UNIVERSITÉ DE BOURGOGNE

UFR Sciences et Techniques

THÈSE

Pour obtenir le grade de

Docteur de l'Université de Bourgogne

Spécialité : Informatique

par

Damien LEPROVOST

30 novembre 2012

Découverte et analyse des communautés implicites par une approche sémantique en ligne : l'outil WebTribe

devant le jury composé de

Bernd AMANN	Professeur (Université Pierre et Marie Curie, Paris)	Rapporteur
Stefano A. CERRI	Professeur (Université Montpellier 2, Montpellier)	Rapporteur
Danielle BOULANGER	Professeur (Université Jean Moulin Lyon 3, Lyon)	Examinateur
Fabien GANDON	Chargé de Recherche – HDR (Inria, Sophia Antipolis)	Examinateur
Lylia ABROUK-GOUAÏCH	Maître de Conférences (Université de Bourgogne, Dijon)	Co-encadrante de thèse
David GROSS-AMBLARD	Professeur (Université de Rennes 1, Rennes)	Co-encadrant de thèse
Nadine CULLOT	Professeur (Université de Bourgogne, Dijon)	Directrice de thèse

Laboratoire Électronique, Informatique et Image – LE2I

Équipe Systèmes d'Information et Systèmes d'Image

Remerciements

Je tiens tout d'abord à remercier Christophe Nicolle pour m'avoir encadré durant les deux premières années de cette thèse, ainsi que Nadine Cullot pour en avoir assumé la direction par la suite. Leurs expériences respectives dont ils ont su me faire profiter ont permis l'accomplissement de ce travail.

Je remercie tout particulièrement mes co-encadrants, Lylia Abrouk-Gouaïch et David Gross-Amblard, pour m'avoir montré la voie depuis le début de cette aventure. Leur rigueur scientifique et leur façon de penser resteront toujours un exemple pour moi. J'espère que Géraldine, Abdelkader, Malik et Samy sauront me pardonner de les avoir parfois accaparés à des heures inavouables.

Je tiens à exprimer mes remerciements aux membres du jury, qui ont accepté d'évaluer mon travail de thèse. Un grand merci à Bernd Amann et Stefano A. Cerri pour avoir accepté d'être rapporteurs de ce travail. Je voudrais exprimer également ma reconnaissance envers Danielle Boulanger et Fabien Gandon pour m'avoir fait l'honneur de participer à mon jury.

Je souhaite exprimer ma profonde reconnaissance envers Serge Abiteboul, Sihem Amer-Yahia, Zohra Bellahsene Cédric de Mouza, Michel Scholl et Fabian Suchanek avec qui j'ai eu la chance de pouvoir échanger. J'ai beaucoup appris à leur contact.

Mes remerciements vont également aux membres du département informatique du Le2i que j'ai eu la chance de côtoyer et dont j'ai pu apprécier les conseils : Jean-Luc Baril, Richard Chbeir, Albert Dipanda, Dominique Faudot, Irène Foucherot, Thierry Grison, Sandrine Lanquetin, Joël Savelli, Marinette Savonnet, Olivier Togni et Kokou Yétongnon. Merci à tous pour votre accueil. Merci également à Mélanie Arnoult, Nadia Bader et Dounia Radi pour leur extrême disponibilité.

Je remercie l'ensemble des doctorants – ou jeunes docteurs à présent – que j'ai pu fréquenter durant cette thèse. Merci donc Mohamed Ryadh Dahimene, Zeinab Hmedeh, Roxana Horincar, Clément Mignard, Romain Picot-Clemente, Perrine Pittet, Eli Raad, Fekade Getahun Taddesse, Nelly Vouzoukidou. Une pensée toute spéciale pour mes compagnons de dispositif JCE : Alban Bajard, Lauriane Boisard, Charles-Henri Hage, Souhaiel Khalfaoui, Nicolas Navoret, Virginie Molinier. Bonne chance pour la suite !

Je suis profondément reconnaissant envers mes parents d'avoir su développer et cultiver ainsi ma curiosité, de m'avoir toujours soutenu et permis de poursuivre aussi loin. Cette thèse est aussi la leur.

La thèse, comme toutes les expériences passionnantes, est grande consommatrice de temps et d'énergie. Merci à Mylène de toujours éclairer notre chemin.

Résumé

Avec l'essor du Web 2.0 et des technologies collaboratives qui y sont rattachées, le Web est aujourd'hui devenu une vaste plate-forme d'échanges entre internautes. La majeure partie des sites Web sont actuellement soit dédiés aux interactions sociales de leurs utilisateurs, soit proposent des outils pour développer ces interactions. Nos travaux portent sur la compréhension de ces échanges, ainsi que des structures communautaires qui en découlent, au moyen d'une approche sémantique. Pour répondre aux besoins de compréhension propres aux analystes de site Web et autres gestionnaires de communautés, nous analysons ces structures communautaires pour en extraire des caractéristiques essentielles comme leurs centres thématiques et contributeurs centraux. Notre analyse sémantique s'appuie notamment sur des ontologies légères de référence pour définir plusieurs nouvelles métriques, comme la *centralité sémantique temporelle* et la *probabilité de propagation sémantique*. Nous employons une approche « en ligne » afin de suivre l'activité utilisateur en temps réel, au sein de notre outil d'analyse communautaire WebTribe. Nous avons implémenté et testé nos méthodes sur des données extraites de systèmes réels de communication sociale sur le Web.

Mots-clés : Communautés, Analyse sémantique, Ontologie, Système de communication, Réseaux sociaux, Web 2.0, Collaboratif.

Abstract

With the rise of Web 2.0 and collaborative technologies that are attached to, the Web has now become a broad platform of exchanges between users. The majority of websites is now dedicated to social interactions of their users, or offers tools to develop these interactions. Our work focuses on the understanding of these exchanges, as well as emerging community structures arising, through a semantic approach. To meet the needs of web analysts, we analyze these community structures to identify their essential characteristics as their thematic centers and central contributors. Our semantic analysis is mainly based on reference light ontologies to define several new metrics such as the *temporal semantic centrality* and the *semantic propagation probability*. We employ an online approach to monitor user activity in real time in our community analysis tool WebTribe. We have implemented and tested our methods on real data from social communication systems on the Web.

Keywords: Communities, Semantic analysis, Ontology, Communication system, Social network, Web 2.0, Collaborative.

Table des matières

Chapitre 1

Introduction

Sommaire

« Je ne connais pas la moitié d'entre vous autant que je le voudrais. Et j'aime moins de la moitié d'entre vous à moitié moins que vous ne le méritez. »

Bilbon Sacquet
La ***communauté*** de l'anneau

1.1 Contexte

Un Web en évolution Dans sa vision traditionnelle, le World Wide Web permet au grand public de consulter des pages aux contenus variés. L'ensemble des utilisateurs pouvaient accéder à des informations mises à disposition par les concepteurs de ces pages Web. Ces concepteurs représentaient nécessairement un groupe restreint d'utilisateurs, ne serait-ce que par les compétences techniques à mobiliser. Bien qu'un utilisateur pouvait appartenir aux deux groupes, nous avions donc un noyau de contributeurs, pour un ensemble plus large de consommateurs. Le Web traditionnel se caractérisait donc ainsi : une plate-forme informative, où le grand public était consommateur de l'information.

Depuis une dizaine d'années, une évolution de l'utilisation du Web s'observe. Cette évolution se caractérise par la simplification des méthodes de publication de contenu, ainsi que par l'interactivité accrue des plates-formes qui manipulent ce même contenu. Le Web se transforme peu à peu en un système ouvert de collaboration, où tout utilisateur peut téléverser des informations en utilisant des outils de publication pensés pour le non-initié. Parmi ces outils, nous pouvons citer les forums de discussion, les blogs, les wikis, ou plus généralement les applications Internet riches («*Rich Internet Applications*»). Ces outils définissent l'interaction utilisateur comme le but même de la navigation sur le Web, et forment ce qui est désormais appelé le « Web 2.0 ». Ils ont permis l'émergence de plates-formes à vocation sociale, telles que Myspace [1], Facebook [2] ou Flickr [3], où tout un chacun peut créer ou annoter des renseignements sur des ressources, sur lui-même, ou

1. http://www.myspace.com, réseau social à vocation musicale.
2. http://www.facebook.com, réseau social généraliste.
3. http://www.flickr.com, partage de photographies.

sur des tiers. De fait, les fournisseurs de contenu ne sont plus sélectionnés par leur familiarité avec les technologies de l'Internet. Les profils de ces fournisseurs peuvent être désormais extrêmement variés, allant de simples visiteurs curieux à des experts d'un sujet de discussion donné.

Un besoin de compréhension La recherche du contenu à travers le Web, passe par la compréhension de ce contenu. Toujours dans la vision traditionnelle du Web, la première étape de cette compréhension fût de trier les contenus connus afin de répondre à des requêtes. Ce travail, pris en charge par les moteurs de recherche, consiste à indexer les contenus disponibles afin de pouvoir les retrouver et satisfaire au mieux les recherches.

Face à l'évolution du Web 2.0, le besoin de compréhension se pose à nouveau. L'indexation n'est plus adaptée, ni à la fréquence d'édition des contenus, ni même à ce qui est maintenant le cœur du système : les utilisateurs eux-mêmes. Alors qu'il n'était auparavant que le consommateur du contenu, l'utilisateur est devenu une ressource active du Web en tant que telle, et c'est désormais à son propos que se pose la nécessité de compréhension. Une bonne maîtrise des interactions du Web 2.0 et des objectifs qu'il supporte, comme la personnalisation de l'expérience utilisateur ou le ciblage précis d'un public donné, requiert une compréhension fine des utilisateurs ainsi que des groupes d'utilisateurs de taille significative. La compréhension de ces groupes, qui portent le nom de **communautés**, représente un fort enjeu de cette décennie du Web.

Dans le monde professionnel, cet enjeu revêt également un poids important. De nombreuses entreprises et institutions gèrent aujourd'hui des plates-formes communautaires, ou prennent en considération l'impact de leurs actions et produits sur les activités des utilisateurs. À ces fins, on observe une forte croissance d'offres d'emplois pour un nouveau profil de poste, appelé « *Community Manager* » qui regroupe les fonctions de modérateur, de gestionnaire et d'analyste Web. Usuellement [4], on attend de lui qu'il :

- identifie et analyse les enjeux, les modèles et les tendances dans les demandes des clients et la performance du produit ;

4. http://conniebensen.com/2008/07/17/community-manager-job-description/

– participe à un réseau en suivant les blogueurs et autres contributeurs éminents ;
– participe à des événements autour du produit.

L'une des tâches usuelles du gestionnaire de communautés est de veiller sur les forums de discussion de son organisation (activité de modération, organisation des sujets, etc.). Cette surveillance peut également cibler des forums extérieurs, afin d'attirer de nouveaux utilisateurs. Des sociétés émergentes [5] proposent d'ailleurs aujourd'hui des services en matière de gestion et de surveillance multi-réseaux des influences sur les réseaux sociaux. Notre approche s'inscrit donc dans ce contexte des besoins du gestionnaire de communautés : analyser et comprendre les communautés du système dont il a la gestion.

1.2 Problématique

Il convient donc de comprendre le comportement des utilisateurs sur le Web, ainsi que les aspects communautaires qui en découlent. Comment se constituent les groupes d'utilisateurs, autour de qui sont-ils organisés et sur quelles thématiques ? Cette problématique se pose sur l'ensemble du Web 2.0, partout où les utilisateurs peuvent échanger du contenu et exprimer leurs opinons. Elle est d'ailleurs aujourd'hui partagée par de nombreux acteurs du Web, comme Google qui a officiellement lancé en mai 2012 son *Knowledge Graph* [6], qui a pour objectif l'amélioration de la pertinence des résultats du moteur de recherche en y ajoutant des informations sémantiques issues de sources diverses telles que Wikipédia ou le *CIA World Factbook*.

Toutefois, trois obstacles majeurs sont à considérer lors de cette recherche :

– la quantité de données : L'ensemble des données échangées entre les utilisateurs est gigantesque. À titre d'exemple, le nombre d'utilisateurs actifs sur Facebook a atteint le milliard le 14 septembre 2012, pour 140 milliards de liens d'amitié ;
– la croissance de cette masse d'informations : l'ensemble des données utilisa-

5. comme Sprout Social (http://sproutsocial.com) ou eCairn (http://ecairn.com).
6. http://www.google.com/insidesearch/features/search/knowledge.html

teur est en augmentation perpétuelle. On estime par exemple qu'en 2011, 35 heures de vidéo sont publiées sur Youtube chaque minute et 250 millions de tweets sont postés chaque jour [7]. La figure 1.1 montre en outre l'évolution du nombre de comptes utilisateurs actifs sur Facebook et Twitter depuis leurs créations [8] ;

– la majeure partie des informations nécessaires à la compréhension sont implicites ; les besoins des utilisateurs doivent être déduits de leurs activités. Ces besoins sont d'ailleurs parfois cachés à l'utilisateur lui-même. La compréhension d'un utilisateur ne peut donc se limiter à lui seul, et doit inclure son environnement.

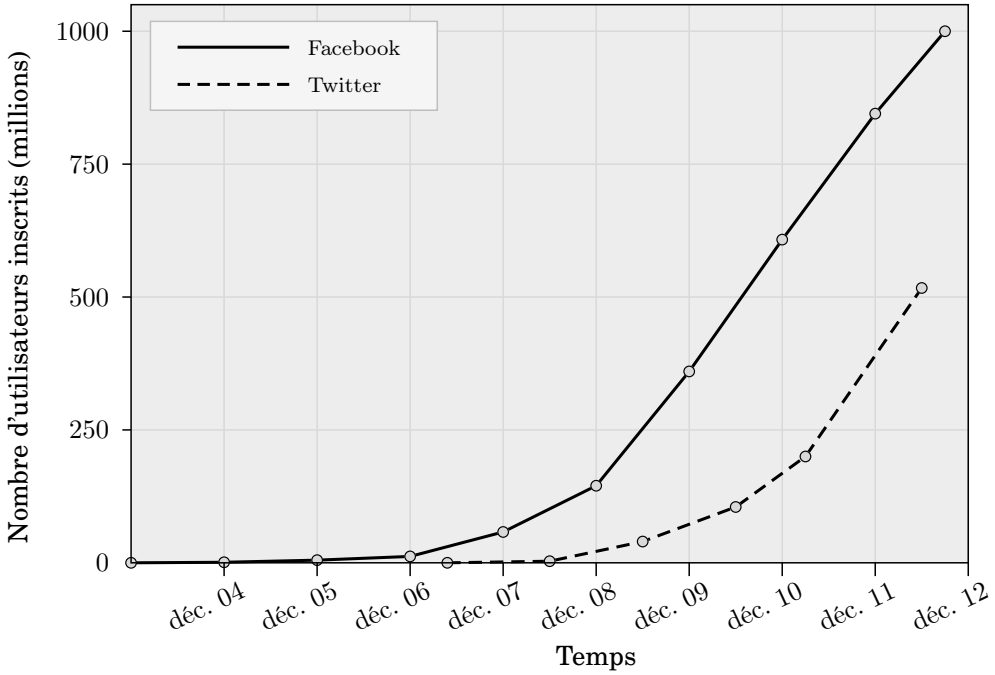


FIGURE 1.1 – Évolution du nombre de comptes utilisateurs actifs sur Facebook et Twitter

La problématique de cette thèse peut alors se résumer par la question suivante : **Comment analyser en temps réel les centres d'intérêt des utilisateurs et la structure de leurs communautés, en prenant en compte la sémantique des échanges ?**

7. http://www.eteamsys.com/content/statistiques-web-de-2011
8. Depuis http://newsroom.fb.com et communiqués de presses de Twitter.

1.3 Méthodologie

Pour répondre à cette problématique, nous proposons une approche sémantique. Cette approche s'inscrit dans une évolution communément observée sur le Web. En effet, la compréhension sémantique apparaît comme le principal moyen d'aller au delà de l'analyse structurelle, basée uniquement sur les liens entre des éléments (pages utilisateurs). La disponibilité grandissante de bases de connaissances sémantiques (telles que Wordnet [51], YAGO [68], DBpedia [10]) nous permettent d'analyser le sens des communications des utilisateurs.

Nous proposons de définir chaque utilisateur en fonction du contenu sémantique des données qu'il produit ou manipule. À partir de ces informations, nous définissons des communautés d'utilisateurs qui partagent les mêmes accointances sémantiques. Et enfin, à partir de ces communautés, nous raisonnons sur l'ensemble du système et sur les relations qui lient les utilisateurs, afin de comprendre au mieux leurs fonctionnement et organisation.

Nous choisissons une approche dite « en ligne » de cette problématique. Un algorithme en ligne est capable de traiter une entrée de manière factionnaire, et continue à traiter en série les entrées reçues ultérieurement. Il s'oppose à l'algorithme hors-ligne, qui nécessite la totalité des entrées pour fonctionner. Cette approche en ligne doit nous permettre d'assimiler le contenu et de pourvoir l'analyser en temps réel. Cette contrainte vise une analyse passant à l'échelle avec le flux grandissant des contributions utilisateurs et permettant de suivre l'évolution permanente et continue des communautés. Nous concentrons notre approche sur les forums de discussion, qui sont un lieu particulier d'expression sur le Web. Nous choisissons d'appréhender les échanges utilisateurs par le biais de ce support en particulier, car il s'agit d'un mode d'expression très largement répandu et relativement ancien — hérité des listes de diffusion publiques et autres systèmes de bulletins électroniques (*bulletin boards*) —, tout en conservant un fort attrait de nos jours.

1.4 Contributions

Nous présentons dans ce qui suit les principales contributions de cette thèse.

Découverte de communautés par la sémantique Par défaut, dans un système de communication quelconque, les utilisateurs sont techniquement indifférenciés. Ils produisent du contenu qu'ils destinent aux autres utilisateurs. Néanmoins, de ces échanges naissent des structures communautaires, qui sont au delà du modèle technique du système. Nous pouvons citer comme exemple le rapprochement d'utilisateurs en fonction de leurs centres d'intérêts. Notre contribution consiste à définir et analyser les échanges entre les utilisateurs, afin de les regrouper en communautés. Pour chaque communauté identifiée, nous identifions les thématiques réunissant les utilisateurs au sein de cette communauté.

Nous proposons trois approches, en variant la quantité et la qualité des informations de classement disponibles :

- nous nous basons tout d'abord sur l'analyse des usages utilisateurs, en raisonnant sur les annotations portées par les ressources que les utilisateurs manipulent ;
- nous avons ensuite recours à des vocabulaires, que nous structurons pour en déduire des distances sémantiques entre les utilisateurs et les termes des vocabulaires ;
- nous utilisons enfin des ontologies légères de référence.

Nous employons chacune de ces sources comme base de connaissances afin de dresser le profil du système étudié et de ses utilisateurs. Nous construisons ainsi des communautés basées sur les similarités sémantiques détectées dans les profils utilisateurs.

Analyse sémantique de communautés Dans sa considération la plus brute, une communauté identifiée d'utilisateurs n'est que la liste des membres qui la composent, potentiellement étiquetée par la cause de ce regroupement. Il est indispensable de mener des investigations sur ce groupe et ses échanges, afin de comprendre sa dynamique. L'analyse de communautés a pour objet de définir et

caractériser les éventuelles structures internes des communautés, leurs évolutions, etc.

Pour expliciter ces caractéristiques, nous proposons deux approches :

- nous définissons et proposons deux nouvelles métriques pour déterminer la centralité d'un utilisateur dans une communauté, à savoir la place qu'il occupe dans l'ensemble des échanges entre les différents utilisateurs. Nous identifions ainsi les utilisateurs les plus importants ;
- nous proposons ensuite une caractérisation des rôles des utilisateurs, basée sur l'influence qu'ils exercent sur la dynamique des communautés (attracteur, répulseur, etc.).

L'outil WebTribe Pour valider nos propositions et réaliser des expérimentations, nous avons développé un prototype de découverte et d'analyse des communautés, appelé WebTribe. Ce système se compose d'un ensemble de modules permettant l'analyse de diverses sources de données (mails, forums, tweets, etc.) et mettant en œuvre l'ensemble des analyses présentées tout au long de cette thèse. WebTribe est conçu pour effectuer des analyses « en ligne », c'est-à-dire en temps réel par rapport aux réseaux de communication qu'il supervise. La figure 1.2 illustre l'architecture générale. L'outil Webtribe :

- normalise les données publiées par les utilisateurs à partir de sources externes diverses ;
- analyse la sémantique des échanges entre les utilisateurs selon des paramètres définis ;
- modélise les utilisateurs et leurs sujets de discussion au sein de communautés thématiques ;
- rend l'ensemble des résultats accessibles et manipulables par le gestionnaire de communautés.

Cette thèse vise à proposer des méthodes pour la découverte et la gestion des communautés d'utilisateurs, méthodes à même de permettre à terme la réalisation d'un outil exploitable pour le monde industriel. Cette approche s'inscrit dans la dynamique portée par le dispositif « Jeunes Chercheurs Entrepreneurs » de la région Bourgogne, qui a financé cette thèse.

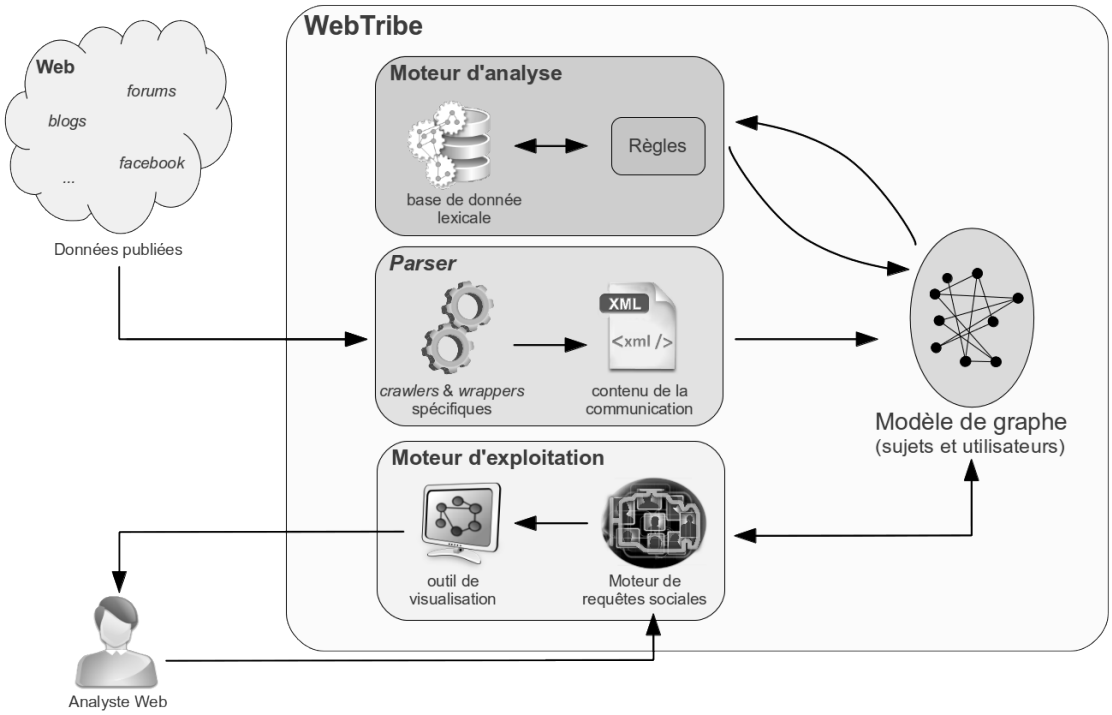


FIGURE 1.2 – Architecture globale du système WebTribe

1.5 Plan de la thèse

Cette thèse est organisée comme suit : le chapitre 2 présente tout d'abord l'état de l'art de notre travail. Cette étude porte sur la notion de communautés et leur découverte, les réseaux sociaux et leur analyse, ainsi que les aspects sémantiques qui y sont attachés. Nous terminons ce chapitre par un bilan de cet état de l'art.

Nous présentons ensuite dans les chapitres suivants l'ensemble de nos propositions en matière de découverte et d'analyse des communautés implicites par une approche sémantique en ligne. Le chapitre 3 donne une vue d'ensemble de notre approche et de nos contributions. Dans le chapitre 4, notre proposition en matière de découverte de communautés est détaillée. Nous y présentons trois approches sémantiques distinctes. La première se rapporte à l'utilisation statistique des systèmes d'annotation, la seconde repose sur une utilisation de vocabulaires et la troisième utilise des ontologies. Enfin, nous présentons nos méthodes d'analyse sémantique de communautés dans le chapitre 5. Nous y traitons en particulier

la notion de centralité sémantique temporelle, ainsi que la question du rôle des utilisateurs au sein des communautés.

Le chapitre 6 est consacré à l'implémentation du système WebTribe et aux diverses optimisations effectuées. Le chapitre 7 conclut sur les perspectives de la thèse.

Chapitre 2

État de l'art

Sommaire

Cet état de l'art présente la notion de communauté et dresse le bilan de son évolution. Sont ensuite explicitées les caractéristiques propres aux réseaux de communication et aux réseaux sociaux. Les principales orientations d'analyse des réseaux sont ensuite dépeintes, en partant de méthodes structurelles pour finir par des méthodes sémantiques. Nous présentons ensuite les outils existants permettant l'extraction de communautés en ligne. La dernière section conclut cet état de l'art.

2.1 Évolution des communautés

Une « communauté » est un terme ambigu. Poplin [57] relève plus de 120 définitions pour ce mot. Selon l'article Wikipedia sur les communautés [1], « Une communauté est une interaction d'organismes partageant un environnement commun. Dans les communautés humaines, l'intention, la croyance, les ressources, les besoins ou les risques sont des conditions communes affectant l'identité des participants et le degré de leur cohésion. ». Dans notre domaine d'étude, nous considérons la communauté comme un groupe virtuel d'entités en interaction sur le Web. Ces entités partagent ou possèdent un élément en commun. Cette notion de communautés d'intérêt est introduite dès la fin des années 1960 par Licklider et Taylor [47], contrastant la définition traditionnelle d'emplacement commun. La nature de ces entités partageantes peut être variée, comme une page Web, un utilisateur, une ressource, etc. L'élément partagé peut être une ressource, un comportement, un intérêt, etc.

Avec l'évolution du Web, la notion de communauté Web a également évolué. Nous dépeignons ici cette évolution par la mise en avant des étapes majeures que sont les communautés hypertextes, les communautés de tags et les communautés sociales.

2.1.1 Communautés hypertextes

La communauté hypertexte est la plus ancienne notion de communauté Web qui ait été formulée, bien avant l'avènement de contenu dynamique et des notions

1. http://en.wikipedia.org/wiki/Community

de Web social et de Web collaboratif. Ces communautés sont donc des communautés de pages Web. En raison de son caractère ouvert, la toile mondiale Web a permis à un grand nombre d'entités — personnes physiques ou organisations — de produire du contenu selon un format convenu, le HTML. Du contenu a ainsi été produit, indépendamment de son classement ou de son accessibilité. Afin de retrouver ces contenus pour répondre à des besoins utilisateurs, la première ébauche de traitement communautaire des pages Web voit le jour par le biais de leur enregistrement : inscrire explicitement un site Web auprès d'une autorité, comme DMOZ [2], Lycos [3], ou Yahoo ! directory [4]. Mais ce genre de référencement simple peut paraître aussi arbitraire qu'imperméable à la pertinence.

