tableau 3.8.
3. **Représentation 3 :** nommée recepD, comme Big5D et Big5SD c'est aussi l'ensemble de vecteurs de scores mais par contre décrits suivant les 59 dimensions Receptiviti.

Dimension Big5	Dimensions sous-jacentes
Openness	Artistic, Intellectual, Liberal, Imaginative, Emotionally Aware, Adventurous
Conscientiousness	Self-assured, Disciplined, Ambitious, Dutiful, Cautious, Organized
Extraversion	Sociable, Friendly, Assertive, Energetic, Cheerful, Active
Agreeableness	Generous, Trusting, Cooperative, Empathetic, Genuine, Humble
Neuroticism	Impulsive, Stressed, Anxious, Aggressive, Melancholy, Self-conscious

Table 3.8: Dimensions sous-jacentes des traits du modèle Big5 dans Receptiviti.

En définitive, chaque représentation Big5D, Big5SD et recepD contient les 1 446 profils utilisateurs du jeu de données SelfDS, respectivement 5, 35 et 59 variables explicatives et chacune une variable à expliquer qui définit les intérêts des utilisateurs. C'est sur ces trois jeux de données qu'on va effectuer plus tard, dans la section évaluation, l'analyse en composantes principales.

3.2.3 Sélection des variables explicatives.

Cette étape consiste à identifier les variables explicatives fortement corrélées à la variable à expliquer. Les variables identifiées sont celles qui sont utilisées pour construire notre modèle de prédiction. Il existe généralement trois approches de sélection de variables qui sont utilisées la plupart du temps : la sélection ascendante, descendante ou mixte.

L'approche *ascendante* ou *forward selection* consiste à inclure progressivement les variables explicatives X_i à un modèle minimaliste, en laissant de côté celles qui n'apportent pas suffisamment d'information au modèle. Quant à l'approche *descendante* ou *backward elimination*, qui est la plus couramment utilisée, elle consiste à inclure au préalable toutes les variables explicatives dans le modèle et par la suite retirer progressivement celles qui ne participent pas suffisamment à la construction du modèle. Par contre, l'approche *mixte* est un mélange des deux approches sus-citées. Autrement dit, certaines variables déjà introduites dans le modèle finale peuvent y être supprimées. Le critère d'introduction ou de suppression d'une variable dans un modèle dépend fortement de plusieurs indications. Nous avons utilisé celui qui est le plus souvent utilisé à savoir le critère d'information d'Akaike (*Akaike Information Criteria* [1] ou AIC). De manière globale, le critère d'information

d'Akaike est une mesure de la qualité d'un modèle statistique basée essentiellement sur le maximum de vraisemblance. L'idée consiste à tester plusieurs modèles (avec de vraies valeurs) construits en retirant ou en regroupant à tour de rôle plusieurs variables. Cependant, il n'y a pas de règle unanime concernant le choix de l'approche de sélection de variables à utiliser.

Globalement, l'étape de sélection des variables se déroule en trois phases. Dans un premier temps, on construit les modèles de prédiction saturés à partir de toutes les variables explicatives de nos données BIG5D, BIG5SD et RECEPD. Comme chacune de nos trois données contient plus d'une variable explicative et que la modalité de notre variable à expliquer (intérêts des utilisateurs) dépasse largement deux valeurs distinctes, d'après la figure 3.5 nous utilisons une régression logistique polytomique multiple pour construire nos modèles de prédiction. Ensuite pour chaque modèle construit, on supprime successivement les X_i faiblement corrélées à la variable à expliquer Y (intérêts). A la fin de la suppression, les variables restantes sont celles que nous utilisons à l'étape suivante pour construire les modèles de prédiction finaux.

Construction du modèle saturé. Nous avons utilisé la fonction *multinom* de la librairie "nnet" avec le logiciel R[6] pour construire nos modèles de régression logistique polytomique. La syntaxe de cette fonction est la suivante :

$$modele \leftarrow multinom(Y \sim X_i,\ data) \tag{3.3}$$

où, la variable *modele* est le modèle ainsi construit à partir des données contenues dans la variable *data*. C'est ainsi que nous obtenons les modèles saturés nommés BIG5MS, BIG5SMS et RECEPMS à partir des données BIG5D, BIG5SD et RECEPD respectivement.

Choix des variables du modèle final. Pour le faire, nous avons utilisé la fonction *stepAIC* de la librairie "MASS" avec le logiciel R. Comme mentionné précédemment, nous avons choisi d'utiliser l'approche descendante et le critère AIC pour sélectionner nos variables d'où l'utilisation de la fonction *stepAIC* qui est défini comme suit:

$$selectVar \leftarrow stepAIC(modele, scope, data, direction) \tag{3.4}$$

où $selectVar$ est la variable qui contient les résultats de la sélection des variables sur *modele* qui représente le modèle initial pour lequel on cherche à minimiser le nombre de variables explicatives tout en réduisant les pertes d'informations. Dans notre cas, la variable *modele* est l'un des trois modèles saturés construits précédemment (BIG5MS, BIG5SMS et RECEPMS) et la variable *scope* les différentes combinaisons de groupe de variables explicatives qui peuvent constituer un modèle. La variable *data* représente l'un des jeux de données BIG5D, BIG5SD et RECEPD. Quant à la variable *direction*, elle représente le type de sélection à appliquer à savoir descendante dans notre cas.

3.2.4 Construction du modèle de prédiction

Cette étape consiste dans un premier temps à diviser chaque jeu de données BIG5D, BIG5SD et RECEPD en deux parties distinctes, l'une pour l'apprentissage (BIG5D_TRAINING, BIG5SD_TRAINING et RECEPD_TRAINING respectivement) et l'autre pour la validation (BIG5D_TEST, BIG5SD_TEST et RECEPD_TEST respectivement) du modèle. Ensuite, les données d'apprentissage sont utilisées pour construire ledit modèle de prédiction. Et enfin, on utilise le modèle construit pour prédire les intérêts des utilisateurs des données de validation. Le tableau 3.9 présente la distribution des utilisateurs par intérêt dans les données d'apprentissage et de validation.

[6] https://www.r-project.org/

	Apprentissage	Validation	**Total**
Politique	175	75	250
Sports	200	86	286
Économie	141	61	202
Jeu vidéo	164	71	235
Gastronomie	153	66	219
Tourisme	177	77	254
Grand total	**1 010**	**436**	**1 446**

Table 3.9: Distribution des utilisateurs selon les données d'apprentissage et de validation.

Pour la construction de nos modèles de prédiction nommés BIG5M, BIG5SM et RECEPM, nous avons utilisé la formule 3.3 dans laquelle la variable *data* représente nos données d'apprentissage BIG5D_TRAINING, BIG5SD_TRAINING et RECEPD_TRAINING respectivement, les variables explicatives X_i sont celles obtenues lors de la sélection des variables et Y représente la variable qui définit les intérêts des utilisateurs.

3.2.5 Évaluation et interprétation des résultats du modèle de prédiction

Cette étape consiste d'une part, à utiliser les mesures de performances telles que la précision, le rappel et la F-mesure, pour évaluer la capacité du modèle construit à classer les utilisateurs par intérêt à partir de leur dimensions psychologiques Receptiviti. Et d'autre part, commenter les résultats obtenus afin de vérifier s'il existe une corrélation entre les traits de personnalité d'un utilisateur et ses intérêts.

Globalement, l'évaluation du modèle consiste à répondre à la question suivante : le modèle construit décrit-il bien les valeurs observées ? Autrement dit, le modèle prédit-il avec exactitude les intérêts des utilisateurs de nos données de validation ? Les valeurs des mesures de performance d'ASCERTAIN (précision, rappel et f-mesure) vont nous permettre de répondre à cette question. La définition détaillée de ces mesures se trouve à la section 4.2.2 du chapitre 2.

Soit $\mathcal{C}$ l'ensemble de nos six intérêts : "Politique", "Économie", "Jeu vidéo", "Sports", "Gastronomie" et "Tourisme". Nous notons Pr_i, la précision d'ASCERTAIN pour un intérêt $i \in \mathcal{C}$. Nous avons :

$$Pr_i = \frac{|TP_i|}{|TP_i| + |FP_i|} \tag{3.5}$$

Où TP_i encore appelé vrais positifs est l'ensemble des utilisateurs dont l'intérêt réel et prédit est i. En d'autres termes, TP_i est le nombre d'utilisateurs qu'ASCERTAIN a classé dans la catégorie représentant l'intérêt i, et qui sont réellement associés à l'intérêt i dans SELFDS. Cependant, FP_i ou faux positifs est l'ensemble des utilisateurs dont l'intérêt prédit et réel sont respectivement i et j avec $i \neq j$. Autrement dit, FP_i est le nombre d'utilisateurs qui sont associés à l'intérêt j dans SELFDS, alors qu'ASCERTAIN a prédit pour ces derniers comme intérêt dominant i.

Par ailleurs, nous notons Ra_i, le rappel d'ASCERTAIN pour un intérêt $i \in \mathcal{C}$. Ainsi, nous avons :

$$Ra_i = \frac{|TP_i|}{|TP_i| + |FN_i|} \tag{3.6}$$

Où FN_i encore appelé faux négatifs est l'ensemble d'utilisateurs dont l'intérêt prévu et réel sont respectivement j et i avec $j \neq i$. En d'autres termes FN_i est le nombre d'utilisateurs qui sont associés à l'intérêt i dans SELFDS, alors qu'ASCERTAIN a prédit pour ces derniers comme intérêt dominant j.

Et enfin, nous notons Fm_i, la f-mesure d'ASCERTAIN pour un intérêt $i \in \mathcal{C}$. Elle est définit comme suit :

$$Fm_i = \frac{2 \times Pr_i \times Ra_i}{Pr_i + Ra_i} \tag{3.7}$$

En définitive, la précision $Pr_{\text{ASCERTAIN}}$, le rappel $Ra_{\text{ASCERTAIN}}$ et la f-mesure $Fm_{\text{ASCERTAIN}}$ globale d'ASCERTAIN sur chacune de nos trois bases de données est définit comme suit :

$$Pr_{\text{ASCERTAIN}} = \frac{\sum Pr_i}{|\mathcal{C}|} \; ; \quad Ra_{\text{ASCERTAIN}} = \frac{\sum Ra_i}{|\mathcal{C}|} \; ; \quad Fm_{\text{ASCERTAIN}} = \frac{\sum Fm_i}{|\mathcal{C}|} \tag{3.8}$$

3.3 Expérimentations et analyses

Dans cette section, nous exécutons la méthodologie ASCERTAIN décrite précédemment sur nos trois jeux de données BIG5D, BIG5SD et RECEPD. Tout d'abord, nous présentons et interprétons les résultats de l'analyse en composantes principales sur nos trois jeux de données. Par la suite, nous découvrons les variables explicatives utilisées pour la construction de nos modèles de prédiction finaux et enfin, nous présentons et interprétons les résultats obtenus à partir de chacune de nos différentes sources de données.

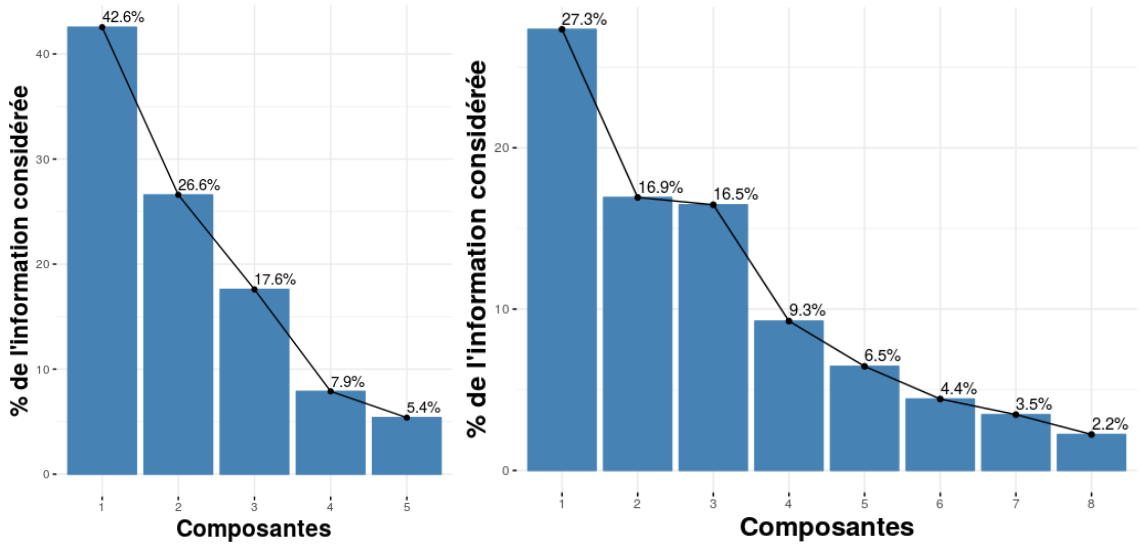
3.3.1 Analyse en composantes principales

Pour faire l'analyse en composante principale, nous avons utilisé la fonction "PCA" de la librairie "FactoMineR" du logiciel R qui se présente comme suit :

$$res.pca \leftarrow PCA(data,\ scale.unit = TRUE) \tag{3.9}$$

où *res.pca* est la variable contenant le résultat de l'analyse en composantes principales sur les données contenues dans la variable *data*. Le paramètre *scale.unit* permet de normaliser ou pas les valeurs des données de *data*. Dans notre étude, comme nous normalisons nos valeurs.

À l'issue de l'exécution de cette fonction "PCA" sur nos trois jeux de données, on peut découvrir les différentes composantes ou facteurs obtenus, qui résument chacune un pourcentage bien précis de l'information initiale. Les figures 3.2a, 3.2b et 3.3 indiquent les différentes composantes principales ainsi que l'information qu'elles restituent, obtenues respectivement à partir des données BIG5D, BIG5SD et RECEPD. Dans le cas de BIG5D, nous voyons que les trois premières composantes sur les 5 obtenues permettent de représenter 86.8% de l'information, par conséquent on peut se contenter d'utiliser uniquement ces 3 premières composantes pour caractériser nos individus. Cependant, dans le cas de BIG5SD, on voit que les sept premières composantes (sur les 8 représentées) permettent de représenter 84.5% de l'information et pour le cas de RECEPD, les quinze premières composantes (sur 16 représentées) permettent de représenter 83.7% de l'information.



(a) Big5D (b) Big5SD

Figure 3.2: Composantes principales de Big5D et Big5SD.

Chaque composante est caractérisée par le pourcentage d'information qu'elle restitue. Huit premières composantes dans le cas de Big5SD. Par exemple, 42.6% de l'information est restitué par la composante1 de Big5D.

Par ailleurs, nous notons que chaque composante regroupe en son sein plusieurs variables explicatives sémantiquement proches les unes des autres. A cet effet, nous avons opté pour une représentation graphique qui met en avant le degré de liaison existant entre les différentes composantes et les variables explicatives X_i de chaque jeu de données. La figure 3.4 illustre les liaisons entre les variables explicatives et les composantes (1, 2 et 3) obtenues à l'issue de l'analyse en composantes principales sur Big5D. Nous notons que les pourcentages indiqués au niveau de chaque dimension représentent les proportions d'informations qui constituent chaque composante cible. De même, plus le nom d'une variable a une couleur proche du bleu et sa flèche proche du périmètre du cercle de corrélation, plus cette variable est fortement corrélée à la composante avec qui son vecteur forme le plus petit angle. A l'inverse, les variables dont leurs noms ont une couleur qui tend vers le rouge et/ou leurs vecteurs éloignés du périmètre du cercle de corrélation ne peuvent pas être interprétées selon les composantes cibles. Cependant, elles peuvent être discriminantes suivant d'autres composantes, c'est le cas de la variable *extraversion* qui est faiblement corrélée aux composantes 1 et 2 d'après la figure 3.4a, alors que sur la figure 3.4b on s'aperçoit qu'elle est fortement corrélée à la composante 3. En résumé, nous pouvons dire que les variables *openness* et *agreeableness* sont positivement corrélées à la composante 1, pendant que *conscientiousness* l'est négativement sur cette même composante. Nous constatons également que c'est la variable *extraversion* qui constitue la composante 3 et *neuroticism* pour le cas de la composante 2.

De même, la figure 3.5, présente les douze premières variables les plus corrélées aux trois premières composantes de l'analyse en composantes principales sur Big5SD. Nous avons choisi douze juste par soucis de lisibilité des figures. Sur la figure 3.5a, on voit que la variable "extraversion" est l'unique variable fortement corrélée à la composante 1, même si on ne peut négliger les autres

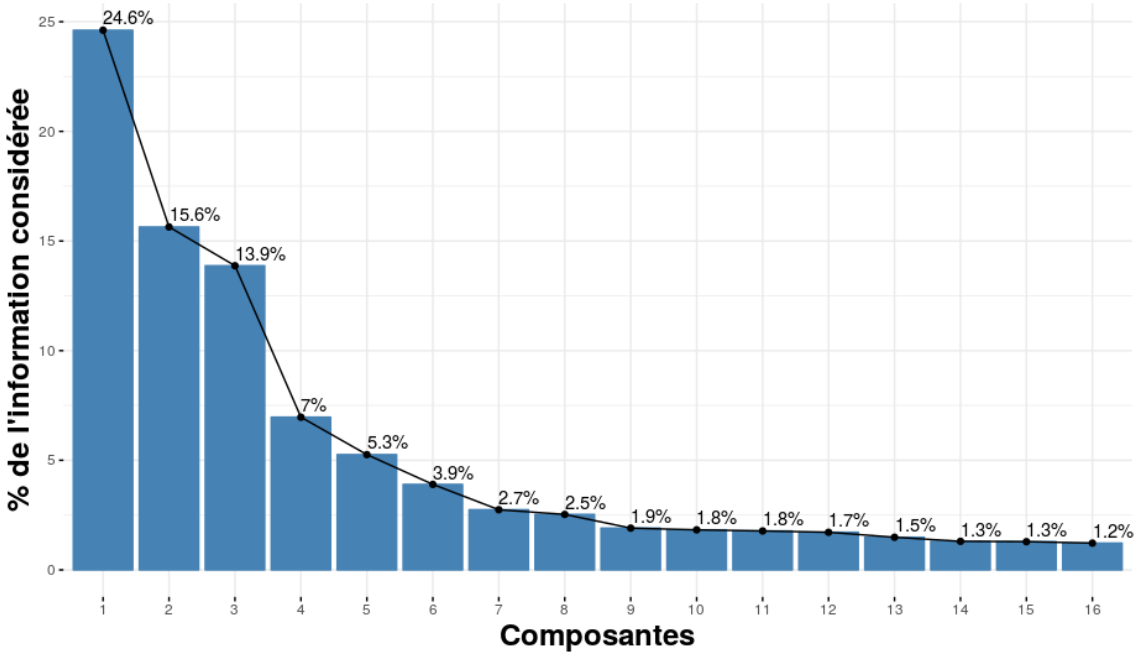


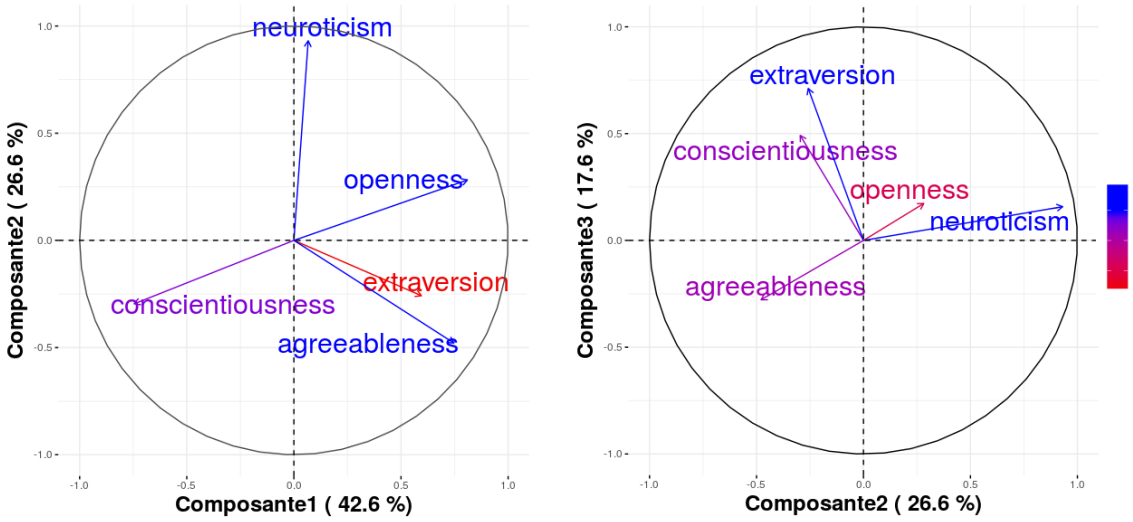
Figure 3.3: Seize premières composantes principales de RECEPD.

Chaque composante est caractérisée par le pourcentage d'information qu'elle restitue. Huit premières composantes dans le cas de RECEPD. Par exemple, 24.6% de l'information est restitué par la composante1 de RECEPD.

variables telles que "cooperative", "cheerful", "energetic" et "agreeableness". Sur la composante 2, on voit plutôt les variables "conscientiouness" et "ambitious". Par contre, sur la figure 3.5b, ce sont les variables "neuroticism" et "stressed" qui sont fortement et voir "self_conscious" légèrement corrélées à la composante 3.

L'analyse en composantes principales sur RECEPD, nous a donné les différentes liaisons de la figure 3.6 entre les variables explicatives et les quatre premières composantes obtenues. Nous notons également que nous choisissons par soucis de lisibilité de ne présenter que les douze premières variables les plus fortement corrélées à ces quatre premières composantes. On constate que plusieurs variables composent de plus en plus les composantes, et par conséquent, on peut dire plus il y a de variables explicatives plus les composantes sont occupées par plusieurs variables. D'après la figure 3.6a, *extraversion*, *happiness*, *persuasive* et *cheerful* contribuent majoritairement à la construction de la composante 1, quant à *conscientiousness*, *insecure* ainsi que *melancholy* elles contribuent plutôt pour la composante 2. De même, sur la figure 3.6b les variables *stressed*, *type_ a*, *neuroticism*, et "intellectual" (respectivement *humble*) sont visiblement corrélées positivement (respectivement négativement) à la composante 3.

À la sortie de cette analyse nous pouvons dire que certaines variables sont sémantiquement proches et se comportent de la même manière. C'est notamment, le cas des variables explicatives fortement corrélées à un même axe. En outre, pour avoir exactement le degré de corrélation entre les variables, on peut déterminer leurs coefficients de corrélation. Plus le coefficient de corrélation entre deux variables est proche des valeurs extrêmes -1 et 1, plus la corrélation entre ces variables est



(a) (b)

Figure 3.4: Corrélation entre composantes principales et variables explicatives de BIG5D.