La notion de communauté hypertexte repose sur la notion de lien. Le Web est vu comme un immense graphe. Les nœuds sont les pages Web, et les arcs les liens hypertextes entre les pages. Le graphe est donc orienté, chaque lien pointe d'une page vers une autre. Définir une communauté hypertexte revient donc à définir un sous-ensemble de ce graphe global du Web [6]. La communauté hypertexte est donc une communauté strictement structurelle, mais représente un pas en avant par rapport au simple référencement. Introduit en 1998 par Page et al. [55, 11], l'algorithme PageRank classe les pages indexées selon leur popularité. Cette popularité est fonction du nombre de pages pointant sur la page considérée, pondérée par la popularité de ces mêmes pages, calculée récursivement. Compte-tenu de la nature du Web, principalement sa taille et sa perpétuelle évolution, un calcul exact du PageRank est illusoire. Des approximations sont donc effectuées [38]. La figure 2.1 illustre le principe de pondération récursive des liens de PageRank, où chaque lien ajoute du poids à la page cible, en fonction du poids de la page source.

Dans une approche connexe à celle de PageRank, l'algorithme HITS de Kleinberg et al. [42, 30] définit les notions de *hubs* et d'*autorities*. Il se base sur l'idée que l'ensemble du Web est naturellement structuré en ensemble bipartite. D'une part, les *autorities* sont des pages qui font autorité dans leur domaine, et sont liées par de nombreux *hubs*. Ces derniers sont définis de façon inverse, comme des pages liant des *autorities*. Ces pages représentent des catalogues de références, sous

2. http://www.dmoz.org
3. http://www.lycos.com
4. http://dir.yahoo.com

forme de liens hypertextes. Une communauté autour d'un domaine donné est alors constituée de deux sous-ensembles fortement liés que sont les *hubs* et *autorities* du domaine. Cette approche sert de base à de nombreuses autres ; Dourisboure et al. [19] y ajoutent un calcul de similarité des contenus entre chaque lien pour raffiner le graphe et ne conserver que les liens pertinents. La figure 2.2 illustre le principe structurel de HITS, où les *hubs* pointent sur les *autorities*.

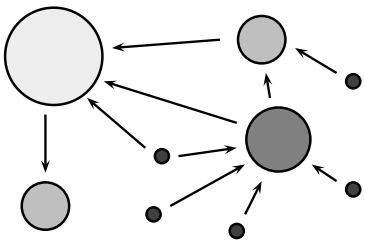


FIGURE 2.1 – Principe de PageRank

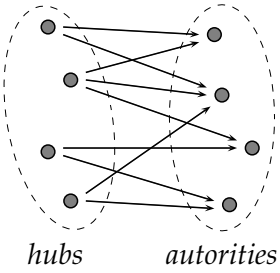


FIGURE 2.2 – Principe de HITS

Toutefois, toutes les communautés identifiables ne sont pas uniquement structurelles. Basées sur le contenu et ses annotations, nous décrivons ci-après les communautés de tags.

2.1.2 Communautés de tags

D'une façon générale, le *tag* est un mot-clé attribué à un élément d'information. Il agit comme un marqueur lexical ou sémantique. Ce type de métadonnées permet de décrire la cible et est utilisé à des fins de navigation et de recherche. Le plus souvent, les *tags* sont librement choisis par l'acteur qui réalise l'annotation, qu'il soit l'auteur du contenu annoté ou non. Ils ne font pas nécessairement partie des *tags* précédemment présents et utilisés sur le système concerné. Dans le cadre de l'analyse des communautés de *tags*, nous faisons référence aux *tags* annotant les pages Web, et par analogie le contenu des pages concernées. Quelques sites Web comme delicious [5] structurent le Web par le biais de *tags*.

Les communautés de *tags* représentent une étape intermédiaire dans l'évolution de la notion de communauté Web. Sans encore représenter un réseau social, ces

5. Site Web social de partage de marque-pages classés par *tags*. http://del.icio.us

communautés se démarquent des approches strictement structurelles, et intègrent l'activité d'annotation opérée sur les ressources. Ces annotations sont associées à des notions de taxonomie ou de folksonomie[6]. La taxonomie est un ensemble de termes avec un lien, généralement hiérarchique. La folksonomie en est une variante collaborative où l'annotation est libre : les utilisateurs peuvent alors ajouter sans contrainte des marqueurs de classification. Ce type d'annotation souffre donc souvent d'un manque significatif de structure. Il n'y a pas ou peu de relations entre les différents éléments, ce qui peut être un obstacle à l'interprétation directe des résultats de recherche. Une compréhension plus sémantique des éléments apparaît donc comme nécessaire. C'est pourquoi la plupart des approches de découverte de communautés de *tags* améliorent des techniques de calcul de similarité entre les ressources.

Par exemple, Simpson [67] utilise la similarité cosinus (*cosine similarity*) des termes des documents qu'annotent les *tags* pour déterminer leur proximité et les regrouper. La similarité cosinus est une métrique qui est fréquemment utilisée en fouille de textes, se basant sur la différence d'angle de deux vecteurs à n dimensions, où les valeurs des vecteurs sont les occurrences ou poids des mots du corpus dans le document. Astrain et al. [4] l'utilisent également pour raffiner les résultats de partitionnement de *tags* d'un automate à logique floue. Toujours à base de techniques héritées de la fouille de texte, Cattuto et al. [12] regroupent les *tags* distincts et indépendants en sous-ensembles proches, en fonction de la pertinence des *tags*. La pertinence de ces *tags* est calculée à partir de TF-IDF, une mesure statistique qui évalue la pertinence d'un terme en fonction de sa présence dans un document et sa rareté dans le reste du corpus [64].

Lorsque les utilisateurs annotent librement les ressources, l'utilisation des *tags* est bien souvent inégale. Les ressources populaires sont très souvent marquées, les autres beaucoup moins. Ce phénomène s'auto-entretient car les ressources les plus marquées ont plus de chance d'être les résultats d'une recherche, de se voir à nouveau marquées, et ainsi de suite. Kim et al. [41] proposent de lutter contre cet effet de longue traîne par le marquage automatique de *tags*. Cette automatisation est basée sur la similarité des ressources, de leurs auteurs et sur les fréquences de

6. Adaptation française de *folksonomy*, mot-valise combinant les mots *folk* (le peuple) et *taxonomy* (la taxinomie).

co-occurences de *tags*. Cet apprentissage « social » des *tags* est également proposé par Giannakidou et al. [29] où ces co-occurences permettent de lier la sémantique d'un *tag* à la nature sociale des ressources qu'il marque.

2.1.3 Communautés sociales

L'avènement du Web collaboratif entraîne l'explosion du volume de communications entre les utilisateurs. Si ces interactions sont bien antérieures à ce phénomène, comme l'usage des listes de diffusion thématique par e-mail, de nouvelles plates-formes dédiées à l'interaction sociale émergent (comme Facebook [7]). Au sein de ces plates-formes, les relations entre les utilisateurs deviennent une fin en soi, un enjeu.

Plusieurs études définissent les communautés d'utilisateurs en fonction de leurs objectifs. Haas [35] définit la notion de communauté épistémique comme une communauté ayant pour but la création de connaissances — comme par exemple la résolution de problèmes — en acceptant une structure commune. Cette notion s'oppose à celle de communauté de pratique présenté par Lave et Wenger [43], communauté auto-organisée orientée vers ses membres et vue comme un moyen de faire valoir les compétences individuelles de ces membres, par le partage de ressources [72].

De nos jours, la quasi-totalité des sites Web les plus fréquentés au monde intègre des outils dédiés à ces interactions sociales (Voir Annexe A). Dans ce contexte, la gestion communautaire change de nature. Au lieu de pages Web identifiées par des URL [8], ces communautés lient des utilisateurs, identifiés par des UID [9]. Les liens entre ces entités ne sont plus des liens hypertextes, directionnels et unifiés, mais des relations entre personnes, qui peuvent être directionnelles ou non, et de natures variables. Les méthodes de détection de ces communautés sont présentées plus loin dans les sections 2.3 et 2.4.

7. http://www.facebook.com
8. *Uniform Resource Locator*, littéralement « localisateur uniforme de ressource ».
9. *User IDentifier*, littéralement « identifiant utilisateur ».

2.2 Réseaux de communication et réseaux sociaux

Afin de s'intéresser aux communautés issues des réseaux sociaux, et de l'activité sociale sur le Web des utilisateurs en général, il convient d'en expliciter le modèle. Un réseau social peut être modélisé sous la forme d'un graphe. Les nœuds sont les éléments du réseau, tels que les utilisateurs ou ressources, et les arêtes — ou arcs si le réseau est orienté — sont les connexions entre ces éléments. Si le Web hypertexte peut être modélisé comme un graphe simple, les graphes issus de réseaux sociaux sont quant à eux de natures variables, avec des arêtes aux caractéristiques diverses, qui sont fonction de celles que fournit le réseau social concerné. Ils peuvent donc être orientés ou non, bipartites ou n-partites, aux arêtes annotées ou non, pondérés ou non, etc. Ces relations peuvent être par exemple des relations d'amitié entre les utilisateurs, des relations d'utilisation ou de notation entre les ressources et les utilisateurs, etc.

En 1999, Albert et al. [2] définissent le graphe du Web comme un graphe du petit monde. Ce concept, issu des études de Travers et Milgram [70], a été popularisé par l'expression « six degrés de séparation » (paradoxe de Milgram). Celle-ci suggère que deux personnes quelconques sont reliées en moyenne par une chaîne de six relations. Cette caractéristique est partagée par de nombreux réseaux sociaux. A titre d'exemple, il a été mesuré un degré moyen de séparation inférieur à 5 entre tous les utilisateurs de Facebook en 2011 — soit 721 millions de personnes, — avec un nombre moyen de 4,74 relations entre deux utilisateurs [71]. Diverses études extraient les caractéristiques propres aux graphes issus de réseaux sociaux [3, 54, 63] :

Graphes creux Les graphes issus des réseaux sociaux sont très loin d'être des graphes denses. Cette analyse transcrit une idée reçue qui veut que, dans tout groupe humain suffisamment important (une université, un pays, le monde), chacun ne connaît qu'une minorité de l'ensemble du groupe.

Faible distance typique Le plus court chemin possible entre deux nœuds dans ces réseaux est très petit par rapport à la taille du graphe. Il augmente de manière

logarithmique avec l'augmentation du nombre de nœuds du graphe. Cela concorde avec le « paradoxe de Milgram » exposé ci-dessus.

Haute transitivité Si un nœud est relié à deux autres nœuds, alors ces deux nœuds ont une forte probabilité d'être reliés entre eux. Cela illustre également le principe de « petit monde ».

Distribution des degrés La distribution des degrés suit globalement soit une loi de puissance (un minimum de nœuds agrège un maximum des liens) soit une loi gaussienne (le graphe présente un faible écart-type du nombre moyen de liens par nœud).

Ces caractéristiques sont similaires à celles du graphe du Web, en raison des caractéristiques sociales équivalentes aux caractéristiques thématiques de la toile mondiale, mais aussi à de nombreux réseaux du vivant, tels que les réseaux neuronaux ou génétiques, le réseau des mots du langage naturel, etc. Ils s'opposent totalement aux réseaux structurels, comme les infrastructures physiques de communication (comme le maillage téléphonique) ou de transport, qui présentent des variations de densité et une connectivité bien plus régulière, ou même aux réseaux aléatoires dans lesquels ne se retrouvent pas les notions de petits-mondes et de transitivité [7, 58].

2.3 Analyse des réseaux sociaux

Dans le contexte actuel du Web 2.0, les réseaux sociaux tendent à devenir omniprésents sur le Web. Outre les sites qui fournissent l'accès à un réseau social (comme Facebook ou Linkedln [10]), une part grandissante des autres sites partagent des fonctionnalités qui exploitent ces mêmes réseaux. En 2012, 18 des 20 sites les plus fréquentés au monde fournissent ces services (voir Annexe A). L'activité globale des utilisateurs du Web est aujourd'hui majoritairement tournée vers les réseaux sociaux. Dans ce contexte, l'analyse des réseaux sociaux (*Social Network Analysis*), domaine d'étude sociologique qui existait bien avant l'Informatique,

10. http://www.linkedln.com, réseau social professionnel.

trouve des applications grandissantes dans notre domaine d'étude [56]. L'analyse des réseaux sociaux a deux objets principaux et complémentaires, que sont la modélisation des utilisateurs et la détection des communautés qu'ils contiennent.

2.3.1 Modélisation des utilisateurs

La modélisation des activités utilisateurs se compose de plusieurs objectifs plus ou moins liés. Cela peut se traduire par la modélisation des interactions de l'utilisateur avec l'ensemble des ressources du système, la compréhension des avis qu'exprime l'utilisateur, ou la prédiction de l'activité future de l'utilisateur (ce dernier objectif étant souvent lié à l'analyse des précédents). Nous explicitons ces objectifs en trois axes que sont la prise en compte de l'utilisation des ressources par les utilisateurs, la compréhension de leurs avis et sentiments, et la prédiction de leurs activités futures.

Utilisation de ressources Dans son expérience du réseau social, l'utilisateur est amené à manipuler des ressources de natures variées. Communément, il s'agit de textes, de documents multimédias, des *tags*, de références à d'autres utilisateurs, etc. L'utilisateur manipule ces ressources, celles qu'il crée ainsi que celles créées par autrui, par un ensemble de transactions (lecture, modification, annotation, etc.). À partir de ces échanges, la totalité des éléments manipulés par le réseau peut alors être localisée dans un graphe unique, afin d'en exploiter les relations. Bertier et al. [8] utilisent par exemple l'activité d'annotation pour évaluer la probabilité de passage d'un tag à un autre. De nombreuses approches analysant des réseaux de communication sociale, tels que les *blogs*, interprètent ces derniers comme des relais sélectifs de communication. Par exemple, Gumbrecht [34] considère les commentaires laissés sur les *blogs* comme des retours des utilisateurs sur les sujets qui y sont exposés, permettant d'en déduire leurs degrés d'intérêt. Au travers de diverses méthodes de *data mining*, Richardson et Domingos [60] modélisent les influences des utilisateurs afin de mettre en lumière le marketing viral qui s'exerce entre les utilisateurs et comment ils s'influencent entre eux.

Avis et sentiments En lien direct avec l'interprétation de l'utilisation des ressources, l'analyse des contributions des utilisateurs — comme les systèmes de billets sur des *blogs* ou des commentaires relatifs aux contenus existants — est souvent utilisé pour déterminer les avis et sentiments des utilisateurs vis-vis des sujets ou contenus concernés. Mishne et al. [52] mesurent la popularité des sujets abordés et détectent les sujets polémiques comme ceux qui génèrent des controverses et des avis contraires entre les utilisateurs. À partir d'un terme donné, Das et Chen [15] extraient des publications des utilisateurs si leurs contenus expriment un sentiment d'un texte vis-à-vis de ce terme, c'est-à-dire de savoir s'il est positif, négatif, ou non-pertinent. D'une façon dérivée, ces analyses permettent de mettre en lumière les motivations des utilisateurs : Trevino [50] en détermine en quoi cette activité influe sur les autres activités en utilisant les mécanismes classiques de *feedback* dans la communication. En outre, Mitrović et al. [53] mettent en évidence que le caractère émotionnel influe sur l'activité. Pour tout intérêt donné, le caractère négatif a préséance sur le positif : un utilisateur se manifeste davantage lorsqu'il est mécontent que lorsqu'il est satisfait. De même il en ressort que l'expression des sentiments permet de passer au delà de l'organisation interne (un peu à la manière d'outrepasser le devoir de réserve sous le coup de la colère dans la vie réelle).

Prédictions En extrapolant les précédentes approches, la compréhension des utilisateurs permet de prévoir leurs activités futures. La proposition la plus fréquente, comme présentée par De Choudhury et al. [16], consiste à réaliser une projection de l'activité actuelle pour prédire l'activité future. Roth et al. [62] utilisent par exemple une interprétation de l'activité utilisateur pour lui proposer de nouveaux amis, ce qui permet d'aller au delà de l'interprétation du graphe social uniquement. Bilenko et Richardson [9] externalisent ce type d'approche, agrégeant différentes sources externes d'expérience utilisateur pour fournir de la recommandation personnalisée, en publicité ou autre.

L'ensemble de ces méthodes, qui permettent de comprendre et de modéliser l'activité des utilisateurs — et à travers cette dernière, les utilisateurs eux-mêmes — est le premier pas vers la détection de communautés.

2.3.2 Détection de communautés

Si l'ensemble des éléments du Web — des pages Web aux utilisateurs — peut être représenté comme autant de sommets dans un graphe, constituer des communautés de ces éléments revient à définir des sous-graphes dans ce graphe global. En d'autres termes, cela revient à mettre en évidence les liens du graphe qu'il convient d'éliminer pour isoler des communautés. Si l'on considère notre graphe comme un graphe de flots, alors notre découpage de communautés s'apparente à un problème classique de coupe minimale. C'est-à-dire à la recherche de la capacité minimale à retirer du graphe pour rendre nul son flot. Les parties ainsi isolées formant des communautés. Cette approche ne peut toutefois être exploitée directement, car la recherche de partitions équilibrées par coupe minimale est un problème NP-complet, comme évoqué par Cormen et al. [14]. En revanche, de très nombreuses méthodes se basent sur le problème de recherche du flot maximum du graphe, en utilisant le théorème de Ford-Fulkson [24] qui veut que cette coupe minimale soit nécessairement égale au flot maximum du graphe. L'approche considérée comme la plus rapide en pratique est proposée par Goldberg et Tarjan [33], mais présente le besoin de connaître tout le graphe pour réaliser le calcul. Ce qui n'est bien évidement pas adapté au Web.

Conjointement basée sur l'approche HITS de Kleinberg (un Web de *hubs* et d'*autorities*) et sur cette problématique de flot précédemment évoquée, Flake et al. [23] proposent une nouvelle définition formelle d'une communauté, comme un ensemble de pages qui sont plus liées entre elles à l'intérieur de la communauté qu'à l'extérieur. A partir des *hubs* et *autorities* de HITS, vus comme autant de sources (*seeds*) et puits (*sinks*) d'un calcul de flot maximum, les auteurs découvrent le graphe par son parcours, ne retenant que les nœuds qui ne mettent pas en péril le flot. Cet algorithme, appelé *Incremental Shortest Augmentation* (ISA), basé sur le *shortest augmentation path algorithm* d'Edmonds et Karp [20], permet une détection en ligne des communautés, et ce sans connaissance de tout le graphe. La figure 2.3 illustre la coupe ainsi effectuée.

Cette approche est largement reprise et améliorée par la suite, notamment en matière de temps d'exécution. Imafuji et Kitsuregawa [36] proposent par exemple de rendre le point de départ mobile pour ne plus être dépendant du choix initial,

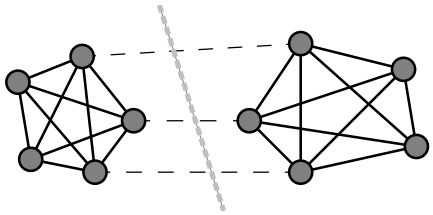


FIGURE 2.3 – Utilisation de max-flow/min-cut pour séparer les communautés

d'éliminer les nœuds « inondant » le graphe, et de contrôler les itérations requises. Ino et al. [37] précisent la définition de la communauté, en distinguant les communautés fortes (strictement plus de liens entrant que sortant) des communautés faibles (au moins autant de liens) et raffinant ainsi l'algorithme.

Plusieurs approches exploitent également le concept de communautés à plusieurs niveaux, c'est-à-dire de communautés dans les communautés. Lozano et al. [49] proposent une méthode récursive de séparation d'un graphe en deux sous-ensembles, eux-mêmes analysés et séparés, jusqu'à atteindre la stabilité. Girvan et Newman [31] décrivent les communautés comme des grappes denses, qui sont reliées entre elles par des connexions plus souples ayant le rôle de « ponts », le tout formant des méta-communautés. Mais des complexités de l'ordre de $O(n^3)$ rendent difficiles leurs applications en ligne.

2.4 Distances sémantiques

Lors du traitement sémantique des contenus, une problématique cruciale est de déterminer la proximité ou non des éléments manipulés, et des concepts qu'ils représentent. De nombreuses contributions permettent de définir la distance entre deux entités, en ayant ou non recours à une base de connaissances tierce. Ces distances ainsi calculées permettent de définir des similarités, ou d'autres types de mesures.

2.4.1 Similarité sémantique

Une première famille d'approches regroupe l'ensemble des méthodes statistiques sur le corpus analysé. Ces méthodes répondent à une problématique forte qui est de pouvoir calculer une distance sémantique sans avoir de base de connaissances ou de comparaison pour effectuer la mesure. Zanardi et Capra [74] calculent par exemple la similarité de *tags* utilisateurs. Cette distance est calculée à partir de la distribution commune des *tags* dans le corpus : les co-occurrences de *tags* pertinents définissent leur proximité. Cattuto et al. [12], précédemment évoqués, mesurent statistiquement d'une façon similaire la distance entre les *tags* à partir de leurs utilisations.

Lorsqu'il est possible d'utiliser une base de connaissances pour calculer les distances sémantiques, la solution la plus native est d'utiliser une ontologie. De très nombreuses approches permettent de définir une distance entre deux concepts au sein d'une ontologie. Parmi elles, nous pouvons citer par exemple Wu et Palmer [73] qui définissent la distance sémantique entre deux concepts comme fonction de leur longueur de chemin dans l'arborescence de l'ontologie, pondérée par leur éloignement de la racine. Ce dernier point valorise la proximité des concepts spécialisés sur les concepts généraux. Cette méthode est couramment utilisée pour réduire le nombre de concepts d'un ensemble représentatif. Desmontils et Jacquin [17] l'utilisent par exemple pour valoriser les concepts dominants d'une page, ou Eirinaki [21] pour relier des mots-clés à des concepts existants dans une ontologie. Une approche hybride proposée par Resnik [59] combine l'utilisation de la structure hiérarchique de thésaurus et les fréquences d'utilisation des termes dans les corpus analysés. Une autre contribution majeure est celle de Lin [48] qui définit cette distance comme la probabilité de chevauchement des concepts qui héritent des concepts en comparaison. Jiang et Conrath [39] analysent la structure taxonomique et comparent les effets de la prise en compte de la distance entre les nœuds (en nombre d'arcs) ou du contenu des nœuds. Enfin, il peut être intéressant de citer l'approche de Cilibrasi et Vitanyi [13] qui, sans base de connaissances dédiée, utilisent le moteur de recherche de Google comme base de connaissances universelle. La similarité entre deux concepts est alors calculée à partir du nombre de résultats des requêtes de co-occurence sur ce moteur de recherche.

2.4.2 Mesures

Toujours dans une optique d'un traitement sémantique des données, il est possible de définir un certain nombre de mesures, qui aideront à la compréhension de ces mêmes données.

Girvan et al. [31] se basent sur les travaux de Freeman [25] qui définissent le concept de « *vertex betweenness* », une mesure de centralité et d'influence d'un nœud dans un réseau. Ils généralisent cette approche à travers la notion d' « *edge betweenness* », définie comme le nombre de plus courts chemins qui passent à travers ces nœuds. Cette mesure est utilisée au sein d'un algorithme permettant de réveler les « ponts », les nœuds de passage, entre les sous-graphes détectés au sein d'un graphe. Mais cette mesure — comme beaucoup d'autres — nécessite une connaissance complète du graphe et souffre d'une complexité de $O(n^3)$. Fuehres et al. [27] utilisent la « *betweenness centrality* » des utilisateurs d'un graphe social pour infléchir la similarité des contenus manipulés par ces derniers. Tang et al. [69] définissent des mesures temporelles sémantiques comme « *Temporal Betweenness Centrality* » et « *Temporal Closeness Centrality* » pour situer un utilisateur dans un réseau de communication social. Nous pouvons également citer les travaux de Jung [40], qui utilise la notion de centralité sémantique appliquée à la réécriture de requêtes, basées sur diverses ontologies.

2.4.3 Autres utilisations de la sémantique

Le Web sémantique représente un large domaine d'études, bien plus vaste que les aspects sémantiques que cette thèse recoupe. Nous présentons donc quelques méthodes issues du Web sémantique dont les applications sont connexes à notre sujet.

Les relations sémantiques qui existent entre les différents éléments des ontologies permettent un apport d'informations supplémentaires sur les interactions de ces éléments. Il est alors possible d'appliquer des raisonnements similaires aux ressources étudiées. Gauch et al. [28] réalisent ainsi l'historique de navigation de l'utilisateur comme autant de concepts détectés dans une ontologie. Sieg et al. [66] utilisent une approche identique pour personnaliser les futures requêtes des utilisateurs.

2.5 Outils d'analyse

Les outils permettant l'extraction de communautés implicites ne sont pas disponibles en tant que tel sur le Web. Toutefois, Dourisboure et al. [19] ont publié une version disponible sur le Web [11] de leur expérimentation, qui permet la visualisation de communautés. Cet outil, baptisé « *Community Watch* » permet de consulter l'implémentation de leur approche de nettoyage de graphe, afin d'obtenir des sous-graphes denses comme communautés-résultat. Cette démonstration fonctionne à partir de cinq jeux de données prédéfinis par les auteurs. Bien qu'il ne soit pas possible d'en changer, il est tout de même possible de faire varier les variables et seuils de l'algorithme, pour comparer différents résultats. Bien que limité dans ses paramètres d'entrée, il s'agit à notre connaissance du seul outil de visualisation d'extraction de communautés actuellement disponible publiquement.

Un outil connexe, Condor [12] mérite d'être mentionné. Issus des travaux de Gloor et Zhao [32], cet outil permet la visualisation des structures et flots de communication dans un réseau. Il est composé de différents modules qui réalisent des extractions et calculs sur différentes sources de données. Fuehres et al.[27] étendent notamment le module sémantique de Condor pour réaliser des communautés par partitionnement selon les éléments sémantiques détectés. Mais cette extension n'est pas disponible sur le Web.

D'une façon connexe, dans le monde de l'entreprise, l'analyse de réseaux sociaux est relativement répandue, et plusieurs outils de visualisation de réseaux sociaux coexistent. Mais l'ensemble de ces solutions s'intéressent aux liens sociaux entre les utilisateurs et ne saurait entrer en l'état dans le domaine de la recherche de communautés implicites.

2.6 Bilan

Dans ce chapitre, nous avons présenté un état de l'art de la notion de communauté Web et des techniques qui y sont liées. De cette étude, nous distinguons

11. http://comwatch.iit.cnr.it/
12. http://www.ickn.org/download.html

deux évolutions :

- du structurel vers le relationnel : partant des méthodes hypertextes traditionnelles, où la structure des pages Web jouait un rôle central dans les approches initiales, la problématique principale des communautés Web est aujourd'hui centrée sur l'utilisateur et ses activités. La compréhension des données est désormais un moyen de compréhension de l'utilisateur.
- du statistique au sémantique : les aspects sémantiques, tels que la compréhension fine des contenus ou encore le calcul de similarités sémantiques entre les ressources, prennent une place grandissante dans les approches contemporaines. Dans le même temps, nous observons la naissance d'une problématique existant entre le besoin de positionnement sémantique de plus en plus précis et le manque de bases de références sémantiques pouvant fournir une telle précision.

Cette étude confirme notre vision d'un Web en mutation, où les outils doivent s'adapter à des enjeux en évolution, faisant désormais de l'utilisateur à la fois l'origine et le but des échanges sur le Web. Toutefois, il persiste un fossé dans les approches contemporaines, axées soit sur les aspects utilisateurs de la notion de communauté, soit sur les considérations sémantiques des contenus. C'est pourquoi nous souhaitons apporter une réponse commune à même de prendre en compte l'ensemble des aspects des échanges entre utilisateurs, tant sémantiques que structurels, afin d'en déduire des communautés. Dans le chapitre suivant, nous présentons notre proposition de système de compréhension et gestion des communautés d'utilisateurs : le système WebTribe.

Chapitre 3

Le système WebTribe : vue d'ensemble

Sommaire

Dans ce chapitre, nous présentons une vue d'ensemble de notre approche pour répondre à la problématique de la compréhension des échanges utilisateurs et de leurs aspects communautaires. Nous définissons tout d'abord les objectifs de notre approche. Nous introduisons ensuite nos propositions en matière de détection de communautés et d'analyse de communautés qui seront détaillées respectivement dans les chapitres 4 et 5. Nous présentons ensuite l'architecture de l'outil Webtribe, notre outil de découverte et d'analyse de communautés, dont les détails d'implémentation sont exposés au chapitre 6.

3.1 Objectifs

Pour répondre à notre problématique, nous nous proposons de fournir un outil d'assistance et d'exploitation à disposition du gestionnaire de communautés d'un système informatique sur le Web à interactions sociales donné. Ce système peut être un forum de discussion, une plate-forme de diffusion de messages électroniques, un réseau social, etc. Cette gestion du système analysé passe par le suivi des activités des utilisateurs, la compréhension de ces activités ainsi que des interactions entre utilisateurs, et la réponse aux besoins en information du gestionnaire de communautés sur le système dont il a la gestion. Ces différents objectifs sont détaillés ci-dessous.

Extraction des activités utilisateurs Le premier objectif de notre approche est d'analyser l'ensemble des échanges de données entre les utilisateurs, afin d'en extraire le contenu essentiel. Cette extraction a pour but de réduire l'information par rapport au flux brut des échanges entre utilisateurs à un résumé compréhensible, ainsi que de fournir une vision utile au gestionnaire de l'activité en temps réel sur son système. Cette extraction doit être menée de façon continue, au fil des contributions des utilisateurs.

Définition des utilisateurs À partir de l'analyse de leurs publications, nous dressons un profil des utilisateurs du système. Ce profilage vise à fournir au gestionnaire une compréhension fine et dynamique de ses utilisateurs, actualisée au

fil des contributions de ces derniers.

Résumé du système Nous raisonnons également sur l'ensemble des profils utilisateurs que nous maintenons, afin de définir les centres d'intérêt principaux des utilisateurs du système et donc les sujets de discussion majoritaires sur l'ensemble du système.

Compréhension des interactions À partir de la définition des utilisateurs ainsi que des centres d'intérêt majoritaires du système, nous identifions les principaux regroupements autour de ces sujets. Ces ensembles, que nous définissons comme étant les communautés thématiques du système, sont explicités au gestionnaire. La prise en compte de leur existence ainsi que le suivi de leur évolution permet de comprendre les interactions entre les utilisateurs et autour de quoi ou de qui elles se structurent.

Réponses aux besoins du gestionnaire L'ensemble des informations collectées ou calculées évoquées plus haut doit être accessibles en temps réel au gestionnaire de communautés. Nous devons donc maintenir et fournir au gestionnaire une vision d'ensemble de son système, tout en permettant l'interrogation sur des points précis (détails sur une communauté, un utilisateur, etc.) qui permettent d'accéder à l'ensemble des informations dont nous disposons. Cet accès doit permettre de répondre à la majorité des besoins du gestionnaire de communautés en matière de gestion de ses utilisateurs, de leurs interactions, intérêts, et groupements communautaires.

Pour remplir ces objectifs, nous développons notre approche en deux étapes que sont la détection de communautés et l'analyse de communautés.

3.2 Découverte de communautés

Au sein du Web 2.0, l'utilisateur a à sa disposition un grand nombre d'outils lui permettant de s'exprimer et d'échanger avec d'autres utilisateurs, comme les

forums, les blogs, les wikis ou encore les réseaux sociaux. Ce nouveau Web collaboratif ou participatif permet de construire et d'entretenir des interactions selon ses relations professionnelles ou ses intérêts. Cependant, lorsque ces sites exigent de chaque utilisateur une description explicite de son réseau social ou de son profil, il n'est possible d'identifier que des communautés ainsi explicitement constituées.

Or un grand nombre de communautés d'utilisateurs existent de façon implicite dans de nombreux domaines. Par exemple, tout site de musique généraliste rassemble une communauté d'utilisateurs ayant des goûts musicaux variés. Toutefois, cette communauté est en fait composée de sous-communautés potentiellement disjointes, toutes liées à la musique (la communauté des amateurs de musique pop, de musique punk, etc.). Découvrir et identifier précisément ces communautés implicites est un gain pour de nombreux acteurs : le propriétaire du site, les régies publicitaires Web et surtout, les utilisateurs du système. C'est en ce sens que nous proposons « WebTribe » pour identifier et expliciter ces communautés implicites.

Dans ce travail, nous nous concentrons sur des forums Internet pour des raisons de simplicité, mais la même technique s'applique pour tout système de communication avec une structure « envoyer / recevoir » de communication interne, comme les messages sur les murs Facebook accessibles, les *tweets* sur Twitter, les mails, etc. Nous proposons trois approches pour définir des communautés en se basant respectivement sur l'activité des utilisateurs, sur un vocabulaire, puis enfin sur une ontologie légère.

3.3 Analyse de communautés

L'une des principales difficultés de la tâche du gestionnaire de communautés est l'énorme quantité de messages à laquelle il est confronté. Si dans un système de taille réduite — avec quelques dizaines d'utilisateurs tout au plus — un gestionnaire peut s'imprégner du comportement des utilisateurs et communautés pour en comprendre la structure interne et les évolutions, ce mode de fonctionnement ne saurait passer à l'échelle et est impensable sur des systèmes bien plus importants.

L'analyse automatique des communautés est donc une étape nécessaire de notre approche, afin de caractériser les communautés que nous détectons précédemment.

Cela revient à extraire les traits essentiels de chaque regroupement d'utilisateur, tels que la mise en évidence des flux principaux de communication qui composent la communauté, ses thématiques, ses utilisateurs principaux, et ainsi de suite. Nous proposons donc que le système WebTribe mène ce genre d'analyse conjointement à la découverte de ces communautés.

De plus, sur le plan pratique, l'analyse des communautés doit s'appuyer sur une structure incrémentale, qui est mise à jour dès que de nouveaux messages sont pris en compte. Il s'agit d'une condition obligatoire pour passer à l'échelle, afin de pouvoir suivre une grande quantité de forums. Pour mettre en place cette approche, nous pensons WebTribe comme un ensemble de services interconnectés pour répondre à ces différentes contraintes. Nous proposons deux approches s'appuyant sur la définition de deux métriques et sur la caractérisation des rôles des utilisateurs.

3.4 L'outil WebTribe

La figure 3.1 présente une vision modulaire du système Webtribe précédemment introduit au chapitre 1. Dans cette vision, nous décrivons le système WebTribe comme un ensemble de tâches distinctes, servant deux missions principales que sont la découverte et l'analyse des communautés. Ces tâches modulaires sont décrites ci-dessous.

Normalisation de la communication À partir de sources Web que déclare le gestionnaire de communautés du système, WebTribe extrait les échanges entre les utilisateurs, et retranscrit la communication sous une forme normalisée. Respectant un format commun et standardisé, WebTribe rend ainsi ces échanges exploitables indépendamment de la nature de leurs sources d'origine. L'annexe C présente un exemple de cette normalisation.

Analyse sémantique Le cœur d'analyse sémantique de WebTribe a pour but de détecter et d'extraire les contenus sémantiques pertinents des données qu'il traite. Pour cela, il s'appuie sur une base de connaissances de référence, qui peut être

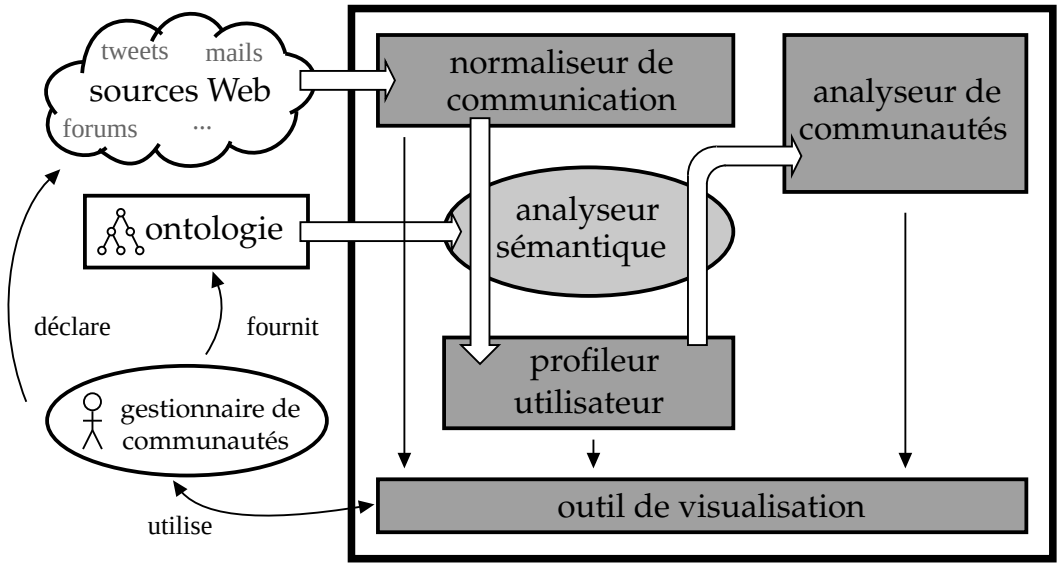


FIGURE 3.1 – Organisation modulaire de WebTribe

un vocabulaire ou une ontologie légère. L'analyse sémantique de WebTribe a une fonction de synthèse, car le contenu sémantique extrait des communications est alors considéré comme le résumé pertinent des données analysées.

Profilage utilisateur Grâce aux résultats de l'analyseur sémantique, WebTribe dresse un profil sémantique des utilisateurs du système, à partir des concepts extraits de la communication. Comme l'analyseur sémantique, le profilage utilisateur a également une fonction de synthèse, car il résume l'utilisateur aux concepts essentiels majoritaires qu'il manipule lors de ces échanges. Cette fonction permet également à WebTribe d'être relativement insensible à la charge, puisque le profil résumant un utilisateur ne grossit pas comme la somme de ses contributions.

Analyse de communautés À partir du profilage utilisateur, WebTribe regroupe les utilisateurs en fonction de leurs similitudes pour construire et maintenir des communautés d'utilisateurs. En fonction des besoins du gestionnaire de communautés, WebTribe est en mesure de réaliser diverses analyses et mesures sur ces communautés et sur les utilisateurs qui les composent.

Visualisation Au moyen de diverses interfaces, WebTribe fournit au gestionnaire une vue globale de son système, ainsi que des vues détaillées des utilisateurs et communautés, selon les besoins. Cette visualisation peut prendre plusieurs formes, en fonction des implémentations utilisées. Ces aspects sont étudiés plus en avant dans le chapitre 6 dédié aux implémentations.

3.5 Conclusion

Dans ce chapitre, nous avons présenté une vue d'ensemble de notre approche en matière de découverte et d'analyse de communauté, les objectifs de notre travail et l'outil WebTribe qui permet la mise en œuvre des méthodes proposées. Nous avons défini l'outil WebTribe comme un système modulable, pouvant répondre à différents besoins du gestionnaire de communautés. Dans les chapitres suivants, nous détaillons les approches introduites dans ce chapitre en matière de découverte de communautés, puis d'analyse de communautés, dont la mise en œuvre est assurée par WebTribe.

Chapitre 4

Découverte de communautés

Sommaire

Dans ce chapitre, nous présentons nos différentes approches pour la découverte de communautés. Ces propositions sont basées sur une analyse des activités des utilisateurs. Nous employons des approches aux caractéristiques variées — des *tags*, références internes non-structurées, aux ontologies externes et structurées — pour tenter de diversifier au mieux nos détections. Nous décrivons tout d'abord notre approche de partitionnement basée sur l'annotation de ressources par les utilisateurs [1]. Reposant sur une analyse en composantes principales des usages (ACP), cette approche est détaillée en section 4.1. Nous présentons une approche basée sur un vocabulaire de référence [44] dans la section 4.2 et sur une ontologie [46] dans la section 4.3. Enfin, la section 4.4 conclut le chapitre.

4.1 Découverte basée sur l'activité

Dans cette section, nous présentons une méthode de détection de communautés générique qui ne s'appuie que sur un étiquetage des ressources et sur l'annotation par les utilisateurs. L'étiquetage consiste en l'apposition d'un *tag* par un utilisateur sur une ressource. L'utilisation d'une ressource est l'interaction quelconque d'un utilisateur avec une ressource.

Exemple 4.1

> Est considéré comme une utilisation de ressource le fait que Bob achète sur une boutique Web un fichier vidéo étiqueté `western`.

Le cœur de notre méthode est une analyse statistique en composantes principales (ACP [22]) des étiquettes des ressources manipulées par les utilisateurs. Cette méthode permet de représenter les données originelles — utilisateurs et étiquettes manipulées — dans un espace de dimension inférieure à celle de l'espace originel, tout en minimisant la perte d'information. La représentation des données dans cet espace de faible dimension en facilite considérablement l'analyse et permet ainsi de regrouper ou d'opposer des communautés. L'analyse en composantes principales est une méthode géométrique et statistique couramment employée dans ce cas de figure où les données initiales — ici les *tags* — sont nombreux et corrélés, pour obtenir des résultats — les communautés — plus concis et décorrélés.

Ce travail a été publié dans le *workshop Web Social* de la conférence *Extraction*

et Gestion des Connaissances 2010, à Hammamet, Tunisie [1].

4.1.1 Définition du modèle

On considère un ensemble d'utilisateurs $U = \{u_1, \ldots, u_n\}$ et un ensemble de ressources $R = \{r_1, \ldots, r_m\}$ sur un site donné, comme par exemple des fichiers de musiques, des vidéos, des nouvelles sur une plate-forme de consultation Web. Nous supposons que les utilisateurs émettent un vote sur un sous-ensemble des ressources du site. Ce vote n'est pas nécessairement explicite et peut être obtenu en se basant sur les usages des utilisateurs.

Exemple 4.2

> L'achat réalisé par Bob dans l'exemple 4.1 est interprété comme un vote implicite en faveur de la ressource, à moins qu'il ne dépose un vote explicite négatif.

Les votes sont illustrés par une matrice $M : |U| \times |R|$ définie comme suit, pour un utilisateur $u_i \in U$ et une ressource $r_j \in R$:

$$M(u_i, r_j) = \begin{cases} 1 & \textit{si } u_i \textit{ a de l'intérêt pour } r_j, \\ 0 & \text{sinon.} \end{cases}$$

Cette matrice est mise à jour dynamiquement lorsque de nouveaux utilisateurs, de nouvelles ressources ou de nouveaux usages apparaissent sur le site. Nous supposons également qu'un ensemble de *tags* $T = \{t_1, \ldots, t_m\}$ est défini (par exemple, films `western`, `animation`, `drame`, etc.), et que chaque ressource est annotée avec un sous-ensemble de ces tags (sous-ensemble potentiellement vide). Ces annotations proviennent des fournisseurs de ressources, qui peuvent être les utilisateurs eux-mêmes, et peuvent s'enrichir au fur et à mesure. Étant donnés les votes des utilisateurs et ces annotations, nous définissons l'ensemble $A(u_i) \subseteq R$ des ressources intéressant l'utilisateur $u_i \in U$, soit :

$$r_j \in A(u_i) \Leftrightarrow M(u_i, r_j) = 1.$$

Nous définissons également l'ensemble $A(u_i, t_j) \subseteq A(u_i) \subseteq R$, où $t_j \in T$, l'ensemble

des ressources intéressant u_i annotées par le tag t_j.

L'objectif principal de l'approche proposée est de scinder les utilisateurs en communautés distinctes, en se basant sur les groupes de tags qu'ils apprécient. Nous calculons le degré d'appartenance x_{ij} d'un utilisateur u_i à un tag t_j :

$$x_{ij} = \frac{|A(u_i, t_j)|}{|A(u_i)|}. \tag{4.1}$$

Plus un coefficient x_{ij} est proche de 1, plus l'utilisateur i manipule des tags de type j.

4.1.2 Communautés de tags

On cherche ensuite à rassembler les tags similaires, de façon statistique. Pour cela, on utilise la technique de l'analyse en composantes principales (ACP). Cette méthode consiste à transformer des variables corrélées en nouvelles variables décorrélées les unes des autres, nommées « composantes principales ». Elle permettent de réduire l'information en un nombre de composantes plus limité que le nombre initial de variables. Dans la suite, l'usage d'une ressource portant un tag donné est vu comme la réalisation d'une variable aléatoire représentant ce tag. Les intérêts de chaque utilisateur sont alors autant de réalisations indépendantes des m variables représentant les m tags possibles. L'objectif de l'ACP est de trouver des combinaisons linéaires des variables représentant les tags pour expliquer au mieux les intérêts des utilisateurs. Ainsi, à chaque utilisateur u_i, nous associons le vecteur X_i de ses degrés d'appartenance à chaque tag, $X_i = (x_{i1}, x_{i2}, ..., x_{im})$. Ce vecteur représente le positionnement de l'utilisateur dans l'espace des tags, et l'ensemble des vecteurs X_i donne ainsi un nuage de points dans l'espace des tags. De la même manière, on peut associer à chaque tag t_j le vecteur V_j, correspondant à ses degrés d'appartenance chez les n utilisateurs : $V_j = (x_{1j}, x_{2j}, .., x_{ij}..., x_{nj})$. Ces nuages de points sont difficiles à analyser, à cause des dimensions considérées (nombre de tags, nombre d'utilisateurs) et de la variabilité des observations. L'analyse en composantes principales va alors :

1. permettre une projection du nuage de points utilisateurs (initialement exprimés dans un espace de dimension k) sur des plans principaux (de dimen-

sion 2) qui reconstituent au mieux la variabilité entre les utilisateurs ;

2. permettre une représentation des variables initiales dans ces plans principaux, la contribution des variables dans la construction des axes principaux n'étant pas la même pour toutes les variables. Par exemple, la figure 4.1 donne une représentation compacte des rassemblements de tags selon leurs usages.

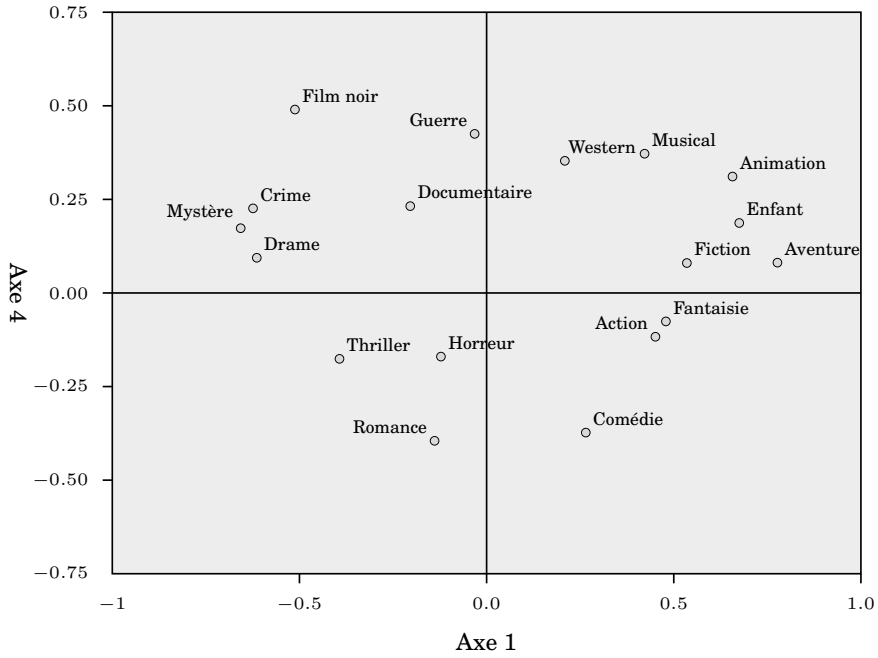


FIGURE 4.1 – Projection des variables de l'ACP sur deux axes

Ainsi, des axes explicatifs sont identifiés, en minimisant la perte d'information effectuée lors de cette simplification. La figure 4.1 représente les variables originales de nos expérimentations — basées sur un jeu de données cinématographiques — sur deux axes significatifs, appelés *composantes principales* (dans cette figure, nommés axes 1 et 4). Cette figure présente la corrélation des variables d'origine avec les composantes principales (une variable est bien représentée sur l'axe si sa corrélation avec la composante principale correspondante est en valeur absolue proche de 1). Selon la composante 1 (Axe 1), on voit que les tags `Animation` et `Enfant` sont très corrélés (corrélation supérieure à 0,6). De même, la composante 4 oppose les tags `Film Noir`, `Guerre` aux tags `Romance`, `Comédie`.

Notre méthode de rassemblement de tags est alors la suivante : l'ACP fournit les composantes principales pertinentes pour l'analyse des usages. Selon chacune de ces composantes, on ignore les tags situés dans la zone de faible corrélation (corrélation entre $-\alpha$ et $+\alpha$, pour un seuil $\alpha \in]0,1]$ fixé). Les tags restants, situés dans les zones de forte corrélation (inférieure à $-\alpha$ ou supérieure à $+\alpha$), sont rassemblés dans une même communauté de tags. Par exemple, `Animation` et `Enfant` seront dans une même communauté. L'algorithme 1 résume notre méthode.

Algorithme 1 : Découverte

entrées : Vecteurs V_j, seuil de décision α
sorties : Communautés de tags $G_1, \ldots, G_K$

```
début
    identifier les composantes principales C = ((c1, c2), (c3, c4)...), expliquant
    la plus grande proportion de la variabilité des données
    tant que (il reste des composantes principales (c, c') dans C) faire
        ignorer les tags non corrélés (|coordonnées selon c et c'| < α)
        rassembler dans une même communauté les tags corrélés selon c
        (|coordonnées selon c| > α)
        rassembler dans une autre communauté les tags corrélés selon c'
        (|coordonnées selon c'| > α)
        supprimer ces tags
    fin tant que
fin
```

4.1.3 Communautés d'utilisateurs

Après avoir décomposé l'ensemble des *tags* de T en K communautés de *tags* $G_1, \ldots, G_K$, nous en déduisons les communautés d'utilisateurs. Pour un utilisateur u_i donné, nous calculons son degré d'appartenance x'_{ij} à chaque communauté de *tag* G_j :

$$x'_{ij} = \sum_{t_k \in G_j} x_{ik}.$$

Sa communauté $c(u_i)$ est alors sa communauté de *tag* majoritaire, c'est-à-dire l'indice j tel que x'_{ij} soit maximal. Chaque utilisateur est alors associé à ce groupe de *tags*. Ce groupe aura comme intitulé l'ensemble des *tags* qui le constitue.

4.1.4 Expérimentations

Contexte Nous avons expérimenté notre méthode sur la base de films MovieLens [1]. Cette base contient 100 000 votes pour 1 682 films appréciés par 943 utilisateurs. Les films sont évalués par une note entre 1 et 5. Nous avons remplacé ces notes par un vote binaire (les notes supérieures à 2 indiquant un intérêt pour le film). Nous avons construit la matrice M avec l'ensemble des utilisateurs U et l'ensemble des films R, et calculé le degré d'appartenance des utilisateurs aux différents tags. Nous présentons les résultats de notre approche sur un ensemble de 18 tags (1 : `Aventure`, 2 : `Enfant`, 3 : `Animation`, 4 : `Mystère`, 5 : `Crime`, 6 : `Drame`, 7 : `Fiction`, 8 : `Film noir` , 9 : `Fantasy`, 10 : `Musical`, 11 : `Action`, 12 : `Thriller`, 13 : `Romance`, 14 : `Comédie`, 15 : `Horreur`, 16 : `Guerre`, 17 : `Documentaire`, 18 : `Western`). Le seuil de décision α a été fixé à 0,6 de façon empirique.

Matrice de corrélation La première étape de l'analyse est de vérifier que les données sont factorisables, c'est-à-dire qu'elles sont corrélées entre elles. Pour cela, on examine la matrice de corrélation :

- si les coefficients de corrélation entre variables sont faibles, il est improbable d'identifier des facteurs communs. On peut éventuellement supprimer les variables qui ont une corrélation faible ;
- un autre paramètre pouvant aider au choix des variables est la qualité de la représentation (*Communalities*) ; QLT_j est le cosinus carré de l'angle formé entre la variable initiale x_j et l'axe principal c (voir exemple 4.3).

La table 4.1 représente la matrice de corrélation entre une partie des variables initiales et les 6 premières composantes principales.

1. http ://www.grouplens.org/node/73

Tag	1	2	3	4	5	6
Aventure	**,777**	,349	-,272	,081	,037	-,056
Enfant	**,675**	-,231	*,465*	,187	-,145	-,147
Animation	**,657**	-,200	*,391*	,311	-,052	-,218
Mystère	**-,657**	,258	*,367*	,173	-,254	-,057
Crime	**-,624**	,265	,094	,226	,237	-,016
Drame	**-,614**	-,561	-,230	,094	,016	-,112
Fiction	,535	,531	-,252	,080	,249	-,152
Film noir	-,512	,066	,209	**,490**	,083	,158
Fantasy	,479	-,108	,197	-,076	,208	-,022
Musical	,422	-,409	,380	,372	-,193	,096
Action	,451	**,746**	-,262	-,117	-,128	,028
Thriller	-,393	**,704**	,314	-,176	-,221	,011
Romance	-,139	**-,685**	-,221	**-,395**	-,231	-,023
Comédie	,265	**-,592**	,161	**-,373**	,225	,242
Horreur	-,122	,424	*,360*	-,170	,369	,179
Guerre	-,032	-,037	**-,633**	**,425**	-,331	-,103
Documentaire	-,204	-,263	-,166	,232	**,639**	-,400
Western	,209	-,105	-,262	,353	,142	**,780**

TABLE 4.1 – Corrélation entre les variables et les composantes de l'ACP

Exemple 4.3

La qualité de la représentation de la variable `Action` est obtenue en élevant au carré les coefficients de corrélation entre cette variable et les 6 axes principaux, puis en les sommant :

$$QLT_{\texttt{Action}} = (0,451)^2+(0,746)^2+(0,262)^2+(0,117)^2+(0,128)^2+(0,028)^2 = 0,859.$$

Ainsi pour chaque variable initiale, nous obtenons la variance prise en compte par l'ensemble des facteurs extraits. Plus cette valeur est proche de 1, plus l'ensemble de l'information contenue dans la variable est prise en compte. Il serait par exemple possible de négliger la variable correspondant au tag `Fantaisie` en raison de sa faible qualité de représentation (nous l'avons cependant conservée lors de nos expérimentations).

Sélection des composantes principales La deuxième étape consiste à déterminer le nombre de facteurs à retenir. On tient compte :

- des facteurs qui permettent d'extraire une quantité d'informations (valeur propre) > 1. Quand on a beaucoup de variables, il y a un grand nombre de facteurs pour lesquels la valeur propre est supérieure à 1. Dans ce cas, on retient beaucoup de facteurs et l'interprétation devient difficile.
- de la distribution des valeurs propres : utilisation du graphique des valeurs propres.

La figure 4.2 représente la variance expliquée par chaque composante principale (valeur propre). Pour savoir combien de composantes principales utiliser, on recherche une rupture de pente sur le graphique. Cette rupture signifie que l'on passe d'un facteur représentant beaucoup d'informations à un facteur en représentant moins. On s'arrête au facteur précédant cette rupture de pente. Dans notre expérimentation, on retient les 6 premières composantes dont la valeur propre est supérieure à 1. Le pourcentage de variance expliquée est de 70%.