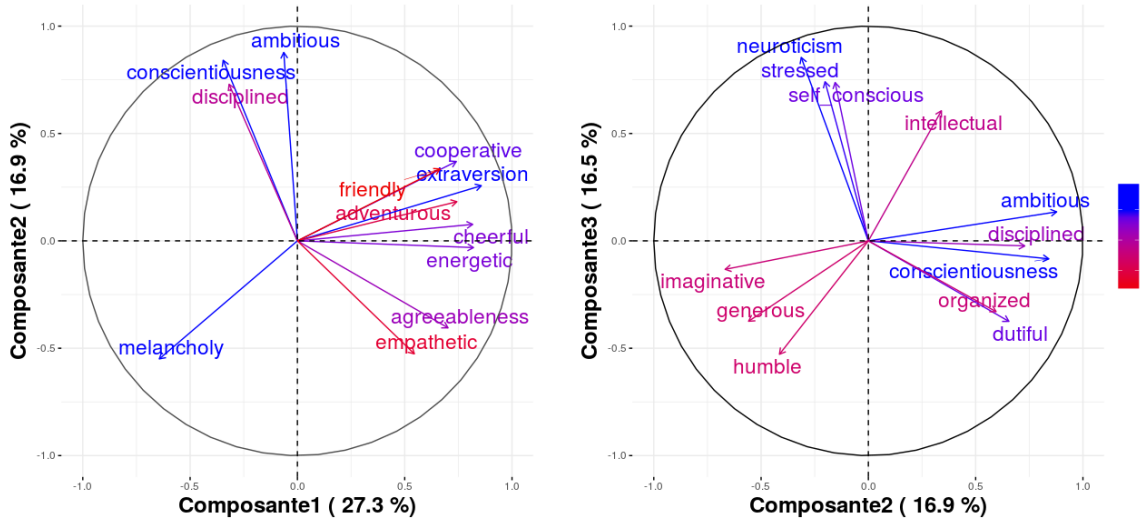
Composante1(42.6%) signifie que la composante1 restitue 42.6% de l'information initiale. Pus le nom d'une variable a une couleur proche du bleu et sa flèche proche du périmètre du cercle de corrélation, plus cette variable est fortement corrélée à à la composante avec qui son vecteur forme un angle minimale. A l'inverse, les variables dont leurs noms ont une couleur qui tend vers le rouge et/ou leurs vecteurs éloignés du périmètre du cercle de corrélation ne peuvent pas être interprétées selon les composantes cibles.

forte. Par contre, une corrélation égale à 0 signifie que les variables ne sont pas corrélées. De manière générale, le coefficient de corrélation est interprété comme l'indique le tableau 3.10. Par exemple, une corrélation positive entre deux variables montre que lorsqu'une variable augmente l'autre augmente également. A l'inverse, une corrélation négative montre que lorsque l'une augmente, l'autre diminue.

Corrélation	Négative	Positive
Faible	de -0.5 à 0.0	de 0.0 à 0.5
Forte	de -1.0 à -0.5	de 0.5 à 1.0

Table 3.10: Interprétation des valeurs du coefficient de corrélation.

Par exemple, pour les données de BIG5D nous avons découvert une seule corrélation forte et négative de -0.51 entre les variables *openness* et *conscientiousness*. L'interprétation qui découle de cette corrélation est qu'on peut dire que le degré auquel les utilisateurs de BIG5D sont ouverts à de nouvelles idées ou expériences s'oppose à leur degré de fiabilité. Le tableau 3.11 montre les coefficients de corrélation dont la valeur absolue est supérieure à 0.8 entre les variables explicatives de BIG5SD. Nous avons choisi un seuil de 0.8 afin de pouvoir avoir une représentation qui tient dans un tableau lisible. On constate que la variable *agreeableness* qui mesure le degré auquel une personne est enclin à satisfaire les autres est positivement et fortement corrélée à la variable *empathetic* qui



(a) (b)

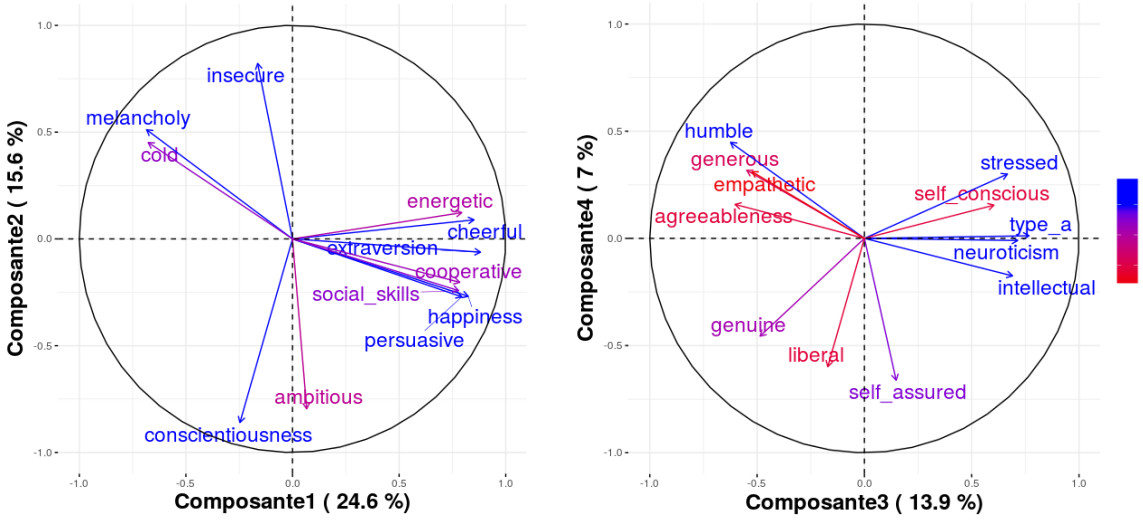
Figure 3.5: Corrélation entre composantes principales et variables explicatives de BIG5SD.

Composante1(27.3%) signifie que la composante1 restitue 27.3% de l'information initiale. Plus le nom d'une variable a une couleur proche du bleu et sa flèche proche du périmètre du cercle de corrélation, plus cette variable est fortement corrélée à la composante avec qui son vecteur forme un angle minimale. A l'inverse, les variables dont les noms tendent vers la couleur rouge ne peuvent pas être interprétées selon les composantes cibles.

mesure le degré selon lequel une personne intériorise les sentiments des autres. De même, on voit que la variable *neuroticism* qui mesure le degré auquel une personne exprime de fortes émotions négatives est également positivement et fortement corrélée à la variable *stressed* qui mesure le degré selon lequel une personne éprouve du stress.

Quant au tableau 3.12, il présente les coefficients de corrélations dont la valeur absolue est supérieure à 0.8 entre les variables explicatives de RECEPD. Nous notons qu'à ces corrélations du tableau 3.12 on ajoute celles du tableau 3.11. Sur ce nouveau tableau on peut lire que la variable "cold" qui mesure le degré auquel une personne est émotionnellement insensible et a du mal à empathiser avec les autres (respectivement la variable *genuine* qui mesure l'authenticité et de l'honnêteté d'une personne) sont négativement et fortement corrélées à la variable *cooperative* qui mesure à quel point une personne prend en compte les besoins des autres (respectivement la variable *money_ oriented* qui mesure le degré auquel une personne pense à l'argent et aux finances).

Une fois que nous avons analysé nos variables, nous pouvons nous pencher sur le cas des individus. Autrement dit, on va chercher à expliquer les groupes d'utilisateurs classés selon leurs intérêts par le biais des composantes principales obtenues à partir de chacun des trois jeux de données. Plus précisément, on va faire une représentation graphique des individus selon les deux premières composantes principales de chacune de nos données. La figure 3.7 est une représentation des utilisateurs par intérêt selon les composantes 1 et 2 sur BIG5D. Nous nous rendons compte que certains groupes d'utilisateurs se comportent sensiblement de la même manière. C'est le cas d'une part, des utilisateurs de la catégorie "Gastronomie" et "Tourisme" qui sont pour la plupart du côté positif de la composante 1 qui à son tour est constitué essentiellement des variables *openness* et *agreeableness*



(a) (b)

Figure 3.6: Corrélation entre composantes principales et variables explicatives de RECEPD.

Composante1(24.6%) signifie que la composante1 restitue 26.6% de l'information initiale. Plus le nom d'une variable a une couleur proche du bleu et sa flèche proche du périmètre du cercle de corrélation, plus cette variable est fortement corrélée à la composante avec qui son vecteur forme un angle minimale. À l'inverse, les variables dont les noms tendent vers la couleur rouge ne peuvent pas être interprétées selon les composantes cibles.

	artistic	organized	cheerful	generous	empathetic	stressed	trusting
openness	0,84	-	-	-	-	-	-
conscientiousness	-	0,81		-	-	-	-
extraversion	-	-	0,8	-	-	-	-
agreeableness	-	-	-	0,87	0,91	-	-
neuroticism	-	-	-	-	-	0,85	-
friendly	-	-	-	-	-	-	0,85
empathetic	-	-	-	0,92	-	-	-

Table 3.11: Coefficient de corrélation dont la valeur absolue est supérieur à 0.8 entre les variables explicatives de BIG5SD.

(voir figure 3.4a). Autrement dit, les utilisateurs s'intéressant à la "Gastronomie" et au "Tourisme" sont ouverts à de nouvelles idées et expériences, ainsi que enclins à satisfaire les autres. D'autre part, les utilisateurs s'intéressant à la "Politique" et à l'"Économie" sont plutôt du côté négatif de la composante 1, c'est-dire qu'ils ont tendance à être moins ouverts et agréables, mais plutôt con-

	money_oriented	generous	agreeableness	social_skills	cooperative
assertive	0,96	-	-	-	-
empathetic	-	0,92	0,91	-	-
trusting	-	-	-	0,81	-
cold	-	-	-	-	-0,8
happiness	-	-	-	0,81	-
extraversion	-	-	-	0,81	-
friendly	-	-	-	0,85	-
genuine	-0,82	-	-	-	-

Table 3.12: Coefficient de corrélation dont la valeur absolue est supérieur à 0.8 entre les variables explicatives de RECEPD.

Il contient également les coefficients de corrélation du tableau 3.11 que nous avons omis par soucis de lisibilité.

sciencieux (variable *conscientiousness*). On peut également lire que la plupart des utilisateurs qui s'intéressent au "Jeu vidéo" sont nerveux car ils sont représentés du coté positif de la composante 2 qui est constitué principalement par la variable *neuroticism*.

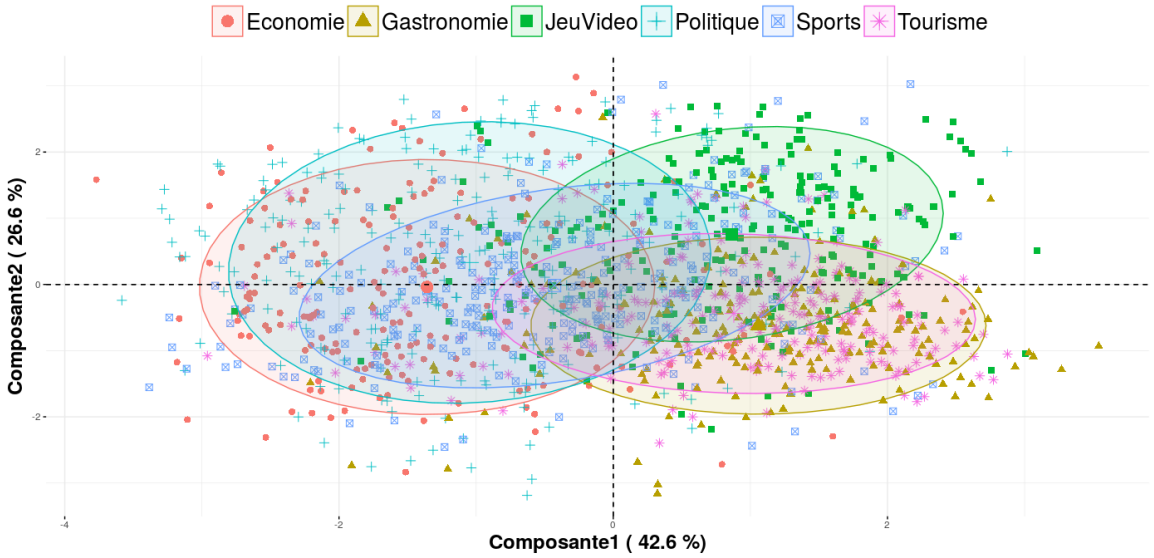


Figure 3.7: Représentation des utilisateurs par intérêt selon les composantes 1 et 2 de BIG5D.

De même, la figure 3.8 représente les utilisateurs par intérêt selon les composantes 1 et 2 de BIG5SD. Le coté positif de la composante 1 qui est constitué principalement des utilisateurs des catégories "Gastronomie", "Tourisme" et voir même "Jeu vidéo" est représenté par les variables *extraversion*, *cooperative*, *cheerful* et *energetic* (voir figure 3.5a). Par contre, le côté négatif de cette même composante qui est constitué des utilisateurs des catégories "Économie" et "Politique" est porté par la variable *melancholy*. Par ailleurs, la composante 2 qui contient du coté positif pas mal d'utilisateurs des catégories "Économie", "Politique" et "Sports" est représentée par les variables *ambitious* et *conscientiousness*. En résumé, au vue de la position des utilisateurs sur cette figure 3.8 on peut dire que les utilisateurs intéressés par la "Gastronomie", le "Tourisme" prennent compte des besoins des autres (*cooperative*), agissent généralement joyeusement (*cheerful*) ou ont tendance à être pour la plupart du temps enthousiastes. A l'inverse, on a les utilisateurs s'intéressant à la "Politique" et à l'"Économie", qui sont plutôt mélancoliques. Sur la composante 2, nous pouvons dire que les utilisateurs de la catégorie "Jeu vidéo" sont ni ambitieux ni consciencieux.

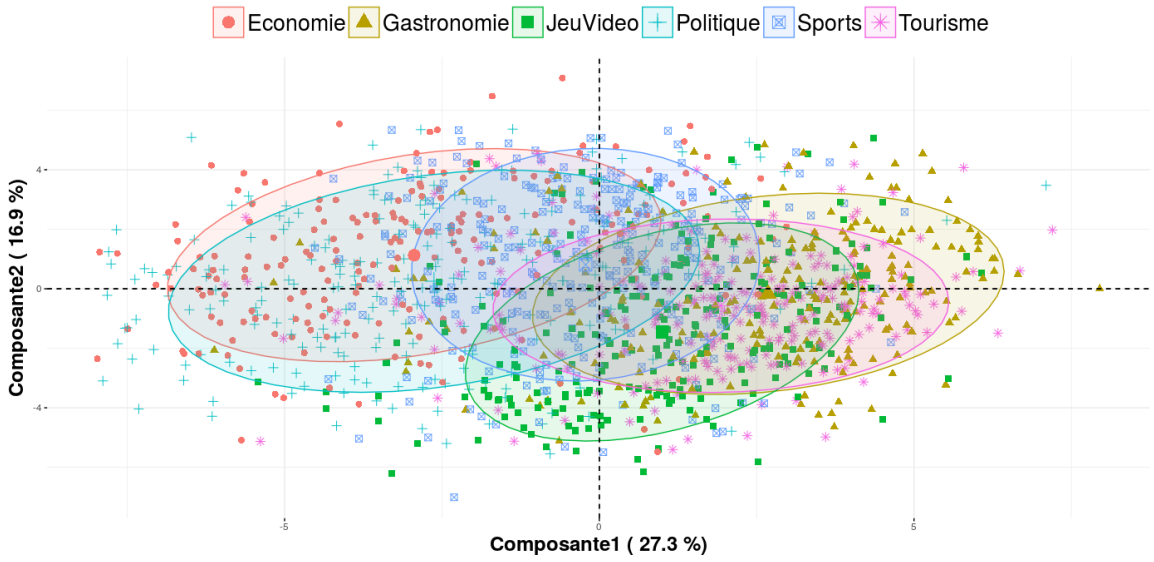


Figure 3.8: Représentation des utilisateurs par intérêt selon les composantes 1 et 2 de BIG5SD.

Pour l'interprétation des utilisateurs selon les deux premières composantes principales de RECEPD, on se réfère à la figure 3.9. Nous observons des phénomènes semblables à ceux de la figure 3.8, c'est-à-dire ce sont les utilisateurs des catégories "Gastronomie" et "Tourisme" qui peuplent principalement le coté positif de la composante 1. A l'inverse, du coté négatif on revoit également les utilisateurs des groupes "Économie" et "Politique". Nous notons que la composante 1 est constituée du coté positif par les variables *extraversion*, *cheerful*, *happiness* et *persuasive* et du coté négatif par la variable *melancholy*. Les utilisateurs intéressés par la "Gastronomie", le "Jeu vidéo" et le "Tourisme" sont des personnes entres autres éveillées, interagissant avec d'autres personnes, s'engageant dans des activités, ainsi que créant des rapports avec l'intention de persuader les autres. De même, les utilisateurs de la catégorie "Jeu vidéo" sont ceux qui manquent le plus de confiance

lorsqu'ils traitent avec d'autres personnes. D'après la composante 2 les économistes sont les plus consciencieux.

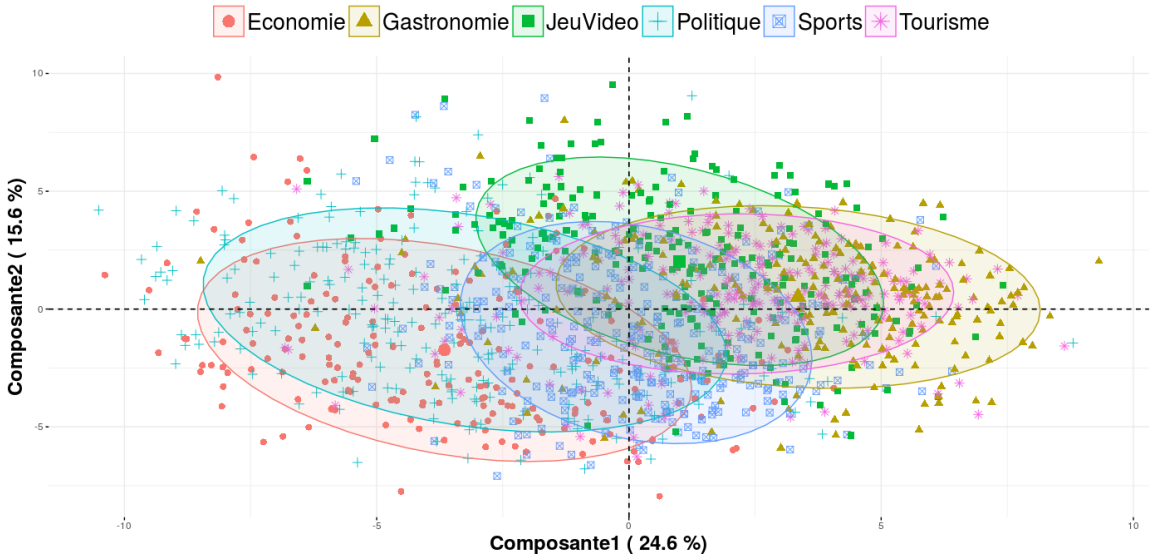


Figure 3.9: Représentation des utilisateurs par intérêt selon les composantes 1 et 2 de RECEPD.

Après l'analyse de nos variables, nous passons à la sélection des variables que nous utilisons par la suite pour construire notre modèle de prédiction.

3.3.2 Prédiction des intérêts des utilisateurs

A la fin de l'exécution de la fonction de sélection des variables sur BIG5D, nous obtenons les mêmes 5 variables de départ (*openness, conscientiousness, extraversion, agreeableness, neuroticism*). Par contre, pour BIG5SD nous sommes passés de 35 variables explicatives à 32. Les trois variables supprimées du modèle saturé BIG5SMS sont *intellectual*, *imaginative* et *persuasive*. Pour le cas de RECEPD nous sommes passés de 59 variables explicatives à 57. Les deux variables supprimées sont *cheerful* et *persuasive*. Dans la suite, nous utilisons les variables obtenues pour chaque jeu de données pour construire le modèle de prédiction des intérêts des utilisateurs.

Pour prédire les intérêts des utilisateurs à partir des modèles construits, nous utilisons la fonction *predict* de la librairie "stats" avec le logiciel R. Cette fonction se définit comme suit :

$$result \leftarrow predict(modele,\ data) \tag{3.10}$$

où la variable *result* est le résultat de la prédiction des intérêts des utilisateurs des données de validation *data*.

Le tableau 3.13 présente la matrice de confusion ainsi que les mesures de performances (précision, rappel et f-mesure) par intérêt du modèle de prédiction BIG5M. On constate que BIG5M a du mal à classer les utilisateurs dans leurs vrais intérêts respectifs. Ceci est visible par les taux de précision et

de rappel atteints pour chaque intérêt qui sont très faibles. Globalement, les mesures de performance d'Ascertain, c'est-à-dire la précision $Pr_{\text{Ascertain}}$, le rappel $Ra_{\text{Ascertain}}$ et la f-mesure $Fm_{\text{Ascertain}}$ avec ce modèle Big5M sont respectivement 0.51, 0.51 et 0.51. Au vue des taux atteints par ce modèle Big5M on peut dire qu'il ne représente pas efficacement les données SelfDS.

	Politique	Économie	Jeu vidéo	Gastronomie	Sports	Tourisme
Politique	39	22	1	2	13	5
Économie	10	21	1	2	10	2
Jeu vidéo	6	4	55	11	14	13
Gastronomie	4	2	3	31	3	15
Sports	10	10	7	4	42	8
Tourisme	6	2	4	16	4	34
Pr_i	0.52	0.34	0.77	0.47	0.49	0.44
Ra_i	0.48	0.46	0.53	0.53	0.52	0.52
Fm_i	0.5	0.39	0.63	0.5	0.5	0.48

Table 3.13: Matrice de confusion et mesures performance du modèle de prédiction Big5M.

Pr_i, Ra_i, Fm_i sont respectivement la précision, le rappel et la f-mesure d'Ascertain pour un intérêt i.

D'un autre côté, la figure 3.14 présente la matrice de confusion ainsi que les mesures de performances par intérêt du modèle de prédiction Big5SM. Nous notons une amélioration des taux de précision, rappel et f-mesure même s'ils ne sont pas encore satisfaisants. Pour les performances globales d'Ascertain, nous sommes passés de 0.51 de précision (respectivement 0.51 de rappel et 0.51 de f-mesure) à 0.71 (respectivement 0.71 et 0.71). Nous observons que parmi tous les intérêts cibles, le modèle Big5SM réussit à mieux prédire les utilisateurs de la catégorie "Sports" avec une f-mesure de 81% suivit de ceux de la catégorie "Économie" (77% de f-mesure).

Le tableau 3.15, quant à lui présente la matrice de confusion ainsi que les mesures de performances par intérêt du modèle de prédiction recepM. Une fois de plus on observe une nette amélioration des taux de précision, rappel et f-mesure. Nous constatons que les performances globales d'Ascertain, en particulier sa précision est passée de 0.71 à 0.78. De plus, ce modèle recepM prédit sensiblement bien les intérêts des utilisateurs de la catégorie "Gastronomie", "Sports" et voir même "Politique" et "Économie".

En résumé, il est visible que le modèle de prédiction recepM obtenu à partir des variables explicatives issues de la sélection des variables sur le jeu de données recepD est le meilleur des trois modèles construits bien qu'il n'atteint pas encore une précision optimale ou acceptable. Nous notons que sans la sélection de variables nous obtenons $Pr_{\text{Ascertain}}$= 0.75 et $Ra_{\text{Ascertain}}$= 0.75. Compte tenu de nos résultats on va dans ce qui suit interpréter plus en détails les résultats obtenus à partir

	Politique	Économie	Jeu vidéo	Gastronomie	Sports	Tourisme
Politique	54	9	1	3	5	5
Économie	4	48	4	4	0	4
Jeu vidéo	2	0	55	7	9	6
Gastronomie	7	3	4	40	3	13
Sports	3	0	4	1	64	0
Tourisme	5	1	3	11	5	49
Pr_i	0.72	0.79	0.77	0.61	0.74	0.64
Ra_i	0.7	0.75	0.7	0.57	0.89	0.66
Fm_i	0.71	0.77	0.73	0.59	0.81	0.65

Table 3.14: Matrice de confusion et mesures performance du modèle de prédiction BIG5SM.

Pr_i, Ra_i, Fm_i sont respectivement la précision, le rappel et la f-mesure d'ASCERTAIN pour un intérêt i.

	Politique	Économie	Jeu vidéo	Gastronomie	Sports	Tourisme
Politique	63	10	1	1	2	7
Économie	5	46	2	0	0	4
Jeu vidéo	3	0	50	3	10	6
Gastronomie	1	0	4	59	1	2
Sports	2	0	8	0	70	5
Tourisme	1	5	6	3	3	53
Pr_i	0.84	0.75	0.7	0.89	0.81	0.69
Ra_i	0.75	0.81	0.69	0.88	0.82	0.75
Fm_i	0.79	0.78	0.69	0.88	0.81	0.72

Table 3.15: Matrice de confusion et mesures performance du modèle de prédiction RECEPM.