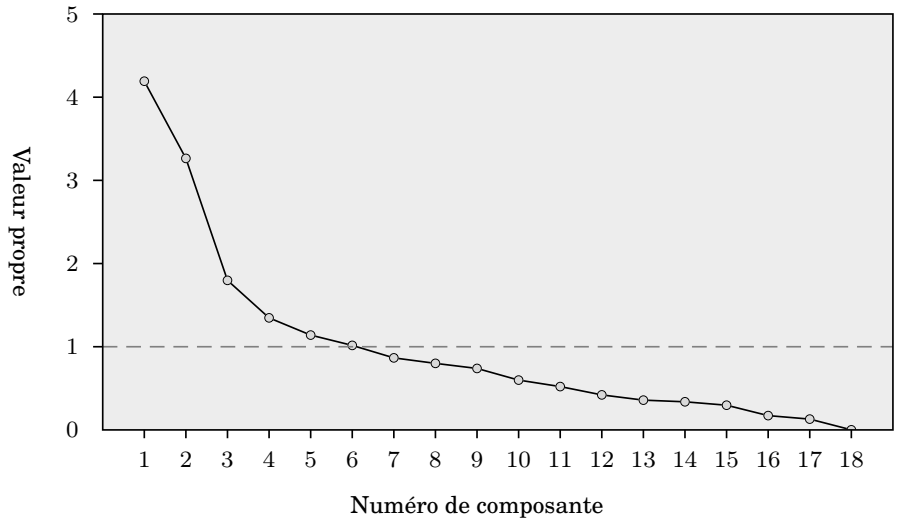


FIGURE 4.2 – Variance expliquée par chaque composante principale de l'ACP

Les composantes obtenues ont la structure suivante :

- la 1re composante principale est la combinaison qui totalise la plus grande quantité de variance ;
- la 2e composante principale est la combinaison qui totalise la 2ème plus grande quantité de variance. On peut déterminer autant de composantes

principales qu'il existe de variables. La valeur propre de la 1re composante principale est 4,192 (soit 23,29% de la variance), celle de la 2e composante est 3,264 (soit 18,13% de la variance),etc. Les composantes principales sont indépendantes les unes des autres.

À partir de la matrice de corrélation, on voit que :

- la 1re composante principale représente essentiellement les variables `Aventure`, `Enfant`, `Animation`, `Mystère`, `Crime` et `Drame` ;
- la 2e composante principale représente essentiellement les variables `Action`, `Thriller`, `Romance` et `Comédie` ;
- la 3e composante principale représente essentiellement la variable `Guerre` et à un moindre degré les variables `Enfant`, `Animation`, `Mystère` et `Horreur` ;
- la 4e composante principale représente essentiellement les variables `Film Noir`,`Guerre` d'une part, et `Romance`, `Comédie` d'autre part ;
- la 5e composante principale représente essentiellement la variable `Documentaire` ;
- la 6e composante principale représente essentiellement la variable `Western`.

Interprétation des axes La dernière étape de l'expérimentation est l'interprétation des axes. On donne un sens à un axe à partir des coordonnées des variables. Ce sont les valeurs extrêmes qui concourent à l'élaboration des axes. Les facteurs avec de larges coefficients (en valeur absolue) pour une variable donnée indiquent que ces facteurs sont proches de cette variable. Nous rapprochons les tags par les degrés d'appartenance des utilisateurs à ces tags en nous basant sur les graphiques générés lors de cette étape :

- le 1er axe (figure 4.3) oppose les tags `Animation`, `Enfant` et `Aventure` aux tags `Mystère`, `Crime` et `Drame`. Ceci correspond à une interprétation naturelle : les personnes qui aiment le premier groupe de films n'aimant en général pas le second. Deux communautés sont ainsi créées ;
- le 2e axe oppose les films de `Romance` et de `Comédie` aux films `Thriller` et `Action`, en créant ainsi deux nouvelles communautés ;
- le 3e axe (figure 4.4) oppose les films de `Guerre` aux films étiquetés `Enfant`, d'`Animation` ou de `Mystère`.

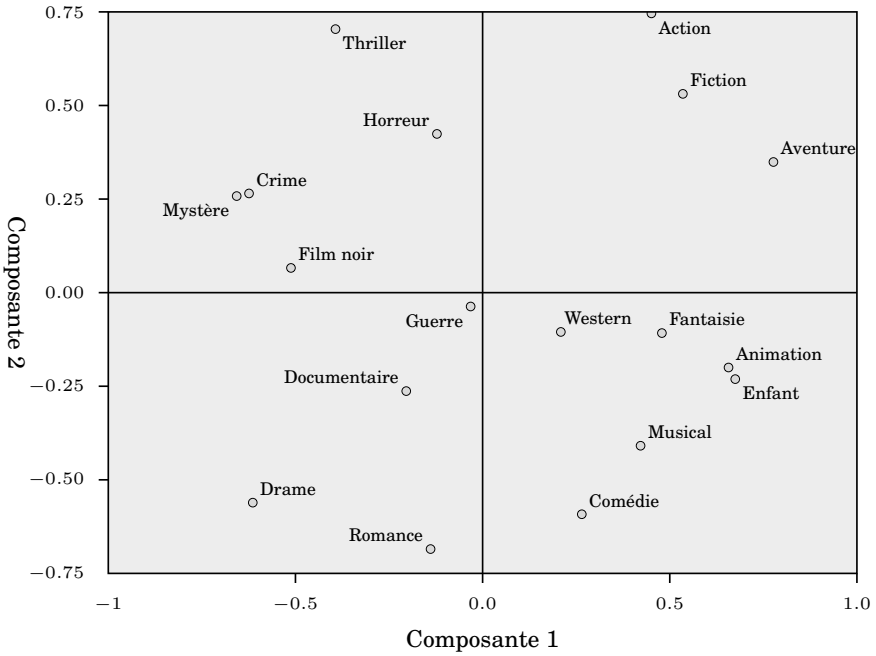


FIGURE 4.3 – Composantes 1 et 2 de l'ACP

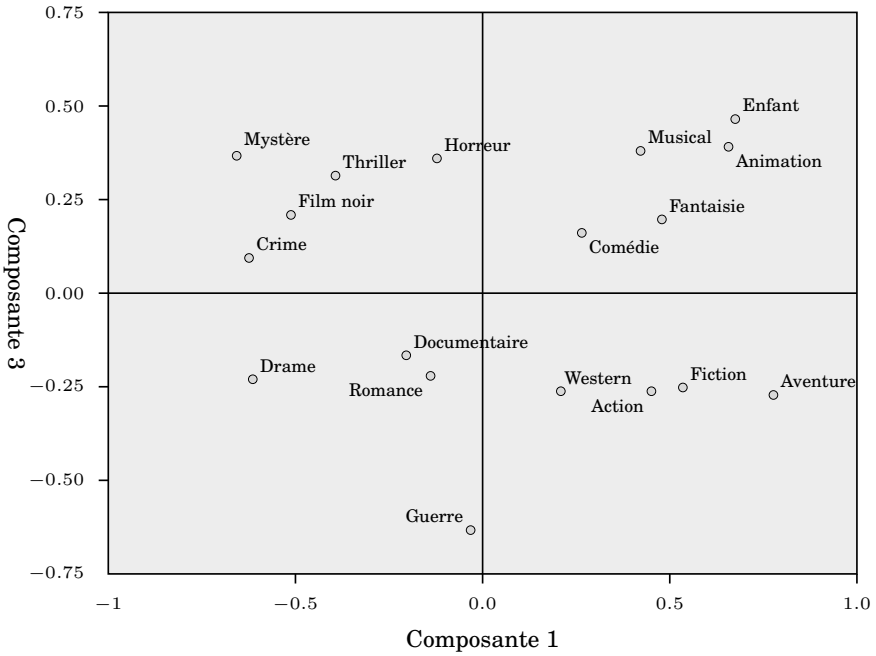


FIGURE 4.4 – Composantes 1 et 3 de l'ACP

Les axes 4, 5 et 6 nous donnent les résultats suivants :

- le 4e axe oppose les films `Film noir` et les films de `Guerre` aux films de `Romance` et de `Comédie` ;
- le 5e axe oppose les films `Documentaire` aux films de `Guerre` ;
- le 6e axe oppose les films `Western` aux films `Documentaire`.

Cette interprétation nous donne 7 groupes de tags, comme indiqué dans la table 4.2. Les groupes qui sont disjoints sont 1 et 2, 3 et 4, 4 et 6 et enfin 6 et 7. Les utilisateurs sont regroupés en fonction de ces communautés de tags. Les tags qui ne sont pas pris en compte par les axes sont expliqués par leur faible occurrence : par exemple le tag `Fantaisie` n'est utilisé que 22 fois sur toute la collection des 1682 films.

communauté	tags associés
1	`Aventure`, `Enfant`, `Animation`
2	`Mystère`, `Crime`, `Drame`
3	`Action`, `Thriller`
4	`Romance`, `Comédie`
5	`Western`
6	`Film noir`, `Guerre`
7	`Documentaire`

TABLE 4.2 – Communautés de tags

4.1.5 Conclusion

Par cette approche, nous avons proposé une première méthode de découverte de communautés d'utilisateurs basée sur l'observation des usages. Nous utilisons une analyse en composantes principales (ACP) pour réaliser des corrélations entre les *tags* utilisés par les utilisateurs, constituant ainsi des communautés. Nous faisons apparaître le fait qu'un *tag* ne peut être bien corrélé — et donc mis en avant par notre méthode — que s'il est suffisamment employé par les utilisateurs. Il convient donc de définir un ensemble de *tags* correspondant au public qui les utilise, ou de gérer ceux créés par ce public. Dans la section suivante, nous présentons une autre approche de découverte de communautés, basée sur un vocabulaire de termes.

4.2 Découverte basée sur les termes

L'objectif de cette approche est d'identifier les communautés implicites à partir des discussions des utilisateurs, en se concentrant sur des sujets spécifiques. Les sujets concernés sont issus d'un vocabulaire. Nous utilisons dans cette approche une première version du système WebTribe, pour extraire et analyser les communications entre utilisateurs.

Ce travail a été présenté dans la conférence *International Conference on Advances in Semantic Processing (SEMAPRO 2010)*, à Florence, Italie [44].

4.2.1 Méthode

Nous présentons brièvement chacune des étapes de notre analyse, qui seront détaillées ensuite.

En premier lieu, un gestionnaire de communautés fournit une liste de sujets qui est utilisée pour construire un graphe des sujets présenté ci-après. Ce graphe sert de base de notre analyse. Il sera potentiellement réduit en retirant les sujets non pertinents, et sera utilisé pour le positionnement sémantique des utilisateurs.

En parallèle, le système recueille diverses publications — comme les messages publiés par les utilisateurs — à partir de diverses sources choisies par l'analyste Web, et associe chaque publication à son auteur.

Ensuite, nous extrayons pour chaque publication ses thèmes principaux, et quantifions « l'attractivité de la publication » par sujet, comme étant le degré d'importance évalué de chaque sujet dans la publication. En analysant toutes les publications trouvées pour un auteur, nous sommes maintenant en mesure de calculer « l'attractivité de l'utilisateur » par sujet, sur le même principe. En utilisant une méthode de calcul de distance sémantique basée sur le Web, nous évaluons la distance entre les sujets et positionnons ainsi les utilisateurs au sein du graphe des sujets.

Enfin, interroger le système revient maintenant à calculer un sous-graphe de nos résultats, en y incluant les utilisateurs qui valident une contrainte de proximité donnée par la requête, basée sur la distance sémantique calculée précédemment.

Nous détaillons dans la suite chacune de ces étapes.

4.2.2 Graphe de sujets

Choix des sujets Notre méthode vise à regrouper les utilisateurs en fonction de leurs affinités avec les sujets définis. L'analyste Web doit définir quels sont les principaux sujets pertinents pour l'analyse de son système. Nous appelons cette liste de sujets le « lexique » du système, le but étant d'avoir suffisamment de sujets pour couvrir l'ensemble des utilisateurs. Mais à l'inverse, avoir trop de sujets n'est pas souhaitable non plus et surchargerait inutilement le système. Nous présentons une méthode pour réduire l'ensemble des sujets de sorte que seuls restent les sujets les plus pertinents, formant ainsi notre vocabulaire optimisé.

Création du graphe Une fois que l'analyste Web a fourni les sujets du système, il nous faut les organiser. Pour cela, nous construisons un graphe sémantique pondéré, que nous basons sur une matrice de distance entre les sujets. Pour définir ces similarités, nous utilisons la méthode de calcul de distance sémantique basée sur le Web de Cilibrasi et Vitanyi [13]. Cette méthode évalue la distance sémantique entre un terme x et un terme y à partir du contenu du Web (plus précisément, à partir du moteur de recherche de Google, ce qui semble être un des meilleurs points de vue disponible). Cette méthode est bien adaptée à notre approche, car elle n'extrait pas les distances sémantiques à partir de bases de connaissances qu'il nous faudrait fournir, comme des taxonomies ou ontologies de référence. De plus, étant donné que nous avons analysé des données échangées entre utilisateurs sur le Web, utiliser le Web comme référence semble approprié au contexte.

La distance sémantique entre x et y définie par Cilibrasi et Vitanyi se calcule comme suit, pour $f(\lambda)$ donnant la fréquence du terme λ et M le nombre total de termes indexés :

$$\textsc{Dist}(x, y) = \frac{max\{\log\ f(x), \log\ f(y)\} - \log\ f(x, y)}{\log\ M - min\{\log\ f(x), \log\ f(y)\}}.$$

En utilisant Google, $f(\lambda)$ correspond au nombre de résultats de la requête « λ » et M le nombre de documents indexés par Google, qui est estimé à 1 000

milliards à la date de [13]. Cette expression calcule la plus basse probabilité de $x|y$ et $y|x$. L'utilisation d'un logarithme négatif permet d'augmenter l'importance des différences, et la division permet de résoudre les problèmes d'échelle. Enfin, notre graphe est un graphe complet avec des sujets comme sommets. Une arête entre les sujets t_i et t_j est annotée par la distance sémantique ainsi calculée.

Réduction des sujets Le graphe résultant du lexique est un graphe complet, avec donc un grand nombre d'arêtes. Afin d'être le plus pertinent possible, mais aussi afin d'être facilement utilisé, il doit être réduit. En effet, plus le nombre de sujets est faible par rapport à la masse de contenu analysé, plus les nuances de distance entre les utilisateurs ont une signification intéressante.

Pour ces raisons, nous écartons les sujets considérés comme non-pertinents. Un sujet est défini comme non-pertinent quand il est sémantiquement trop proche d'un autre. Autrement dit, lorsque leur distance est plus petite qu'un seuil δ_s, fourni par l'analyste Web. Dans ce cas, le sujet de plus basse fréquence est supprimé. Cela permet également d'effectuer un retour auprès de l'analyste, l'informant de la non-pertinence de certains des sujets qu'il a lui-même choisi.

Exemple 4.4

> L'analyste Web d'un forum de fans de voiture soumet le lexique suivant : `ferrari`, `porsche`, `tuning`, `essence`, `concessionnaire`, `moteur` et `carburant`.
> Par la réduction présentée, le sujet `essence` a été retiré, considéré comme ayant une proximité avec `carburant` inférieure au seuil δ_s. Les distances calculées par WebTribe à partir de ce lexique sont exposées dans la table 4.3 et illustrées par la figure 4.5.

	`carburant`	`moteur`	`concess.`	`tuning`	`porsche`
`ferrari`	1,3478	1,6431	1,0418	1,3010	0,4195
`porsche`	1,1140	1,4399	0,9475	1,1301	-
`tuning`	1,3064	1,4529	0,7161	-	-
`concessionnaire`	0,8998	1,1027	-	-	-
`moteur`	1,0774	-	-	-	-

TABLE 4.3 – Matrice des distances calculées avec le lexique d'exemple

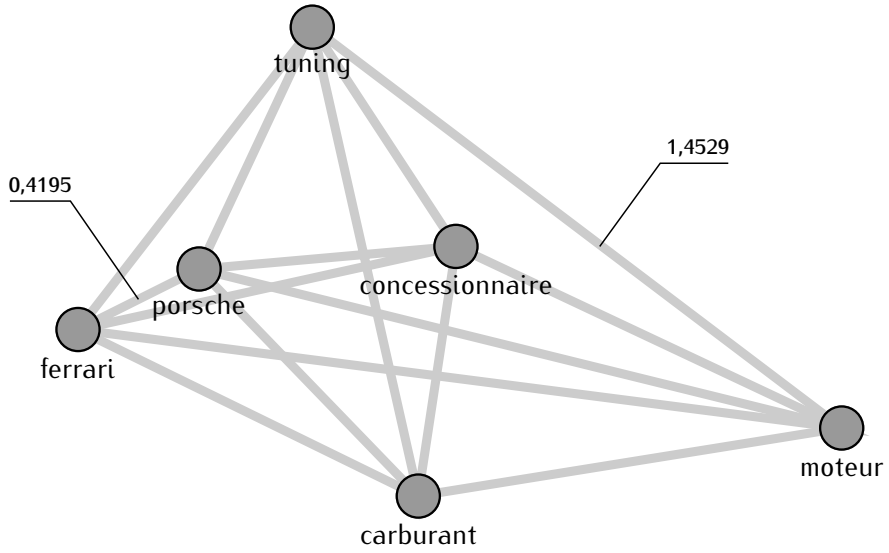


FIGURE 4.5 – Exemple de graphe des sujets réduit

4.2.3 Attractivités et interrogation

Acquisition de données utilisateur À partir de la source de données que nous analysons, nous récupérons les différentes publications des utilisateurs du système. Le système source peut être un site, un blog, un réseau social, la version Web d'un journal permettant de déposer des commentaires, ou toute plate-forme permettant aux utilisateurs de publier du contenu. Cette opération peut être faite par un *wrapper* spécialement conçu pour la source donnée, incluant un *parser* spécifique et fournissant des données dans un format normalisé. Il est également possible de s'appuyer sur une API classique pour extraire des informations, comme l'API Facebook [2] par exemple.

Attractivité des publications Pour chaque contenu analysé, nous cherchons à en extraire les sujets principaux et pour chacun d'entre eux, définir l'attractivité de la publication par sujet. Nous utilisons le lexique précédemment optimisé $t_1, \ldots, t_n$, qui contient tous les sujets pertinents à rechercher. S'il y a n sujets dans le lexique, nous pouvons définir que chaque sujet t est assigné à une dimension dans un espace vectoriel ; le lexique est alors la base d'un hypercube à n dimensions. Chaque

2. http://developers.facebook.com

publication p peut être considérée comme un vecteur $\bar{p}$ de sujets dans cet espace, de sorte que $\bar{p} = (p_{t_1}, \ldots, p_{t_n}) \in \mathbb{R}^{+^n}$.

Pour déterminer les sujets abordés dans une publication, nous utilisons une approche issue des travaux de Das et Chen [15]. Cette méthode consiste à analyser le document avec cinq différents algorithmes pour déterminer à la majorité simple si le texte contient un sentiment sur le sujet — positif ou négatif — ou non. Étant donné que nous cherchons à prendre en compte l'intérêt et non l'opinion, nous interprétons les deux sentiments comme un vote positif en intérêt. Inversement, nous interprétons l'absence de sentiment comme un vote négatif en intérêt. Sur cette base, nous construisons un vecteur pour chaque publication. Cette méthode, a l'avantage de fournir des résultats pertinents et fiables en ne retenant pas les contenus qui ne remportent pas la majorité. En d'autres termes, la qualité des contenus détectés prime sur la quantité de détections. Comme effet secondaire intéressant, il nous permet également d'éliminer les spams : n'ayant pas remporté la majorité des tests, ils sont tout simplement ignorés.

Exemple 4.5

> Considérant le lexique de l'exemple 4.4, la publication « `Revue et test de ma nouvelle Carrera` » est analysée comme suit :
>
> $$p = (0, 5, 0, 0, 3, 1).$$
>
> Cela signifie que le sujet `porsche` est considéré comme hautement pertinent avec un score de 5 (la similarité avec le terme « Carrera » est ressortie de l'analyse du Web précédemment évoquée, « Carrera » étant une marque déposée portée par plusieurs voitures produites par Porsche). Le sujet `moteur` est identifié comme un sujet d'importance moyenne dans le document avec un score de 3, et `carburant` comme un sujet mineur avec un score de 1. Les sujets `ferrari`, `tuning` et `concessionnaire` ont été considérés comme non-pertinent pour ce document.

Attractivité des utilisateurs À partir de toutes les attractivités calculées depuis les publications collectées, nous sommes maintenant en mesure de calculer

l'attractivité des utilisateurs comme autant de vecteurs du même type que précédemment. Nous définissons le vecteur $\bar{u}$, avec $\bar{u} \in \mathbb{R}^{+^n}$, tel que $\bar{u} = \sum \bar{p}$ avec $\bar{p}$ étant un vecteur de publication de l'utilisateur.

Nous utilisons une somme plutôt que de normaliser ces résultats, afin de préserver le caractère indépendant du taux de participation. Il ne faut pas en effet que l'activité d'un utilisateur dans un domaine baisse à elle seule la valorisation de son activité dans d'autres domaines. Par exemple, si un utilisateur est l'auteur de nombreuses contributions relatives à un sujet donné, normaliser les résultats impliquerait de réduire son importance dans cette communauté, dès lors qu'il publierait également d'autres documents relatifs à un autre sujet indépendant. Ce calcul n'aurait aucun sens dans ce cas.

Graphe Nous avons maintenant un graphe sémantique du lexique basé sur la matrice de distance sémantique, et un vecteur d'attractivité $\bar{u}$ par utilisateur. Nous traduisons ces vecteurs en distance sémantique afin d'uniformiser nos mesures avec celles du graphe, comme suit :

$$\textsc{Dist}(u, t_i) = \frac{1}{\log u_{t_i}}$$

Enfin, les utilisateurs sont positionnés dans le graphe, en fonction de leurs attractivités.

Exemple 4.6

> La figure 4.6 montre le positionnement des utilisateurs en se basant sur les mêmes données que précédemment.

Ce graphe est la simple traduction de la matrice des distance précédemment calculée. Cependant, sa visualisation — par des outils classiques de type Tulip [5] — constitue une aide pour le gestionnaire de communautés.

4.2.4 Aspects incrémentaux

Nouvelle publication Le système est prévu pour une exploration continue des sites ciblés, et une analyse évolutive. Quand une cible reçoit de nouveaux messages,

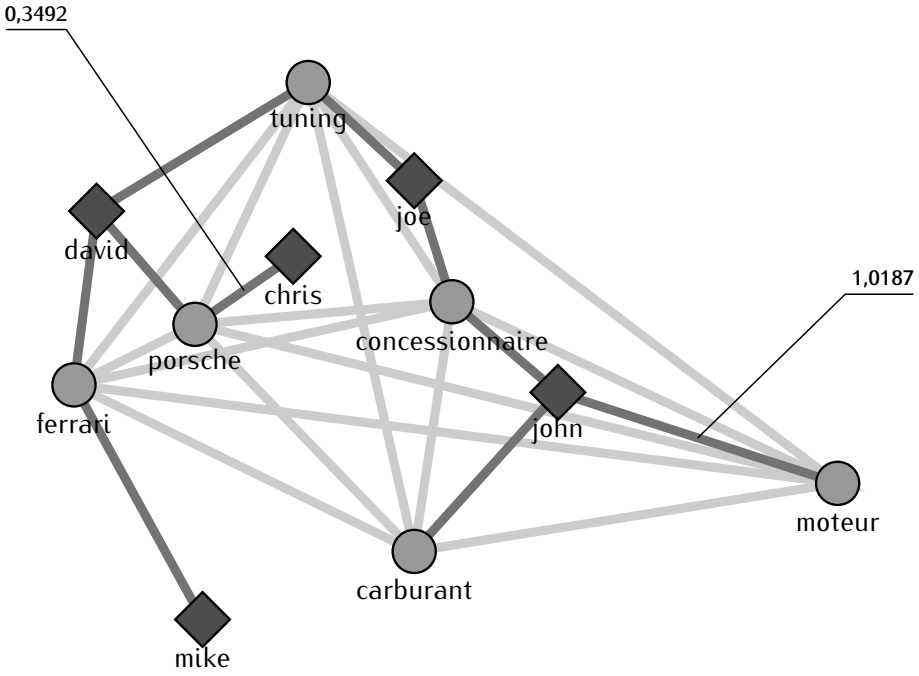


FIGURE 4.6 – Exemple de graphe de sujets après le positionnement des utilisateurs

ces derniers doivent être ajoutés à l'analyse. La formule de la distance sémantique entre un utilisateur et un sujet est inversible, nous n'avons pas besoin de stocker des vecteurs d'attractivité utilisateur (voir ci-dessus). Pour chaque nouvelle publication, il est juste nécessaire d'en extraire l'ensemble des attractivités par sujets concernés.

Pour chaque sujet t_i évalué avec une attractivité a, nous mettons à jour la distance sémantique entre l'utilisation u, auteur de la publication, et le sujet t_i comme suit :

$$\text{DIST}'(u, t_i) = \frac{1}{\log\left(10^{\frac{1}{\text{DIST}_{(u,t_i)}}} + a\right)}$$

Nouveau sujet Pour diverses raisons, telles qu'une modification de la politique du site, l'apparition de nouveaux comportements, etc., l'analyste Web peut avoir besoin d'ajouter de nouveaux sujets au lexique. Si un sujet est pertinent (voir plus haut), nous pouvons le situer dans le graphe comme précédemment. Après cela, nous devons évaluer la distance entre tous les utilisateurs et ce nouveau

sujet. Si l'ensemble des archives de publications est en mémoire ou accessible, nous pouvons calculer l'attractivité du nouveau sujet pour chaque publication, et définir une nouvelle distance sémantique, comme précédemment. Mais si ces archives ne sont pas accessibles, nous devons approximer les nouvelles distances. Comme nous connaissons les distances sémantiques entre les anciens sujets et le nouveau, nous évaluons la distance de chaque utilisateur par rapport au nouveau sujet comme la valeur du plus court chemin entre les deux, passant par un sujet préexistant (le plus court chemin ne peut nécessairement pas passer par deux sujets préexistants, le graphe des sujets étant complet et leurs distances nécessairement supérieures à un seuil strictement positif).

4.2.5 Conclusion

Dans cette approche, nous avons proposé une méthode pour regrouper les utilisateurs en fonction d'un vocabulaire fourni. Pour cela, nous définissons un ensemble de distances entre les utilisateurs et les termes du vocabulaire, basé sur les données des utilisateurs, et les fréquences de co-utilisation des termes sur le Web. Nous proposons également une méthode pour augmenter la concision du vocabulaire en supprimant les termes redondants. Dans la section suivante, nous présentons une approche pour produire des communautés d'utilisateurs en se basant sur des ontologies de référence externes.

4.3 Découverte basée sur une ontologie

Dans cette approche, nous présentons une méthode incrémentale et passant à l'échelle pour l'analyse sémantique des communications Web. Quelques précédents travaux considéraient l'analyse des communications en ligne [65], mais avec une approcher purement statistique, et sans une connaissance *a priori* du vocabulaire de la cible. À l'inverse, notre méthode s'appuie sur une ou plusieurs ontologies sélectionnées par le gestionnaire de communautés pour leurs adéquations avec le contenu qu'il a en gestion.

La contribution principale de cette approche est un modèle de profil qui tient

compte de la sémantique des messages au cours de la communication, et relie chaque utilisateur aux concepts de l'ontologie. Nous proposons une méthode évolutive pour résumer les contributions des utilisateurs par des généralisations sur les concepts en fonction de l'ontologie. Notre détection de concept est enrichie par une propagation de contexte, basée sur la structure des discussions du forum analysé.

Ce travail a été publié en janvier 2012 dans la revue *Web Intelligence and Agent Systems : An International Journal* [46].

4.3.1 Analyse sémantique des communications

Abstraction des communications en ligne

Comme lors de nos précédentes approches, notre contribution est applicable à toutes communications textuelles structurées où les utilisateurs sont identifiés, ce qui est le cas de la grande majorité des systèmes de communication sur le Web. Il est bien entendu nécessaire d'avoir accès à l'ensemble des communications du système, ce qui implique qu'elles soient publiques ou que le gestionnaire dispose d'un accès privilégié.

Exemple 4.7

> À titre d'exemple, nous considérons le gestionnaire de communautés d'une entreprise de produits pharmaceutiques et para-pharmaceutiques, et supposons qu'il surveille les forums des entreprises concurrentes pour identifier les communautés intéressantes et les utilisateurs importants. La figure 4.7 montre un exemple du genre de communications relatives à ce travail de veille. Nous supposons l'existence d'un robot d'indexation (*crawler*) destiné à surveiller ces forums. Ces derniers peuvent être découverts par une recherche classique sur le Web, ou plus spécifiquement en se concentrant sur les plates-formes les plus répandues qui les hébergent. Par exemple, la requête « *phpBB health* » est susceptible de retourner des résultats intéressants pour notre gestionnaire de communautés, et hébergés par le système de forums PhpBB [3], permettant d'analyser plusieurs sources avec un même robot.

3. http://www.phpbb.com

Alice	I feel a pain in my left arm.
Joey	Is there a physician on this WebSite ?
Bob	"I feel a pain in my left arm." Could you be more precise ?
Alice	@Bob : My shoulder hurts.
Bob	Did you perform a strong move ?
Alice	Yes, I played tennis yesterday.

FIGURE 4.7 – Messages sur un forum de discussion

Nous considérons un forum comme un ensemble P de publications produites par un ensemble U d'utilisateurs. Les utilisateurs sont identifiés de façon unique par un identifiant propre visible (comme une adresse mail, un UID, etc.). Nous définissons par $author(p) \in U$ l'auteur unique d'une publication p donnée, et $post(u) \in P$ l'ensemble des publications de l'utilisateur u.

Il existe plusieurs techniques d'annotations ou conventions d'écriture pour indiquer la réponse à un message donné. En ce qui concerne les mails et les tweets, l'utilisateur à qui est adressé la réponse est explicitement indiqué. Pour les messages purement textuels (comme le *chat*, le commentaire ou le message de forum), il existe des situations courantes et des pratiques tacites pour exprimer la notion de réponse. Un schéma classique consiste à commencer la réponse par le motif "@u", ce qui indique que le message est adressé à — ou en réponse à — u. Une autre convention est de citer par recopiage tout ou partie du message auquel la réponse s'adresse. Des exemples de ces motifs de réponse sont visibles dans la figure 4.7. Afin de conserver cette information, nous définissons $cite(p) \subseteq P$ l'ensemble des publications auxquelles la publication p répond.

Afin de définir aussi finement que possible l'axe d'analyse sémantique du fo-

rum, nous nous appuyons sur une ontologie, qui peut être de domaine ou générale. Son choix est néanmoins essentiel : une ontologie de domaine spécifique et détaillée devrait être choisie pour l'analyse des forums spécialisés, tandis que des ontologies plus générales pourraient se révéler plus adaptées pour des forums génériques ou pour toute exploration initiale sans *a priori* sur le contenu sémantique. Une ontologie spécialisée pourrait par exemple contenir les noms précis de produits de marques d'un domaine, avec leurs relations (exemple de contenu : un « iPhone 4 - 32Go » est une sorte de « iPhone » qui est un « SmartPhone »). Les ontologies génériques sont nombreuses : Wordnet [51], YAGO [68], DBpedia [10], pour n'en nommer que quelques-unes. Du point de vue du gestionnaire de communautés, le juste choix de l'ontologie est une problématique très intéressante en elle-même — et cruciale — mais cela n'est pas l'objet de nos travaux.

Exemple 4.8

> Pour notre exemple, l'ontologie choisie peut être une ontologie des informations médicales comme MESH [18], une ontologie anatomique comme FMA [61] ou une coupe thématique dans une ontologie générique. Notre exemple présente des communications en anglais étant donné qu'il s'agit de la langue de l'ontologie utilisée.

Plus formellement, soit une ontologie légère $(C, is-a)$, où $C = \{c_1, \ldots, c_n\}$ est un ensemble de concepts et $is-a$ est la relation directe de sous-concept structurant l'ontologie. La notation $is-a(c, c')$ signifie que c est un hyponyme direct de c'. L'ensemble des concepts de C qui sont manipulés par une publication p donnée, est noté $concept(p)$. Cet ensemble peut être calculé par racinisation [4] des termes de p et en supprimant ses mots vides [5], puis en les comparant avec les termes de l'ontologie, eux aussi racinisés. Nous notons par $occurrence_p(c) \in \{0, 1\}$ la présence

4. Transformation en radical par suppression des préfixes et suffixes (*stemming* en anglais).
5. L'ensemble des mots trop communs (*stop-words* en anglais).

d'un concept c dans une publication p.

Exemple 4.9

Selon l'ontologie de la figure 4.8, pour p le second message d'Alice dans la figure 4.7, nous avons :

$$occurrence_p(\text{"shoulder"}) = 1,$$

et aucun autre concept pertinent n'est détecté dans cette publication.

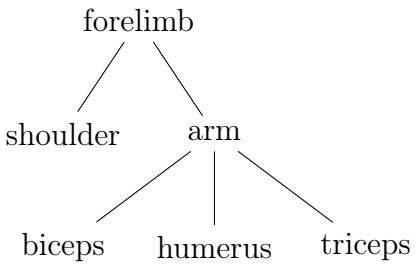


FIGURE 4.8 – Exemple partiel d'ontologie utilisée

4.3.2 Profils sémantiques et généralisation

Notre analyse sémantique d'un forum se décompose en deux étapes. Premièrement, nous associons à chaque utilisateur un profil sémantique, qui représente ses contributions sur le forum au travers de l'ontologie. Ces profils sémantiques prennent en compte les structures de discussions précédemment évoquées. Ensuite, nous agrégeons les contributions de l'utilisateur dans ce profil, afin de résumer son activité en quelques concepts.

Profils utilisateurs Cette méthode doit être vue comme une méthode incrémentale : dès qu'une nouvelle publication p d'un utilisateur u est détectée par le robot d'indexation, le système met à jour son profil. Le profil de l'utilisateur u pour un concept c, $profile_u(c)$ est pour une première définition le nombre total d'occurrences de c dans $post(u)$.

Mais cette première définition n'est pas satisfaisante : un message peut couvrir un champ sémantique plus large que les seuls mots qui le composent, en fonction de son contexte. Bien sûr, nous ne prétendons pas une compréhension globale de tout le contexte sémantique, mais nous considérons le contexte de discussion. Quand un utilisateur répond à un autre, il place en effet son message dans le contexte de la publication initiale.

Joe	I underwent several heart operations in the past, and I just moved to Manhattan. I seek the address of a good cardiologist.
Dave	@Joe : There is one good at 1546, 7th Avenue.

FIGURE 4.9 – Réponse avec contexte hérité

Exemple 4.10

> Nous considérons les deux publications p_1 et p_2 de la figure 4.9.
> Même si le second message ne contient pas explicitement de contenu sémantique en rapport avec le domaine analysé, nous pouvons propager le contenu sémantique du premier message, identifié comme étant le contexte du second.

Il est à noter que, pour assurer le passage à l'échelle, l'ensemble des publications ne peut être maintenu en mémoire. Il n'est donc pas possible d'évaluer la relation *cite* dans tous les cas possibles dans la pratique. Une fenêtre temporelle pertinente doit être choisie, à partir de la date du message analysé. Une réponse @u dans une publication p est alors interprétée comme une citation du dernier message de l'utilisateur u dans la fenêtre temporelle, ou éliminée si aucun message concordant n'est trouvé. De même, le texte d'un message en cours est comparé à tous les messages dans la même fenêtre de temps. Si une partie importante du texte se révèle être un extrait d'un message précédent dans la fenêtre de temps, il est considéré comme une citation de ce dernier. La figure 4.10 illustre ce fonctionnement.

Ainsi, nous utilisons la relation *cite* pour enrichir le profil sémantique. On définit alors $profile_u(c)$ comme la somme de toutes les occurrences du concept c

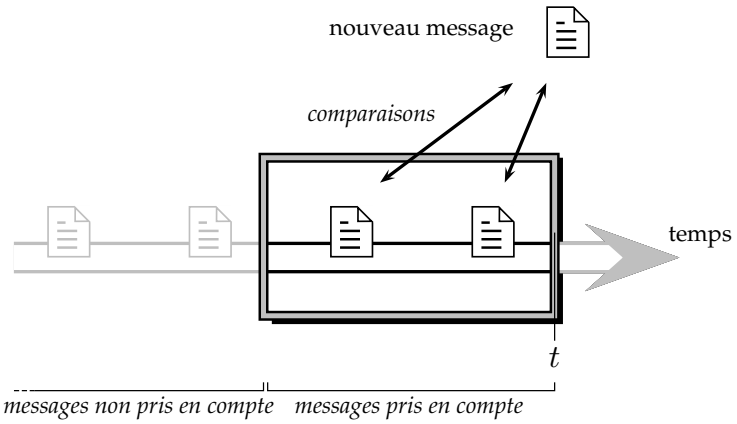


FIGURE 4.10 – Fenêtre temporelle de prise en compte des nouveaux messages

dans les messages de l'utilisateur u et dans les messages qu'il cite.

$$profile_u(c) = \sum_{p \in post(u)} \left(occurence_c(p) + \sum_{p' : p \in cite(p')} occurrence_c(p') \right)$$

Il est important de remarquer que nous avons choisi une définition non-récursive, afin d'évaluer les messages uniquement dans une fenêtre temporelle donnée, là encore pour des raisons de passage à l'échelle. Quand une nouvelle publication p d'un utilisateur u est analysée, les messages qu'il cite dans la fenêtre temporelle sont extraits, et $profile_u$ est incrémenté pour les concepts pertinents.

Exemple 4.11

> De part le message p_2 et sa citation implicite du message p_2 dans la figure 4.9, le profil sémantique de l'utilisateur « Dave » voit les poids de ses concepts *cœur* et *cardiologue* augmenter. Ces ajouts sont réalisés grâce à la prise en compte du contenu du message p_1 comme *contexte sémantique* du message p_2.

Sans la contrainte de passage à l'échelle, une définition récursive du contexte pourrait être envisagée : l'ordre temporel des messages garantit une récursivité sans boucle, étant donné qu'une publication du passé ne peut pas citer un message dans le futur (bien que certains systèmes de forums permettent la modification des messages, cette modification doit être considérée comme une nouvelle publication, à la date de modification).

Résumé des utilisateurs Les profils utilisateurs peuvent être enrichis progressivement au cours des futures contributions. Toutefois, le profil, qui décrit l'ensemble des activités, contient souvent des informations qui ne sont pas pertinentes, comme les concepts utilisés seulement dans de rares cas. Mais si ils ne sont pas pertinents aujourd'hui, nous ne pouvons pas les retirer pour autant, car ils peuvent obtenir un rôle modéré — ou mieux — plus tard. En effet, un utilisateur peut modifier progressivement son activité. Nous introduisons donc le résumé de l'utilisateur, $abstract_u$, le résumé actuel des intérêts sémantiques de l'utilisateur u. Pour un concept c, $abstract_u(c)$ est le poids du concept de c dans la vue résumée de u. Cette synthèse est composée de deux opérations distinctes :

- ajout des concepts couverts par généralisation ;
- suppression des concepts non-pertinents.

Généralisation La première étape, la généralisation, permet de mettre en évidence la couverture d'un concept par un utilisateur qui manipule ses sous-concepts. Pour un concept feuille c de l'ontologie, le résumé correspond au profil. Aucune généralisation n'intervient alors, à savoir :

$$abstract_u(c) = profile_u(c).$$

Pour les autres nœuds, nous considérons qu'un utilisateur u qui manipule une « part significative » des sous-concepts directs $c_1, \ldots, c_k$ d'un concept c, est considéré comme manipulant c. Le seuil de signification est matérialisé par $\delta_{coverage} \in [0, 1]$. Ensuite, si :

$$\frac{|\{c_i : is-a(c_i, c) \text{ and } abstract_u(c_i) > 0\}|}{|\{c_i : is-a(c_i, c)\}|} \geq \delta_{coverage},$$

le résumé de c est la moyenne des résumés de tous les sous-concepts de c :

$$abstract_u(c) = \frac{1}{|c' : is-a(c', c)|} \sum_{c' : is-a(c', c)} abstract_u(c') + profile_u(c).$$

Si la couverture des sous-concepts n'est pas suffisante, alors simplement :

$$abstract_u(c) = profile_u(c).$$

Suppression La deuxième étape de la construction de notre résumé supprime simplement les concepts dont le poids est inférieur à un seuil de poids minimum. Ce poids minimum est relatif à la somme des poids des contributions de l'utilisateur et définie par le seuil $\delta_{pertinence}$. De cette manière, un concept c est supprimé si :

$$\frac{abstract_u(c)}{\sum_{c' \in C} abstract_u(c')} < \delta_{pertinence}.$$

Il est à noter que nous commençons par généraliser les concepts, avant de supprimer ceux qui n'ont pas de pertinence. Cela permet de découvrir des concepts bien couverts, bien qu'ils ne soient pas explicitement utilisés ni que leurs sous-concepts pris individuellement ne soient pertinents pour l'utilisateur.

Exemple 4.12

Nous considérons une portion locale de l'ontologie et le $profile_u$ qui y est lié, comme visible dans la figure 4.11. Pour un seuil de couverture $\delta_{couverture}$ de 0.66 et un seuil de pertinence $\delta_{pertinence}$ de 0.5, le résumé $abstract_u$ en résultant est visible en figure 4.12.

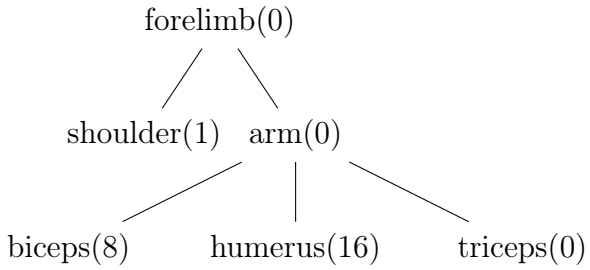


FIGURE 4.11 – Exemple de profil utilisateur

Comme le souligne l'exemple précédent, la généralisation impacte les poids sur des concepts supérieurs, sans retirer les poids précédents des sous-concepts. Ce faisant, la prise en compte d'une généralisation n'efface pas les spécialisations de l'utilisateur. La méthode basique pour calculer le résumé se déroule d'une manière

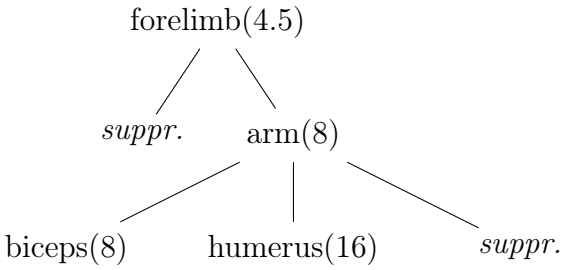


FIGURE 4.12 – Exemple de résumé utilisateur, avec $\delta_{coverage} = 0.66$ and $\delta_{relevance} = 0.5$

ascendante. L'algorithme 2 présente un calcul global sur l'ensemble de l'ontologie, alors que l'algorithme 3 est incrémental : seuls les concepts impactés sont testés et éventuellement modifiés. Cette méthode débute avec des $profile$ et $abstract$ à 0 pour tout utilisateur.

Algorithme 2 : Généralisation initiale du résumé utilisateur

entrées : $profile_u, ontologie, \delta_{coverage}, \delta_{relevance}$
sorties : $abstract_u$

```
1  début
2      abstract_u = profile_u
3      pour chaque profondeur d de l'ontologie, partant du bas faire
4          pour chaque concept c à la profondeur d faire
5              abstract_u(c) = profile_u(c)
6              si c n'est pas une feuille alors
7                  S = sous-concept de c
8                  SU = {c' ∈ S : abstract_u(c') ≠ 0}
9                  si |SU|/|S| > δ_couverture alors
10                     abstract_u(c) += (Σ_{c'∈S} abstract_u(c')) / |S|
11                 fin si
12                 si abstract_u(c) / (Σ_{c'∈C} abstract(c')) < δ_pertinence alors
13                     abstract_u(c) = 0
14                 fin si
15             fin si
16         fin pour
17     fin pour
18 fin
```

Algorithme 3 : Généralisation incrémentale du résumé utilisateur

entrées : $abstract_u, post, ontologie, \delta_{coverage}$
sorties : $abstract_u$

début
  pour chaque *profondeur d de l'ontologie, partant du bas* **faire**
    si *c n'est pas une feuille* **alors**
      $previous = 0$
      $S =$ sous-concepts de c
      $SU = \{c' \in S : abstract_u(c') \neq 0\}$
      si $\frac{|SU|}{|S|} > \delta_{coverage}$ **alors**
        // Sauvegarder les valeurs précédentes
        $previous = \frac{\Sigma_{c' \in S} abstract_u(c')}{|S|}$
      fin si
    fin si
    $abstract_u(c) += post(c)$
    si *c n'est pas une feuille* **alors**
      $SU = \{c' \in S : abstract_u(c') \neq 0\}$
      si $\frac{|SU|}{|S|} > \delta_{coverage}$ **alors**
        // Mettre à jour par ajout de la différence avec le précédent
        $abstract_u(c) += \frac{\Sigma_{c' \in S} abstract_u(c')}{|S|} - previous$
      fin si
    fin si
  fin pour
fin

Une troisième version de l'algorithme a été produite, récursive cette fois. L'algorithme 4 a l'avantage de ne traiter un concept que s'il est affecté par la modification d'un de ses sous-concepts. Toutefois, pour ces raisons de performances, nous utilisons pour les expérimentations présentées dans le chapitre 6 la version incrémentale. En effet cette version permet de neutraliser le surcoût grâce à l'analyse de l'arbre de l'ontologie, et d'en réaliser une abstraction relationnelle. La profondeur de chaque concept étant connue durant l'abstraction et la méthode incrémentale permet alors de « naviguer » dans l'arbre, sans le coût traditionnel de ce type d'opérations.

Algorithme 4 : Généralisation récursive du résumé utilisateur

entrées : $abstract_u, concept, value, ontologie, \delta_{couverture}$
sorties : $abstract_u$

début
  $valueurPere = 0$
  si *concept n'est pas la racine* **alors**
    $S =$ sous-concepts de $parent(concept)$
    $SU = \{c' \in S : abstract_u(c') \neq 0\}$
    si $\frac{|SU|}{|S|} > \delta_{coverage}$ **alors**
      $valueurPere = \frac{\Sigma_{c' \in S} abstract_u(c')}{|S|}$
    fin si
  fin si
  $abstract_u(c) += value$
  si *concept n'est pas la racine* **alors**
    $SU = \{c' \in S : abstract_u(c') \neq 0\}$
    si $\frac{|SU|}{|S|} > \delta_{coverage}$ **alors**
      $valueurPere = \frac{\Sigma_{c' \in S} abstract_u(c')}{|S|} - valueurPere$
    fin si
    si $valueurPere \neq 0$ **alors**
      récursivement Généralisation$(parent(concept), valueurPere)$
    fin si
  fin si
fin

Partitionnement en communautés Les calculs précédents permettent de déduire des communautés au sein du forum cible. Nous effectuons cette tâche en deux étapes :

- détection des principaux concepts couverts par les utilisateurs ;
- partitionnement des utilisateurs autour de ces concepts.

Concepts principaux Nous faisons la somme de tous les *abstracts* calculés, et y appliquons le même raffinement que pour un *abstract* utilisateur. Le « résumé global » en résultant peut être vu comme l'*abstract* de l'ensemble du forum, après généralisation de l'ensemble des contributions et application du seuil de pertinence $\delta_{relevance}$ pour ne concerner que les concepts majoritaires (Algorithme 5).

Algorithme 5 : Résumé Global du système

entrées : $abstracts, ontologie, \delta_{relevance}$
sorties : $globalAbstract(C)$

début
  pour chaque *concept* $c \in C$ **faire**
    $globalAbstract(c) = 0$
  fin pour
  pour chaque *user* $u \in U$, *concept* $c \in C$ **faire**
    si $\frac{abstract_u(c)}{\sum_{c' \in C} abstract(c')} \geq \delta_{relevance}$ **alors**
      $globalAbstract(c) += abstract_u(c)$;
    fin si
  fin pour
  pour chaque $c \in C$ **faire**
    si $\frac{globalAbstract(c)}{\sum_{c' \in C} globalAbstract(c')} < \delta_{relevance}$ **alors**
      $globalAbstract(c) = 0$
    fin si
  fin pour
fin

Communautés Nous relions à présent concept majeur identifié à ses contributeurs principaux (Algorithme 6).

Algorithme 6 : Construction des communautés

entrées : $abstracts, globalAbstract, \delta_{relevance}$
sorties : $Communities$

début
  $Communities = \emptyset$
  pour chaque *user* $u \in U$ **faire**
    pour chaque $c \in C$ *s.t.* $globalAbstract(c) > 0$ **faire**
      si $abstracts_u(c) > 0$ **alors**
        si $\frac{globalAbstract(c)}{\sum_{c' \in C} globalAbstract(c')} \geq \delta_{relevance}$ **alors**
          $Communities(c) += u$
        fin si
      fin si
    fin pour
  fin pour
fin

Analyse sociale En exploitant les étapes précédentes, on peut obtenir divers renseignements sur le comportement des utilisateurs et leurs activités sociales :

- principaux centres d'intérêt : les intérêts principaux d'un utilisateur u sont simplement la liste des concepts présents dans $abstract_u$, triés par poids décroissant ;
- utilisateurs principaux d'une communauté : Les « *top-users* » d'une communauté représentent le groupe d'utilisateur qui en sont le centre, ceux qui fournissent abondamment le forum en contenus les plus proches possible du sujet de la communauté. En d'autres termes, les utilisateurs u qui ont les plus hautes valeurs de $abstract_u$ pour les concepts de la communauté.

Ces informations sont illustrées dans nos expérimentations ci-dessous.

4.3.3 Expérimentations

Jeu de données

Afin d'expérimenter notre approche, nous utilisons comme source de données le site Web « *USA Today* », version Web du journal le plus diffusé des États-Unis, et plus précisément sur sa section « *Health News* » regroupant les actualités liées au domaine de la santé. Comme chaque actualité du site, les publications de cette section sont ouvertes aux commentaires des utilisateurs. Chaque utilisateur est identifié de façon unique, et peut utiliser des techniques de citation et de réponse aux commentaires des autres utilisateurs. À partir de celà, nous réalisons une étude spécialisée sur ce domaine.

Nous avons développé un robot d'indexation en ligne (*web crawler*) spécialisé ainsi qu'adaptateur (*wrapper*) incluant des analyseurs syntaxiques (*parsers*) HTLM et JSON. Nous avons extrait de la sorte environ 15 000 commentaires utilisateurs. Toutes ces contributions sont signées par leurs auteurs, par le biais de leur authentification. Nous normalisons ces publications et les représentons sous forme de fichiers XML standardisés, dont la structure est déclarée dans la DTD fournie en Annexe B. Le flux des contributions est traité au fil de l'eau.

L'analyse statistique des données collectées, résumées dans la table 4.4, montre que plus de la moitié des utilisateurs du forum sont seulement identifiés par la pu-

Nombres d'articles collectés	200
Nombre de commentaires collectés	14978
Nombre d'utilisateurs détectés	3682
Commentaires par utilisateur (minimum)	1
Commentaires par utilisateur (moyenne)	4
Commentaires par utilisateur (maximum)	447
Nombre d'utilisateurs à commentaire unique	1848
Commentaires par article (minimum)	0
Commentaires par article (moyenne)	75
Commentaires par article (maximum)	1642

TABLE 4.4 – Mesures des articles collectés sur USA Today

blication d'un unique commentaire. Inversement, comme illustré par la figure 4.13, une minorité des utilisateurs est responsable de la majorité des contributions.

Ces caractéristiques sont typiques d'un forum ouvert, où chacun peut participer sans être personnellement impliqué dans le forum (comportement du « passager clandestin »). L'investissement et la participation de la majorité des utilisateurs sont faibles. En conséquence, un large volume de données est nécessaire pour obtenir des déductions valides et pertinentes relatives aux communautés, la grande majorité des contributions n'étant pas significative. En revanche, dans un forum plus dense — comme une plate-forme collaborative de développement logiciel — les membres se sentent fortement impliqués ; le nombre de contributions requis serait alors plus bas.

Détection de concepts

Pour analyser ces discussions et en accord avec notre modèle, nous utilisons une ontologie. Toutefois, les ontologies médicales communes, comme MESH [18] or FMA [61] ne nous apparaissent pas adaptées dans ce cas précis. En effet, elles utilisent des termes précis et spécialisés : ce niveau de vocabulaire est rarement utilisé par le grand public. Pour surmonter ce problème, nous réalisons une coupe thématique dans WordNet [51], une base de données lexicale de référence pour la

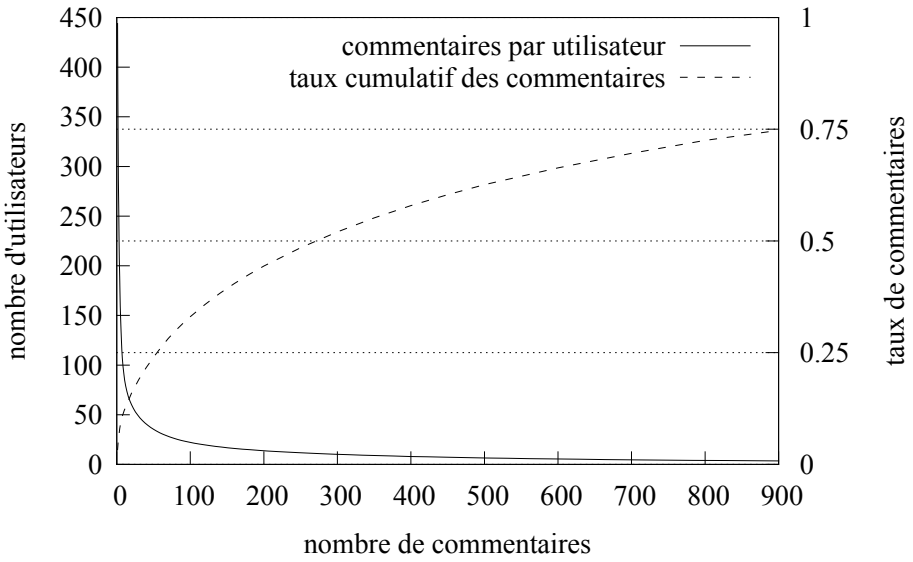


FIGURE 4.13 – Distribution des messages par utilisateur

langue anglaise. Nous prenons comme concept racine le *synset*[6] *body part*, ce qui nous fournit une ontologie de termes utilisés pour décrire les différents éléments du corps humain.

Concepts disponibles	1824
Détections de concepts	55875
Nombre de concepts moyen par message	3.73
utilisateurs sans concepts détectés	125
Taux d'utilisateur sans concepts	3.4%
Concepts non couverts	994
Taux de concepts non couverts	6.64%

TABLE 4.5 – Mesures sémantiques

Nous appliquons alors notre méthode de détection de concepts précédemment décrite. Comme visible dans la table 4.5, les commentaires utilisateurs sont souvent pauvres en contenu sémantique (nous abordons plus loin comment améliorer cette détection de concept), mais pratiquement tous les utilisateurs peuvent être rattachés à un concept de notre ontologie.