Pr_i, Ra_i, Fm_i sont respectivement la précision, le rappel et la f-mesure d'ASCERTAIN pour un intérêt i.

du modèle RECEPM uniquement.

L'interprétation des résultats consiste à déterminer les variables fortement liées à chaque groupe d'utilisateurs. Pour le faire, on va utiliser la fonction *odds.ratio* de la librairie "questionr" avec le logiciel R qui prend en entrée un modèle de prédiction et qui retourne pour chaque variable explicative contenue dans le modèle leurs rapport des chances et intervalles de confiance.

Le rapport des chances ou l'*odds ratio* (OR) est une mesure statistique exprimant le degré de dépendance entre deux variables, l'une la cible et l'autre la référence. Plus nettement, il est illustré sous forme d'une valeur numérique toujours supérieure ou égale à zéro. Traditionnellement, l'*odds ratio* s'interprète toujours par rapport à une modalité de référence qui est sélectionnée au préalable avant la construction du modèle. Ainsi, on peut comprendre l'impact du choix d'une modalité en fonction des variables explicatives et en comparaison avec une modalité fixée. Lors de la construction de nos modèles de prédiction nous avons sélectionné l'intérêt "Politique" comme catégorie de référence. Par exemple, un *odds ratio* de 1.4 pour les utilisateurs de la catégorie "Économie" sur la variable *neuroticism* s'interprète comme suit : "les utilisateurs qui s'intéressent à l'économie ont 1.4 fois plus de chances que ceux qui s'intéressent à la politique d'être nerveux". Autrement dit, les économistes sont plus nerveux que les politiciens.

Quant à l'intervalle de confiance, elle permet de vérifier s'il existe une relation entre une variable explicative X_i précise et la variable expliquée Y. En particulier, dans notre cas elle nous permet d'identifier les variables explicatives discriminantes selon deux groupes d'utilisateurs distincts. Généralement, en fonction du test statistique utilisé si l'intervalle de confiance ne contient pas une valeur précise notamment 1 dans notre cas la variable cible est considérée comme discriminante. Par exemple, avec les données BIG5SD on obtient un intervalle de confiance de $[0.84; 0.93]$ sur la dimension *disciplined* entre les utilisateurs des catégories "Jeu vidéo" et "Politique". Ainsi, comme $1 \notin [0.84; 0.93]$ on peut dire que la dimension *disciplined* est une variable discriminante pour ces deux catégories et nous pouvons quantifier la différence à l'aide de la valeur de l' *odds ratio* associée.

En somme, le tableau 3.16 présente pour chaque catégorie d'intérêt, les variables qui leurs sont les plus corrélées. Les valeurs entre parenthèses représentent les *odds-radio* des variables explicatives entre la catégorie cible et la catégorie de référence "Politique". Les symboles ***, **, * et . signifient que la variable associée discrimine respectivement très fortement, fortement, faiblement et pas du tout les utilisateurs des catégories cible et référence. De ce tableau, nous pouvons dire que les utilisateurs qui s'intéressent à l'"Économie" ont 0.45 fois plus de chance que ceux qui s'intéressent à la "Politique" d'être agréables (*odds ratio* de 0.45 sur la variable *agreeableness*). De ce fait, on peut dire que les économistes sont moins enclins à satisfaire les autres par rapport aux politiciens. De même, les utilisateurs intéressés par le "Tourisme" font nettement plus confiance aux personnes que les politiciens (*odds ratio* de 1.28 sur la variable "*trusting*).

En définitive, au vue des résultats obtenus sur nos données nous pouvons dire qu'il existe statistiquement une corrélation entre certaines dimensions psychologiques Receptiviti et les intérêts des utilisateurs. Par ailleurs, nous constatons qu'ASCERTAIN réussit à bien prédire les intérêts des utilisateurs de certaines catégories, notamment "Gastronomie" et "Sports". A cet effet, nous avons tenté d'explorer une autre démarche qui consiste à utiliser plutôt la régression logistique multiple pour prédire les intérêts des utilisateurs. Autrement dit, au lieu de faire une régression logistique polytomique multiple comme précédemment nous faisons plutôt une régression logistique multiple, dans laquelle la variable expliquée Y est transformée en une variable binaire.

Catégories cibles	Variables corrélées
Économie	melancholy (0.83**), trusting (1.12*), impulsive (0.91·), thinking_style (0.87·), cold (0.91**), social_skills (0.92·), happiness (0.91**), depression (1.07·), imaginative (1.21**), assertive (1.12·), workhorse (1.11**), sexual_focus (1.05*), **empathetic (1.27***)**, **humble (1.24***)**, **neuroticism (1.37***)**, **generous (1.53***)**, **agreeableness (0.45***)**, adjustment (0.84·), **genuine (1.57***)**, disciplined (1.08*), money_oriented (1.19*), friendly (1.16*), religion_oriented (0.95**), independent (0.92**), health_oriented (0.96*)
Jeu vidéo	**melancholy (0.78***)**, reward_bias (1.1*), emotionally_aware (1.11·), cold (0.93·), liberal (1.1*), **happiness (0.89***)**, **ambitious (0.83***)**, workhorse (1.11*), friend_focus (1.04·), **empathetic (1.44***)**, **humble (1.28***)**, **neuroticism (1.59***)**, power_driven (1.09**), **generous (1.66***)**, **agreeableness (0.39***)**, leisure_oriented (1.14*), **genuine (1.75***)**, sociable (1.11*), family_oriented (0.95·), food_focus (0.96*), money_oriented (1.17*), stressed (0.83**), work_oriented (0.95*), **religion_oriented (0.92***)**, **independent (0.86***)**, anxious (0.83**), health_oriented (1.04*)
Gastronomie	**melancholy (0.79***)**, **trusting (1.22***)**, impulsive (0.89·), assertive (1.19**), workhorse (1.09·), sexual_focus (1.05*), **empathetic (1.48***)**, **humble (1.22***)**, neuroticism (1.24*), **generous (1.93***)**, **agreeableness (0.29***)**, adjustment (0.81*), leisure_oriented (1.07·), **genuine (1.94***)**, family_oriented (0.95*), **food_focus (1.12***)**, **religion_oriented (0.92***)**, independent (0.91**)
Sports	**melancholy (0.81***)**, **reward_bias (1.14***)**, **trusting (1.18***)**, impulsive (0.86**), insecure (1.1**), netspeak_focus (0.91*), **body_focus (1.07***)**, **empathetic (1.47***)**, humble (1.18**), neuroticism (1.24*), power_driven (1.05·), **generous (1.72***)**, **agreeableness (0.35***)**, adjustment (0.84·), leisure_oriented (1.1·), **genuine (1.87***)**, **food_focus (0.94***)**, **religion_oriented (0.92***)**, independent (0.92**), artistic (1.09·)
Tourisme	adventurous (1.22**), type_a (0.94**), **trusting (1.28***)**, happiness (1.08*), ambitious (0.91·), **workhorse (1.15***)**, sexual_focus (1.06**), active (1.1·), **empathetic (1.23***)**, **humble (1.5***)**, power_driven (1.06*), **generous (1.95***)**, **agreeableness (0.26***)**, cautious (0.91·), **genuine (1.53***)**, extraversion (0.81*), aggressive (0.86*), food_focus (0.96**), disciplined (0.91**), stressed (1.2**), work_oriented (0.94**), cooperative (1.13*), artistic (0.85**), health_oriented (0.97*)

Table 3.16: Variables explicatives discriminantes entre une catégorie cible et la catégorie de référence "Politique" obtenue à partir du modèle de prédiction RECEPM.

Les valeurs entre parenthèses désignent l'*odds ratio* des variables associées pour la catégorie cible. Les symboles ***, **, * et . signifient que la variable associée discrimine respectivement très fortement, fortement, faiblement et pas du tout les utilisateurs des catégories cible et référence.

3.3.3 Exploration de la régression logistique multiple

Comme mentionné précédemment, les différentes étapes de cette nouvelle démarche sont semblables à celles précédemment décrites à l'exception du type de la méthode statistique utilisée pour construire

le modèle de prédiction. Autrement dit, toutes les étapes d'Ascertain restent inchangées, sauf la régression logistique polytomique multiple qui est remplacée par une régression logistique multiple. Pour le faire on va d'une part, fixer un intérêt cible $\mathcal{T}_{int}$ et une catégorie $\mathcal{C}at_{\mathcal{T}}$ qui représente les utilisateurs s'intéressant à $\mathcal{T}_{int}$, et d'autre part, noter $\mathcal{C}at_{\mathcal{O}}$ la classe des utilisateurs s'intéressant à l'un des intérêts de $\mathcal{O}_{int}$ qui représente nos six intérêts initialement choisis à l'exception de l'intérêt cible. Ensuite, le jeu de données SelfDS est transformé en un nouveau SelfDS_New dans lequel seule les valeurs de la variable "intérêt" sont modifiées, de telle sorte qu'ont attribue la valeur 1 à tous les utilisateurs appartenant à la catégorie cible et 0 aux autres. Et enfin, on va construire et évaluer le modèle de prédiction construit à partir des données de SelfDS_New.

Pour construire notre modèle de prédiction on va comme précédemment diviser notre jeu de données SelfDS en 2 parties: l'une pour l'apprentissage SelfDS_NewTr et l'autre pour la validation SelfDS_NewTe. le tableau 3.17 montre le nombre d'utilisateurs contenu dans SelfDS_NewTr et SelfDS_NewTe. Par soucis d'équitabilité, nous avons fixé le nombre total d'utilisateur de la catégorie cible à 200 car c'est le nombre minimum d'utilisateurs qu'on peut avoir dans chacune de nos six catégories (voir figure 3.6). Pour déterminer les utilisateurs des cinq autres catégories tout en tenant compte d'une répartition équilibrée entre catégories nous avons fixé le nombre d'utilisateurs à 40 pour chacune des cinq autres catégories. Ce qui fait un total de 200 utilisateurs également. Ainsi, les données d'apprentissage contiennent 140 utilisateurs de la catégorie cible (soit 70 des 200 utilisateurs de départ) et 28 de chacune des cinq catégories (soit 70 également des 40 utilisateurs de départ) pour un total de $28 \times 5 = 140$ utilisateurs. Au final on a 280 utilisateurs pour l'apprentissage et 120 pour la validation ce qui nous fait un total de 400 utilisateurs.

	Apprentissage	Validation	Total
Catégorie cible	140	60	200
5 autres catégories	$28 \times 5 = 140$	$12 \times 5 = 60$	200
Grand total	**280**	**120**	**400**

Table 3.17: Répartition des utilisateurs entre SelfDS_NewTr et SelfDS_NewTe.

Nous notons que cette nouvelle démarche utilise toutes les 59 dimensions psychologiques Receptiviti comme variables explicatives. Après la construction du modèle de prédiction à partir des variables explicatives issues de l'étape de sélection des variables, nous obtenons pour chaque modèle de prédiction les mesures de performances illustrées dans le tableau 3.18. Nous constatons que les résultats sont assez prometteurs puisque nous atteignons dans certains cas un taux de 92% voir 97% de précision. Les modèles construits en prenant comme intérêt cible "Sports" , "Jeu vidéo" et "Gastronomie" présentent des meilleurs résultats par rapport aux autres intérêts.

Par ailleurs, pour l'évaluation des modèles de prédiction construits à partir des régressions logistique simple et multiple, il existe une représentation simple et efficace : la courbe ROC (*Receiver Operating Characteristic*) ou courbe sensibilité/spécificité qui est utilisée pour mesurer les performances des systèmes de classification binaire (classe les éléments dans deux groupes distincts). La sensibilité d'un modèle est sa capacité à moins se tromper lors du classement des utilisateurs de la catégorie cible $\mathcal{C}at_{\mathcal{T}}$, tant dis que sa spécificité est sa faculté à moins se tromper lors du classement des utilisateurs de la catégorie $\mathcal{C}at_{\mathcal{O}}$. Plus clairement, la courbe ROC se présente sous la forme d'une courbe qui expose le taux de vrais positifs (fraction d'utilisateurs affectés à leur vrai catégorie) en

	$\mathcal{T}_{int}$	$\mathcal{O}_{int}$
$\mathcal{C}at_{\mathcal{T}}$	48	11
$\mathcal{C}at_{\mathcal{O}}$	12	49
Précision	0.8	0.82
Rappel	0.81	0.8

(a) **Politique**

	$\mathcal{T}_{int}$	$\mathcal{O}_{int}$
$\mathcal{C}at_{\mathcal{T}}$	54	8
$\mathcal{C}at_{\mathcal{O}}$	6	52
Précision	0.9	0.87
Rappel	0.87	0.9

(b) **Économie**

	$\mathcal{T}_{int}$	$\mathcal{O}_{int}$
$\mathcal{C}at_{\mathcal{T}}$	52	5
$\mathcal{C}at_{\mathcal{O}}$	8	55
Précision	0.87	0.92
Rappel	0.91	0.87

(c) **Jeu vidéo**

	$\mathcal{T}_{int}$	$\mathcal{O}_{int}$
$\mathcal{C}at_{\mathcal{T}}$	56	2
$\mathcal{C}at_{\mathcal{O}}$	4	58
Précision	0.93	0.97
Rappel	0.97	0.94

(d) **Gastronomie**

	$\mathcal{T}_{int}$	$\mathcal{O}_{int}$
$\mathcal{C}at_{\mathcal{T}}$	55	8
$\mathcal{C}at_{\mathcal{O}}$	5	52
Précision	0.92	0.87
Rappel	0.87	0.91

(e) **Sports**

	$\mathcal{T}_{int}$	$\mathcal{O}_{int}$
$\mathcal{C}at_{\mathcal{T}}$	46	10
$\mathcal{C}at_{\mathcal{O}}$	14	50
Précision	0.77	0.83
Rappel	0.82	0.78

(f) **Tourisme**

Table 3.18: Mesures de performances de SelfDS_New par intérêt cible.

$\mathcal{C}at_{\mathcal{T}}$ est la classe des utilisateurs s'intéressant à l'intérêt cible $\mathcal{T}_{int}$. $\mathcal{C}at_{\mathcal{O}}$ est la classe des utilisateurs s'intéressant à l'un des intérêts de $\mathcal{O}_{int}$ qui représente nos six intérêts initialement choisis à l'exception de l'intérêt cible.

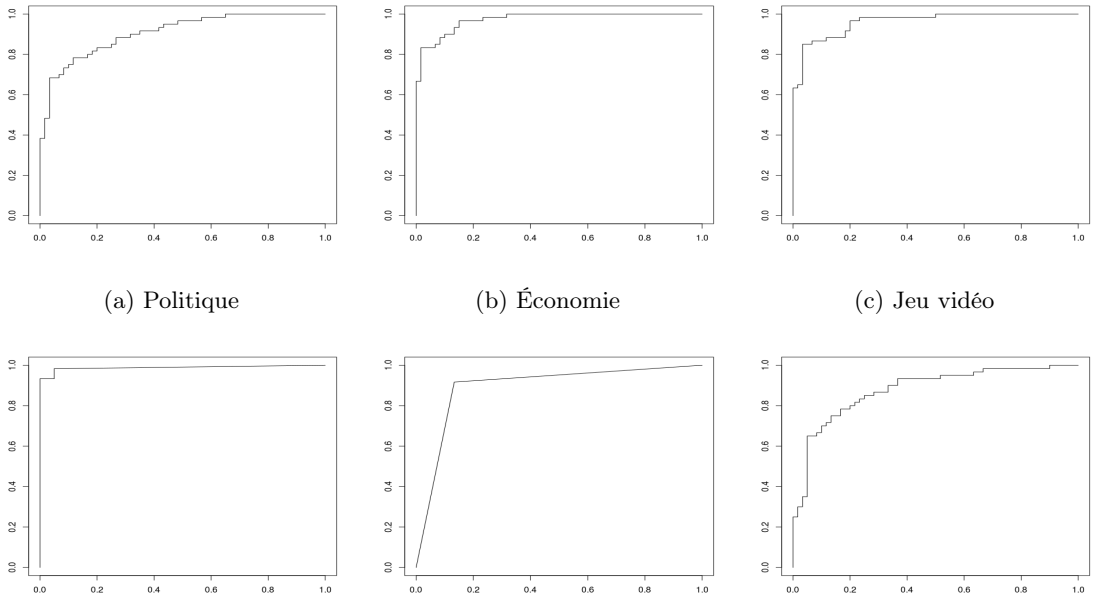
fonction du taux de faux positifs (fraction d'utilisateurs appartenant à la catégorie $\mathcal{C}at_{\mathcal{O}}$ alors qu'ils appartiennent réellement à la catégorie $\mathcal{C}at_{\mathcal{T}}$ et vice verca) obtenus à l'issue de la classification. Plus la courbe est proche de l'axe de sensibilité (axe des ordonnées) plus le modèle de prédiction qui le représente est meilleur. L'évaluation d'un modèle de régression à l'aide d'une courbe ROC consiste à calculer l'aire de la surface existante entre la courbe et la droite reliant les points de coordonnées (0,0) et (1,1). Traditionnellement, il existe cinq interprétations différentes en fonction des valeurs obtenues. Si l'aire est comprise entre :

1. $]0.90, 1]$: modèle excellent, c'est-à-dire prédit bien les valeurs attendues.
2. $]0.80, 0.90]$: modèle bien
3. $]0.70, 0.80]$: modèle assez bien
4. $]0.60, 0.70]$: modèle médiocre
5. $]0.50, 0.60]$: modèle d'échec.

Le tableau 3.19 présente les valeurs des aires sous la courbe ROC des modèles de prédiction obtenus en prenant respectivement comme intérêt cible "Politique", "Économie", "Jeu vidéo", "Gastronomie", " Sports" et "Tourisme". Au vue des résultats ont peu dire que ces modèles construits sont statistiquement "bien" et voir "excellent" pour le cas de l'intérêt cible "Gastronomie". Les courbes ROC associées à chaque modèle de prédiction sont visibles sur la figure 3.10. En abscisse on a la spécificité et en ordonné la sensibilité. Ces courbes confirment nos interprétations puisque la courbe qui est beaucoup plus proche de l'axe de sensibilité est celle de l'intérêt cible "Gastronomie"

	Politique	Économie	Jeu vidéo	Gastronomie	Sports	Tourisme
Aire ROC	0.81	0.88	0.89	0.95	0.89	0.8

Table 3.19: Aire sous la courbe ROC de chaque modèle de prédiction et par intérêt cible



(a) Politique (b) Économie (c) Jeu vidéo

(d) Gastronomie (e) Sports (f) Tourisme

Figure 3.10: Courbes ROC de chaque catégorie cible obtenues à partir de SelfDS_New.

Plus la courbe est proche de l'axe de sensibilité plus le modèle de prédiction qui le représente est meilleur. En abscisse on a la spécificité et en ordonné la sensibilité.

qui correspondent à 0.95 d'aire. De même, visuellement on constate que c'est la courbe de l'intérêt "Tourisme" qui est la moins bonne ce correspond à l'aire 0.8.

Au vue des performances des modèles de prédiction par intérêt construits sur SelfDS_New on peut reconfirmer notre hypothèse qu'il existe une relation entre les dimensions psychologiques Receptiviti et les intérêts d'un utilisateur. Dans ce qui suit on va essayer de comparer Ascertain avec les méthodes de classification les plus populaires.

3.3.4 Comparaison d'Ascertain avec les méthodes de classification

Pour la comparaison d'Ascertain avec les méthodes de classification existantes dans la littérature, nous sélectionnons tout d'abord deux classifieurs de texte les plus répandus, à savoir les machines à vecteurs de support ou SVM, la classification naïve bayésienne ou tout simplement Naive Bayes. Ensuite, nous utilisons les données construites précédemment à savoir Big5D_Training, Big5SD_Training et recepD_Training pour l'apprentissage (à partir de laquelle les classi-

fieurs se basent pour construire leurs modèles respectifs) et BIG5D_TEST, BIG5SD_TEST et RECEPD_TEST pour la prédiction. La distribution des utilisateurs selon les données d'apprentissage et de validation est visible dans le tableau 3.9. Et enfin, nous avons comparé les résultats obtenus à ceux d'ASCERTAIN. Nous notons que le modèle utilisé ici par ASCERTAIN pour prédire les intérêts des utilisateurs est le modèle saturé, c'est-à-dire celui qui tient compte de toutes les variables explicatives de chaque jeu de données. Nous choisissons le modèle saturé pour éviter de pénaliser les méthodes de classification.

La figure 3.20 montre les résultats d'ASCERTAIN et les méthodes SVM et Naives Bayes sur les données BIG5D. Même si les taux de performances sont très faibles (moins de 75%) pour toutes les trois méthodes et que la f-mesure d'ASCERTAIN est comparable à ceux de SVM, nous constatons que SVM affecte mieux les intérêts aux utilisateurs par rapport aux autres.

		Politique	*Économie*	*Jeu vidéo*	*Gastronomie*	*Sports*	*Tourisme*	*Avg*
ASCERTAIN	**P**	0.48	0.44	**0.58**	0.50	0.44	0.49	0.49
	R	0.40	0.47	**0.74**	**0.48**	0.44	0.44	0.49
	F	0.44	0.45	**0.65**	0.49	0.44	0.46	0.49
SVM	**P**	**0.56**	**0.46**	**0.58**	**0.52**	**0.51**	**0.56**	**0.54**
	R	**0.44**	**0.52**	0.71	**0.48**	**0.56**	**0.48**	**0.53**
	F	**0.50**	**0.49**	0.64	**0.50**	**0.54**	**0.52**	**0.53**
Bayes	**P**	0.49	0.37	0.53	0.46	0.43	0.45	0.46
	R	0.32	0.51	0.63	0.40	0.39	**0.48**	0.45
	F	0.39	0.43	0.58	0.42	0.41	0.47	0.45

Table 3.20: ASCERTAIN Vs. SVM et Naives Bayes sur BIG5D.

Quant au tableau 3.21, elle représente les résultats d'ASCERTAIN et les méthodes SVM et Naives Bayes sur les données BIG5SD. Nous constatons que les résultats s'améliorent pour chaque méthode. Bien que SVM est la meilleur méthode on voit qu'elle est comparable à ASCERTAIN.

Cependant, le tableau 3.22 montre les résultats d'ASCERTAIN et les méthodes SVM et Naives Bayes sur les données RECEPD. Nous constatons que plus on a de variables explicatives plus les performances de SVM diminuent par rapport à ceux d'ASCERTAIN. Par contre, ceux de Naives Bayes sont stables.

		Politique	*Économie*	*Jeu vidéo*	*Gastronomie*	*Sports*	*Tourisme*	*Avg*
Ascertain	**P**	0.70	**0.75**	**0.77**	0.61	**0.79**	0.65	0.71
	R	**0.70**	0.79	**0.75**	0.61	0.77	0.66	0.71
	F	0.70	**0.77**	**0.76**	0.61	0.78	0.65	0.71
SVM	**P**	**0.74**	0.73	0.69	**0.64**	0.78	**0.71**	**0.72**
	R	**0.70**	**0.81**	0.66	**0.62**	**0.82**	**0.69**	**0.72**
	F	**0.72**	**0.77**	0.68	**0.63**	**0.80**	**0.70**	**0.72**
Bayes	**P**	0.62	0.64	0.56	0.46	0.77	0.50	0.60
	R	0.57	0.73	0.58	0.48	0.65	0.54	0.59
	F	0.59	0.68	0.57	0.47	0.71	0.52	0.59

Table 3.21: Ascertain Vs. SVM et Naives Bayes sur Big5SD.