6. Le *synset* (contraction de *synonym set*), est la composante atomique de WordNet. Il représente un groupe de mots interchangeables, dénotant un sens ou un usage particulier.

Communautés

En appliquant l'algorithme présenté plus haut, nous construisons des profils utilisateurs, obtenons des résumés sémantiques et les regroupons en communautés. Une communauté est caractérisée par son nombre d'utilisateurs et son poids sémantique (voir les tables 4.6, 4.7 et 4.8, commentées ci-dessous).

Prendre en compte le contexte Nous avons effectué trois calculs différents, avec trois définitions distinctes de la notion de contexte. Sur la première, nous avons calculé les communautés en prenant en compte la propagation du contexte du sujet (méthode « TC » pour *Thead Context*) : les messages sont contextualisés selon le fil auquel ils répondent. Dans notre cas, il s'agit du contenu sémantique de l'actualité sur laquelle se situe le fil de commentaires. La figure 4.14 illustre le fait que prendre en compte le contexte initial permet d'augmenter significativement la détection de concepts. Les communautés résultant de ce calcul sont listées dans la table 4.6.

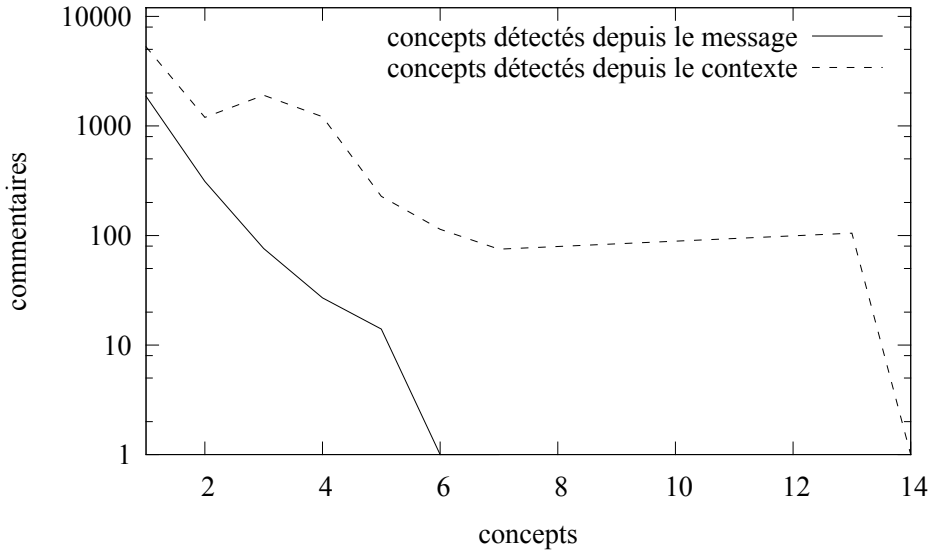


FIGURE 4.14 – Détection de concepts en fonction de la source

Pour le second calcul, nous prenons en compte uniquement le contexte local des réponses directes (méthode « AC » pour *Anwser Context*). Chaque message prend

position	concept principal	utilisateurs	poids
1	Heart	338	5098
2	Brain	159	2227
3	Lung	153	1952
4	Heart	23	1104
5	Belly	93	982
6	Neck	110	917
7	Skin	131	913
8	Large inestine	25	168
9	Liver	21	167
10	Heel	3	167

TABLE 4.6 – Communautés détectées (méthode TC)

en contexte sémantique le message auquel il répond. Les communautés résultant de ce calcul sont listées dans la table 4.7.

position	concept principal	utilisateurs	poids
1	Belly	225	1677
2	Heart	172	653
3	Lung	121	541
4	Brain	117	413
5	Side	123	395
6	Heel	27	263
7	Liver	48	155
8	Skin	47	139

TABLE 4.7 – Communautés détectées (méthode AC)

Pour le troisième calcul, nous utilisons les deux contextes simultanément (méthode « T&AC » pour *Thread and Anwser Contexts*). Les communautés résultant de ce calcul sont listées dans la table 4.8.

L'influence de ces différents contextes est liée à la structure du système analysé. Dans notre exemple, nous travaillons sur un système de commentaires sur les actualités publiées. Nous constatons que les actualités ont un volume d'informations plus important que les commentaires fréquents usuellement courts laissés par les utilisateurs. De plus, les utilisateurs répondent habituellement d'avantage aux articles initiaux qu'aux autres commentaires. C'est la raison pour laquelle, dans

position	concept principal	utilisateurs	poids
1	Heart	306	5192
2	Brain	167	2368
3	Lung	145	2081
4	Belly	112	1582
5	Neck	111	940
6	Skin	115	877
7	Side	64	699
8	Heel	5	213
9	Liver	21	195
10	Large intestine	25	175
11	Eye	18	151
12	Knee	15	123
13	Hand	13	112

TABLE 4.8 – Communautés détectées (méthode T&AC)

le cas d'un site d'actualités, l'influence du contexte de fil (méthode TC) est dominant, tandis que le contexte de réponse (méthode AC) n'est qu'une faible source d'informations supplémentaires.

En conséquence, les communautés calculées uniquement à partir de la méthode AC ont un poids sémantique affiché relativement faible. Les communautés calculées à partir de la méthode TC sont plus robustes, et la méthode T&AC peut être vue comme une amélioration, mais faiblement significative. Mais dans un système totalement différent — comme un forum de discussion où les réponses sont souvent plus importantes que les sujets initiateurs — ce comportement serait probablement inversé, avec une importance prépondérante de la méthode AC sur le méthode TC. Leur utilisation conjointe permet de lisser ces disparités.

Utilisateurs principaux À partir de la plus grande communauté détectée, relative au concept *Heart*, nous recherchons ses utilisateurs principaux. La table 4.9 présente les résultats de ce calcul (les noms d'utilisateurs ont été anonymisés).

Les utilisateurs principaux sont classés en fonction du poids qu'ils occupent dans la communauté, c'est-à-dire en fonction de leur poids sémantique propre dans le domaine de la communauté. Cependant, nous pouvons distinguer deux profils

position	**utilisateur**	**part de l'utilisateur dans**	
		la commun.	**le site**
1	li	28,12 ‰	**37,61** ‰
2	gre	27,54 ‰	**35,99** ‰
3	xiu	**20,22** ‰	16,25 ‰
4	BAL	**18,68** ‰	12,63 ‰
5	uknowi	**15,60** ‰	11,98 ‰
6	popo	**14,83** ‰	4,35 ‰
7	brokena	**14,06** ‰	3,41 ‰
8	zoila	**13,29** ‰	3,37 ‰

TABLE 4.9 – Utilisateurs principaux de la communauté *Heart*

de membres de la communauté :

- les principaux contributeurs du site : à moins qu'ils négligent le domaine précis de la communauté, les principaux contributeurs de l'ensemble du site doivent nécessairement afficher un score élevé dans cette communauté. Leur rang est une conséquence logique de leur forte implication dans l'ensemble du système. Cela se ressent principalement dans un système ouvert avec des proportions inégales de contribution, comme expliqué précédemment ;
- les contributeurs dédiés : leur implication dans le sujet de la communauté est plus importante que dans le reste de l'ensemble du système. Les contributeurs dédiés sont principalement axés sur le domaine en question de la communauté et n'ont pas nécessairement besoin d'un grand nombre de contributions à figurer en bonne place dans la communauté. Plus le système se développe et devient plus dense, plus ce type d'utilisateur a tendance à se retrouver dans les communautés.

Dans notre exemple analysé, la densité moyenne du système fait apparaître les contributeurs dédiés légèrement après les principaux utilisateurs du site, comme illustré par la table 4.9.

Principaux sujets utilisateurs Comme présenté dans le modèle, la détection des principaux sujets abordés par un utilisateur est une simple coupe réalisée dans le résumé de son profil. La table 4.10 en est un exemple.

position	concept	part
1	Heart	26.25%
2	Lung	9.75%
3	Side	7.00%
4	Skin	3.75%
5	Liver	3.00%
6	Hand	1.75%

TABLE 4.10 – Sujets principaux de l'utilisateur *xiu*

Ordre de grandeur

Nous avons cherché à savoir quel est l'ordre de grandeur du nombre d'utilisateurs que nous pourrions gérer grâce à cette méthode. Nous prenons comme exemple un serveur aux capacités relativement courantes de nos jours, avec 10 Gio de RAM. Nous considérons une ontologie de 1024 termes et le stockage d'entiers à deux octets non signés.

D'une part, nous calculons l'espace occupé par les données utilisateurs — profils et résumés — et d'autre part, par les données de traitement, comme le message en cours d'analyse, les contextes à propager, etc. La taille des données des utilisateurs apparaît comme dépendante du nombre moyen de concepts que les utilisateurs manipulent. Nous considérons deux cas : le pire des cas possibles, où tous les utilisateurs manipulent toujours la totalité des concepts de l'ontologie, et le cas moyen tel qu'il est observé sur *USA Today*.

Sur la base de ces hypothèses, le pire cas possible nous permet de maintenir autour de 268 millions de comptes utilisateurs. Avec les moyennes d'utilisation d'*USA Today*, ce nombre s'élève à 11 milliards de comptes utilisateurs. Cela est dû à la faible couverture de la majorité des utilisateurs, typique d'un système très ouvert. Nous déduisons également de ces chiffres que, quel que soit le système analysé, la difficulté technique devrait donc toujours se situer au niveau du temps de calcul par rapport à la vitesse de publications des utilisateurs, et non au niveau du nombre de ces utilisateurs.

Impact de l'ontologie

Afin d'analyser l'impact du choix de l'ontologie, nous avons également testé notre ensemble de données avec une base de connaissances plus petite. Comme expliqué précédemment, nous voulons éviter le problème de la spécialisation du langage. Donc, nous avons construit une esquisse d'ontologie descriptive du corps humain que nous appelons *HumanAnatomyBasics*[7], basée sur la description du corps de Wikipédia. Elle représente une première approche utilisant des noms communs de la langue de tous les jours, y compris les parties du corps, les muscles et les os. Il est intéressant de noter qu'avec une telle ontologie — beaucoup plus petite que la coupe réalisée sur WordNet — les résultats sont globalement similaires. Le taux de détection est beaucoup plus faible, mais l'aspect général des collectivités est similaire. Cela confirme que la densité de l'ontologie permet d'affiner les résultats, mais que la principale contribution réside dans l'opération de généralisation. En effet, les relations entre les concepts peuvent préserver la cohérence sémantique, quel que soit le niveau de précision.

Inversement, l'utilisation de l'ensemble de WordNet, en plus d'une dégradation de la performance, présente une dispersion des concepts. Nous identifions les mots les plus utilisés dans la langue anglaise, mais ceux-ci ne sont pas conformes à l'objet de l'étude dirigée sur le forum : les communautés perdent alors de leur intérêt. Cela confirme notre besoin d'une relation forte entre le système étudié et l'ontologie choisie.

4.3.4 Conclusion

Nous avons présenté ici une méthode d'analyse pour extraire les communautés implicites à partir d'un système de communication donné, et illustré l'importance de la propagation de contexte pour comprendre la sémantique de communication. Nous utilisons pour ce faire une ontologie. Cette ontologie nous sert de base de référence sémantique pour analyser les échanges. Nous mettons en lumière qu'obtenir des résultats pertinents est conditionné au fait qu'il existe une adéquation thématique entre le système analysé et l'ontologie choisie. Ces résultats permettent au

7. http://www.damien-leprovost.fr/webtribe/HumanAnatomyBasics.owl

gestionnaire de communautés de suivre les sujets abordés, les profils utilisateurs, etc. Il peut ensuite par exemple décider de diviser les fils de discussion selon les communautés identifiées, de communiquer directement avec les utilisateurs principaux pour améliorer les rétro-actions, et ainsi de suite.

4.4 Conclusion générale

Dans ce chapitre, nous avons présenté nos approches et contributions en matière de découverte de communautés. Nous regroupons des utilisateurs à partir du contenu de ce qu'ils échangent au sein de leurs communications. Nos approches se sont concentrées successivement sur :

- la compréhension des utilisateurs par la recherche de similarités et corrélations entre les éléments qu'ils manipulent. Nous raisonnons pour cela sur les *tags* liés aux ressources manipulées par les utilisateurs ;
- le positionnement d'utilisateurs par rapport à un vocabulaire de référence. Nous travaillons sur la recherche de la pertinence optimale du vocabulaire par la suppression des termes redondants ;
- le rapprochement d'utilisateurs en fonction de la taxonomie des concepts issue d'une ontologie. Nous mettons en avant la prise en compte du contexte de la communication pour maximiser les informations acquises lors de son analyse.

Quelle que soit la source de référence sémantique utilisée — un vocabulaire ou une ontologie — nous soulignons la dépendance qu'il existe entre le choix de cette source et la pertinence des résultats de l'analyse des échanges entre les utilisateurs. Le degré de finesse et de prise en compte des éléments non-explicites de la communication joue également un rôle fondamental dans la compréhension de ces échanges, et donc des partitionnements en communautés qui en découlent.

Dans le chapitre suivant, nous nous intéressons aux analyses qui peuvent être menées sur les utilisateurs au sein des communautés ainsi identifiées.

Chapitre 5

Analyse de communautés

Sommaire

Dans ce chapitre, nous présentons nos différentes approches pour l'analyse des communautés telles que nous les détectons dans le précédent chapitre. Notre objectif est d'extraire des informations utiles et pertinentes pour comprendre et caractériser les communautés précédemment définies. Lors de ces approches, nous nous appuyons tout d'abord sur une méthode de détection de communautés telles qu'exposées au chapitre précédent. Ces thèmes ont été identifiés dans les messages des utilisateurs à l'aide d'une ontologie. Nous décrivons tout d'abord notre proposition de détermination des utilisateurs centraux au sein d'une communauté. Se basant sur la définition d'une nouvelle métrique appelée *centralité sémantique temporelle*, cette approche est détaillée en section 5.1. Nous présentons en section 5.2 notre proposition de caractérisation des rôles des utilisateurs sur la dynamique des communautés. Enfin, la section 5.3 conclut le chapitre.

5.1 Centralité sémantique et temporelle

Dans cette approche, nous présentons une méthode pour la découverte des utilisateurs centraux qui jouent un rôle important dans le flux de communication de chaque communauté. Un utilisateur central est un utilisateur dont l'attractivité est forte. Il y a de très nombreuses façon de considérer la centralité [26], mais toutes devraient s'accorder sur le fait que notre utilisateur central est majoritairement présent dans le graphe de communication de la communauté (un graphe complet n'a pas de centre, alors qu'un graphe en croix a le centre qu'on imagine). Nous proposons une nouvelle mesure, la probabilité de propagation sémantique (abrégée *SPP* pour « *Semantic Propagation Probability* »), qui caractérise la capacité de l'utilisateur à propager un concept sémantique à d'autres utilisateurs, d'une manière rapide et ciblée. La sémantique des messages est analysée selon une ontologie donnée. Nous utilisons cette mesure pour obtenir la centralité sémantique temporelle (abrégée *TSC* pour « *Temporal Semantic Centrality* ») d'un utilisateur dans une communauté. Nous proposons et évaluons cette mesure, en utilisant une ontologie et des données réelles issues du Web.

Ce travail a été présenté en version courte dans la conférence *International Conference on Web Engineering (ICWE 2012)*, à Berlin, Allemagne [45]. La version

étendue a été présentée le 25 octobre 2012 dans la conférence *Bases de Données Avancées (BDA)*, à Clermont-Ferrand, France.

5.1.1 Initialisation des communautés

Vue d'ensemble

Nous basons notre raisonnement sur une ontologie légère $O = (C, is-a)$, où C est un ensemble de concepts et $is-a$ la relation de subsomption. Nous équipons C avec une mesure de similarité sémantique $d_C(c, c')$ entre deux concepts c and c' de C.

Nous considérons un réseau de communication $G = (U, S)$, où U est un ensemble d'utilisateurs et $S \subseteq U \times U \times \mathbb{N}$ est la relation directionnelle et temporelle *envoi* d'un message $m = (u, v, t)$, d'un utilisateur u à un utilisateur v à l'instant t. Nous prenons les entiers naturels $\mathbb{N}$ comme horloge pour des raisons de simplicité. Les messages parfaitement simultanés sont possibles dans ce modèle, et leurs occurrences sont prises en compte. Par ailleurs, en présence d'un très fort trafic, beaucoup de messages sont susceptibles d'être simultanés, quelle que soit la précision temporelle choisie [1]. Ce modèle simple suppose que l'émetteur et le récepteur d'un message donné soient connus. Bien que réaliste pour les réseaux de communication basés sur la messagerie électronique, son applicabilité à d'autres réseaux de communication, comme les forums est discutée en fin de section. La fonction *content* associe un message $m = (u, v, t)$ à son contenu textuel $content(m)$. Afin de se concentrer sur les concepts de C, la fonction $content_C$ associe m à l'ensemble de concepts de C qui apparaît dans $content(m)$. Cette fonction peut être réalisée par le biais d'une racinisation comme évoquée au précédent chapitre.

Le but de cette approche est d'identifier les utilisateurs centraux agissant sur les sujets principaux du réseau de communication. Nous analysons ces réseaux en utilisant des robots d'indexation et adaptateurs spécifiques, puis nous extrayons les concepts de messages des utilisateurs selon l'ontologie prédéfinie. Nous commençons par identifier les thèmes principaux du réseau de communication. Ensuite,

1. Par exemple, Twitter connaît un débit moyen de 3 000 tweets par seconde, avec des pointes à 25 000 tweets par seconde.

nous identifions les communautés sémantiques associées, et enfin nous appliquons la méthode de centralité sémantique temporelle proposée pour identifier les utilisateurs centraux de ces communautés. La figure 5.1 présente une vue globale de cette méthode.

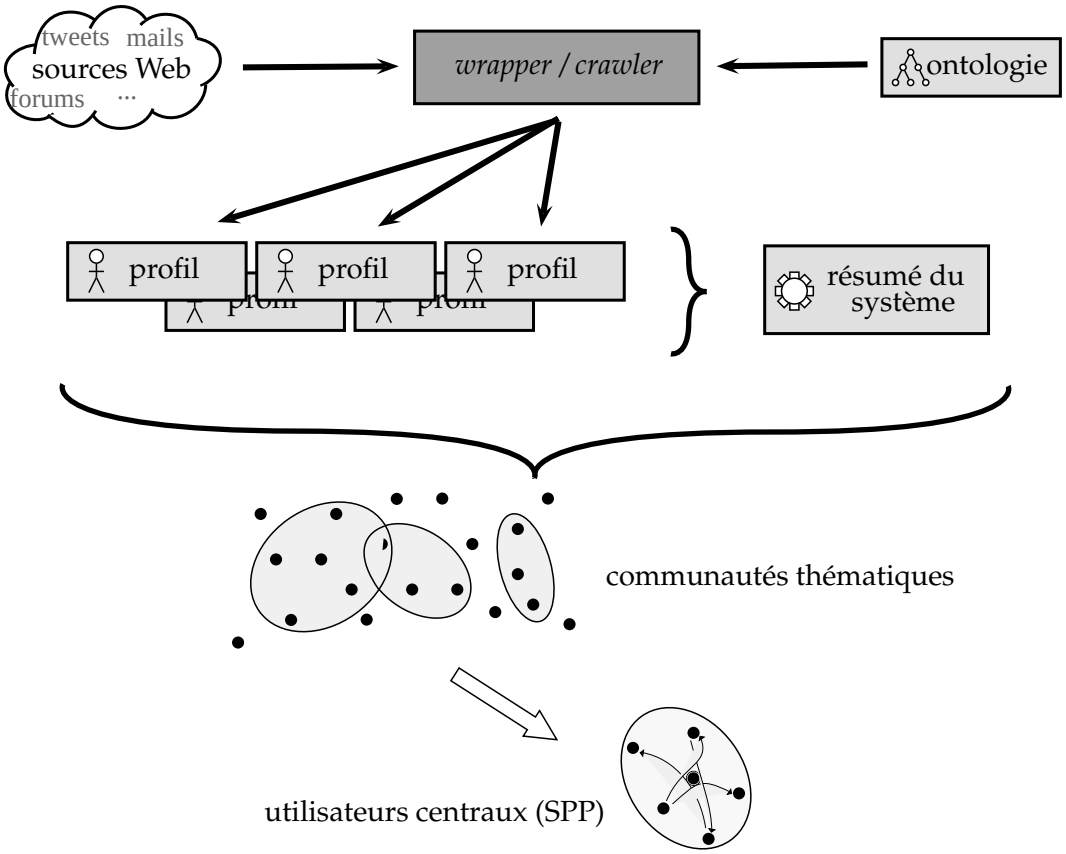


FIGURE 5.1 – Vue générale de la méthode de centralité sémantique temporelle

Identification des sujets principaux

La première étape de notre méthode est de déterminer les sujets principaux du réseau de communication. Comme précédemment, l'ensemble de ces sujets est vu comme un sous-ensemble des concepts de l'ontologie O utilisée. Nous nous appuyons sur la méthode de découverte de communautés via des ontologies précédemment évoquée au chapitre 4, pour construire des profils utilisateurs et définir

les concepts recouvrant le mieux ces profils comme étant les sujets principaux du système.

Distances sémantiques

Une fois les sujets principaux identifiés, notre but est de diviser le réseau de communication G en k communautés sémantiques $G_1 \ldots, G_k$, où chaque communauté G_i est labellisée avec un ensemble de concepts $L_i \subseteq C$. Nous allons filtrer les utilisateurs en fonction de leurs profils sémantiques. Ces profils englobent déjà une déduction sémantique grâce à l'ajout des concepts plus généraux correctement couverts, comme également décrit au chapitre 4. Afin de contrôler le nombre de communautés sémantiques, nous regroupons les utilisateurs en fonction de leurs concepts communs ou similaires. La similarité entre deux concepts d'une ontologie O est mesurée en utilisant une distance sémantique. Il existe de nombreuses implémentations de ce type de distance (voir chapitre 2). Nous utilisons ici la distance de Wu et Palmer [73], limitée à l'arbre hiérarchique $is-a$ des concepts, suivant une méthode déjà appliquée à des cas similaires [17]. Nous utilisons cette méthode afin de valoriser la proximité des concepts spécialisés sur celle des concepts généraux. La similarité est définie en fonction de la distance structurelle séparant les deux concepts dans la hiérarchie de l'ontologie, mais en prenant également en compte leur profondeur relative par rapport à la racine de l'ontologie. La similarité sémantique entre deux concepts c_1 et c_2 est donc

$$sim_{Wu\&Palmer}(c_1, c_2) = \frac{2 * depth(c)}{depth(c_1) + depth(c_2)} ,$$

où c est le plus proche ancêtre commun de c_1 et c_2, et $depth(x)$ le nombre de nœuds entre x et la racine. La similarité maximale, si $c_1 = c_2$ est donc de 1 et tend vers 0 avec l'éloignement des concepts. Deux concepts c_1 et c_2 sont considérés comme similaires si $d_C(c_1, c_2) \leq \delta$, où δ est un seuil de similarité.

$$d_C(c_1, c_2) = 1 - sim_{Wu\&Palmer}(c_1, c_2).$$

Construction des communautés sémantiques

Nous passons ensuite à la construction des communautés sémantiques. Nous définissons $N_i^+(G_i)$ comme étant le degré entrant de la communauté G_i, ce qui correspond au nombre de communications des membres de G_i en direction des membres de G_i qui ont pour profil des concepts de L_i (ou des concepts similaires à ces derniers, en accord avec le seuil δ précédemment défini). À l'inverse, nous définissons $N_i^-(G_i)$ comme étant le degré sortant de G_i, ce qui correspond au nombre de communications des membres de G_i en direction des membres extérieurs à G_i qui ont pour profil des concepts de L_i (ou similaires). Nous pouvons maintenant définir une communauté sémantique :

Définition 1 *Un ensemble $G_i \subseteq G$ est une communauté sémantique sur les concepts $L_i \subseteq C$ si, en limitant G_i aux messages qui contiennent un concept similaire à un concept dans L_i, le degré entrant de G_i est plus grand que son degré sortant, c'est-à-dire, $N^+(G_i) > N^-(G_i)$.*

Les approches traditionnelles proposées par Flake et al. [23] et ses diverses optimisations [36, 37, 49, 19] nous permettent de regrouper de manière efficace au sein de communautés les utilisateurs liés par une relation binaire. Nous prenons exemple sur cette approche pour définir une méthode de coupe, en appliquant le partitionnement entre utilisateurs aux messages remplissant une condition sémantique. Cette simplification donne la définition 1.

Pour chaque communauté G_i, nous maintenons pour chaque utilisateur u, deux ensembles de messages $N_i^+(u)$ et $N_i^-(u)$, représentant respectivement les communications de u au sein de G_i et les communications vers l'extérieur de G_i, selon des concepts similaires à L_i. Un message m_k est initialement considéré par défaut dans $N_i^-(u)$. Chaque message m_k destiné à un utilisateur u est considéré initialement comme non traité par son destinataire [2]. En conséquence, nous ajoutons le message à $N_i^-(u)$. Ensuite, si un message m_l est émis par u, avec $d_C(m_l, m_k) \leq \delta$, m_k est retiré de $N_i^-(u)$ et ajouté à $N_i^+(u)$.

2. Jusqu'à preuve du contraire, il n'y a pas répondu.

À tout moment, les communautés sont $G_i = (U_i, S_i)$, où

$$U_i = u \in U, N_i^+(u) \leq N_i^-(u)$$

et

$$S_i \subseteq U_i \times U \times \mathbb{N} .$$

Les algorithmes 7 et 8 présentent ce partitionnement en communautés.

Algorithme 7 : Classification des messages

entrées : message m, concepts $L_1, \ldots, L_i, \ldots, L_k$, seuil δ

```
début
    pour chaque i ... k faire
        pour chaque c ∈ L_i, c ∈ context(m) faire
            si m est entrant alors
                N_i^-(u) = N_i^-(u) ∪ m
            sinon
                // m est sortant
                pour chaque m_λ pour u avec d(m, m_λ) ≤ δ faire
                    N_i^+(u) = N_i^+(u) ∪ m ∪ m_λ
                    N_i^-(u) = N_i^-(u) − m
                fin pour
            fin si
        fin pour
    fin pour
fin
```

Une fois ces communautés définies, nous pouvons rechercher les utilisateurs qui en sont les centres.

5.1.2 Centralité, dispersion sémantique et temps de latence

Motivation

A l'intérieur d'une communauté sémantique étiquetée par les concepts L_i, tous les utilisateurs sont reconnus comme discutant fréquemment des sujets de L_i ou de sujets similaires. Nous souhaitons classer ces utilisateurs en fonction de leur cen-

Algorithme 8 : Définition des communautés

Entrées : $G = (U, S)$, $L_1, \ldots, L_i, \ldots, L_k$

```
1 début
2     pour chaque G_i ∈ G faire
3         pour chaque u ∈ U faire
4             si N_i^+(u) ≤ N_i^-(u) alors
5                 U_i = U_i ∪ u
6             fin si
7         fin pour
8     fin pour
9 fin
```

tralité, c'est-à-dire identifier les participants pourvoyeurs d'informations les plus importantes à l'intérieur de la communauté. Dans cette proposition, nous basons notre classement à la fois sur la sémantique et la temporalité du message. Nous définissons la *centralité sémantique temporelle*, abrégée TSC pour *Temporal Semantic Centrality*. Pour cela, nous définissons et utilisons une mesure basée sur les concepts que nous appelons *probabilité de propagation sémantique*, abrégée SPP pour *Semantic Propagation Probability*. Globalement, cette mesure vise à capturer :

- combien les réponses d'un utilisateur sont ciblées sémantiquement par rapport à un message entrant ;
- combien ces réponses sont rapides, relativement au rythme général de la communauté.