		Politique	*Économie*	*Jeu vidéo*	*Gastronomie*	*Sports*	*Tourisme*	*Avg*
Ascertain	**P**	**0.77**	**0.76**	**0.74**	0.86	**0.84**	0.75	**0.79**
	R	**0.74**	0.78	**0.75**	**0.88**	0.84	0.73	**0.79**
	F	**0.76**	**0.77**	**0.74**	**0.87**	**0.84**	**0.74**	**0.79**
SVM	**P**	0.73	0.73	0.68	**0.87**	0.81	**0.77**	0.77
	R	0.69	0.78	0.72	0.86	**0.85**	0.69	0.77
	F	0.71	0.76	0.70	0.86	0.83	0.73	0.77
Bayes	**P**	0.70	0.69	0.59	0.86	0.82	0.54	0.70
	R	0.59	**0.79**	0.62	0.65	0.67	**0.76**	0.68
	F	0.64	0.74	0.61	0.74	0.74	0.63	0.68

Table 3.22: Ascertain Vs. SVM et Naives Bayes sur recepD.

Conclusion

Nous avons consacré ce chapitre à la description de notre méthodologie nommée Ascertain qui vise à faire une étude sur la corrélation entre les intérêts et les traits de personnalité des individus. Pour le faire, nous avons défini une méthode incrémentale basée essentiellement sur la régression logistique, qui se déroule en cinq étapes : la collecte des données, l'exploration des données, la sélection des variables, la construction et l'évaluation du modèle de prédiction et enfin l'interprétation des résultats obtenus.

L'étape de collecte de données consiste à construire notre base de données que nous avons nommée SelfDS à laquelle nous appliquons la méthodologie Ascertain. SelfDS contient d'une part, 1 446 vecteurs de scores des utilisateurs du réseau social Twitter sélectionnés à partir du moteur de recherche Twitter, et d'autre part, une variable expliquée qui représente les intérêts des utilisateurs. Chaque vecteur de scores contient 59 variables explicatives qui correspondent au 59 dimensions psychologiques de Receptiviti. Nous avons utilisé d'une part, l'API Twitter pour extraire 500 tweets maximum par utilisateur sélectionné, et d'autre part, l'API Receptiviti pour calculer les valeurs des dimensions psychologiques de chaque utilisateur sélectionné à partir de son document de tweets. Pour évaluer la contribution de chaque dimension psychologique, nous avons construit trois jeux de données Big5D, Big5SD et recepD qui contiennent chacune les mêmes utilisateurs que SelfDS, mais par contre 5, 35 et 59 variables explicatives respectivement.

L'étape d'exploration des données permet de s'assurer que les données sont plus ou moins complètes et de mesurer les corrélations existantes entre variables explicatives par le biais d'une analyse en composantes principales. Cette analyse en composantes principales nous a permis de découvrir les variables qui sont fortement corrélées entre elles comme les variables "extraversion", "cheerful", "happinness" et "persuasive".

Pour la sélection des variables, nous avons utilisé la technique de sélection descendante basée essentiellement sur le critère AIC. En fait, cette étape nous a permis d'identifier le sous-ensemble réduit de variables explicatives qui permet de caractériser les utilisateurs tout en minimisant les pertes d'informations.

Quant à l'étape de construction du modèle de prédiction, elle consiste à utiliser la régression logistique polytomique multiple avec comme variables explicatives celles issues de l'étape de sélection de variables pour construire le modèle que nous utilisons pour prédire les intérêts des utilisateurs.

Et enfin, l'étape d'évaluation et d'interprétation des résultats du modèle construit consiste d'une part, à estimer la capacité du modèle construit à bien prédire les intérêts des utilisateurs, et d'autre part, à déduire des observations ou tendances que présentent les résultats. Pour le faire, nous nous sommes appuyés sur les mesures de performances telles que la précision et le rappel. La précision globale obtenue pour le modèle construit à partir des données de Big5D est de 0.49, pour ceux de Big5SD est de 0.71 et pour ceux de recepD est de 0.79. Nous constatons une nette amélioration de la capacité des modèles à prédire les intérêts des utilisateurs. Les résultats obtenus sur les données recepD sont meilleurs et montrent que les utilisateurs qui s'intéressent à l'économie sont moins enclins à satisfaire les autres par rapport à ceux qui s'intéressent à la "Politique" et que les utilisateurs qui s'intéressent à la catégorie "Gastronomie" apprécient de discuter de nourriture ou de boissons avec d'autres personnes que les politiciens. Nous avons également exploré la régression logistique multiple à la place de la régression logistique polytomique multiple. A l'issue de cette analyse nous avons obtenu des résultats très prometteurs puisque certaines catégories cibles "Gastronomie" et "Sport" ont atteint une précision de 0.95 et 0.89, respectivement.

En définitive, au vue des résultats obtenus sur notre jeu de données SelfDS nous pouvons dire

qu'il existe certaines dimensions psychologiques Receptiviti qui sont statistiquement corrélées aux intérêts d'un individu.

Chapitre 4

RÉCONCILIATION DES PROFILS UTILISATEURS DANS LES RÉSEAUX SOCIAUX

Introduction

Les évolutions technologiques de l'Internet, avec notamment le développement du Web 2.0, sont à l'origine de l'émergence des services de réseautage social, et par la suite ont fait d'eux des outils facilement accessibles sur le Web. A cet effet, on assiste de plus en plus à une prolifération de ces services, qui mettent l'utilisateur, en tant que créateur de contenu, au centre de leurs préoccupations. Un service de réseautage social se rapporte à l'ensemble des moyens virtuels mis en œuvre pour réunir des personnes physiques ou morales, tels que les réseaux sociaux.

En effet, un *réseau social* est un ensemble d'individus connectés par des liens, leurs permettant de partager des ressources comme des vidéos (YOUTUBE), des photos (FLICKR) ou des ressources annotées (Del.icio.us), d'échanger des informations et de construire des relations personnelles ou professionnelles (FACEBOOK, LINKEDIN) ou encore de diffuser des news (TWITTER, blogs). Chaque utilisateur dans un réseau social possède un espace personnel ou *profil utilisateur*, qui contient ses informations personnelles telles que son nom et prénom, son lieu de résidence, son âge, ses adresses électroniques, ses numéros de téléphone, ses relations, ses institutions fréquentées, ses intérêts, etc [30, 44, 78].

Selon la politique d'accès aux données définie par chaque réseau social, certaines informations du profil d'un utilisateur sont soient publiques, c'est-à-dire accessibles à tous, soient semi-publiques c'est-à-dire accessibles uniquement aux utilisateurs du réseau social qui le contient, soient privées c'est-à-dire visibles uniquement par cet l'utilisateur. Cependant, un utilisateur peut choisir de restreindre l'accès à ses informations, uniquement, à ses contacts, ou les laisser disponibles à tout le monde. Avec le foisonnement des services de réseautage, les utilisateurs possèdent de plus en plus différents profils, dans plusieurs réseaux sociaux distincts.

Par ailleurs, la diversité de profils utilisateurs appartenant à une même personne physique à travers le Web est également due au fait que chaque réseau social a son orientation personnelle. Ainsi, en fonction des objectifs que l'utilisateur cherche à atteindre, il est amené à créer de nouveaux profils. Par exemple, lorsqu'une personne cherche un emploi, elle s'intéresse aux moyens lui permettant soit d'entrer en contact avec des recruteurs, soit d'avoir accès aux offres d'emplois qui l'intéressent. Et par conséquent, elle peut se tourner vers les sites tels que *LinkedIn*, qui est un réseau social purement professionnel permettant, non seulement à ses utilisateurs de mettre en ligne leur curriculum vitae, mais aussi aux recruteurs d'y avoir accès. De même, si une personne cherche plutôt ses anciens amis du collège ou du lycée, elle va se tourner vers le réseau social *Copains d'avant*, dont le but est de permettre aux participants de retrouver d'anciens camarades avec qui, elles ont partagé leur scolarité (à l'école primaire, au collège, au lycée et dans les cursus universitaires), ainsi que leurs activités associatives et professionnelles (entreprises, administrations). Ainsi de proche en proche, on observe sur la toile l'apparition d'une situation selon laquelle, une même personne possède plusieurs profils dans différents réseaux sociaux.

Les différents profils d'un même utilisateur sont des sources de données insoupçonnables qui contiennent des ressources diverses, variées, redondantes ou complémentaires très précieuses. L'exploitation de ces ressources permet de construire, pour chaque utilisateur, un profil global très utile pour des applications, telles que les systèmes de recommandations. Il existe déjà, sur le Web, des services permettant d'agréger différents profils sociaux comme FRIENDFEED[1], PLAXO[2] ou SPOKEO[3], qui brassent un large éventail d'informations concernant une personne et fournissent des liens vers ses profils ou ses ressources Web classées par exemple par vidéos, photos, blogs ou tweets. Cependant, ces informations sont souvent mêlées avec celles d'autres personnes, comme par exemple des

[1] www.friendfeed.com [2] www.plaxo.com [3] www.spokeo.com

personnes portant le même nom.

Dans ce chapitre, nous nous intéressons à la réconciliation des profils utilisateurs dans plusieurs réseaux sociaux. Autrement dit, nous proposons une méthode automatique, de découverte des différents profils d'un même utilisateur à travers plusieurs réseaux sociaux. Le travail de ce chapitre a été fait en collaboration avec notre collègue Mohammad Ghufran [65]. La mise en œuvre de notre approche passe par un ensemble de challenges à relever tels que :

1. Compte tenu de la dimension internationale des réseaux sociaux, lors de la création de leurs profils, certains utilisateurs ont tendance à être le plus discret possible, dans le but de se préserver contre certaines attaques éventuelles comme l'usurpation d'identité. Par conséquent, ces utilisateurs s'arrangent à ne renseigner exactement que le strict minimum nécessaire pour la génération de leurs profils respectifs, ce qui est parfois à l'origine de l'existence des profils utilisateurs contenant des informations incomplètes, non mises à jour, et parfois fausses. Par exemple, lors des différentes variations des états de la vie d'un utilisateur telles qu'un changement d'école ou d'emploi, ce dernier ne les mentionne parfois pas dans ses multiples profils. De même, il existe certains utilisateurs qui décident d'omettre, de leurs profils, leur année de naissance et ne renseignent que leur jour et mois de naissance. D'autres vont même plus loin en indiquant des informations erronées. Tous ces agissements ont une répercussion négative sur le processus de réconciliation des utilisateurs puisqu'ils rendent l'analyse des données plus complexe.

2. Par ailleurs, les utilisateurs éditent généralement leurs profils en langage naturel dans un vocabulaire libre. De ce fait, certains mots ou expressions employés sont souvent ambigus. Par ailleurs, ils peuvent avoir plusieurs significations ou interprétations possibles. Par exemple, prenons un utilisateur qui renseigne comme ville de résidence "Paris". L'expression "Paris" est ambiguë, car elle peut faire référence, entre autres, soit à la capitale de la France, soit à une ville des États-Unis, située dans le nord-est du Texas, ou soit à une ville du sud de l'Ontario au Canada, dans le comté de Brant. Si l'utilisateur n'a pas associé à sa ville "Paris" son pays de résidence, alors choisir la bonne interprétation dont il veut faire référence en mentionnant "Paris" n'est pas une tâche évidente.

3. De plus, au vu des moyens rapides et faciles mis en œuvre pour accéder aux réseaux sociaux à l'échelle internationale, plusieurs individus dans le monde possèdent au moins un profil utilisateur. Le nombre d'utilisateurs des réseaux sociaux ne cesse de croître à une vitesse impressionnante. De ce fait, dénicher parmi la masse de profils existants dans un réseau, celui appartenant à un utilisateur dont on dispose de son profil dans un autre réseau social, devient une opération complexe.

Nous proposons une solution automatique Liaison[4] [65] qui va au delà des difficultés sus-citées. Nos contributions sont les suivantes :

1. Nous proposons une approche qui utilise une stratégie remarquable pour sélectionner automatique un sous-ensemble de paires de profils candidats susceptibles d'être réconciliés, ce qui réduit considérablement le nombre de candidats par rapport à une comparaison à l'ensemble des utilisateurs d'un réseau social. La stratégie employée est basée essentiellement sur la topologie de l'ensemble des réseaux sociaux interconnectés ciblés.

[4] reconciLIAtion of Individuals profiles across SO- cial Networks

2. Nous présentons aussi une collection de règles combinant un ensemble d'informations extraites des profils utilisateurs, afin de capturer parmi les couples de profils candidats sélectionnés ceux faisant référence à une même personne. Plus précisément, chaque règle exprime la contribution d'un ou plusieurs éléments d'informations ou attributs utilisés pour mettre en correspondance deux profils. Plus le nombre d'attributs impliqués dans une règle est important, plus la probabilité que les deux profils désignent une même personne est forte.
3. Nous décrivons également une méthode de désambiguisation des localités, afin de rapprocher les valeurs des lieux géographiques de chaque paire de profils utilisateurs. La méthode proposée utilise principalement la base de données géographiques libre OpenStreetMap[5].
4. Et enfin, nous définissons une approche qui s'applique sur n réseaux sociaux sans avoir à faire des comparaisons par paires de réseaux. En d'autres termes, nous ne faisons pas de rapprochement de profils par couples de réseaux successifs, nous travaillons tout d'un coup sur l'ensemble des réseaux interconnectés.

Le plan de ce chapitre se présente comme suit :

Tout d'abord, nous présentons dans la section 1, un aperçu sur les travaux de recherche portant sur le problème de réconciliation des utilisateurs dans les réseaux sociaux. Nous poursuivons dans la section 2 avec, d'une part, la description des notations permettant la formalisation du problème posé, et d'autre part, une présentation des différents attributs extraits des profils utilisateurs et qui nous ont servi à la réconciliation. Dans la section 3, nous détaillons les différentes étapes de notre approche de réconciliation Liaison, ainsi que les expérimentations menées sur une collection de données réelles provenant de quatre réseaux sociaux et enfin, nous concluons.

1 État de l'art

Étant donné plusieurs réseaux sociaux interconnectés, le problème de réconciliation des profils utilisateurs entre ces différents réseaux consiste à une recherche des divers profils utilisateurs appartenant à une même personne. Ce problème a été à l'origine de nombreux travaux, et grâce à différents chercheurs d'important solutions ont vu le jour.

Parmi les propositions, deux d'entre elles [63, 90] se concentrent uniquement sur l'attribut pseudonyme pour faire de la réconciliation des profils. En effet, l'attribut pseudonyme est un élément d'information renseigné par l'utilisateur qui permet de le distinguer des autres utilisateurs. L'idée est basée sur l'observation selon laquelle les individus ont tendance à utiliser des pseudonymes similaires ou identiques à travers différents réseaux sociaux. Même si, dans notre évaluation, nous confirmons cette observation, nous considérons également d'autres attributs, afin de découvrir les profils utilisateurs qui choisissent d'utiliser des pseudonymes dissimilaires.

L'utilisation des attributs pour faire la réconciliation des profils utilisateurs à travers différents réseaux sociaux a été largement étudiée [66, 72, 17, 46, 13, 23, 25, 51]. Deux de ces approches [46, 51] représentent chaque paire de profils comme des vecteurs de scores, qui décrivent la similarité entre les valeurs des attributs cibles et utilisent les techniques d'apprentissage pour déterminer si elles peuvent être réconciliées. Même si les résultats sont encourageants, ces deux approches nécessitent des données d'apprentissage, qui ne sont toujours pas faciles à construire. Les techniques d'apprentissage nécessitent une analyse minutieuse des données disponibles pour créer un échantillon assez représentatif de toutes les situations possibles où les paires de profils peuvent être réconciliées

[5] http://www.openstreetmap.org/

ou pas. De plus, un modèle d'apprentissage entraîné sur un ensemble de paires de profils de deux réseaux sociaux n'est pas généralisable sur des couples de profils de deux autres réseaux distincts quelconques. Ce qui implique une construction des données d'apprentissage pour chaque couple de réseaux sociaux.

Certains réseaux sociaux permettent l'exportation des profils décrits à partir de l'ontologie Friend of a Friend (FOAF), afin d'utiliser les techniques du Web sémantique, telles que le raisonnement OWL, pour réconcilier les profils [25, 72]. Cependant, ces techniques sont appliquées seulement à une collection limitée d'attributs, et en particulier à ceux tels que l'adresse électronique, qui sont des éléments d'informations susceptibles d'identifier de façon unique un individu.

Comme nous, Carmagnola et al. [13] déterminent les attributs qui sont les plus susceptibles d'identifier un individu de manière unique. En réalité, ils attribuent aux valeurs d'attributs de deux profils un facteur d'importance, qui est un score de similarité qui représente le degré de ressemblance existant entre ces valeurs. Notre approche va plus loin en utilisant les paires de profils découverts pour réconcilier itérativement de nouveaux profils. De plus, notre évaluation est basée sur un jeu de données réels provenant de plusieurs réseaux sociaux interconnectés, tandis que le leur utilise différents systèmes fermés. La principale différence entre ces deux types de données est que, dans les réseaux sociaux, les utilisateurs sont souvent très réticents quand il s'agit de révéler leurs identités réelles, alors que dans les systèmes fermés, ces derniers estiment que leur vie privée n'est pas mise en avant (publique) et sont donc moins discrets. Ainsi, les données des réseaux sociaux sont susceptibles d'être erronées, incomplètes et voire même non mises à jour, ce qui constitue un véritable challenge pour nous. Certains chercheurs [17, 66] proposent également le calcul d'une similarité sémantique entre les valeurs des attributs issus des profils. Bien que, ces approches soient originales, ils fournissent peu (50 profils d'utilisateurs [66]) ou aucune évaluation.

Certains auteurs vont au-delà de l'utilisation des attributs provenant des profils et examinent plutôt la possibilité d'utiliser les propriétés du réseau [3, 10, 33, 52]. L'approche proposée par Buccafurri et al.[10] considère que deux profils sont similaires, et donc susceptibles de faire référence à un même individu, s'ils ont des pseudonymes similaires et les profils auxquels ils sont liés sont récursivement similaires. En d'autres termes, deux profils appartiennent à une même personne si leur voisinage contient des couples de profils des mêmes personnes de manière récursive. Cette approche présente deux inconvénients majeurs. Tout d'abord, les profils associés à des pseudonymes dissimilaires sont ignorés sans aucune autre analyse, bien qu'ils peuvent très bien appartenir à une même personne. Deuxièmement, les profils découverts ne sont pas utilisés pour retrouver de nouveaux profils faisant référence à la même personne. Notre approche passe au travers de ces deux limites.

Quant à Jain et al.[33], ils tiennent compte aussi de la structure du réseau et proposent en plus d'utiliser le contenu sous forme de textes courts que chaque utilisateur publie. Plus clairement, cette approche exploite d'une part le contenu textuel généré par les utilisateurs, et d'autre part, leurs connexions avec d'autres utilisateurs pour faire de la réconciliation. Cependant, les expériences faites révèlent que cette manière de procéder n'est pas très efficace, car seulement 4 profils sur 543 sont réconciliés correctement.

Une autre approche, celle de Bartunov et al.[3] combine les attributs des profils et la structure du réseau en se basant sur une classe de modèles statistiques utilisés en reconnaissance des formes et plus généralement en apprentissage statistique, les champs aléatoires conditionnels. L'avantage principal de cette approche est qu'elle est robuste, car elle peut être appliquée en absence d'une partie ou de la totalité des informations sur les profils, mais est à l'origine d'une baisse significative du rappel dû principalement au fait que certains attributs ne sont pas renseignés. L'inconvénient ici est

que le modèle proposé nécessite des données d'apprentissage qui, comme mentionné précédemment, peuvent ne pas être faciles à obtenir.

Narayanan et al.[52] travaillent sur les réseaux anonymes où les profils contiennent peu ou pas d'attributs mais la structure du réseau est disponible. Ils proposent une méthode qui sélectionne tout d'abord un petit échantillon de paires de profils dans deux réseaux qui sont très susceptibles d'appartenir à des mêmes personnes. Ensuite, ils découvrent de nouveaux profils qui sont propagés itérativement en utilisant l'échantillon initial. Cette approche est semblable à la nôtre, mais comme elle n'utilise que la structure du réseau, sa précision est assez faible par rapport à la nôtre.

Enfin, nous avons les systèmes d'agrégation des réseaux sociaux, tels que FriendFeed [21] Ou Plaxo [64], qui fournissent une plate-forme pour que leurs utilisateurs puissent gérer leurs différents profils. Ils n'essayent pas de découvrir automatiquement des profils utilisateurs liés à une même personne à travers plusieurs réseaux sociaux. Quant à Spokeo [75], il semble être assez précis en trouvant des informations personnelles provenant de différentes sources (pas nécessairement celles des réseaux sociaux), néanmoins il présente des limites quand il s'agit de les agréger.

Dans la suite, nous présentons les notions utilisées pour formaliser le problème posé, ainsi que les différents attributs utilisés dans notre approche.

2 Préliminaires

Le problème de réconciliation des profils, qui est l'objectif de ce chapitre, fait intervenir d'une part, plusieurs services de réseautage social interconnectés, et d'autre part, divers profils utilisateurs appartenant à une ou plusieurs personnes. La formalisation du problème se présente donc ainsi.

2.1 Notations et formalisation

Les services de réseautage social utilisés dans notre approche sont une collection de n réseaux sociaux distincts, que nous représentons par un graphe orienté étiqueté, où les nœuds représentent les profils ou plus précisément les utilisateurs, et les arcs, les différents types de liens existants entre les utilisateurs. Chaque profil utilisateur possède un identifiant ou *uri* permettant d'y avoir accès sur le Web et est constitué d'un ensemble d'attributs généralement décrits par l'utilisateur. En effet, les attributs sont des éléments d'informations tels que le pseudonyme, nom, prénom, adresse email, etc. En outre, nous considérons deux types de liens : les *liens d'amitié* existant entre les utilisateurs d'un même réseau social et les *liens transversaux* reliant deux profils d'un même utilisateur appartenant à deux réseaux différents.

Plus formellement, un ensemble de n réseaux sociaux est un graphe défini comme suit :

$$\mathcal{G} =< \bigcup_{i=1}^{n} V_i \ , \ \bigcup_{i=1}^{n} E_i \ , \ \bigcup_{i,j=1}^{n} E_{i,j} > \tag{4.1}$$

où:

- V_i est l'ensemble des nœuds du réseau social i. Comme les réseaux sociaux sont distincts, on a $\forall i, j,\ i \neq j,\ V^i \cap V^j = \phi$. Chaque nœud $v_i \in V_i$ représente le profil d'un utilisateur dans le réseau social i. Nous notons A, l'ensemble des attributs, qui peuvent être à valeurs multiples, définis dans un profil utilisateur, et $P_a(v_i)$, la(les) valeur(s) associée(s) à un attribut $a \in A$ d'un profil v_i.

- E_i est l'ensemble des arcs étiquetés avec le label $friend$. Plus clairement, chaque arc noté $(v_i, friend, u_i) \in E_i$ représente un lien d'amitié existant entre un utilisateur associé au profil v_i et un autre utilisateur lié au profil u_i dans le réseau social i. Nous notons $friends(v_i) = \{u_i \mid (v_i, friend, u_i) \in E_i \vee (u_i, friend, v_i) \in E_i\}$ l'ensemble des utilisateurs reliés à v_i par des liens d'amitié dans le réseau social i.
- $E_{i,j}$ est l'ensemble des arcs transversaux c'est-à-dire étiquetés avec le label me. Plus formellement, nous notons (v^i, me, v^j) un lien transversal entre les profils v_i et v_j d'un même utilisateur appartenant, soit à un même réseau social (lien transversal intra-réseau), soit à deux réseaux distincts (lien transversal extra-réseau). Par définition, les liens transversaux sont symétriques et transitifs. Par exemple, considérons l'utilisateur *Bob* qui déclare dans son profil v_{fk} du réseau social FLICKR (fk) l'uri de son profil v_{lj} dans le réseau LIVEJOURNAL (lj) et sur cette page il déclare l'uri de son profil v_{tw} dans TWITTER. Dans ce cas, on a:

$$E_{fk,lj} = \{(v_{fk}, me, v_{lj}), (v_{lj}, me, v_{fk})\} \text{ et } E_{tw,lj} = \{(v_{tw}, me, v_{lj}), (v_{lj}, me, v_{tw})\}$$
$$\Rightarrow \quad E_{tw,fk} = \{(v_{tw}, me, v_{fk}), (v_{fk}, me, v_{tw})\}$$

Le problème de réconciliation des profils d'un même utilisateur à travers différents réseaux sociaux se ramène à un problème de détermination des liens transversaux ou me manquants dans une collection de réseaux sociaux interconnectés et formalisé comme suit :

$$\textbf{Entrée: } \mathcal{G} =< \bigcup_{i=1}^{n} V_i \ , \ \bigcup_{i=1}^{n} E_i \ , \ \bigcup_{i,j=1}^{n} E_{i,j} >$$

$$\textbf{Sortie: } \mathcal{G} =< \bigcup_{i=1}^{n} V_i \ , \ \bigcup_{i=1}^{n} E_i \ , \ \bigcup_{i,j=1}^{n} E_{i,j} \ , \ \bigcup_{i,j=1}^{n} D_{i,j} >$$

où $D_{i,j} = \{(v_i, me, v_j) \mid v_i \in V_i \ , v_j \in V_j, (v_i, me, v_j) \notin E_{i,j}\}$ est l'ensemble des liens transversaux découverts. Pour plus de simplicité, nous écrivons dans la suite E_{me} et D_{me} pour faire référence successivement à $\bigcup_{i,j=1}^{n} E_{i,j}$ et à $\bigcup_{i,j=1}^{n} D_{i,j}$.