Les utilisateurs avec un SPP élevé sont davantage susceptibles de répondre ou de transmettre des messages au reste de la communauté, sémantiquement pertinents et dans un temps raisonnable.

Prenons une communication orientée

$$u \rightarrow_t u' \rightarrow_{t'} u'' \ ,$$

qui signifie qu'il existe dans le graphe de communication G un message $m = (u, u', t)$ de u vers u' au temps t, et un message $m' = (u', u'', t')$ de u' vers u'' au temps t'. Pour $t' > t$, m' peut être vu comme un relais de m dans un sens très large. Globalement, l'utilisateur u' est affecté — de diverses manières — par la

réception de m avant l'envoi m'. En outre, le contenu de m' peut être lié à m ou complètement indépendant de celui-ci. Nous allons mesurer cette relation afin qu'elle dépende de la *dispersion sémantique* du message envoyé, et son *temps de latence*.

Dispersion sémantique

Nous notons $dispersion_c(m)$ la dispersion sémantique d'un message m par rapport à un concept c. Il s'agit du ratio entre la distance sémantique minimale entre c et les concepts de m, et la distance sémantique maximale entre c et les concepts de l'ontologie :

$$dispersion_c(m) = \frac{\min_{c' \in content(m)} d_C(c, c')}{\max_{c' \in C} d_C(c, c')}.$$

Si le message utilise le concept c ($c \in content(m)$) alors $dispersion_c(m) = 0$. Il est à noter que la dispersion maximale est de 1, si aucune similarité n'est relevée. Pour le cas spécial où aucun concept n'est détecté dans le message, et où donc $content(m)$ est vide, nous considérons que $dispersion_c(m) = 1$.

Temps de latence

De façon similaire, nous définissons le temps de latence entre un message reçu par u_i au temps t_{i-1} et un message envoyé par u_i au temps t_i comme la durée les séparant, relativement au rythme naturel moyen de la communauté. En effet, certaines communautés dédiées à l'actualité ou professionnelles supposent un rythme d'échange rapide de ses utilisateurs — qui se compte en heures, tout au plus 2 jours —, tandis que certaines communautés techniques peuvent considérer un temps de

réponse de l'ordre du mois comme une durée naturelle pour un sujet spécifique.

Exemple 5.1

> Les systèmes de messageries instantanées sont parmi les systèmes de communication entre utilisateurs possédant le rythme naturel le plus élevé. Ce rythme influe sur la péremption des messages : une demande restant sans réponse plus d'une heure aura de fortes chances d'être oubliée, tandis qu'une réponse arrivant après ce même délai risquera fort d'être obsolète. A l'inverse, une demande sur un forum de développement collaboratif de logiciels peut trouver une réponse dans les semaines qui suivent sa publication, sans que la satisfaction du demandeur ou la pertinence de la réponse soit impactée.

Nous définissons $meanpace_{L_i}$, relatif à une communauté labellisée par L_i, comme le temps moyen de transmission d'un message entre les utilisateurs de la communauté :

$$meanpace_{L_i} = avg_{m=(u,u',t),m'=(u',u'',t') \text{ avec } u,u',u'' \in G_i, t'>t}(t'-t).$$

Le temps de latence lag entre deux messages $m = (v, u, t)$ et $m' = (u, v', t')$, relatif au temps moyen $meanpace_{L_j}$ de la communauté G_j labellisée par les concepts L_j est définie par :

$$lag(m, m') = \begin{cases} \infty \text{ si } t' \leq t, \\ \frac{t'-t}{meanpace_{L_j}} \text{ sinon.} \end{cases}$$

Il est à noter que le temps de latence lag infini est utilisé pour créer des chaînes de communication avec un horodatage (*timestamp*) strictement croissant. Deux messages simultanés ($t = t'$) ne peuvent donc pas, par définition, appartenir à une même chaîne de communication.

5.1.3 Probabilité de propagation sémantique et centralité sémantique temporelle

Nous pouvons maintenant définir de la *probabilité de propagation sémantique (SPP)*. Le SPP d'un utilisateur u, basé un message entrant m et un message

sortant m' est défini par :

$$SPP_c(u, m, m') = \frac{(1 - dispersion_c(m) \times dispersion_c(m'))}{1 + lag(t, t')}.$$

En conséquence, un utilisateur recevant un message traitant de c et envoyant un message traitant également de c immédiatement après — correspondant à t' arbitrairement proche de t —, possède un SPP_c arbitrairement proche de 1.

Finalement, la centralité sémantique temporelle $TSC_{L_i}(u)$ d'un utilisateur u au sein de la communauté labellisée par L_i est calculée sur tous les messages entrants et sortants de u :

$$TSC_{L_i}(u) = avg_{c \in L_i}\Big(\sum_{m=(u,u',t) \in G} \; \sum_{m'=(u',u'',t') \in G, t' > t} SPP_c(u, m, m')\Big).$$

Pour un concept donné, nous additionnons plutôt que de normaliser le SPP_c de u, afin de mettre en avant les utilisateurs avec de nombreuses communications ciblées. En effet, un utilisateur ne va pas nécessairement répondre ou faire suivre un message immédiatement, mais va probablement générer des réponses à plusieurs messages. Pour résumer l'ensemble sémantique L_i, nous prenons la moyenne des SPP_c, afin de mettre en avant les utilisateurs qui couvrent le mieux les concepts de L_i. Ce modèle est illustré dans nos expérimentations décrites ci-dessous.

5.1.4 Expérimentations

Jeu de données

Pour valider cette approche, nous avons utilisé comme source de données le *Enron Email data set*[3] en raison de sa disponibilité en tant que réseau complet de communication sociale, fournissant une relation d'envoi et des horodatages précis. Ce jeu de données est composé de courriers électroniques collectés auprès d'environ 150 utilisateurs — principalement des cadres supérieurs d'Enron — et rendus publics par les autorités fédérales américaines lors de leurs investigations sur le

3. Disponible à http://www.cs.cmu.edu/~enron/

« scandale Enron » [4]. Le jeu de données a été préalablement purgé par ces mêmes autorités des courriers électroniques à caractère privé et contient un total d'environ 500 000 messages.

Nous avons réalisé un nettoyage initial du jeu de données, dans le but d'effacer tous les messages ayant un horodatage incorrect. Si 99,87 % des courriers électroniques du jeu de données est marqué comme datant de 1997 à 2002 — date des enquêtes fédérales américaines —, le jeu contient également quelques milliers de messages marqués avec des horodatages sujets à caution, allant de 1970 à 2044. Ces messages, qui sont des *spams*, des notifications serveur ou d'autres messages générés automatiquement et malformés, sont écartés par notre nettoyage. La taille finale de notre jeu de données est alors de 494 910 courriers électroniques. La figure 5.2 montre la dispersion temporelle du jeu de données.

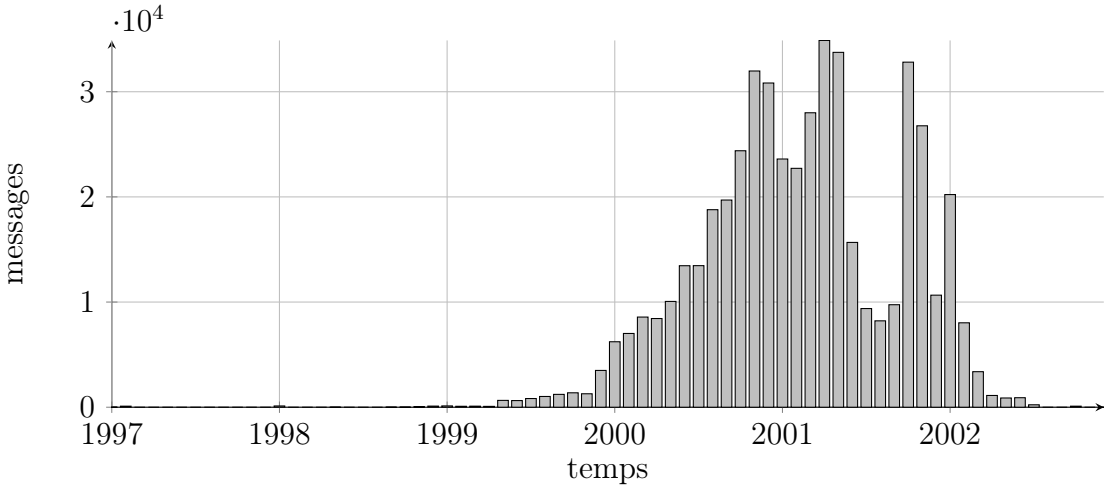


FIGURE 5.2 – Quantité d'emails par mois

Ontologie

Pour comprendre le contenu sémantique des messages, nous utilisons WordNet [51] comme une ontologie légère. Nous prenons la relation d'*hyperonymie* pour assumer le rôle de relation $is-a$ structurant l'ontologie, ainsi que le *synset* *entity* comme racine de l'ontologie. Nous réalisons une traduction relationnelle de

4. http://fr.wikipedia.org/wiki/Enron

l'ontologie résultant de cette interprétation. Cela nous permet de naviguer dans l'ontologie et de calculer des distances sémantiques en un temps contant [5]. De plus, l'utilisation des *synsets* de WordNet nous permet de lever la majeure partie des ambiguïtés de sens, comme illustré par la table 5.1 : le plus proche ancêtre commun détecté, `digit`, n'est pas une source de confusion bien que `thumb` et `seven` soient des fils de `digit`. Ils ne sont pas issus du même sens du mot, et ne sont donc pas reliés par notre détection de similarité.

concept 1	**concept 2**	**ancêtre commun**	**similarité**
`dog`	`cat`	`animal`	`0.571`
`Persian cat`	`Egyptian cat`	`domestic cat`	`0.888`
`thumb`	`little finger`	`digit`	`0.778`
`seven`	`two`	`digit`	`0.857`
`seven`	`little finger`	`entity`	`0`

TABLE 5.1 – Exemple de calcul de similarité sémantique via WordNet

Communautés

Comme expliqué dans le modèle, nous analysons chaque message et en extrayons leurs concepts principaux. Nous généralisons et résumons ces concepts, afin d'obtenir les sujets principaux de l'ensemble du système. À partir des similarités sémantiques que nous calculons, nous définissons et partitionnons les communautés liées aux sujets principaux. La table 5.2 présente les communautés identifiées à partir de notre jeu de données, ainsi que les concepts qui les structurent.

Centralité

À partir de ces regroupements, nous calculons les valeurs de SPP et de TSC pour chaque communauté. Les tables 5.3 et 5.4 montrent les résultats obtenus pour deux d'entre elles, avec les identifiants des utilisateurs, leurs valeurs calculées des différentes métriques présentées dans le modèle, ainsi que leurs postes ou fonctions affichées.

5. Cette opération est présentée de façon détaillée lors du chapitre 6.

position	**concepts**
#1	{market,services,providence,questioning,management}
#2	{forward,informant,attache,reporter}
#3	{pleasing,contraction}
#4	{subjectivity}
#5	{energy,gas}
#6	{time,change}
#7	{company,business}
#8	{newness}
#9	{thanks}
#10	{power}

TABLE 5.2 – Regroupement de concepts par communautés sémantiques

identifiant	$\mathbf{N^+ - N^-}$	**centralité**	**poste ou fonction**
kate.symes	4310	**5438**	Employee
kay.mann	14332	**3208**	Assistant General Counsel
vince.kaminski	8432	**1170**	Managing Director for Research
		...	
steven.kean	4571	**348**	Vice President & Chief of Staff
		...	
enron.announcements	7284	**0**	Mailing list

TABLE 5.3 – Centralités de la communauté #1{market,services,... }

identifiant	$\mathbf{N^+ - N^-}$	**centralité**	**poste ou fonction**
kay.mann	1884	**2810**	Employee
vince.kaminski	2456	**1335**	Managing Director for Research
tana.jones	650	**810**	Employee
		...	
steven.kean	1203	**272**	Vice President & Chief of Staff
		...	
enron.announcements	2477	**0**	Mailing list

TABLE 5.4 – Centralités de la communauté #5{energy,gas}

Il est intéressant de remarquer que la centralité n'apparaît pas comme directement liée au niveau d'activité des utilisateurs au sein de la communauté. Le meilleur exemple est sans doute l'adresse de diffusion enron.announcements. Malgré une très forte activité dans chacune des communautés identifiées, il n'obtient aucune valeur de centralité. Cela reflète le fait que si cette adresse communique

auprès de tous les autres utilisateurs du système, personne n'échange avec elle. En conséquence, elle est absente de tout chemin de communication identifié, et ne témoigne donc d'aucune centralité.

Dans un second temps, il est également intéressant de relever le rôle des cadres supérieurs. Bien que leurs communications soient importantes et leur centralité honorable, ils ne sont que rarement en tête de classement. Cela peut s'expliquer par leur position dans le mode de communication interne de l'entreprise. En tant que dirigeants, ils sont souvent en tête ou en queue de la chaîne de la communication. C'est la raison pour laquelle les meilleures centralités sont souvent assumées par des postes de secrétariat ou équivalents.

5.1.5 Discussion

Commentaires généraux

L'expérimentation de notre modèle sur le jeu de données Enron nous permet de comparer nos résultats avec la réalité organisationnelle de la compagnie et de son réseau de communication. Un point intéressant est également à relever : bien que le jeu de données contienne une large proportion de *spams*, pas un seul contenu de ce style n'a émergé de notre analyse. C'est un grand avantage de la prise en compte de la centralité sémantique, comparativement aux méthodes simples de fréquences brutes : bien que ces messages soient diffusés en très large quantité, le désintérêt total des utilisateurs vis-à-vis de ces contenus les rend inexistants au sein du contenu « utile » de la communication que nous extrayons.

En outre, notre analyse tend à dépeindre la réalité de la communication de l'entreprise. Si les dirigeants sont bien sûr toujours présents dans les discussions concernant leurs centres d'activité et de responsabilité, ils ne sont pas, cependant, au cœur de la communication. Nous avançons que les employés centraux dans ce modèle semblent être les responsables de secrétariat et de supervision des tâches externalisées : nécessitant une forte communication à double-sens, ils deviennent rapidement les centres de la communication dédiée à ces activités. Mais le manque de données sur les affectations de personnel dans le jeu de données ne nous permet pas de valider entièrement cette autre conclusion.

Propriétés de TSC

La centralité sémantique temporelle (TSC) possède plusieurs propriétés intéressantes. Premièrement, il peut être observé qu'un utilisateur faisant suivre un courrier électronique (fonction *Forward*) reçoit systématiquement une valeur de TSC élevée. En effet, cette centralité ne mesure pas l'ajout d'information à un message, mais la probabilité qu'a l'utilisateur de transmettre l'information de façon efficace.

Deuxièmement, nous ne favorisons pas explicitement les co-occurrences des concepts liés aux e-mails. Par exemple, il semble naturel de pondérer davantage un utilisateur qui transmet les concepts $\{a, b\} \in L_i$ dans un message unique m_1 plutôt qu'à un utilisateur qui transmet a puis b dans deux messages distincts m_2 et m_3. Mais la définition de SPP prend cette co-occurrence en compte. De fait, m_1 contribuera deux fois pour un temps de latence unique, alors que m_2 — respectivement m_3 — contribuera une fois, avec un temps de latence supplémentaire. Cet avantage est mathématiquement annulé si m_2 et m_3 sont simultanés, mais ce cas de figure est peu probable.

Aspects incrémentaux

Notre approche peut être interprétée à la fois comme en ligne ou comme en mode hors-ligne. L'interprétation hors-ligne nous permet à partir d'un jeu de données issu d'un système de communication donné, d'extraire ses sujets principaux, d'identifier les utilisateurs et leurs communautés, et finalement de les ordonner en fonction de leur centralité sémantique et temporelle. Cette approche permet la détection des sujets principaux qui sont représentatifs de l'ensemble du jeu de données, et de réaliser une analyse globale de ces communautés.

Mais il est intéressant de noter que nos algorithmes peuvent être mis en application de manière incrémentale : lorsqu'un nouveau message est acquis par le système — comme la soumission d'un message ou la réception d'un e-mail — le profil de l'utilisateur et la liste actuelle des sujets principaux peuvent être mis à jour, sans recalcul complet sur l'ensemble des archives de publication. En outre, les calculs de SPP et TSC peuvent être mis à jour uniquement pour les utilisateurs concer-

nés par les nouveaux messages. Cette approche implique qu'un nouveau concept c prochainement considéré comme « principal » peut apparaître à un instant t au cours de l'analyse des messages, et que la centralité selon c doit être comprise comme « après l'instant t ».

Exemple 5.2

> Le sujet « *federal investigation* » pour le jeu de données Enron peut apparaître comme principal à un temps donné. Mais les utilisateurs qui pouvaient contribuer sur ce sujet jusqu'alors ne sauraient être considérés comme centraux qu'à partir du moment où ce sujet est un sujet principal du système.

Le principal avantage de cette interprétation en ligne est de permettre une vision tant des sujets d'actualité que de la notion de centralité en temps réel. De plus dans ce contexte, un message ne requiert pas d'être stocké après son traitement, ce qui peut s'avérer crucial pour des tâches de supervision de systèmes aux taux de publications intensifs, comme Twitter par exemple.

Applicabilité à d'autres types de données

Notre méthode reposant sur des chaînes de communications entre plusieurs utilisateurs, nous nous interrogeons sur la capacité d'appliquer cette méthode à d'autres types de données que les courriers électroniques. Pour que ce genre d'adaptation soit possible, nous devons impérativement identifier pour chaque message son expéditeur et son destinataire. Tout type de jeu de données répondant à ce critère peut alors être pris en compte.

Un cas particulier se pose pour les forums de discussion. En effet, tout message diffusé sur ce support ne connaît pas de destinataire particulier, mais vise tout lecteur potentiel. En l'état le destinataire, étant à la fois tout le monde et personne, ne peut être défini pour chaque message. Nous pouvons toutefois définir des destinataires *a posteriori*, en définissant toute personne répondant à un message comme assumant le rôle de destinataire de ce dernier. Dans le même temps, nous pouvons conclure que l'auteur du message initial est considéré comme le destinataire de la réponse. Néanmoins, cette prise en compte retardée et évolutive de la notion de destinataire ne permet pas une analyse au fil de l'eau des contributions des

utilisateurs.

5.1.6 Conclusion

Nous avons présenté ici une approche visant à détecter les utilisateurs centraux dans un réseau de communication sociale, par le biais de la construction de communautés sémantiques et l'évaluation de la qualité sémantique et temporelle des messages. Pour ce faire, nous avons introduit une nouvelle mesure, la *probabilité de propagation sémantique* et la *centralité sémantique temporelle* pour évaluer et prendre cette qualité, et avons aussi montré expérimentalement cette pertinence.

Nous présentons ci-après une approche liminaire pour la détection des rôles des utilisateurs dans la vie de la communauté.

5.2 Vers une détermination des rôles utilisateurs

Dans cette approche, nous visons à prendre du recul par rapport à l'expérience utilisateur et à raisonner sur les communautés en tant que telles, afin d'observer leurs interactions et évolutions, et d'en détecter les causes. Nous proposons un modèle qui capture la dynamique communautaire. Cette dynamique se caractérise par des transformations des communautés telles que l'apparition, la croissance, la division, la fusion, la disparition, etc. Nous visons à établir une classification des rôles des utilisateurs en fonction de cette dynamique. Nous distinguons les rôles micro-communautaires — tels qu'attracteur ou répulseur — qui influent sur le comportement des utilisateurs au sein d'une communauté, des rôles macro-communautaires — comme diviseur ou fusionneur — qui influencent les interactions entre communautés.

5.2.1 Un modèle pour la dynamique des communautés

À partir de nos travaux précédents axés sur la détection de communautés et présentés au chapitre 4, étant donné un ensemble d'utilisateurs U et une ontologie

légère $O(C, is - a)$, nous définissons des communautés sémantiques $G_{L_i} \subseteq U$ avec $L_i \subseteq C$, ensemble de concepts labellisant la communauté.

Nous considérons à présents plusieurs états des communautés, à des instants différents $t_0, t_1, \ldots$. La durée de ces intervalles est discutable. Elle peut être régulière, mais nécessite alors d'évaluer la vitesse d'évolution propre du système. Elle peut être adaptative, mais nécessite dans ce cas des algorithmes pour suivre les changements de rythmes du système. En prenant en compte le temps, nous utilisons $G_L(t)$ comme la communauté G_L à l'instant t.

Nous définissons le graphe d'évolution des communautés comme un graphe orienté :

- les sommets sont des communautés $G_{L_i}(t_j)$;
- il existe un arc $G_{L_i}(t) \rightarrow G_{L_j}(t+1)$ si un nombre significatif d'utilisateurs (par exemple 30 %) de $G_{L_i}(t)$ apparaissent dans $G_{L_j}(t+1)$.

Il ne peut y avoir d'arc uniquement qu'entre communautés consécutives dans le temps. En revanche, ces arcs ne symbolisent pas forcement uniquement des transferts d'utilisateurs, car une communauté peut changer l'étiquetage sémantique L_i. La figure 5.3 montre un exemple de ce graphe d'évolution des communautés.

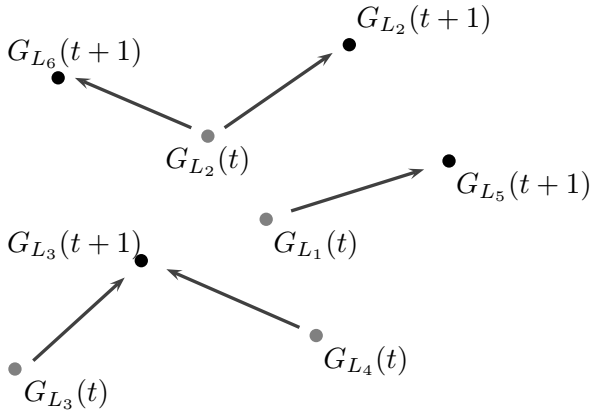


FIGURE 5.3 – Graphe d'évolution des communautés

Notion de dynamique Le point important à propos de ces communautés est que ce sont des groupes d'utilisateurs construits à partir de la communication existant entre les utilisateurs. Cette communication, ainsi que les intérêts des uti-

lisateurs, sont par nature des éléments changeants. C'est pourquoi au fil du temps, les communautés sont nécessairement appelées à évoluer. Pour une communauté donnée, cette tendance se reflète par deux indicateurs : les utilisateurs entrants, ceux qui rejoignent la communauté, et les utilisateurs sortants, ceux qui la quittent.

Nous avons choisi d'étudier cette dynamique. Il s'agit donc de définir les causes de ces changements dans la composition de la communauté. Comme ces regroupements sont basés sur la communication entre les utilisateurs, nous cherchons les causes du comportement des utilisateurs. Nous définissons ainsi la notion de *rôles utilisateurs*. Ces rôles décrivent les actions qui conduisent à des mouvements au sein des communautés. Nous distinguons deux types d'analyse des rôles : l'analyse micro-communautaire, qui met l'accent sur l'évolution de la composition des communautés, et l'analyse macro-communautaire, qui s'intéresse à l'interaction globale des communautés entre elles. Nous décrivons ces deux aspects ci-dessous.

5.2.2 Analyse micro-communautaire des rôles

L'analyse micro-communautaire des rôles utilisateurs se concentre sur l'influence que les utilisateurs ont sur la composition d'une communauté donnée. Ils sont divisés en deux groupes : les rôles incrémentiels et les rôles décrémentiels. L'action des premiers se traduit par l'arrivée de nouveaux utilisateurs dans la communauté, tandis que l'action des seconds provoque le départ d'utilisateurs de la communauté.

Comme précédemment évoqué lors de la découverte de communautés au chapitre 4, nous utilisons la notion de contexte sémantique d'un message m, notée $context(m)$, pour extraire et comparer aux concepts des communautés la sémantique des échanges. Pour déterminer si un utilisateur u exerce un rôle sur un utilisateur u' entre l'instant t et l'instant $t+1$ pour une communauté G_{L_i} donnée, nous prenons en compte l'appartenance ou non de u et u' à la communauté, ainsi que le fait que le contexte sémantique des messages de u vers u' appartient ou non à L_i. Ces rôles ainsi détectés sont exposés ci-après, et la table 5.5 résume ces interprétations.

u envoie un(des) message(s) m à u'		$\mathbf{u' \notin G_{L_i}(t) \rightarrow u' \in G_{L_i}(t+1)}$	$\mathbf{u' \in G_{L_i}(t) \rightarrow u' \notin G_{L_i}(t+1)}$
$\mathbf{u \in G_{L_i}(t)}$	$\mathbf{context(m) \subseteq L_i}$	Attracteur	Répulseur
	$\mathbf{context(m) \not\subseteq L_i}$	*non-pertinent*	Détourneur
$\mathbf{u \notin G_{L_i}(t)}$	$\mathbf{context(m) \subseteq L_i}$	Aiguilleur	Détracteur
	$\mathbf{context(m) \not\subseteq L_i}$	*non-pertinent*	Détourneur

TABLE 5.5 – Rôles micro-communautaires identifiés

Rôles incrémentiels

Les rôles incrémentiels pertinents sont définis comme ceux pour lesquels l'utilisateur concerné amène l'utilisateur cible à utiliser le contexte sémantique de la communauté. Nous ne considérons pas le cas où un utilisateur soit joint à une communauté suite à des événements en dehors des échanges du système : nous ne pouvons relier cet événement à un utilisateur du système qui en assumerait alors le rôle. Nous définissons deux rôles incrémentiels que sont l'*Attracteur* et l'*Aiguilleur*. Ils sont décrits ci-dessous et la figure 5.4 en fait une représentation visuelle.

- **Attracteur** : il attire de nouveaux membres au sein de la communauté dont il fait partie, en utilisant le contexte sémantique de la communauté ;
- **Aiguilleur** : il est un utilisateur qui oriente les utilisateurs vers d'autres communautés sans en faire lui-même partie.

Exemple 5.3

- Le gestionnaire de communautés gérant les plates-formes de discussions d'Apple pour promouvoir les produits de la marque et provoquer les discussion est un *Attracteur*.
- Le « bon samaritain » de Yahoo ! Questions/Réponses [6], ou un modérateur redirigeant un flux de discussion sur un réseau de forums, sont des *Aiguilleurs*.

Pour ces deux rôles, nous ne prenons en compte que les interactions liées au contexte sémantique de la communauté. En effet, rien ne peut prouver qu'une communication sans rapport avec la communauté ait un lien avec la nouvelle adhésion de l'utilisateur à la communauté. Privilégiant le faux-négatif au faux-positif, nous

6. Service communautaire collaboratif de réponses aux questions des utilisateurs par les utilisateurs. http://fr.answers.yahoo.com/

définissions donc ces autres cas comme non-pertinents.