Pour mettre ensemble deux profils, nous avons fait une analyse des différents attributs permettant de caractériser un utilisateur. La description des attributs sélectionnés, ainsi que les mesures de similarité utilisées pour comparer les valeurs de chaque attribut sont l'objet de la section suivante.

2.2 Description des attributs

Les profils utilisateurs contiennent beaucoup d'informations renseignées par l'utilisateur et décrites sous forme d'attributs associés à des valeurs. Dans la majorité des réseaux sociaux, certains attributs sont rendus publics soit par défaut, soit l'utilisateur choisit de les publier. Des chercheurs [41] ont identifié, à partir de 12 réseaux sociaux les plus répandus, un ensemble d'attributs provenant des profils utilisateurs qui sont pour la plupart du temps publics. L'idée est de comparer pour chaque paires de profils candidats, les valeurs de leurs attributs deux à deux. Plus précisément, les valeurs $P_a(v_i)$ et $P_a(v_j)$ d'un attribut $a \in A$ sont comparés à l'aide d'une mesure de similarité, qui renvoie un score compris entre 0 (valeurs d'attributs dissimilaires) et 1 (valeurs d'attributs identiques). Pour les scores intermédiaires, on dit que deux valeurs d'attributs correspondent si leur score de similarité est supérieur à un seuil fixé θ_a. Nous discutons des valeurs des seuils de similarité de chaque attribut dans la partie expérimentation. A l'issue de notre analyse, nous avons décidé de nous focaliser essentiellement sur les attributs suivants : PSEUDONYME, NOMS, LOCALITE, EMAILS, PROFILES et WEBSITES.

2.2.1 L'attribut Pseudonyme

Noté u, c'est l'identifiant d'un profil utilisateur dans un réseau social et sa valeur est toujours accessible publiquement. Généralement, il fait parti de l'*uri* de la page du profil sur le Web. Des études comme dans [10, 63, 90] ont montré que les utilisateurs des réseaux sociaux ont tendance à utiliser des pseudonymes identiques ou comportant des sous-chaînes identiques dans leurs différents profils. De ce fait, la similarité entre deux pseudonymes est mieux représentée par la distance de *Levenshtein*.

La distance de *Levenshtein* permet de capturer les différentes variations d'une sous-chaîne de caractères en calculant le nombre de modifications de caractères possibles dans un mot pour qu'il soit identique à un autre mot. Plus formellement, la similarité entre deux pseudonymes $P_u(v_i)$ et $P_u(v_j)$ se calcule comme suit :

$$sim_u(P_u(v_i), P_u(v_j)) = 1 - \frac{LevenshteinDistance(P_u(v_i), P_u(v_j))}{length_{max}(P_u(v_i), P_u(v_j))} \tag{4.2}$$

où $length_{max}$ est le nombre de caractères du plus long mot entre $P_u(v_i)$ et $P_u(v_j)$.

Prenons, par exemple, les pseudonymes de deux profils référençant un même utilisateur, l'un *cospics* du réseau social Flickr `www.flickr.com/photos/cospics` et l'autre *cos* du réseau social Livejournal `www.livejournal.com/users/cos/profile`. La distance de *Levenshtein* entre *cospics* et *cos* est de 4, car il faut supprimer du premier pseudonyme l'expression "pics" composé de quatre caractères pour obtenir le second pseudonyme. Et par conséquent, leur score de similarité est de $0,43$.

2.2.2 L'attribut Noms

Noté n, c'est un attribut dont la valeur représente les noms et/ou prénoms renseignés par l'utilisateur. Ces valeurs ne correspondent généralement pas à des champs bien identifiés et ne sont pas souvent renseignés avec le même niveau de détail d'un réseau social à un autre. De ce fait, le nom est considéré comme un attribut ambigu, en particulier, lorsque les valeurs correspondent à des noms et prénoms communs. De même, il est également sensible dans la mesure où les utilisateurs préfèrent le plus souvent rester anonymes. Toutefois, même s'ils le renseignent, ils ne révèlent pas leur véritable identité. En pratique dans certains réseaux sociaux, tels que Livejournal, les profils utilisateurs sont presque entièrement publics, en conséquence les utilisateurs ne se sentent pas en sécurité s'ils révèlent leur vraie identité. Il est donc clair que l'utilisation seule de cet attribut est insuffisante pour réconcilier deux profils, si oui en combinaison avec d'autres attributs.

Pour mesurer la similarité de deux valeurs de l'attribut n, nous utilisons la mesure de *Jaccard*, qui permet de calculer le nombre de mots communs sans prendre en compte leur ordre d'occurrence dans un ensemble d'expressions. Plus formellement, la mesure de *Jaccard* entre deux noms $P_n(v_i)$ et $P_n(v_j)$ se calcul comme suit :

$$sim_n(P_n(v_i), P_n(v_j)) = \frac{|N_i \cap N_j|}{|N_i \cup N_j|} \tag{4.3}$$

où, N_i et N_j sont deux ensembles de mots contenus respectivement dans $P_n(v_i)$ et $P_n(v_j)$. Par exemple, si $P_n(v_i)$ est "Barack Obama" et $P_n(v_j)$ est "Barack Hussein Obama", alors $N_i = \{Barack, Obama\}$, $N_j = \{Barack, Hussein, Obama\}$ et leur similarité de Jaccard est de $\frac{2}{3}$. Nous avons choisi pour l'attribut Noms la mesure de Jaccard au lieu de la distance de *Levenshtein* comme

précédemment, car les réseaux sociaux ne forcent généralement pas leurs utilisateurs à spécifier d'abord leurs prénoms suivis de leur nom ou vice versa. Autrement dit, il n'y a généralement pas d'ordre précis lors du renseignement des noms et prénoms d'un utilisateur. De ce fait, certains peuvent préciser soit leurs prénoms suivis de leur nom ou vice versa, soit tout simplement leur nom en omettant leur prénoms ou vice versa. Le score de similarité entre "Barack Obama" et "Obama Barack" en utilisant la distance de *Levenshtein* est de 10, tant bien que, ces deux expressions soient équivalentes. Tandis qu'avec la mesure de Jaccard on a un score de 1, qui est celui attendu.

2.2.3 L'attribut Localite

Noté l, il représente l'emplacement actuel ou le lieu de naissance, souvent renseigné par les utilisateurs dans leurs profils. Même si cet attribut est très ambigu par rapport aux autres, il constitue quand même un indicateur utile à la réconciliation de deux profils utilisateurs. Son ambiguïté provient du fait que les utilisateurs ont tendance à le renseigner sous forme de toponyme qui est un nom propre désignant un lieu géographique. Notamment, "Paris" qui fait référence à plusieurs emplacements dans le monde tels que "Paris, France", "Paris, Texas, USA", "Paris, Ontario, Canada". A cet effet, intuitivement, nous disons que deux toponymes indiquent un même lieu géographique, s'ils sont similaires, c'est-à-dire comportent une sous-chaîne de caractères identique qui a une faible quantité d'interprétations. En occurrence, la paire de toponymes ("Paris", "Paris, France") partagent la sous-chaîne "Paris" entre eux, par contre le couple ("Paris, Île-de-France", "Paris, France") partagent plutôt la chaîne de caractères "Paris France". Nous constatons que dans le second cas, la chaîne commune "Paris France" fait référence à un seul emplacement géographique, qui est la capitale de la France et pourtant dans le premier cas le toponyme "Paris" possède plusieurs interprétations précédemment sus-citées. De ce fait, le score de similarité du second couple de toponymes doit être élevé par rapport au score du premier couple. Globalement, deux toponymes dont la sous-chaîne de caractères communs représentent un lieu géographique précis et leurs interprétations ont un fort taux de recouvrement au sein de leurs divisions administratives (pays, état, ville) sont considérés comme très similaires. Contrairement, deux toponymes ayant une sous-chaîne commune qui possède plusieurs interprétations et/ou qui ont un faible taux de recouvrement au niveau de leurs divisions administratives sont jugés peu semblables. Par conséquent, la similarité de deux toponymes doit tenir compte du taux de recouvrement de leurs différentes interprétations, ainsi que leur ambiguïté.

Pour quantifier la similarité existant entre les toponymes, nous nous sommes retournés vers le service Web de requêtage des toponymes OpenStreetMap (OSM). En effet, OSM est un projet qui a pour but de construire une base de données géographiques libre du monde. Pour une requête donnée portant sur un emplacement, son service renvoie en plus des différentes interprétations possibles (tous les lieux dans le monde portant le nom de l'emplacement contenu dans la requête), une répartition hiérarchique des divisions administratives (pays, état, ville, code postal, etc.) de ces interprétations classées par ordre de pertinence ou d'importance. En fait, cette importance est représentée par une valeur numérique qui est fonction du type de requête. Nous exploitons ce service pour déterminer les différentes interprétations possibles de nos toponymes, que nous utilisons par la suite pour mesurer leur similarité.

Analyse de la sortie des requêtes de OSM. Notons R_l le résultat obtenu sur un toponyme l à partir de OSM. En effet, les résultats sont présentés sous forme d'arbre pondéré possédant une profondeur maximale de trois (pays, état et ville), que nous appelons arbre d'interprétations, comme l'indique la figure 4.1. Étant donné que le service OSM possède une répartition plus fine des toponymes, certains d'entre eux sont fusionnés, notamment la ville et la municipalité sont représentées

sous la catégorie ville. Chaque branche représente un lieu géographique unique et possède une importance notée i (indiquée entre parenthèses sur la figure 4.1). De même, chaque niveau administratif, possède une force représentée par un poids empirique w, que nous utilisons dans le calcul de la similarité des toponymes. Par exemple, le fait que deux toponymes indiquent le même pays ne fournit pas une preuve aussi forte qu'ils sont similaires que s'ils mentionnent plutôt la même ville. A cet effet, le pays reçoit un poids inférieur à celui de la ville.

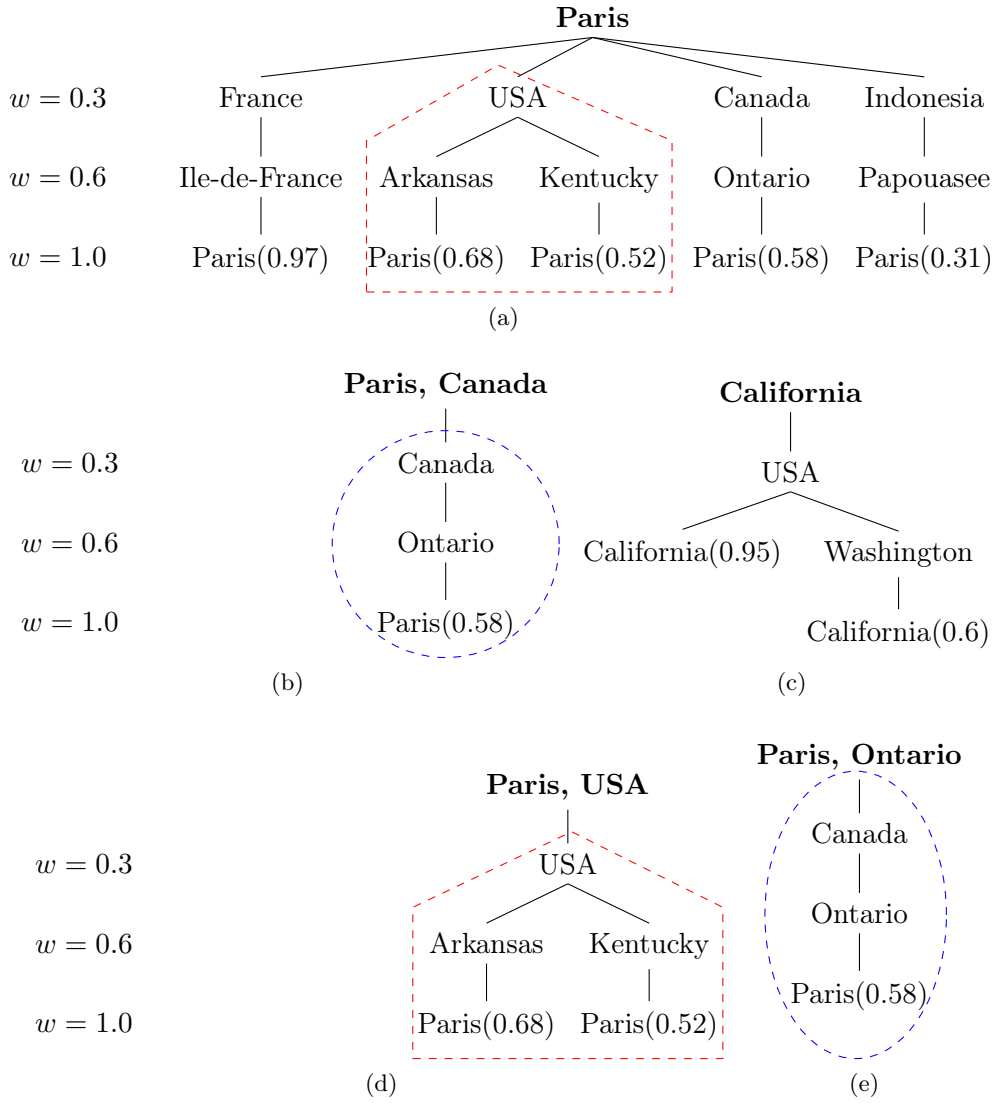


Figure 4.1: Représentation des résultats des toponymes: (a) "Paris"; (b) "Paris, Canada"; (c) "California"; (d) "Paris, USA"; et, (e) "Paris, Ontario".

Par ailleurs, l'arbre d'interprétations est utilisé pour calculer la certitude d'une interprétation d'un toponyme. La certitude ou la netteté d'une interprétation est la combinaison entre son im-

portance (i_r) et sa granularité (w_r) issue d'une branche r (celle contenant l'interprétation cible) de l'arbre d'interprétations de la requête R_l. Formellement, le score de certitude d'une interprétation s'exprime ainsi $w_r \times i_r$. En l'occurrence, sur la figure 4.1a, l'interprétation "Paris, France" du toponyme "Paris" possède un score de certitude plus élevé que "Paris, Canada" d'après OSM. Par contre, sur la figure 4.1c, la première branche de l'arbre d'interprétations portant sur la requête du toponyme "Californie" obtient un score de certitude $(0,57)$ inférieur à celui de la deuxième branche "Californie" dans l'état de Washington$(0,6)$ puisque la deuxième branche est plus précise.

Mesure de similarité de l'attribut LOCALITE. Pour comparer deux toponymes l_1 et l_2, nous nous basons sur leurs arbres d'interprétations obtenus respectivement à partir des requêtes R_{l_1} et R_{l_2}. A cet effet, nous mesurons le taux de recouvrement des deux arbres d'interprétations, noté $R^{\cap}_{l_1,l_2}$. Ainsi, nous notons par n_{l_1}, n_{l_2}, $n_{l_1-l_2}$ et n le score de certitude de R_{l_1} , R_{l_2}, $R^{\cap}_{l_1-l_2}$ et $R^{\cup}_{l_1-l_2}$, respectivement, que nous définissons comme suit:

$$n_{l_1} = \sum_{r \in R_{l_1}} \frac{w_r \times i_r}{I} \qquad n_{l_2} = \sum_{r \in R_{l_2}} \frac{w_r \times i_r}{I}$$

$$n_{l_1,l_2} = \sum_{r \in R^{\cap}_{l_1-l_2}} \frac{w_r \times i_r}{I} \qquad n = \sum_{r \in R^{\cup}_{l_1-l_2}} \frac{w_r \times i_r}{I}$$

où $I = \sum_{r \in R^{\cup}_{l_1-l_2}} i_r$ est le coefficient de normalisation calculé en considérant l'union des branches des arbres d'interprétations issues des requêtes R_{l_1} et R_{l_2}. Le recouvrement des arbres d'interprétations issues des requêtes $R_{\text{"}Paris\text{"}}$ et $R_{\text{"}Paris,\ USA\text{"}}$ des figures 4.1a et 4.1d est représenté par la ligne pointillée rouge. Plus précisément, les scores de certitude entre le lieu "Paris" et les lieux "Paris, USA" et "Paris, Canada" notés $R^{\cap}_{\text{"}Paris-Paris,USA\text{"}}$ et $R^{\cap}_{\text{"}Paris-Paris,Canada\text{"}}$ sont respectivement $0,39$ et $0,18$. Ainsi, comme indiqué précédemment, la mesure de similarité entre deux toponymes l_1 et l_2 dépend fortement de leur taux de recouvrement. A cet effet, nous nous sommes penchés sur deux mesures de similarité à savoir le *Support* pondéré et la mesure *Ochiai* [59], une variation de la similarité cosinus.

$$S_{W-Support} = \frac{1}{\sqrt{n}} \times \frac{n_{l_1,l_2}}{n}$$

$$S_{Ochiai} = \frac{n_{l_1,l_2}}{\sqrt{n_{l_1} \times n_{l_2}}}$$

La valeur de similarité de ces mesures est comprise entre 0, si le recouvrement R_{l_1} et R_{l_2} est nul, et 1, si $R_{l_1} = R_{l_2}$ pour la mesure Ochiai, par contre le Support pondéré est plus stricte, il faut que les toponymes d'entrée aient une interprétation qui correspond (identique). Cette condition pénalise la similarité si les toponymes d'entrée sont ambigus.

Le tableau 4.1 présente les valeurs des différentes mesures de similarité obtenues à partir des arbres d'interprétations issues des requêtes OSM pour certaines paires de toponymes collectées à partir d'une base de données réelle [46]. Plusieurs observations peuvent être faites à partir des résultats de ces exemples. Tout d'abord, les mesures proposées sont capables d'une part, de gérer les différences orthographiques entre les toponymes, et d'autre part, prédire une correspondance exacte même si les entrées sont orthographiquement différentes, mais en fait non ambigu puisqu'ils font référence à un même lieu géographique, c'est le cas des lignes 5, 11 et 13. Ensuite, il est possible d'attribuer une faible similarité aux toponymes possédant une division administrative différente (et

donc ambigus), on peut le constater sur les lignes 3, 7 et 9. De même, il est également possible de donner un faible score de similarité aux toponymes ayant une même division administrative mais faisant référence à des lieux géographiques différents, comme présenté sur les lignes 10 et 12. Et enfin, nous observons que la mesure de similarité Support pondéré donne des valeurs de similarité faibles lorsque les toponymes en entrée ont plusieurs d'interprétations. En fait, cela est dû au terme $\frac{1}{\sqrt{n}}$ de sa formule, qui désavantage son résultat final tant que le nombre d'interprétations des localités cibles augmente.

	Localité 1	Localité 2	Support.P	Ochiai
1.	San Diego ,usa	San Diego	0.4	0.87
2.	Houston, Texas ,usa	Houston	0.56	0.77
3.	Canada	Toronto, Canada	0.11	0.28
4.	Orlando, Florida ,usa	Florida	0.23	0.63
5.	Wausau, Wisconsin ,usa	Wausau, WI	1	1
6.	Los Angeles ,usa	Los Angeles	0.4	0.85
7.	Argentina	Argentina, Buenos Aires, Junín	0.08	0.19
8.	Montreal, Canada	Montreal, Quebec	0.76	0.69
9.	United States,usa	Puerto Rico	0.13	0.3
10.	Apeldoorn , Netherlands	Deventer	0.13	0.3
11.	Bengaluru , India	Bangalore, India	1	1
12.	Utrecht , Netherlands	Amersfoort, the Netherlands	0.26	0.6
13.	New York City ,usa	Brooklyn, NY, USA	1	1

Table 4.1: Mesures de similarité des paires de toponymes à partir des réponses de OSM.

2.2.4 L'attribut Emails

Noté e, c'est un attribut multivalué, dont les valeurs correspondent aux différentes adresses électroniques d'un utilisateur. L'adresse électronique est un attribut très sensible car elle est généralement associée à une personne unique. Autrement dit, si deux profils contiennent une même adresse e-mail, il y a des fortes chances qu'ils fassent référence à une même personne. Néanmoins, il est possible que deux personnes partagent une même adresse e-mail, c'est le cas par exemple des personnes qui travaillent au sein d'une même organisation. En fait ce sont des situations particulières, car seul un faible pourcentage d'utilisateur donne un accès public à leurs adresses mails.

Pour comparer les valeurs de l'attribut e de deux profils utilisateurs, il suffit de déterminer si une des adresses électroniques d'un profil est identique à l'une des adresses électroniques d'un autre profil. Dans le cas positif la similarité est égale à 1, sinon, elle est égale à 0.

2.2.5 Les attributs Websites et Profiles

Websites noté w et Profiles noté p, sont également deux attributs à valeurs multiples, dont les valeurs sont respectivement des URLs vers des pages Web et des pages correspondant à des profils

utilisateurs dans d'autres réseaux sociaux. Bien qu'étant tous deux des adresses Web, nous avons voulu les séparer afin de pouvoir analyser séparément la contribution de chacun des deux.

Habituellement, les profils utilisateurs sont beaucoup plus personnels par rapport aux pages Web, qui sont plutôt génériques. Autrement dit, le fait qu'une paire de profils partagent un même lien vers un troisième profil utilisateur est une indication forte pour la réconciliation, que s'ils partagent un même lien vers plutôt une page Web quelconque. En fait, les liens vers les profils utilisateurs peuvent être des liens transversaux, et ceux vers les pages Web, des liens vers des ressources que l'utilisateur souhaite partager avec d'autres personnes. Ces ressources peuvent correspondre à des sites administrés par l'utilisateur. Nous utilisons l'égalité mathématiques pour comparer les valeurs des attributs w et p. Plus précisément, nous nous limitons à rechercher si les valeurs des attributs w ou p, pour deux profils distincts, ont au moins une *url* commune sans analyser leur contenu.

Dans la section suivante, nous expliquons plus en détails les différentes étapes de notre approche proposée Liaison.

3 LIAISON: reconciLIAtion of Individuals profiles across SOcial Networks

L'approche de réconciliation, que nous proposons, exploite essentiellement les valeurs des attributs issues des profils utilisateurs pour découvrir les profils faisant référence à une même personne. En effet, notre idée se base sur une observation selon laquelle si deux profils v_i et v_j font référence à une même personne, alors leurs différentes valeurs d'attributs sont susceptibles d'être égales ou similaires.

3.1 Approche proposée

Une solution intuitive au problème posé consiste à comparer chaque paire de profils utilisateurs $(v_i, v_j) \mid v_i \in V_i,\ v_j \in V_j$ non reliée par un lien transversal *me*. Dans ce cas, nous avons $\sum_{i,j} |V_i| \times |V_j| - |E_{i,j}|$ paires de profils à comparer. Cette solution est loin d'être envisageable, car les réseaux sociaux contiennent en réalité des millions de nœuds. Par exemple, si nous prenons l'échantillon de quatre réseaux sociaux sur lesquels nous avons évalué notre approche Liaison, il contient environ 2 millions de nœuds, donc environ 4×10^{12} paires de nœuds à comparer. De plus, si nous supposons que la comparaison de chaque paire prend 0.1ms, alors Liaison [65] va s'exécuter en 12 années: ce qui n'est pas concevable.

Au vu de la complexité de la solution intuitive, nous proposons une solution nettement meilleure qui se déroule en deux étapes principales :

- La première étape consiste à sélectionner un sous-ensemble de paires de profils candidats à la réconciliation. Cette étape exploite la topologie du graphe représentant les réseaux interconnectés, particulièrement les liens d'amitié *friend* reliant deux profils dans un même réseau social et les liens transversaux *me* connectant les profils d'un même utilisateur sur différents réseaux sociaux.

- La seconde étape consiste à déterminer parmi les candidats sélectionnés ceux qui appartiennent aux mêmes utilisateurs. En revanche, cette étape exploite un ensemble de valeurs d'attributs à travers une collection de règles que nous avons définis.

De manière générale, l'algorithme 7 met en avant les différentes étapes de LIAISON. Il présente une phase d'initialisation, ligne 2 et 3, pendant laquelle on crée deux variables : D_{me} qui va contenir les liens transversaux *me* découverts et $\mathcal{Q}ueue$ qui est une file d'attente contenant les liens *me* existants pas encore visités ou traités. Initialement, $\mathcal{Q}ueue$ contient tous les liens transversaux E_{me} du graphe construit. Tant que la file $\mathcal{Q}ueue$ contient encore des liens transversaux non traités, le principe est le suivant :

1. A l'aide de la fonction *selectCandidates* qui prend en entrée le graphe et la file $\mathcal{Q}ueue$, nous sélectionnons les paires de profils candidats à la réconciliation qu'on stocke dans la variable $\mathcal{S}_c$.
2. Ensuite, à partir de la fonction $discoveryNewLinksMe$, nous appliquons à chaque paire de profils candidats, un ensemble de règles que nous définissons à la section 3.1.2 afin de vérifier s'il existe un lien *me* entre les profils cibles. Si oui, les profils découverts sont sauvegardés dans la variable D_{me}. Nous notons que cette étape modifie la file $\mathcal{Q}ueue$ en enfilant les paires de profils découverts.
3. Si $\mathcal{Q}ueue$ contient encore des liens *me* non traités, les étapes 1 et 2 sont réitérées jusqu'à ce qu'on ne trouve plus de nouveaux liens *me* c'est-à-dire la file $\mathcal{Q}ueue$ soit vide.

Algorithm 7 LIAISON

```
1: function LIAISON(G =< ∪_i V_i , ∪_i E_i , E_me >): G' =< ∪_i V_i , ∪_i E_i , E_me ∪ D_me >
2:     D_me ← ∅
3:     enqueue(E_me, Queue)
4:     while Queue is not empty do
5:         S_c ← selectCandidates(Queue, G)
6:         D_me ← discoveryNewLinksMe(S_c, (E_me ∪ D_me), Queue) ∪ D_me
7:     end while
8: end function
```

Dans la suite, nous détaillons respectivement les deux principales étapes de LIAISON.

3.1.1 Sélection des paires de profils candidates

Pour choisir les couples de profils susceptibles de faire référence à une même personne, nous proposons une démarche plus objective, par rapport à la solution intuitive précédente. Cette démarche se base sur des travaux qui montrent que certains utilisateurs qui ont plusieurs profils sont connectés généralement avec d'autres utilisateurs qui ont aussi plusieurs profils. Plus clairement, deux amis (utilisateurs reliés par un lien d'amitié) dans un réseau social, sont communément connectés dans leurs différents réseaux [25].

Par ailleurs, dans un réseau social, il existe des utilisateurs qui déclarent explicitement les liens entre leurs différents profils, ce qui permet de déterminer les amis communs déclarés dans deux profils appartenant à deux réseaux différents. En d'autres termes, s'il existe $(v_i, me, v_j) \in E_{i,j}$, $v\prime_i \in friend(v_i)$, $v\prime_j \in friend(v_j)$ alors $(v\prime_i, v\prime_j)$ est une paire de profils candidats, avec $friend(v_i) = \{v\prime_i / (v\prime_i, friend, v_i) \vee (v_i, friend, v\prime_i) \in E_i\}$ qui représente l'ensemble des profils amis de l'utilisateur

v_i. L'ensemble des paires de profils candidats, pour les réseaux sociaux i et j, est formellement défini comme suit :

$$\begin{aligned}\mathcal{S}_c = \{(v\prime_i, v\prime_j)/\exists v\prime_i \in friend(v_i) \wedge v\prime_j \in friend(v_j) \\ \wedge (v_j, me, v_i) \in E_{i,j} \wedge (v\prime_i, me, v\prime_j) \notin E_{i,j}\}\end{aligned} \tag{4.4}$$

L'algorithme 8 présente globalement le processus de sélection des paires de profils candidats. Il prend en entrée une file $\mathcal{Q}ueue$, qui contient l'ensemble des liens transversaux me qu'il va parcourir ainsi que la topologie du graphe de réseaux interconnectés à travers les liens. Pour chaque lien qu'on défile de $\mathcal{Q}ueue$, on appelle la fonction $getCandidates$ qui calcule les couples de profils candidats à l'aide de la formule 4.4, jusqu'à ce que la file soit vide.

Algorithm 8 Sélection des paires de profils candidats

```
1: function selectCandidates(Queue, G): S_c
2:     S_c ← ∅
3:     while Queue is not empty do
4:         S_c ← S_c ∪ ← getCandidates(dequeue(Queue))
5:     end while
6: end function
```

3.1.2 Découverte de nouveaux liens transversaux

Elle consiste à analyser chaque paire de profils candidats de $\mathcal{S}_c$ afin de distinguer ceux qui font référence à un même utilisateur. Pour déterminer si deux profils v_i et v_j référencent une même personne, on a défini un ensemble de règles qui exploitent les attributs présentés à la section 2.2. Chaque règle prend en considération la contribution d'un ou plusieurs attributs. Étant donné que certains attributs ne constituent pas l'identité de l'utilisateur, pour deux profils nous supposons que plus il y a des valeurs d'attributs qui correspondent (selon une mesure de similarité définie), plus la probabilité qu'ils font référence à un même utilisateur est forte. De ce fait, les règles définies sont classées par ordre de pertinence ou confiance noté k. La règle la plus pertinente, d'ordre maximale, est celle qui détermine que tous les attributs correspondent, dans ce cas $k = |A|$. La règle la moins pertinente est celle qui détermine qu'un seul attribut correspond, dans ce cas $k = 1$.

Soit le prédicat noté $match(P_a(v_i), P_a(v_j))$ qui retourne vrai si les valeurs de l'attribut a pour les profils v_i et v_j correspondent selon la mesure de similarité définie pour cet attribut a. Une règle d'ordre k, noté $\mathcal{R}^k$ est définie comme suit:

$$\mathcal{R}^k(v_i, v_j) = \begin{cases} \bigwedge_{a \in A} \ match(P_a(v_i), P_a(v_j)) & \text{si } k = |A| \\ \bigvee_{B \in [A]^k} \bigwedge_{a \in A \setminus B} \ match(P_a(v_i), P_a(v_j)) & \text{si } 1 \leq k < |A| \end{cases} \tag{4.5}$$

où $[A]^k$ représente l'ensemble de tous les sous-ensembles de A de taille k.

Ainsi, si $\bigvee_{1 \leq k \leq |A|} \mathcal{R}^k(v_i, v_j)$ est vrai, alors v_i et v_j sont deux profils faisant référence à un même utilisateur. Les règles sont appliquées par ordre de pertinence décroissante, c'est-à-dire de la plus pertinente ($k = |A|$) vers la moins ($k = 1$). Si pour un couple de profils candidats (v_i, v_j), une règle

$\mathcal{R}^k(v_i, v_j)$ retourne *vrai*, alors les règles d'ordre inférieur à k ne sont plus appliquées. Dans le pire des cas, un couple (v_i, v_j) atteint la règle d'ordre $\mathcal{R}^1(v_i, v_j)$. Les nœuds de tous les couples (v_i, v_j) vérifiant au moins une règle sont réconciliés, c'est-à-dire sont reliés par un arc portant le label *me*.

Globalement, le processus de découverte des liens transversaux est indiqué à l'algorithme 9. Pour chaque paire de profils candidat (ligne 3), on vérifie s'il n'est pas déjà connecté par un lien *me* (ligne 4). Si oui, on applique les règles, du plus grand vers le plus petit ordre (ligne 5). S'il existe au moins une règle d'ordre k qui correspond (ligne 6), on vient de découvrir un nouveau lien *me*, qu'on sauvegarde dans l'ensemble D_{newMe}. Après la phase de découverte des couples de profils à réconcilier contenus dans $\mathcal{S}_c$ (variable en entrée de cette fonction), on va appliquer au graphe $\mathcal{G}$ la fonction *transitiveClosure* (ligne 11), qui effectue l'opération mathématique de fermeture transitive en utilisant tous les liens *me* existants dans $\mathcal{G}$. En effet, cette opération nous permet de découvrir de nouveaux liens *me* qu'on ajoute à l'ensemble D_{newMe}. Tous les nouveaux liens *me* découverts sont enfilés en queue de file et sont traités plus tard (ligne 12). En fait, ces liens sont exploités pour sélectionner de nouveaux couples de profils candidats. Par la suite, les règles sont ré-exécutées sur ces nouveaux couples, ainsi de suite jusqu'à ce qu'on ne découvre plus de nouvelles paires de profils *me*, c'est-à-dire $D_{newMe} = \emptyset$.

Algorithm 9 Découverte de nouveaux liens transversaux

```
1:  function DISCOVERYNEWLINKSME(S_c, D_allMe, Queue): D_newMe
2:      D_newMe ← ∅
3:      for each c ∈ S_c do
4:          if c ∉ D_allMe then
5:              k ← applyRules(c)
6:              if k ≥ 1 then
7:                  D_newMe ← D_newMe ∪ {c};
8:              end if
9:          end if
10:     end for
11:     D_newMe ← transitiveClosure(D_allMe ∪ D_newMe) ∪ D_newMe
12:     enqueue(D_newMe, Queue);
13: end function
```

Le nombre de paires de profils candidats obtenu par LIAISON à partir d'un lien transversal (v, me, w) peut être très grand ($O(|friends(v)| \times |friends(w)|)$). A cet effet, pour réduire ce nombre de profils candidats, nous stockons les valeurs a des attributs de chaque nœud $x \in friends(v)$ dans une structure de données I_a qui permet une récupération rapide des nœuds de $y \in friends(w)$ qui possèdent des valeurs d'attributs identiques ou similaires à ceux stockés dans I_a. La structure de données I_a peut être soit une table de hachage dans le cas d'une comparaison exacte des valeurs d'attributs de x et y (attribut WEBSITES), soit un BK-tree [11] pour le cas des comparaisons approximatives (attribut PSEUDONYME). Les BK-trees sont une structure de données basée sur la distance de *Levenshtein* et utilisée pour comparer les valeurs de chaînes de caractères.

3.2 Expérimentations et évaluations

Pour l'évaluation de LIAISON, nous nous sommes appuyés sur la base de données utilisée par [10] dans leurs expériences. Nous avons constaté l'existence de certaines données manquantes. De ce fait, nous avons opté, dans un premier temps, à une mise à jour de cette base de données, ensuite nous avons exécuté LIAISON sur les données obtenues, et enfin nous avons comparé les résultats de LIAISON aux méthodes existantes dans la littérature.

3.2.1 Mise à jour des données

A l'origine, la base de données [10] contient des profils utilisateurs de quatre réseaux sociaux, à savoir LIVEJOURNAL, FLICKR, TWITTER et YOUTUBE [6]. De plus, elle est constituée de 93 169 nœuds ou profils utilisateurs, 145 580 liens d'amitié et 497 liens transversaux, dont 468 inter-réseaux et 29 intra-réseau. Les données provenant de FLICKR et LIVEJOURNAL sont de loin les plus nombreuses, avec 55 117 (83 993) et 28 008 (41 244) nœuds (liens) respectivement et celles provenant de TWITTER et YOUTUBE comportent seulement 8 842 (19 008) et 1 210 (1367) nœuds (liens) respectivement. Cependant, le nombre de liens transversaux déclarés par [10] est de 745, mais cela inclut des doublons, que nous avons supprimés.

Une analyse minutieuse des données, plus spécifiquement par une vérification des profils utilisateurs sur la toile à l'aide de leur pseudonyme, nous a révélé un grand nombre de liens de type *friend* manquants, qui ont été probablement ajoutés après la construction de la base de données initiale. De plus, les profils de cette base initiale sont décrits uniquement à partir de l'attribut pseudonyme. Ainsi, nous avons cherché à ajouter les informations manquantes dont nous avons besoin, par une mise à jour des liens *friend* dans un premier temps des deux plus importants réseaux FLICKR et LIVEJOURNAL, et d'autre part, par l'ajout des valeurs des autres attributs manquants pour chaque profil des quatre réseaux. Plus clairement, pour chaque nœud existant dans les réseaux FLICKR et LIVEJOURNAL, on extrait en utilisant l'API appropriée du réseau tous les nœuds voisins par le lien *friend*. Ces nœuds sont rajoutés au graphe initial, ainsi que les liens qu'ils ont avec les autres nœuds existants dans la base initiale.

Après enrichissement de la base, nous avons au total, plus de 2 millions de nœuds, plus de 21 millions de liens *friend* et 29 liens transversaux intra-réseau (liens existants entre les profils d'une personne dans un même réseau social). Les caractéristiques du graphe des quatre réseaux sociaux interconnectés sont détaillées dans le tableau 4.2. Le nombre de liens transversaux inter-réseaux (liens existants entre les profils d'une personne dans différents réseaux sociaux) est passé de 468 à 474 suite à la fermeture transitive du graphe final. La distribution des liens transversaux à travers les différents réseaux sociaux interconnectés est indiqué dans le tableau 4.3. Pour l'implémentation de notre approche, nous avons utilisé Neo4j[7], une base de données puissante dédiée pour représenter et manipuler des graphes de grande taille.

En effet, Neo4j est un système de gestion de base de données au code source libre basé sur les graphes, développé en Java et existant depuis 2000. Il permet de représenter les données en tant que "nœuds" reliés par un ensemble "d'arcs". Les nœuds possèdent des propriétés, qui sont constituées d'un couple de clé-valeurs de type simple tel que chaînes de caractères ou numériques et qui peuvent êtres indexées.

Passons à la présentation des résultats de LIAISON sur notre nouvelle base de données.

[6] http://www.ursino.unirc.it/pkdd-12.html [7] www.neo4j.org/

Réseaux	Nœuds	Liens		
		friend	*intra − me*	Total
Flickr	1 814 405	15 415 083	0	15 415 083
Livejournal	211 044	5 628 509	1	5 628 510
Twitter	8 842	19 008	13	19 021
Youtube	1 210	1 367	15	1 382
Total	**2 035 501**	**21 063 967**	**29**	**21 063 996**

Table 4.2: Caractéristiques des réseaux sociaux interconnectés après enrichissement

Network	Flickr	Livejournal	Twitter	Youtube
Flickr	0	148	29	12
Livejournal	148	1	11	2
Twitter	29	11	13	272
Youtube	12	2	272	15

Table 4.3: Liens transversaux entre chaque paire de réseaux sociaux

3.2.2 Résultats de Liaison

Nous avons fait une évaluation préliminaire [5] visant à identifier d'une part, les attributs qui sont les plus efficaces à la réconciliation de deux profils, et d'autre part, les valeurs des seuils de similarité θ_a de chaque attribut $a \in A$. Cette évaluation a montré que :

- Toutes les k-règles, avec $k \geq 2$, découvrent les liens transversaux avec une grande précision. En d'autres termes, si deux profils ont au moins deux valeurs d'attributs qui correspondent, alors ils ont une forte probabilité de faire référence à la même personne.
- De même, la 1-règle basée sur l'attribut Pseudonyme découvre les liens transversaux avec également une grande précision si le seuil θ_u est fixé à $0,9$.
- Par contre, la 1-règle basée sur l'attribut Noms conduit à un taux d'erreur élevé, peu importe la valeur du seuil θ_n fixé. Par conséquent, deux profils où les noms correspondent ne sont pas considérés comme faisant référence à la même personne, à moins que d'autres valeurs d'attributs correspondent.

Pareillement, nous avons réitéré une évaluation similaire sur les données précédemment construites, afin d'avoir une idée sur la valeur du seuil θ_l, qui est utilisée pour comparer les valeurs de l'attribut Localite. Comme dans le cas de l'attribut Noms, le fait que deux profils indiquent des lieux géographiques identiques ou similaires n'est pas une information suffisante pour conclure qu'ils font référence à une même personne. De ce fait, il doit également être utilisé en combinaison avec d'autres attributs. De plus, nous avons également observé que les meilleurs résultats sont obtenus en fixant θ_l à 0,7.

Sur la base de ces observations, nous avons exécuté Liaison sur la base de données enrichie, à l'exception des 1-règles portant sur les attributs Noms et Localite, tout en fixant $\theta_u = \theta_n = 0,9$

et $\theta_l = 0,7$. Le nombre de liens transversaux découverts à chaque itération, pour chaque valeur de k en utilisant soit les règles (R), soit la fermeture transitive (Ft) est indiqué dans le tableau 4.4.

Itération	Méthode	k=1	k=2	k=3	k=4	k=5	**Total**	**Grand Total**
1	R	3,792	853	84	4	0	**4,733**	**4,907**
	Ft	161	13	0	0	0	**174**	
2	R	1,104	69	47	20	4	**1,244**	**1,620**
	Ft	373	2	1	0	0	**376**	
3	R	19	0	1	0	0	**20**	**45**
	Ft	25	0	0	0	0	**25**	
Total		**5,474**	**937**	**133**	**24**	**4**	**6,572**	**6,572**

Table 4.4: Liens transversaux découverts par LIAISON par itération et valeur de k.

Le nombre total de liens transversaux découverts (6.572) ne comprend pas les 6 liens découverts par la fermeture transitive des liens existants.

Comme supposé, le nombre de liens transversaux découverts par LIAISON diminue au fur et à mesure que le nombre d'itérations progresse. Par exemple, à l'itération 3, LIAISON découvre 19 (respectivement 25) liens transversaux grâce à l'exécution des règles (respectivement la fermeture transitive). Ces liens découverts sont utilisés à l'itération 4 pour sélectionner de nouveaux profils candidats, qui ne font pas référence à la même personne d'où l'arrêt de l'exécution. De même, nous observons que la plupart des liens transversaux sont découverts lors de la première itération. Plus clairement, en utilisant la 1-règle, ce qui montre effectivement que deux profils créés par un même individu ont généralement quelques informations qui se chevauchent. Cependant, malgré l'utilisation des valeurs d'un seul attribut (1-règle), LIAISON découvre quand même 5 474 liens transversaux, dont la plupart sont corrects. Nous discutons plutard sur la précision. Ce résultat est particulièrement remarquable si l'on considère que LIAISON débute son exécution avec uniquement 474 liens transversaux, parmi lesquels seulement 239 relient deux nœuds ayant chacun des liens d'amitié avec d'autres nœuds. En d'autres termes, juste 239 liens *me* peuvent être utilisés à la première itération pour sélectionner des paires de profils candidats.

Par ailleurs, nous notons également que le nombre total de liens transversaux découverts (6 572) ne tient pas compte des 6 liens transversaux découverts par fermeture transitive sur les liens *me* existants initialement avant la première itération. Au final, LIAISON découvre 6 578 liens transversaux en 2 heures, 11 minutes et 58 secondes. Les liens découverts par fermeture transitive sont moins que ceux découverts à partir des règles. Une explication possible est que la structure de nos quatre réseaux sociaux qui, bien que large, reste toujours un échantillon limité de 4 réseaux constitués de plus de 500 millions de profils utilisateurs. Il est clair que notre jeu de données ne contient pas tous les profils réels de ces réseaux.

Afin d'évaluer la précision des différentes règles, nous avons déterminé une vérité terre en étiquetant manuellement chaque lien transversal $(v, me, w) \in D_{me}$ avec, soit le label *correct*, si les profils v et w font référence réellement à la même personne, soit le label *incorrect*, s'ils ne le font pas, ou soit le label *indéterminé*, si aucune décision ne peut être prise. De ce fait, nous avons divisé D_{me} en quatre sous-ensembles indépendants de taille égale pour chacun des quatre auteurs de LIAI-

son. A partir d'une inspection manuelle sur le Web des différents profils utilisateurs, chaque auteur va étiqueter (correct, incorrecte, indéterminé) les liens *me* de son ensemble. L'inspection visuelle consiste à examiner tous les aspects possibles des profils utilisateurs à l'exception des valeurs des attributs utilisés par Liaison, plus nettement : les photos (en particulier dans Flickr), le contenu textuel (en particulier dans Livejournal), des informations sur les pages Web liées au profil ou récupérées à partir d'autres réseaux sociaux. La plupart du temps, les informations trouvées ont été suffisantes pour décider si deux profils indiquent ou pas une même personne. Cependant, dans certains cas, les informations sont rares qu'on ne peut conclure, c'est pourquoi nous avons inclus le label "indéterminé", afin d'éviter d'introduire des erreurs dans notre vérité terre, qui peuvent avoir des répercutions sur nos résultats lors du calcul de la performance de Liaison. Au final, nous avons catégorisé les liens *me* étiquetés comme suit :

- $\mathcal{C}$: l'ensemble des liens transversaux contenant le label correct,
- $\mathcal{I}$: l'ensemble des liens transversaux contenant le label incorrect et
- $\mathcal{N}$: l'ensemble des liens transversaux contenant le label indéterminé.

Il nous a fallu environ 10 jours pour étiqueter tous les liens transversaux D_{me} découverts par Liaison. Sur la base de cette vérité terre, nous avons calculé la précision de Liaison comme suit :

$$P_{liaison} = \frac{|\mathcal{C}|}{|\mathcal{C}| + |\mathcal{I}|} \tag{4.6}$$

Par contre, pour obtenir le rappel, qui est le rapport entre les liens transversaux corrects et le nombre total de paires de profils qui font référence à la même personne, nous avons besoin d'étiqueter toutes les paires possibles de profils de notre jeu de données, ce qui est visiblement pas envisageable puisque nous avons environ 2 millions de nœuds. De ce fait, nous discutons du rappel de Liaison sur un autre jeu de données beaucoup plus petit, où les informations nécessaires sont déjà connues.

Précision de Liaison. La précision globale de Liaison sur notre jeu de données après les trois itérations précédentes est de 94%, qui est un résultat très encourageant. La figure 4.2 montre pour chaque valeur de confiance k, la précision obtenue par Liaison. Comme prévu, on constate que la précision augmente avec la confiance, et atteint 100% pour $k \geq 2$. En ce qui concerne la confiance $k = 1$, les règles atteignent une précision largement supérieure à 90%, alors que avec la fermeture transitive elle, est sensiblement plus faible (73%), par opposition avec le cas $k = 2$, due au fait que certains liens transversaux découverts avec les 1-règles ($k = 1$) sont erronés (6% d'entre eux) et malheureusement propagés lors de la fermeture transitive.

3.2.3 Comparaison de Liaison avec les méthodes existantes

Dans cette section, nous comparons notre approche Liaison d'une part, à celle proposée par Buccafurri et al. [10] nommée Bucc, qui sont les auteurs de la base de données initiale que nous avons enrichie, et d'autre part, à celle proposée par Malhotra et al. [46], qui par contre ont évalué leur méthode nommée Mal sur un petit jeu de données de 60 000 nœuds issus uniquement de deux réseaux sociaux.

Comparaison avec Bucc. L'approche Bucc est évaluée par ses auteurs en sélectionnant au hasard 160 liens transversaux existants, qu'ils utilisent pour découvrir de nouveaux liens transversaux. Au final ils ont eu 22 liens transversaux, dont 16 corrects, 2 faux et 2 indéterminés, ce qui

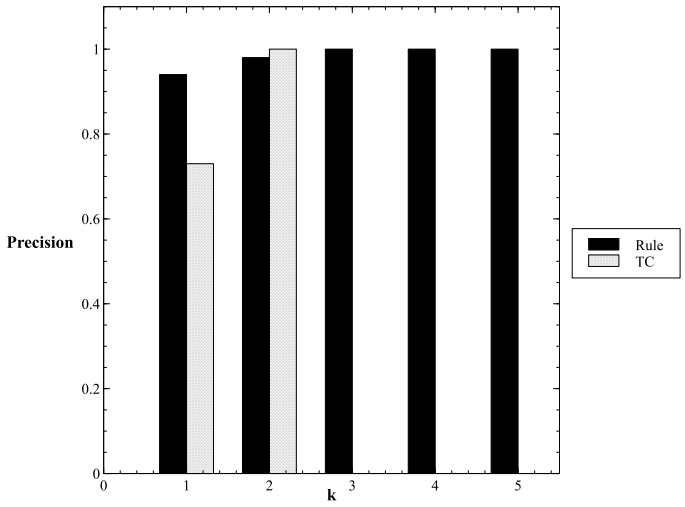


Figure 4.2: La précision de Liaison en fonction des règles et des valeurs de k.

donne une précision de 80%. Nous notons que leur algorithme retourne également un ensemble de 133 paires de nœuds, qui sont des couples de profils ne référençant pas la même personne. De plus ils ont également en leur possession le nombre de vrais et faux négatifs, ce qui leur permet de calculer le rappel global, qui est de 85%.

Notre approche découvre un ensemble beaucoup plus important de liens transversaux avec en plus une meilleure précision. De plus, notre jeu de données est une version de la leur plus enrichie c'est-à-dire avec plus de nœuds et de liens. Liaison repose sur un ensemble de règles qui combine la contribution d'une collection d'attributs, tandis que Bucc exploite uniquement l'attribut Pseudonyme et la topologie du réseau. Et enfin, Liaison utilise les liens transversaux qu'il découvre pour obtenir de nouveaux candidats de manière itérative.

Comparaison avec Mal. L'approche Mal utilise des techniques d'apprentissage basées sur plusieurs valeurs d'attributs pour catégoriser les paires de profils [46]. Pour la comparaison avec Liaison, nous utilisons exactement le même jeu de données, qui est constitué de profils utilisateurs provenant de deux réseaux sociaux populaires, Twitter et LinkedIn. Chaque réseau possède 29 129 nœuds contenant plusieurs valeurs d'attributs, 29 129 liens transversaux inter-réseaux et par contre aucun lien d'amitié. A la place des liens d'amitié absents, chaque nœud possède plutôt un attribut indiquant son nombre de connexions c'est-à-dire le nombre de nœuds existants dans son voisinage, qui est une information utilisée par Mal. De ce fait, l'étape de sélection de paires de profils candidats de Liaison ne peut être exécutée puisqu'il n'y a pas de liens d'amitié. On va donc considérer comme profils candidats à une éventuelle réconciliation, toutes les paires de nœuds (t, l), avec t un profil de Twitter et l celui de Livejournal ayant au moins un attribut similaire ou identique. Pour éviter une comparaison entre tous les paires de profils possibles, nous indexons les valeurs des attributs des profils de Twitter en utilisant les tables de hachage et les BK-trees.

Les attributs utilisés par Mal sont le pseudonyme, le nom, une mini description (en anglais "about me") généralement disponible dans les profils des réseaux sociaux, la localité, la photo de profil et le nombre d'amis ou de connexions. Par ailleurs les auteurs de Mal disposent de 29 219

liens transversaux existant dans leur base de données, que nous utilisons plutôt comme vérité terre. Étant donné qu'il n'y a que deux réseaux, nous ne pouvons pas faire de fermeture transitive. Nous exécutons uniquement une seule itération de notre approche avec les valeurs des seuils θ_u, θ_n, et θ_l définies comme précédemment. LIAISON découvre 9 210 liens transversaux, dont 9 134 sont corrects, en 3 minutes et 24 secondes; la précision globale est de 99%, avec un rappel de 31%. La figure 4.3 indique les précisions obtenues pour différentes valeurs de k. Nous observons également comme précédemment une augmentation de la précision lorsque la confiance croît.

La précision de MAL sur le même jeu de données est de 64%, ce qui est considérablement inférieur à celle obtenue par LIAISON. Les auteurs de cette approche ne font pas allusion au rappel. Le faible taux de rappel de LIAISON est due au fait que nos règles sont exécutées avec des valeurs de seuils de similarité θ_a très élevées, ce qui est extrêmement important pour notre approche car nous propageons les liens découverts à partir de la fermeture transitive. Nous constatons que différentes valeurs de précision et de rappel peuvent être obtenues en réglant les seuils de similarité des attributs pseudonyme, noms et localité. Comme les valeurs des attributs noms et localité sont généralement ambiguës et ne conviennent pas à une réconciliation de profils, nous ne les modifions pas. Le cas de l'attribut pseudonyme est différent, parce que les mêmes individus ont tendance à utiliser des pseudonymes similaires dans leurs différents profils. Nous avons donc fait varier uniquement la valeur de θ_u et nous avons observé un meilleur résultat lorsqu'elle vaut $0,7$ avec une précision de 86%, un rappel de 49% et une f-mesure est de 62%.

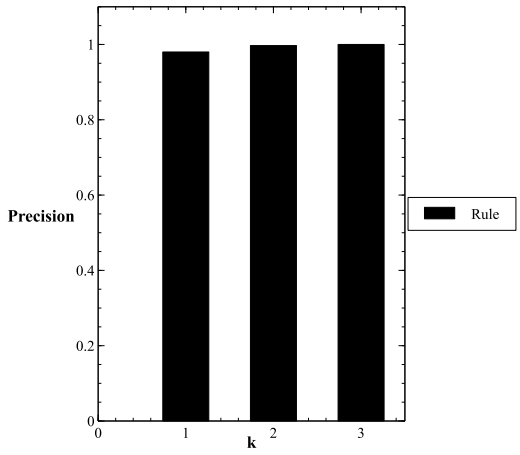


Figure 4.3: Précision de LIAISON par rapport aux valeurs de k sur les données de MAL .

Conclusion

Dans ce chapitre, nous avons présenté notre approche de réconciliation des profils utilisateurs dans différents réseaux sociaux, nommée LIAISON. Cette approche exploite la topologie du graphe et les attributs publics définis dans les profils utilisateurs.

Plus précisément, LIAISON se déroule en deux étapes principales. La première, qui prend en entrée une collection de réseaux sociaux interconnectés contenant des liens d'amitié et transversaux,

et retourne un ensemble $\mathcal{S}_c$ constitué des paires de profils candidats à une éventuelle réconciliation. En fait, ces réseaux sociaux sont représentés sous forme de graphe dans lequel les nœuds représentent les profils utilisateurs et les arcs des liens étiquetés par les labels $friend$ ou me en fonction du type de relation existant entre les nœuds. L'idée de cette étape est basée sur l'observation selon laquelle certains utilisateurs qui possèdent plusieurs profils, sont généralement connectés avec d'autres utilisateurs qui ont également plusieurs profils, d'où l'utilisation de la formule 4.4 pour déterminer les éléments de l'ensemble $\mathcal{S}_c$. La formule utilisée pour la sélection des couples de profils candidats est une démarche très objective face à la solution intuitive, qui consiste à identifier tous les paires de profils existants dans notre graphe de réseaux.

La seconde étape, quant à elle, porte sur la découverte de nouveaux liens transversaux et prend en entrée l'ensemble $\mathcal{S}_c$. Elle est basée essentiellement sur un ensemble de règles qui combinent les valeurs d'une collection d'attributs. Chaque règle possède un degré de confiance k, qui représente le nombre d'attributs utilisé dans la règle. Plus une règle contient d'attributs, plus sa confiance est élevée. Plus précisément, à chaque itération nous sélectionnons des profils candidats, ensuite nous appliquons nos règles sur ces profils, et enfin nous utilisons l'opération de fermeture transitive pour propager les liens découverts sur d'autres réseaux dans l'optique de découvrir de nouveaux liens me, qui sont utilisés pour sélectionner de nouveaux paires de profils candidats de manière itérative.

Les résultats obtenus sur une grande collection de données contenant environ 2 millions de nœuds issus de quatre réseaux sociaux Flickr, Youtube, Livejournal et Twitter ont montré d'une part, la pertinence des attributs considérés, et d'autre part, l'efficacité des règles que nous avons définies. Vu la taille de notre jeu de données, et la manière avec laquelle les utilisateurs éditent leurs différents profils, les informations contenues dans ces profils ont une forte probabilité d'être incomplètes, non mises à jour voire même erronées. Malgré tous ces obstacles, Liaison réussit a atteindre une précision de 94%, en débutant avec uniquement 239 liens transversaux. De plus, cette précision peut être contrôlée en appliquant les règles dont l'ordre est supérieure à une valeur k fixée, contraignant ainsi la similarité d'au moins k attributs. Elle peut également être contrôlée en triant, de manière croissante, les liens me découverts (v_i, v_j), selon l'ordre de la règle et le nombre de liens $friend$ communs entre (v_i, v_j). Après trois itérations le nombre de liens transversaux découverts par Liaison est de 6 572. L'évaluation et la comparaison de Liaison avec deux approches existantes ont montré sa robustesse (au regard de sa précision), ainsi que prouvé son efficacité dans la découverte d'un grand nombre de liens transversaux avec une performance de temps très satisfaisante.

Chapitre 5

CONCLUSIONS ET PERSPECTIVES

Le passage d'Internet vers le Web 2.0 est à l'origine de la création de plusieurs services de réseautage social, qui aujourd'hui occupent quotidiennement une place de plus en plus grandissante dans la vie d'un nombre important d'individus. Ces réseaux sociaux permettent aux utilisateurs de communiquer, de partager, et d'échanger des ressources telles que les photos, les vidéos, les données personnelles, et même leurs activités au sein de leurs profils. Ces profils utilisateurs sont une source de données importante pour les applications telles que les systèmes de recommandation cherchant à découvrir les intérêts professionnels et personnels des utilisateurs afin de les caractériser au mieux.

L'objectif de cette thèse a été d'exploiter les ressources textuelles que les utilisateurs publient dans leurs différents réseaux sociaux afin de construire un profil élargi exploitable pour la recommandation. Nous avons proposé des solutions automatiques et non supervisées en nous focalisant sur deux problématiques : (i) découvrir les intérêts des utilisateurs et leurs traits de personnalité (ii) réconcilier les différents profils des utilisateurs à travers différents réseaux sociaux.

Nous avons proposé une approche FRISK automatique, multilingue et non supervisée pour construire le profil d'intérêts d'un utilisateur. De nombreuses expérimentations, comparaisons et évaluations ont été effectuées. L'évaluation de FRISK a été faite sur 1 347 profils utilisateurs de TWITTER édités en anglais, français, espagnol et italien pour les intérêts "Politique", "Économie", "Jeu vidéo", "Gastronomie" et "Sports". Les résultats obtenus sont très encourageants puisque nous avons atteint une précision de 88% pour en moyenne 238 tweets par utilisateur. De plus, nous avons comparé FRISK avec quatre méthodes de classification : Machine à vecteur de support (SVM), Naives Bayes, Random Forest et l'allocation de dirichlet latente (LDA). Les résultats ainsi obtenus ont montré que FRISK dépasse largement toutes ces méthodes tout en restant stable par rapport au nombre de tweets par utilisateur. Ces résultats ont également montré que la capacité des méthodes de classification à prédire les intérêts des utilisateurs dans un contexte multilingue est très limitée. En particulier, SVM qui donne de meilleurs résultats par rapport aux autres méthodes de classification n'arrive pas à dépasser 60% de précision même si le nombre de tweets est élevé. Par ailleurs, nous avons également comparé FRISK à ces méthodes de classification dans un contexte monolingue. Dans la version monolingue de FRISK, nous avons introduit une variante de FRISK (FRISKTM) qui exploite la méthode de désambiguïsation TagMe pour déterminer les interprétations de chaque mot provenant des ressources textuelles exploitées. Les résultats ont montré que les méthodes de classification à l'exception de Random Forest sont meilleures que FRISK et sa variante FRISKTM. Néanmoins, FRISK est stable quelque soit la langue utilisée, or FRISKTM pour la langue anglaise est comparable à Naives Bayes et pour la langue italienne sa précision chute considérablement jusqu'à ce qu'elle se rapproche de la précision de Random Forest. En effet, TagMe est plus adaptée à la langue anglaise.

Une perspective possible pour introduire une désambiguïsation plus performante quelque soit la langue dans l'algorithme FRISK est d'exploiter les n-grammes. L'algorithme FRISK exploite séparément les mots des tweets, l'idée est de tenir compte des différentes interprétations des mots apparaissant dans le même contexte pour construire l'ensemble $\mathcal{BOA}^u$ à partir des tweets d'un utilisateur donné u.

Les résultats obtenus pour chaque intérêt séparément ont également montré que FRISK donne des précisions plus faibles 81% et 80% pour les intérêts "Politique" et "Économie". En effet, ces deux intérêts sont liés. Une autre perspective de notre travail de thèse consiste à étudier la prise en compte des relations sémantiques telles que la relation de subsomption (par exemple, "Sports" et "Football") entre les intérêts et leur impact sur le score affecté pour la classification des intérêts pour un utilisateur.

De même, des expérimentations et des évaluations pourraient être réalisées en incluant davantage

d'intérêts cibles et de langues dans notre jeu de données.

Nous avons également étudié le problème de la découverte des traits de personnalité des utilisateurs à partir de leurs tweets. A cet effet, nous avons défini la méthodologie ASCERTAIN qui utilise des outils existants d'une part, pour déterminer les dimensions psychologiques Receptiviti à partir des tweets des utilisateurs, et d'autre part, pour construire un modèle de prédiction des intérêts en utilisant les dimensions Receptiviti. Deux modèles de prédiction ont été étudiés à savoir le modèle utilisant la régression logistique polytomique multiple et le modèle utilisant la régression logistique multiple. De nombreuses analyses ont été réalisées afin de déterminer la corrélation qui peut exister entre les traits de personnalité des utilisateurs et différents intérêts cibles. Ces analyses nous ont permis de découvrir les traits de personnalité liés entre eux, notamment les dimensions *extraversion*, *cheerful*, *happiness* et *persuasive*. Les résultats obtenus montrent que la dimension *genuine* qui signifie authentique et honnête est le trait de personnalité qui caractérise le plus les utilisateurs qui ont comme intérêts "Gastronomie", "Sports" et "Jeu vidéo". Aussi, certains traits de personnalité permettent de représenter de manière distincte les utilisateurs ayant un intérêt précis. Par exemple, les résultats obtenus pour le modèle ayant "Politique" comme catégorie de référence montrent que les utilisateurs dont les intérêts sont "Gastronomie" et "Sports" ont moins d'empathie (*empathetic*) envers les autres par rapport à ceux qui ont comme intérêt "Politique". Les mesures de précision pour le modèle de prédiction basé sur la régression logistique polytomique multiple varient de 69% pour l'intérêt "Tourisme" à 89% pour l'intérêt "Gastronomie". Pour les mesures de rappel, la valeur la plus faible a été obtenue pour l'intérêt "Jeu vidéo" (69%). Les valeurs des mesures de précision et rappel les plus élevées sont également obtenues avec le modèle de prédiction utilisant la régression logistique multiple où l'intérêt "Gastronomie" est la catégorie cible (93% précision et 97% rappel).

Nous avons exploré d'autres méthodes de classification à savoir les machines à vecteur de support (SVM) et Naïves Bayes pour prédire les intérêts des utilisateurs. Les valeurs obtenues pour la précision sont de 18% pour SVM, 72% pour Naives Bayes et 75% pour ASCERTAIN. Une autre idée d'analyse serait de construire le modèle de prédiction avec les dimensions Receptiviti qui constituent les composantes principales obtenues lors de l'analyse en correspondance principale. De plus, le travail que nous avons fourni porte essentiellement sur les profils utilisateurs édités en anglais, il serait intéressant d'explorer d'autres langues afin de voir l'impact sur notre méthodologie.

Deux autres perspectives communes à FRISK et ASCERTAIN sont tout d'abord d'exploiter d'autres ressources que le texte telles que le graphe des amis et les communautés d'un utilisateur. Par exemple, analyser les différentes communautés d'un utilisateur afin de confirmer ou de découvrir de nouveaux intérêts. L'analyse des communautés d'un utilisateur consiste entre autres, à déterminer automatiquement les différentes thématiques portées par ses communautés ainsi que ses différents degrés d'implications dans chacune de ses communautés. Ainsi, on pourra dire que l'utilisateur s'intéresse à l'intérêt porté par la thématique d'une de ses communautés avec une importance qui est fonction de son degré d'implication dans ladite communauté. Par contre, l'utilisation du graphe des amis peut aider à étudier les relations d'un utilisateur, notamment les individus avec lesquels l'utilisateur cible est fortement en relation et qui peuvent avoir une influence sur son comportement.

La seconde perspective est de construire un jeu de données à partir des utilisateurs volontaires afin d'avoir les intérêts et traits de personnalité validés par ces utilisateurs. L'idée ici est de créer une base de données contenant les profils des utilisateurs volontaires qui pourront nous donner explicitement leurs tweets, intérêts ainsi que leurs traits de personnalité (à partir des questionnaires en ligne) que nous utiliserons plus tard dans les expérimentations et évaluation des approches FRISK et ASCERTAIN.

Nous avons proposé une méthode LIAISON de réconciliation des profils utilisateurs à travers

plusieurs réseaux sociaux à savoir LIVEJOURNAL, FLICKR, TWITTER et YOUTUBE. De nombreuses expérimentations, comparaisons et évaluations ont également été faites. Les résultats obtenus sont également très encourageants car LIAISON a atteint une précision de 94%. La comparaison avec deux approches de la littérature BUCC et MAL ont montré la pertinence de LIAISON. La valeur de précision obtenue pour l'approche BUCC sur le jeu de données que nous avons enrichi est de 80% qui est faible par rapport à celle de LIAISON. Par contre, la valeur de précision obtenue pour l'approche MAL sur un jeu de données construit par ses auteurs est de 64%, tandis que LIAISON présente 86% de précision sur le même jeu de données. Les valeurs de rappel des approches BUCC et MAL n'ont pas été mentionnées par leurs auteurs.

Deux perspectives pour améliorer la réconciliation des profils sont dans un premier temps d'exploiter d'autres attributs tels que les communautés. Par exemple, analyser les communautés de deux profils utilisateurs lors de la réconciliation afin de déterminer le nombre de communautés de chacun de ces profils qui portent des thématiques similaires (ont des objectifs similaires).

La seconde perspective est de tenir compte du contenu des pages WEBSITES mentionnées dans les profils utilisateurs lors de la comparaison des profils candidats. Actuellement, dans LIAISON les URLs des attributs WEBSITES extraites des profils candidats sont comparées en utilisant l'égalité mathématique sans tenir compte du contenu des pages associées. Or un utilisateur peut indiquer dans un de ces profils une URL qui est une redirection vers une page Web dont l'adresse est explicitement indiquée dans un autre de ses profils.

Il peut également être intéressant d'explorer plus en détails l'utilisation de la topologie du réseau afin de généraliser notre algorithme aux réseaux où les valeurs d'attributs sont anonymisées. Par exemple, en incluant dans la topologie les communautés et les ressources externes indiquées par les utilisateurs dans leurs profils respectifs.

Les travaux de recherche réalisés dans cette thèse vont permettre à Wepingo de perfectionner leur système de recommandation grâce à un meilleur profilage de ses utilisateurs. Wepingo a accès aux données publiées par ses utilisateurs sur Facebook et/ou Twitter, données exploitées par les algorithmes proposés dans cette thèse.

Bibliographie

[1] H. Akaike. Likelihood of a model and information criteria. *Journal of econometrics*, 16(1):3–14, 1981. (Cited on page 87.)

[2] Y. Bachrach, M. Kosinski, T. Graepel, P. Kohli, and D. Stillwell. Personality and patterns of facebook usage. In *Proceedings of the 4th Annual ACM Web Science Conference*, pages 24–32. ACM, 2012. (Cited on pages 73 and 74.)

[3] S. Bartunov, A. Korshunov, S. Park, W. Ryu, and H. Lee. Joint Link-attribute User Identity Resolution in Online Social Networks. In *SNA-KDD Workshop*, 2012. (Cited on page 115.)

[4] A. M. Benis. Npa personality theory. *Speculations in Science & Technology*, 13(3):167–175, 1990. (Cited on page 72.)

[5] N. Bennacer, C. N. Jipmo, A. Penta, and G. Quercini. Matching user profiles across social networks. In *International Conference on Advanced Information Systems Engineering*, pages 424–438. CAISE Springer, 2014. (Cited on page 128.)

[6] P. Bhattacharya, M. B. Zafar, N. Ganguly, S. Ghosh, and K. P. Gummadi. Inferring User Interests in the Twitter Social Network. In *RecSys*, pages 357–360, 2014. (Cited on pages 12 and 14.)

[7] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet Allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003. (Cited on pages 13 and 74.)

[8] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, 2008. (Cited on pages 3 and 4.)

[9] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. *Computer Networks*, 30(1-7):107–117, 1998. (Cited on pages 12, 15 and 27.)

[10] F. Buccafurri, G. Lax, A. Nocera, and D. Ursino. Discovering Links among Social Networks. In *Machine Learning and Knowledge Discovery in Databases*, volume 7524 of *Lecture Notes in Computer Science*, pages 467–482. Springer Berlin Heidelberg, 2012. (Cited on pages 38, 39, 115, 118, 127 and 130.)

[11] W. A. Burkhard and R. M. Keller. Some Approaches to Best-match File Searching. *Commun. ACM*, 16(4):230–236, apr 1973. (Cited on page 126.)

[12] I. Cantador, I. Fernández-Tobías, and A. Bellogín. Relating personality types with user preferences in multiple entertainment domains. In *CEUR Workshop Proceedings*. Shlomo Berkovsky, 2013. (Cited on page 72.)

[13] F. Carmagnola and F. Cena. User Identification for Cross-system Personalisation. *Inf. Sci.*, 179(1-2):16–32, 2009. (Cited on pages 114 and 115.)

[14] J. Chen, O. Zaïane, and R. Goebel. Local community identification in social networks. In *Social Network Analysis and Mining, 2009. ASONAM'09. International Conference on Advances in*, pages 237–242. IEEE, 2009. (Cited on page 3.)

[15] R. L. Cilibrasi and P. M. Vitanyi. The google similarity distance. *IEEE Transactions on knowledge and data engineering*, 19(3), 2007. (Cited on page 27.)

[16] A. Clauset, M. E. Newman, and C. Moore. Finding community structure in very large networks. *Physical review E*, 70(6):066111, 2004. (Cited on pages 3 and 4.)

[17] K. Cortis, S. Scerri, I. Rivera, and S. Handschuh. Discovering Semantic Equivalence of People Behind Online Profiles. In *In Proceedings of the Resource Discovery (RED) Workshop, ser. ESWC*, 2012. (Cited on pages 114 and 115.)

[18] J. M. Digman. Personality structure: Emergence of the five-factor model. *Annual review of psychology*, 41(1):417–440, 1990. (Cited on page 76.)

[19] Y. Ding and J. Jiang. Extracting Interest Tags from Twitter User Biographies. In *Information Retrieval Technology*, pages 268–279. 2014. (Cited on page 12.)

[20] P. Ferragina and U. Scaiella. Tagme: On-the-Fly Annotation of Short Text Fragments (by Wikipedia Entities). In *CIKM*, pages 1625–1628, 2010. (Cited on page 66.)

[21] FriendFeed. `friendfeed.com`, 2007. (Cited on page 116.)

[22] M. Gamon, A. Aue, S. Corston-Oliver, and E. Ringger. Pulse: Mining Customer Opinions from Free Text. In *Advances in Intelligent Data Analysis VI*, pages 121–132. Springer, 2005. (Cited on page 4.)

[23] O. Goga, H. Lei, S. H. K. Parthasarathi, G. Friedland, R. Sommer, and R. Teixeira. Exploiting Innocuous Activity for Correlating Users across Sites. In *Proceedings of the 22nd international conference on World Wide Web*, pages 447–458. International World Wide Web Conferences Steering Committee, 2013. (Cited on page 114.)

[24] J. Golbeck, C. Robles, M. Edmondson, and K. Turner. Predicting personality from twitter. In *Privacy, security, risk and trust (passat), 2011 ieee third international conference on and 2011 ieee third international conference on social computing (socialcom)*, pages 149–156. IEEE, 2011. (Cited on pages 73 and 74.)

[25] J. Golbeck and M. Rothstein. Linking Social Networks on the Web with FOAF: A Semantic Web Case Study. In *AAAI*, volume 8, pages 1138–1143, 2008. (Cited on pages 114, 115 and 124.)

[26] L. R. Goldberg. An alternative" description of personality": the big-five factor structure. *Journal of personality and social psychology*, 59(6):1216, 1990. (Cited on pages 72, 75 and 76.)

[27] L. R. Goldberg. The structure of phenotypic personality traits. *American psychologist*, 48(1):26, 1993. (Cited on page 76.)

[28] L. A. Gottschalk. The unobtrusive measurement of psychological states and traits. *Text analysis for the social sciences: Methods for drawing statistical inferences from texts and transcripts*, pages 117–129, 1997. (Cited on page 75.)

[29] L. A. Gottschalk and G. C. Gleser. *The measurement of psychological states through the content analysis of verbal behavior*. Univ of California Press, 1969. (Cited on page 75.)

[30] R. Gross and A. Acquisti. Information Revelation and Privacy in Online Social Networks. In *Proceedings of the 2005 ACM Workshop on Privacy in the Electronic Society*, WPES '05, pages 71–80, New York, NY, USA, 2005. ACM. (Cited on page 112.)

[31] W. He, H. Liu, J. He, S. Tang, and X. Du. Extracting Interest Tags for Non-famous Users in Social Network. In *CIKM*, pages 861–870. ACM, 2015. (Cited on page 12.)

[32] J. B. Hirsh and J. B. Peterson. Personality and language use in self-narratives. *Journal of research in personality*, 43(3):524–527, 2009. (Cited on page 75.)

[33] P. Jain, P. Kumaraguru, and A. Joshi. @i Seek 'fb.me': Identifying Users Across Multiple Online Social Networks. In *WWW (Companion Volume)*, pages 1259–1268, 2013. (Cited on page 115.)

[34] D. Jannach, M. Zanker, A. Felfernig, and G. Friedrich. *Recommender Systems: An Introduction*. Cambridge University Press, 2010. (Cited on page 4.)

[35] C. N. Jipmo, G. Quercini, and N. Bennacer. Catégorisation et désambiguïsation des intérêts des individus dans le web social. In *EGC*, pages 523–524, 2016. (Cited on pages 12 and 29.)

[36] C. N. Jipmo, G. Quercini, and N. Bennacer. Catégorisation et Désambiguïsation des Intérêts des Individus dans le Web Social. In *EGC*, pages 523–524, 2016. (Cited on page 51.)

[37] M. Joshi and N. Belsare. BlogHarvest: Blog Mining and Search Framework. In *COMAD*, pages 226–229, 2006. (Cited on page 4.)

[38] C. G. Jung. Psychological types: or the psychology of individuation. 1923. (Cited on page 72.)

[39] R. Kannan, S. Vempala, and A. Vetta. On clusterings: Good, bad and spectral. *Journal of the ACM (JACM)*, 51(3):497–515, 2004. (Cited on page 4.)

[40] A. D. Kramer and K. Rodden. Word usage and posting behaviors: modeling blogs with unobtrusive data collection methods. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1125–1128. ACM, 2008. (Cited on pages 73 and 74.)

[41] B. Krishnamurthy and C. E. Wills. On the Leakage of Personally Identifiable Information via Online Social Networks. In *Proceedings of the 2nd ACM Workshop on Online Social Networks*, pages 7–12. ACM, 2009. (Cited on page 117.)

[42] J. Lafferty, A. McCallum, and F. C. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. 2001. (Cited on page 12.)

[43] X. Li, L. Guo, and Y. E. Zhao. Tag-based Social Interest Discovery. In *WWW*, pages 675–684, 2008. (Cited on page 12.)

[44] L. Little, P. Briggs, and L. Coventry. Who Knows about Me?: An Analysis of Age-related Disclosure Preferences. In *Proceedings of the 25th BCS Conference on Human-Computer Interaction*, BCS-HCI '11, pages 84–87, Swinton, UK, UK, 2011. British Computer Society. (Cited on page 112.)

[45] B. Liu. Sentiment Analysis and Opinion Mining. *Synthesis Lectures on Human Language Technologies*, 5(1):1–167, 2012. (Cited on page 4.)

[46] A. Malhotra, L. Totti, W. Meira, P. Kumaraguru, and V. Almeida. Studying User Footprints in Different Online Social Networks. In *International Workshop on Cybersecurity of Online Social Network (ACM ASONAM 2012)*, 2012. (Cited on pages 114, 121, 130 and 131.)

[47] M. Michelson and S. A. Macskassy. Discovering Users' Topics of Interest on Twitter: A First Look. In *4th Workshop on Analytics for Noisy Unstructured Text Data*, pages 73–80. ACM, 2010. (Cited on pages 12 and 14.)

[48] R. Mihalcea and P. Tarau. Textrank: Bringing order into text. In *EMNLP*, volume 4, pages 404–411, 2004. (Cited on page 13.)

[49] D. Milne and I. H. Witten. An effective, low-cost measure of semantic relatedness obtained from wikipedia links. In *In Proceeding of AAAI Workshop on Wikipedia and Artificial Intelligence: an Evolving Synergy*, pages 25–30. AAAI Press, 2008. (Cited on pages 12, 15 and 24.)

[50] A. Moro, A. Raganato, and R. Navigli. Entity Linking meets Word Sense Disambiguation: a Unified Approach. *TACL*, 2:231–244, 2014. (Cited on page 70.)

[51] M. Motoyama and G. Varghese. I Seek You: Searching and Matching Individuals in Social Networks. In *Proceedings of the Eleventh International Workshop on Web Information and Data Management*, pages 67–75. ACM, 2009. (Cited on page 114.)

[52] A. Narayanan and V. Shmatikov. De-anonymizing Social Networks. In *30th IEEE Symposium on Security and Privacy*, pages 173–187. IEEE, 2009. (Cited on pages 115 and 116.)

[53] J. Nerbonne. The secret life of pronouns. what our words say about us. *Literary and Linguistic Computing*, 29(1):139–142, 2014. (Cited on page 75.)

[54] M. E. Newman. Fast algorithm for detecting community structure in networks. *Physical review E*, 69(6):066133, 2004. (Cited on pages 3 and 4.)

[55] M. E. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical review E*, 69(2):026113, 2004. (Cited on pages 3 and 4.)

[56] N. P. Nguyen, T. N. Dinh, Y. Xuan, and M. T. Thai. Adaptive algorithms for detecting community structure in dynamic social networks. In *INFOCOM, 2011 Proceedings IEEE*, pages 2282–2290. IEEE, 2011. (Cited on page 3.)

[57] M. Obschonka and C. Fisch. Entrepreneurial personalities in political leadership. *Small Business Economics*, pages 1–19, 2017. (Cited on page 75.)

[58] M. Obschonka, C. Fisch, and R. Boyd. Using digital footprints in entrepreneurship research: A twitter-based personality analysis of superstar entrepreneurs and managers. *Journal of Business Venturing Insights*, 8:13–23, 2017. (Cited on page 75.)

[59] A. Ochiai. Zoogeographic Studies on the Soleoid Fishes Found in Japan and its Neighbouring Regions. *Bull. Jpn. Soc. Sci. Fish*, 22(9):526–530, 1957. (Cited on page 121.)

[60] A. Odic, M. Tkalcic, J. F. Tasic, and A. Kosir. Personality and social context: Impact on emotion induction from movies. In *UMAP Workshops*, 2013. (Cited on page 72.)

[61] D. Parra and P. Brusilovsky. Collaborative Filtering for Social Tagging Systems: an Experiment with CiteULike. In *Proceedings of the Third ACM Conference on Recommender Systems*, pages 237–240. ACM, 2009. (Cited on page 4.)

[62] D. M. Pennock, E. Horvitz, S. Lawrence, and C. L. Giles. Collaborative filtering by personality diagnosis: A hybrid memory-and model-based approach. In *Proceedings of the Sixteenth conference on Uncertainty in artificial intelligence*, pages 473–480. Morgan Kaufmann Publishers Inc., 2000. (Cited on pages 73 and 74.)

[63] D. Perito, C. Castelluccia, M. A. Kaafar, and P. Manils. How Unique and Traceable are Usernames? In *Privacy Enhancing Technologies*, pages 1–17. Springer, 2011. (Cited on pages 114 and 118.)

[64] Plaxo. http://www.plaxo.com, 2002. (Cited on page 116.)

[65] G. Quercini, N. Bennacer, M. Ghufran, and C. N. Jipmo. Liaison: reconciliation of individuals profiles across social networks. In *Advances in Knowledge Discovery and Management*, pages 229–253. Springer, 2017. (Cited on pages 113 and 123.)

[66] E. Raad, R. Chbeir, and A. Dipanda. User Profile Matching in Social Networks. In *Network-Based Information Systems (NBiS), 2010 13th International Conference on*, pages 297–304. IEEE, 2010. (Cited on pages 114 and 115.)

[67] M. Raghuram, K. Akshay, and K. Chandrasekaran. Efficient User Profiling in Twitter Social Network Using Traditional Classifiers. In *Intelligent Systems Technologies and Applications*, pages 399–411. 2016. (Cited on pages 12 and 13.)

[68] M. S. Raje and A. Singh. Personality detection by analysis of twitter profiles. In *International Conference on Soft Computing and Pattern Recognition*, pages 667–675. Springer, 2016. (Cited on page 75.)

[69] D. Rawlings and V. Ciancarelli. Music preference and the five-factor model of the neo personality inventory. *Psychology of Music*, 25(2):120–132, 1997. (Cited on page 72.)

[70] P. J. Rentfrow, L. R. Goldberg, and R. Zilca. Listening, watching, and reading: The structure and correlates of entertainment preferences. *Journal of personality*, 79(2):223–258, 2011. (Cited on page 72.)

[71] P. J. Rentfrow and S. D. Gosling. The do re mi's of everyday life: the structure and personality correlates of music preferences. *Journal of personality and social psychology*, 84(6):1236, 2003. (Cited on page 72.)

[72] M. Rowe. Interlinking Distributed Social Graphs. In *Linked Data on the Web Workshop, WWW2009*, 2009. (Cited on pages 114 and 115.)

[73] H. A. Schwartz, J. C. Eichstaedt, M. L. Kern, L. Dziurzynski, S. M. Ramones, M. Agrawal, A. Shah, M. Kosinski, D. Stillwell, M. E. Seligman, et al. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, 8(9):e73791, 2013. (Cited on pages 73 and 74.)

[74] N. Spasojevic, J. Yan, A. Rao, and P. Bhattacharyya. LASTA: Large Scale Topic Assignment on Multiple Social Networks. In *KDD*, pages 1809–1818, 2014. (Cited on pages 12 and 13.)

[75] Spokeo. http://www.spokeo.com, 2006. (Cited on page 116.)

[76] K. Steinhaeuser and N. V. Chawla. Identifying and evaluating community structure in complex networks. *Pattern Recognition Letters*, 31(5):413–421, 2010. (Cited on page 3.)

[77] J. Stoyanovich, S. Amer-Yahia, C. Marlow, and C. Yu. Leveraging Tagging to Model User Interests in del. icio. us. In *AAAI Spring Symposium: Social Information Processing*, pages 104–109, 2008. (Cited on page 12.)

[78] F. Stutzman. An Evaluation of Identity-Sharing Behavior in Social Network Communities. *iDMAa Journal*, 3(1), 2006. (Cited on page 112.)

[79] Y. R. Tausczik and J. W. Pennebaker. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54, 2010. (Cited on page 73.)

[80] K. H. Tso-Sutter, L. B. Marinho, and L. Schmidt-Thieme. Tag-aware Recommender Systems by Fusion of Collaborative Filtering Algorithms. In *Proceedings of the 2008 ACM Symposium on Applied Computing*, pages 1995–1999. ACM, 2008. (Cited on page 4.)

[81] E. C. Tupes and R. E. Christal. Recurrent personality factors based on trait ratings. *Journal of personality*, 60(2):225–251, 1992. (Cited on page 76.)

[82] E. Viennet et al. Community detection based on structural and attribute similarities. In *ICDS 2012, The Sixth International Conference on Digital Society*, pages 7–12, 2012. (Cited on page 4.)

[83] T. Vu and V. Perez. Interest Mining from User Tweets. In *CIKM*, pages 1869–1872, 2013. (Cited on pages 12 and 13.)

[84] T. Wang, H. Liu, J. He, and X. Du. Mining User Interests from Information Sharing Behaviors in Social Media. In *Advances in Knowledge Discovery and Data Mining*, pages 85–98. 2013. (Cited on pages 12 and 13.)

[85] X. Wang, H. Liu, and W. Fan. Connecting Users with Similar Interests via Tag Network Inference. In *CIKM*, pages 1019–1024. ACM, 2011. (Cited on pages 12 and 14.)

[86] Z. Wen and C.-Y. Lin. On the Quality of Inferring Interests from Social Neighbors. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '10, pages 373–382, New York, NY, USA, 2010. ACM. (Cited on page 12.)

[87] Z. Wen and C.-Y. Lin. Improving User Interest Inference from Social Neighbors. In *CIKM*, pages 1001–1006, 2011. (Cited on page 12.)

[88] J. Weng, E.-P. Lim, J. Jiang, and Q. He. TwitterRank: Finding Topic-sensitive Influential Twitterers. In *WSDM*, pages 261–270, 2010. (Cited on pages 12, 13 and 64.)

[89] Z. Xu, R. Lu, L. Xiang, and Q. Yang. Discovering User Interest on Twitter with a Modified Author-topic Model. In *WI-IAT*, volume 1, pages 422–429, 2011. (Cited on pages 12 and 13.)

[90] R. Zafarani and H. Liu. Connecting Corresponding Identities across Communities. In *Third International AAAI Conference on Weblogs and Social Media*, 2009. (Cited on pages 114 and 118.)

[91] F. Zarrinkalam, H. Fani, E. Bagheri, M. Kahani, and W. Du. Semantics-Enabled User Interest Detection from Twitter. In *WI-IAT*, volume 1, pages 469–476, 2015. (Cited on pages 12 and 14.)

[92] Z.-K. Zhang, T. Zhou, and Y.-C. Zhang. Personalized Recommendation via Integrated Diffusion on User–Item–Tag Tripartite Graphs. *Physica A: Statistical Mechanics and its Applications*, 389(1):179–186, 2010. (Cited on page 4.)

Annexes

1 Description des dimensions LIWC

La description des différentes catégories et dimensions LIWC est la suivante :

Word count

Summary Language Variables

Analytical thinking - Clout - Authentic (evaluates the degree to which a person actively filters what they're saying for the audience) - Emotional tone (positive/negative) - Words/sentence - Words > 6 letters - Dictionary words.

Linguistic Dimensions

Total function words - Total pronouns - Personal pronouns - 1st pers singular- 1st pers plural - 2nd person - 3rd pers singular - 3rd pers plural - Impersonal pronouns - Articles - Prepositions - Auxiliary verbs - Common Adverbs - Conjunctions - Negations.

Other Grammar

Common verbs - Common adjectives - Comparisons - Interrogatives - Numbers - Quantifiers

Psychological Processes

Affective processes - Positive emotion - Negative emotion - Anxiety - Anger - Sadness - Social processes - Family - Friends - Female references - Male references - Cognitive processes - Insight - Causation - Discrepancy - Tentative - Certainty - Differentiation - Perceptual processes - See - Hear - Feel - Biological processes - Body - Health - Sexual - Ingestion - Drives - Affiliation - Achievement - Power - Reward - Risk - Time orientations - Past focus - Present focus - Future focus - Relativity - Motion - Space - Time.

Personal concerns

Work - Leisure - Home - Money - Religion - Death - Informal language - Swear words - Netspeak - Assent - Nonfluencies - Fillers.

2 Description des dimensions Receptiviti

La description des différentes catégories et dimensions Receptiviti est la suivante :

Cognitive/Thinking Style Insights

Thinking Style : Measures the degree to which the person is an analytical thinker who relies on facts and data or instinct and feelings when making decisions.

Persuasive : Measures the degree to which a person is able to create rapport with the intention of persuading others.

Reward Bias : Measures the degree to which a person weighs risks vs. rewards when making decisions.

Big 5 Insights

Openness : Measures the degree to which a person is open to new ideas and new experiences.

Artistic : Measures how much a person appreciates and enjoys the arts.

Intellectual : Measures how strongly a person is inclined toward intellectual and academic learning.

Liberal : Measures how socially and ideologically liberal a person is.

Imaginative : Measures to what degree a person is imaginative.

Emotionally Aware : Measures to what degree a person is conscious of and connected with their feelings and emotions.

Adventurous : Measures the degree to which a person enjoys and seeks out adventure.

Conscientiousness : Measures the degree to which a person is reliable.

Self-assured : Measures how much confidence a person has in themselves.

Disciplined : Measures a person's propensity to follow routines and rules.

Ambitious : Measures the degree to which a person is ambitious or driven by the desire for achievement.

Dutiful : Measures a person's sense that they should respect expectations and authority.

Cautious : Measures how cautiously a person tends to act.

Organized : Measures how organized and orderly a person is.

Extraversion : Measures the degree to which a person feels energized and uplifted when interacting with others or engaging in activity.

Sociable : Measures how much a person seeks out and enjoys social situations.

Friendly : Measures how friendly a person generally is and how positive they are when interacting with others.

Assertive : Measures how assertive a person is and how comfortable a person is with expressing their ideas and needs.

Energetic : Measures how much energy and enthusiasm a person tends to have.

Cheerful : Measures how happy and cheerful a person generally acts.

Active : Measures how strongly a person feels the need for activity and engagement in their life.

Agreeableness : Measures the degree to which a person is inclined to please others.

Generous : Measures how much a person enjoys spending their time and money on others.

Trusting : Measures how easily a person trusts others.

Cooperative : Measures how well a person takes into account the needs of others.

Empathetic : Measures how strongly a person internalizes the feelings of others.

Genuine : Measures how genuine and honest a person is.

Humble : Measures how humble and modest a person is.

Neuroticism : Measures the degree to which a person expresses strong negative emotions.

Impulsive : Measures how inclined a person is to act impulsively.

Stressed : Measures the degree to which a person is experiencing stress and how strongly affected they are by it.

Anxious : Measures the degree to which a person is experiencing anxiety and how strongly affected they are by it.

Aggressive : Measures the degree to which a person exhibits anger or aggression.

Melancholy : Measures how much a person is expressing sadness.

Self-conscious : Measures how likely a person is to feel embarrassed or anxious about themselves or their skills.

Social Style Insights

Social Skills : Measures the degree to which a person feels at ease with others and is able to navigate social situations.

Insecure : Measures the degree to which a person lacks confidence when dealing with others.

Cold : Measures the degree to which a person is emotionally unresponsive and has difficulty empathizing with others.

Family Orientation : Measures the degree to which a persons values and behaviors are rooted in their sense of family.

Emotional Style Insights

Adjustment : Measures the degree to which a person is grounded, is able to maintain quality relationships with others, and establishes healthy life goals.

Happiness : Measures the degree to which a person is optimistic, upbeat, and happy.

Depression : Measures the degree to which a person may have difficulty finding joy in their life.

Working Style Insights

Independent : Measures the degree to which a person is a non-conformist.

Power Driven : Measures the degree to which a person is driven by the desire for power.

Type-A : Measures the degree to which a person is driven and competitive.

Workhorse : Measures the degree to which a person has a strong work ethic vs. preference for leisure and non-work activity.

Interests and Orientations

Friendship Focused : Measures the degree to which a person focuses on friends and friendship, and likely spends time thinking about their social connections.

Body Focus : Measures the degree to which a person focuses attention on their body or other people's bodies.

Health Oriented : Measures the degree to which a person is focused on health, likely spends time thinking about their own health or the health of others.

Sexual Focus : Measures the degree to which a person focuses on sexuality, sex-related themes, concepts and ideas.

Food Focus : Measures the degree to which a person focuses thoughts on eating or drinking, and likely enjoys discussing food or drinks with others.

Leisure Oriented : Measures the degree to which a person thinks about leisure activities such as sports, entertainment, travel, or organized events.

Money Oriented : Measures the degree to which a person thinks about money and finances. May be focused on personal finances, the broader economy or both.

Religion Oriented : Measures the degree to which a person focuses on religion, and likely spends time discussing religion, religious themes, ideas and topics.

Work Oriented : Measures the degree to which a person is focused on, or preoccupied with work or school.

Netspeak : Measures the degree to which a person is comfortable communicating with Internet shorthand and instant messaging slang, abbreviated words, acronyms and special characters.

université PARIS-SACLAY
ÉCOLE DOCTORALE
Sciences et technologies de l'information et de la communication (STIC)

Titre : Intégration du Web Social dans les Systèmes de Recommandation

Mots Clés: Web social, Text mining, Traitement multilingue, Wikipédia, Personnalité.

Résumé. Le Web social croît de plus en plus et donne accès à une multitude de ressources très variées, qui proviennent de sites de partage tels que del.icio.us, d'échange de messages comme Twitter, des réseaux sociaux à finalité professionnelle, comme LinkedIn, ou plus généralement à finalité sociale, comme Facebook et LiveJournal. Un même individu peut être inscrit et actif sur différents réseaux sociaux ayant potentiellement des finalités différentes, où il publie des informations diverses et variées, telles que son nom, sa localité, ses communautés, et ses différentes activités. Ces informations (textuelles), au vu de la dimension internationale du Web, sont par nature, d'une part multilingue, et d'autre part, intrinsèquement ambiguë puisqu'elles sont éditées par les individus en langage naturel dans un vocabulaire libre. De même, elles sont une source de données précieuses, notamment pour les applications cherchant à connaître leurs utilisateurs afin de mieux comprendre leurs besoins et leurs intérêts. L'objectif de nos travaux de recherche est d'exploiter, en utilisant essentiellement l'encyclopédie Wikipédia, les ressources textuelles des utilisateurs extraites de leurs différents réseaux sociaux afin de construire un profil élargi les caractérisant et exploitable par des applications telles que les systèmes de recommandation. En particulier, nous avons réalisé une étude afin de caractériser les traits de personnalité des utilisateurs. De nombreuses expérimentations, analyses et évaluations ont été réalisées sur des données réelles collectées à partir de différents réseaux sociaux.

Titre : Social Web Integration in Recommendation Systems

Keywords: Social web, Text mining, Multilingual processing, Wikipédia, Personality.

Abstract. The social Web grows more and more and gives through the web, access to a wide variety of resources, like sharing sites such as del.icio.us, exchange messages as Twitter, or social networks with the professional purpose such as LinkedIn, or more generally for social purposes, such as Facebook and LiveJournal. The same individual can be registered and active on different social networks (potentially having different purposes), in which it publishes various information, which are constantly growing, such as its name, locality, communities, various activities. The information (textual), given the international dimension of the Web, is inherently multilingual and intrinsically ambiguous, since it is published in natural language in a free vocabulary by individuals from different origin. They are also important, specially for applications seeking to know their users in order to better understand their needs, activities and interests. The objective of our research is to exploit using essentially the Wikpédia encyclopedia, the textual resources extracted from the different social networks of the same individual in order to construct his characterizing profile, which can be exploited in particular by applications seeking to understand their users, such as recommendation systems. In particular, we conducted a study to characterize the personality traits of users. Many experiments, analyzes and evaluations were carried out on real data collected from different social networks.