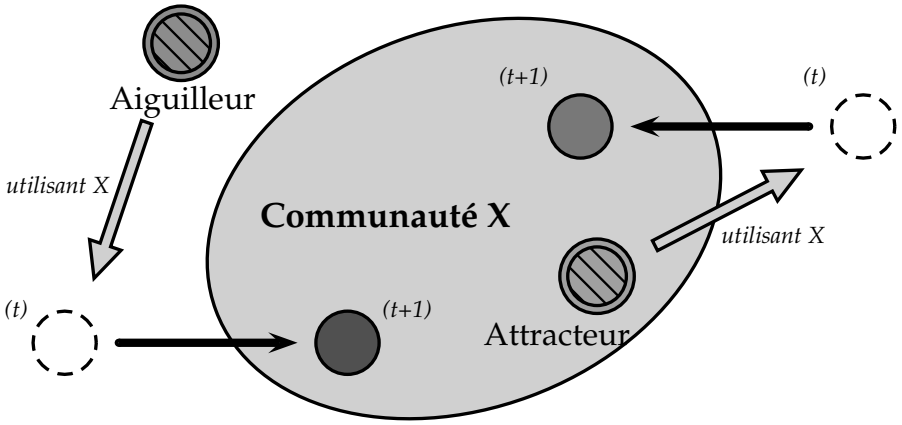


FIGURE 5.4 – Rôles incrémentiels

Rôles décrémentiels

À l'opposé des rôles incrémentiels, les rôles décrémentiels sont définis comme ceux qui font qu'un ou plusieurs utilisateurs quittent la communauté à leur contact. Dans ce cas présent, nous séparons ces rôles en deux groupes, ceux qui utilisent le contexte sémantique de la communauté pour conduire un utilisateur à la quitter, et ceux qui utilisent un contexte sémantique extérieur.

Avec le contexte de la communauté En utilisant le contexte sémantique de la communauté, nous identifions deux rôles décrémentiels, que nous nommons *Répulseur* et *Détracteur*. Ils sont décrits ci-dessous et la figure 5.5 en fait une représentation visuelle.

- **Répulseur** : Le *Répulseur* provoque le départ d'utilisateurs de sa propre communauté ;
- **Détracteur** : Le *Détracteur* fait quitter les membres d'une communauté dont il ne fait pas lui-même partie.

Sans le contexte de la communauté En utilisant un contexte sémantique qui est extérieur au contexte de la communauté, nous définissons un unique rôle

décrémentiel que nous nommons *Détourneur*. Ce rôle est unique étant donné que ses deux sous-cas — que le titulaire du rôle soit intérieur ou extérieur à la communauté — produisent le même effet. En effet, comme il n'utilise pas le contexte de la communauté, sa propre position ne change pas l'effet obtenu. Le *Détourneur* est un utilisateur extérieur à la communauté qui cause le départ d'un utilisateur. La figure 5.5 le représente également.

Exemple 5.4

- Un membre insupportable — qui utilise le *flood*, le *spam*, etc. — peut être un exemple de *Répulseur*. C'est également le cas d'un modérateur qui peut bannir des utilisateurs !
- Un utilisateur de Linux venant argumenter sur des forums Windows peut jouer le rôle de *Détourneur*.

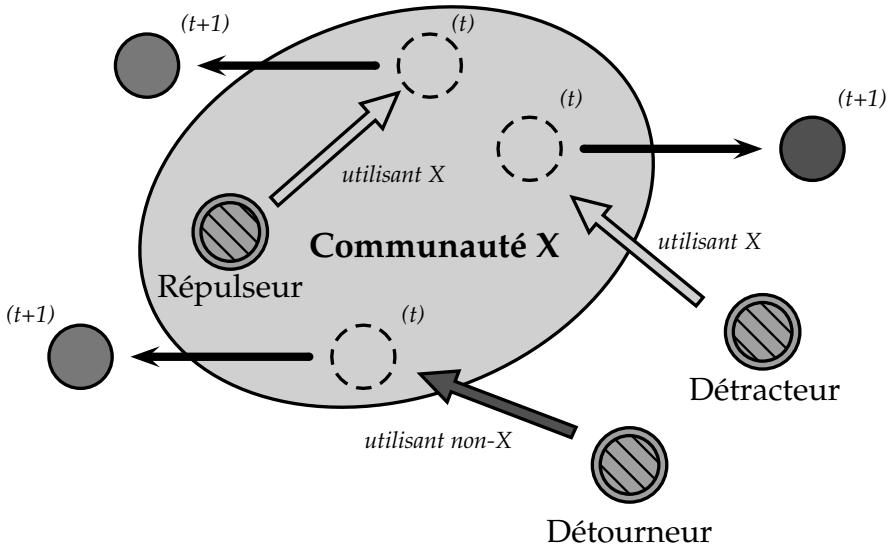


FIGURE 5.5 – Rôles décrémentiels

5.2.3 Analyse macro-communautaire des rôles

L'analyse macro-communautaire des rôles utilisateurs se concentre sur les influences des utilisateurs dans les interactions entre communautés. Basés sur les rôles micro-communautaires, ils sont utilisés pour expliquer comment les communautés évoluent les unes avec les autres. Un rôle macro-communautaire est endossé

par un groupe d'utilisateurs partageant le même rôle micro-communautaire. Ce groupe est identifié comme étant à l'origine du changement macro-communautaire observé entre les communautés. La dynamique macro-communautaire la plus basique est le *Transfert*, qui est la base des autres dynamiques observées.

Transfert Une dynamique de transfert concerne les mouvements d'utilisateurs d'une communauté à une autre. Nous nommons le rôle macro-communautaire relié *Transfert*. Le groupe d'utilisateurs Transfert est responsable du départ d'utilisateurs d'une communauté précise vers une autre communauté donnée. Ce groupe peut être composé de *Répulseurs* et de *Détracteurs* de la communauté de départ, ainsi que d'*Aiguilleurs* et d'*Attracteurs* de la communauté d'arrivée. Les utilisateurs assumant ces rôles micro-communautaires peuvent endosser le rôle de *Transfert* si leur activité caractérise une partie importante du transfert observé.

Apparition et disparition Un cas spécial de transfert apparaît lorsque la communauté de départ ou d'arrivée est un ensemble vide. En d'autres termes, lorsqu'une nouvelle communauté se forme sans source communautaire identifiée, ou qu'une communauté se dissout sans communauté de destination identifiée. Nous en déduisons deux rôles macro-communautaires que sont *Créateur* et *Destructeur*. En tant que sous-rôles de *Transfert*, le premier est composé d'*Aiguilleurs* et d'*Attracteurs*, le second de *Répulseurs* et de *Détracteurs*.

Fusion et division La fusion et la division de communautés sont des cas spéciaux où plus d'une source ou destination est identifiée lors d'un transfert. Les rôles macro-communautaires qui en découlent sont le *Fusionneur* et le *Diviseur*.

5.2.4 Conclusion

Dans cette approche, nous modélisons des rôles utilisateurs en fonction de la dynamique temporelle des communautés. Nous situons notre analyse à deux niveaux : le premier, micro-communautaire, pour identifier les utilisateurs à l'origine de l'évolution d'une communauté ; et le second, macro-communautaire, pour cibler ceux qui influent sur les interactions entre les communautés.

Ce travail n'a pas donné lieu à des expérimentations. Dans notre modèle actuel, l'appartenance à une communauté n'est pas exclusive. Un utilisateur peut avoir des centres d'intérêts divers et variés, ce qui peut se caractériser par l'appartenance à autant de communautés. Or, une dynamique de communautés n'est observable qu'à partir d'une certaine granularité dans les regroupements qui sont réalisés sur les utilisateurs. Dans le cadre de la discussion entre les utilisateurs, cela implique un très vaste champ d'interaction, tant en nombre d'utilisateurs qu'en sujets. Il en résulte que, pour toute étude de forums de discussion ou plate-formes assimilées, notre étude est prisonnière du contexte global du forum en lui-même. Toutes les « sous- » communautés alors détectées, bien qu'aidant à la compréhension du système, ne présentent pas de disparités suffisantes pour établir clairement des notions de rôles telles que définies dans cette section. Ces déductions ne seront possibles qu'avec une large surveillance de nombreux forums distincts, où les utilisateurs présents sur différentes plate-formes sont identifiés et alignés. N'étant pas en mesure de mettre en place ce type d'infrastructure, nous n'avons pas été en mesure de réaliser une expérimentation et un affinage de ce modèle pour le moment.

5.3 Conclusion générale

Dans ce chapitre, nous avons présenté nos approches et contributions en matière d'analyse de nos communautés préalablement détectées. Nous y définissons différents axes d'analyse permettant d'identifier :

- Les utilisateurs centraux des communautés, par le biais du calcul de leur *centralité sémantique temporelle* en prenant en compte tant la sémantique que leur temps de réponse lors de la communication entre les utilisateurs du système ;
- Les utilisateurs responsables de la dynamique des communautés, en identifiant les utilisateurs qui provoquent des changements dans le comportement sémantique des utilisateurs et à travers eux, dans la structuration des communautés.

Dans le chapitre suivant, nous présentons les caractéristiques spécifiques à notre implémentation des modèles précédemment décrits.

Chapitre 6

Implémentation

Sommaire

Dans ce chapitre, nous présentons les détails de l'implémentation du système WebTribe dans la section 6.1. Nous détaillons ensuite dans la section 6.2 les diverses optimisations, ajustements et détails techniques intéressants spécifiques à ces implémentations. Enfin, la section 6.3 conclut le chapitre.

L'implémentation du système WebTribe a donné lieu à plusieurs publications sous forme de démonstrations lors de conférences. Lors de ces démonstrations, nous présentons une implémentation du module *Front-End* en tant qu'applet Java, pour visualiser la découverte des communautés et le profilage des utilisateurs. Ces démonstrations ont eu lieu le 25 octobre 2011 lors de conférence *Bases de Données Avancées*, à Rabat, Maroc, et le 25 juillet 2012 lors de la conférence *International Conference on Web Engineering (ICWE)*, à Berlin, Allemagne. L'annexe D présente quelques captures d'écran de l'applet utilisée pour ces démonstrations [1].

6.1 Implémentation de WebTribe

Le système WebTribe, tel qu'il est décrit dans le chapitre 3, sert de base à la réalisation de l'ensemble de nos expérimentations. Développée en Java, l'implémentation se compose de plusieurs modules techniques distincts et inter-opérants. Ce partage des tâches nous permet d'implémenter successivement plusieurs approches sans avoir à redéfinir l'ensemble des tâches communes, comme par exemple l'extraction des données des systèmes analysés. La figure 6.1 illustre l'organisation globale de ces implémentations.

Les principaux modules utilisés sont :

– **des *parsers* Web** : à partir d'une source donnée, le code des pages Web est analysé. En s'appuyant sur sa structure ainsi qu'un ensemble de règles d'extractions propres à chaque type de source, nous extrayons les communications entre les utilisateurs. Nous stockons cette communication en mémoire le temps de l'analyse, sous forme de fichiers XML légers. Ces fichiers respectent la DTD fournie en Annexe B pour uniformiser les échanges et normaliser la communication ;

1. Une présentation est disponible à : http://www.damien-leprovost.fr/webtribe.

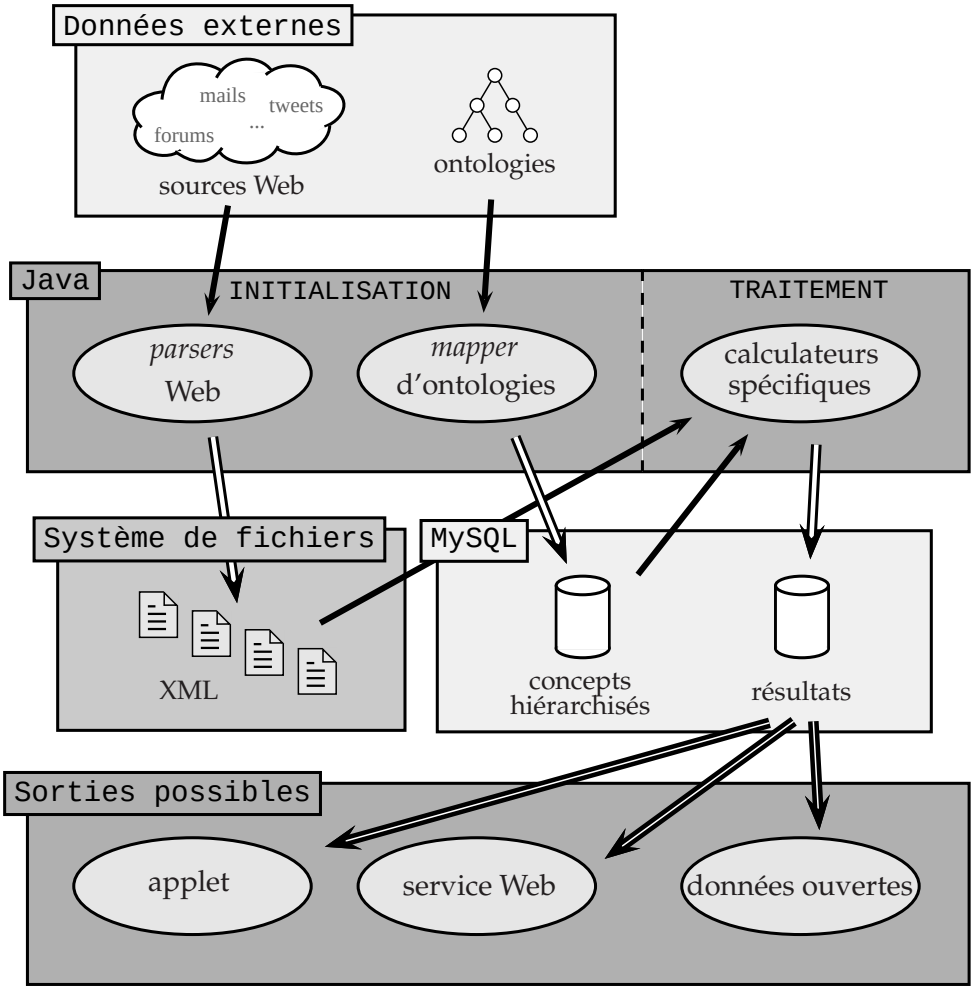


FIGURE 6.1 – Vue générale de l'implémentation de WebTribe

- **un *mapper* d'ontologie** : cet outil réalise une extraction des hiérarchies de concepts de l'ontologie ciblée et les stocke dans une base de données relationnelle. La motivation de cette étape ainsi que ses considérations techniques sont détaillées dans la section suivante ;
- **des calculateurs spécifiques** : cette étape représente le code propre à chaque expérimentation que nous injectons dans le système. De cette manière, les entrées et sorties du système WebTribe sont toujours identiques en nature, facilitant tant l'interopérabilité que la rapidité et faisabilité à faire évoluer le système face à de nouvelles évolutions ;
- **des sorties possibles** : de la même façon, plusieurs méthodes sont possibles

pour exploiter les données calculées qui sont stockées dans la base de données de destination. Outre l'interrogation brute des données, nous avons proposé le développement d'applet, de services Web ou la publication de données ouvertes. Des utilisations au sein de *mashup* sont notamment envisagées.

6.2 Optimisations et ajustements

Cette section présente divers détails techniques relevés lors du passage des modèles à leurs implémentation. Ces ajustements sont motivés par le caractère en ligne de notre approche, qui implique un calcul en temps réel à un coût limité, afin de pouvoir suivre la montée en charge du système. Ces points concernent soit des optimisations, vues comme une méthode plus efficace de parvenir au même résultat, soit des ajustements par le biais d'approximations, qui sont effectuées si elles ne remettent pas en cause les conclusions qui en découlent.

6.2.1 Extraction des hiérarchies de concepts

Lors de l'utilisation d'ontologies, spécialement si ces dernières sont de taille importante, la recherche et l'accès aux concepts avec divers outils comme Jena [2] est une opération coûteuse. Le temps de réponse, par exemple pour rechercher l'occurrence d'un terme dans une ontologie, augmente de façon quadratique avec la taille des ontologies ciblées. Or, nos modèles requièrent de très nombreuses requêtes aux ontologies, notamment en terme de détection de concepts et de calcul de distance. De plus, nos requêtes sont souvent répétitives.

Afin de supprimer ces coûts, ou tout au moins les réduire à un coût constant, nous avons choisi de tirer parti des avantages des bases de données relationnelles, ainsi que des index dont elles disposent. Pour cela, nous ajoutons une phase d'initialisation, réalisée une fois par ontologie, avant l'exploitation des données utilisateurs. Cette phase consiste à transcrire chaque concept dans une table de la base de données, avec la référence de son père, et sa profondeur dans l'ontologie, basée

2. Apache Jena est un framework Java libre pour le Web sémantique. http://jena.apache.org/

sur la relation $is - a$. La table 6.1 présente un exemple de ce stockage.

identifiant	concept	parent	pronfondeur
1	`body part`	-	0
4	`belly`	3	3
119	`cubitus`	101	4
143	`claw`	86	6

TABLE 6.1 – Exemple de stockage d'ontologie légère (extrait)

L'ensemble des termes des concepts et leur racinisation est indexé. Nous maintenons des vues permettant de quantifier et retrouver les concepts fils, afin de réaliser rapidement des calculs de couverture et pertinence, comme détaillé dans le chapitre 4. L'utilisation de ces index et vues permet de réduire drastiquement les temps de réponses à nos requêtes, et surtout d'obtenir une variation du temps de réponse quasi-constant avec la montée en charge et l'augmentation de la taille des ontologies. La table 6.2 montre quelques mesures réalisées pour illustrer l'impact de cette méthode.

Opération	Via accès direct à l'ontologie	Via base relationnelle
Initialisation d'une ontologie (1 727 concepts)	-	< 1 s
Initialisation d'une ontologie (61 102 concepts)	-	20 s
1000 détections de la présence d'un terme donné	50 s	9 s
1000 calculs de la distance entre deux concepts	870 s	4 s

TABLE 6.2 – Temps de calcul des opérations sur l'ontologie (approximations)

6.2.2 Limitation des concepts candidats

Lors de l'interprétation des publications des utilisateurs, nous recherchons dans les termes utilisés des concepts connus et nous les comparons aux concepts qui

structurent les communautés identifiées. En conséquence, si l'ontologie utilisée comporte beaucoup de concepts, le nombre de détections de concepts dans les publications peut être extrêmement élevé, le nombre de comparaisons avec les concepts de référence augmentant alors très rapidement. Pourtant, le nombre de concepts de référence étant lui relativement constant, il en résulte que de nombreuses comparaisons ne retournent aucune proximité exploitable. Nous définissons donc ci-dessous une méthode pour limiter ces comparaisons aux concepts uniquement nécessaires.

Lors des comparaisons de concepts, par le biais de calcul de distance sémantique, nous utilisons comme détaillé dans le chapitre précédent un seuil de similarité. Au delà de ce seuil, la distance est jugée trop importante et la comparaison non pertinente. Pour optimiser ces comparaisons, nous identifions les « concepts candidats », comme l'ensemble des concepts de l'ontologie qui sont à une distance sémantique inférieure au seuil d'au moins un des concepts de référence. Cet ensemble représente le voisinage sémantique des concepts de référence, et sont de fait les seuls pour lesquels le calcul de distance sémantique avec l'un des concepts de référence peut être concluant. À partir de là, nous ne relevons dans les publications des utilisateurs que les termes que nous pouvons relier à ces concepts, et non plus à tous les concepts de l'ontologie. Nous obtenons alors exactement les mêmes résultats que précédemment, mais en réalisant énormément moins de détections et calculs, et pouvant donc les maintenir en temps réel. Le nombre de concepts candidats dépend du nombre de concepts de référence et de la structure de l'ontologie. En moyenne lors de nos expérimentations, ces voisinages excèdent rarement la dizaine de concepts par concept de référence.

Nous pouvons envisager le stockage d'une matrice de distance entre la totalité des concepts candidats et concepts de référence, pour accélérer encore les traitements au prix d'une phase d'initialisation et du coût du stockage qui y est lié. Tout calcul de distance ne serait alors que la lecture d'une donnée précédemment indexée. Cette approche reporte donc la charge de calcul sur le coût de stockage. C'est pourquoi nous sommes restés à l'implémentation médiane décrite ci-dessus.

6.2.3 Optimisation du calcul de la probabilité de propagation sémantique

Dispersion « court-circuit » En première partie du chapitre 5, nous calculons une mesure appelée *probabilité de propagation sémantique* (SPP). Cette mesure représente la probabilité qu'un utilisateur soit un bon relais d'un concept ou ensemble de concepts, et est basée sur ses précédents échanges. Toutefois en l'état, l'application brute de notre modèle ne permet pas de passer à l'échelle et de suivre en temps réel l'évolution de larges communautés au sein de systèmes Web ayant un fort taux de publication. Pour contourner cette limitation et réduire le temps de calcul, nous procédons à un approximation, ayant pour but de supprimer les valeurs avant leur calcul, si nous pouvons prédire que le futur résultat du calcul ne pourrait être signifiant. Nous analysons pour cela la formule du SPP, qui est :

$$SPP_c(u, m, m') = \frac{(1 - dispersion_c(m) \times dispersion_c(m'))}{1 + lag(t, t')}.$$

SPP varie entre 0 et 1, où 1 est d'une importance maximale, et 0 d'une importance nulle. Nous cherchons donc à considérer les cas où SPP est approximativement 0, afin de faire l'économie de ces calculs. SPP tend vers 0 dès qu'une seule des deux dispersions sémantiques ($disp_c(m)$) tend vers 1. La figure 6.2 illustre ce fait, avec un seuil arbitraire de 0.7.

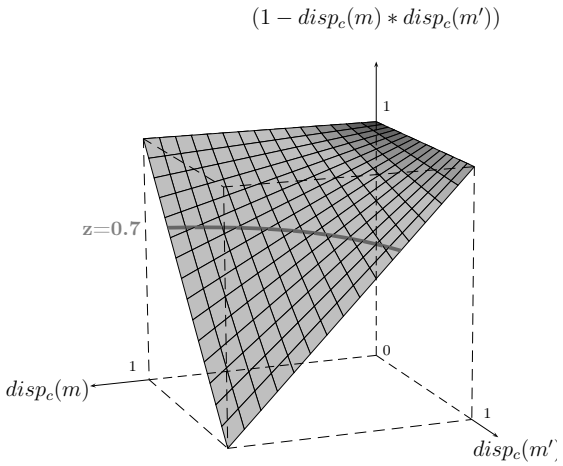


FIGURE 6.2 – Valeurs possibles de SPP, et seuil à 0.7

En conséquence, dès qu'une dispersion trop importante est calculée, nous interrompons le calcul du SPP et l'approximons à 0.

Fenêtre temporelle D'un raisonnement similaire, nous observons que SPP tend vers 0 quand le temps de latence augmente, quelles que soient les valeurs de dispersion des contenus des messages. En conséquence, deux messages trop éloignés dans le temps ne peuvent faire émerger un SPP de valeur. Dans le contexte d'analyse en ligne en temps réel précédemment décrit, nous définissons donc une fenêtre temporelle dépendant du rythme moyen du système. Cette fenêtre temporelle définit un intervalle de temps entre le moment présent et la limite d'analyse potentiellement pertinente, avant que le temps de latence ne devienne nécessairement trop élevé. En conséquence, nous ne conservons en mémoire que les contenus sémantiques des messages précédemment reçus contenus dans le fenêtre temporelle. Quand un nouveau message est traité, il n'est comparé qu'avec les données présentes dans la fenêtre temporelle. Au fur et à mesure que le temps avance, les données relatives aux messages passés sont purgées. Cela limite donc les comparaisons effectuées, tout en préservant une utilisation constante de la mémoire de stockage. L'algorithme 9 décrit cette méthode.

Algorithme 9 : Calcul incrémental de TSC

entrées : $G = (U, S)$, message m, seuil de pertinence δ

début
  pour chaque *nouveau message* m **faire**
    pour chaque $u \in recipient(m)$ **faire**
      ajouter m à $INBOX(u)$
      supprimer de $INBOX(u)$ les messages m' avec $lag(m', m) \leq \delta$
    fin pour
    $s = sender(m)$
    pour chaque $m' \in INBOX(s)$ **faire**
      calculer chaque$SPP_c(s, m, m')$
      mettre à jour $TSC(s)$
    fin pour
  fin pour
fin

6.3 Conclusion

Dans ce chapitre, nous avons présenté les particularités de l'implémentation des diverses approches présentées au cours de cette thèse au sein du système WebTribe. En raison des contraintes techniques comme des besoins de performance inhérents au caractère « en ligne » de nos approches, nous réalisons diverses optimisations et approximations pour maintenir un système opérationnel lors de la montée en charge. En revanche, ces modifications nous permettent un traitement des communautés et utilisateurs en coût constant, nous autorisant — sauf bouleversement majeur — un suivi théoriquement non-limité dans le temps du système analysé.

Chapitre 7

Conclusion et perspectives

Sommaire

7.1 Synthèse

Les travaux présentés dans cette thèse se situent dans le contexte de la recherche de communautés d'utilisateurs, en se focalisant sur les réseaux de communication. Les principales contributions de cette thèse sont des méthodes de détection et d'analyse de communautés en utilisant une approche sémantique. Nous avons présenté pour cela l'outil de détection et d'analyse WebTribe, permettant d'assister un gestionnaire de communautés.

Dans ce travail de thèse, nous avons abordé principalement le problème de la compréhension des échanges entre les utilisateurs, ainsi que les possibilités de regroupement de ces utilisateurs en ensembles homogènes. Afin de situer nos travaux, nous avons présenté au chapitre 2 un état de l'art de la notion de communauté Web et des travaux relatifs à la compréhension des utilisateurs dans des réseaux de communication et réseaux sociaux. Nous avons également présenté un résumé des techniques sémantiques permettant d'assister cette compréhension. Cet état de l'art nous conduit au bilan d'un Web en mutation où les techniques de gestion à la fois sémantiques, relationnelles et en temps réel (au sens humain) des utilisateurs sont un besoin grandissant.

Nous avons présenté au chapitre 3 notre approche et ses objectifs, ainsi que WebTribe, notre outil de découverte et d'analyse des communautés. Nous avons détaillé au chapitre 4 nos contributions sur la détection de communautés, en utilisant plusieurs approches sémantiques, des systèmes de *tags* aux ontologies. Nous avons ensuite exposé nos contributions sur l'analyse des communautés au chapitre 5, en proposant une détection de la place des utilisateurs dans les communautés, et des influences qu'ils y exercent.

Les approches de cette thèse ont été implémentées dans notre outil WebTribe et nous avons présenté dans le chapitre 6 les caractéristiques techniques et optimisations des expérimentations réalisées tout au long des différentes approches.

7.2 Contributions

Les travaux de cette thèse utilisent des bases de connaissances sémantiques de référence, comme des ontologies légères. Nous regroupons en communautés les utilisateurs qui partagent les mêmes centres d'intérêts sémantiques, dans la mesure où ceux-ci présentent un intérêt suffisant en terme de représentativité pour le système analysé. Nous dégageons ensuite des valeurs clés pour caractériser les communautés.

Nos contributions se divisent en deux étapes successives et complémentaires, que sont la découverte de communautés d'utilisateurs, et l'analyse de ces mêmes communautés.

Découverte des communautés Nous avons présenté trois approches pour regrouper les utilisateurs en fonction de leurs activités, avec divers niveaux de connaissance du contenu du système et de structure de la base de connaissances employée. Nous avons introduit successivement une découverte basée sur la manipulation de *tags* par les utilisateurs, sur un vocabulaire librement fourni par le gestionnaire, ou sur une ontologie servant de référentiel de pertinence sémantique.

Analyse des communautés Nous avons proposé plusieurs métriques et caractéristiques pour interpréter les communautés ainsi identifiées. Après avoir sémantiquement labellisé les communautés d'utilisateurs, nous avons défini la *centralité sémantique temporelle* et la *probabilité de propagation sémantique* pour identifier les utilisateurs centraux des communautés. Nous avons proposé une méthode pour détecter et caractériser les utilisateurs ayant un rôle sur la dynamique des communautés et les interactions de leurs membres.

L'ensemble de ces contributions, regroupées au sein de l'outil WebTribe, se veut un moyen de dégager en temps réel d'un large ensemble de communication entre utilisateurs, des informations synthétiques et exploitables par un gestionnaire de communautés sur le Web. Ces considérations permettent à WebTribe de répondre à son objectif « en ligne » pour détecter et analyser des communautés d'utilisateurs en temps réel, pour des réseaux de communication de grande taille.

7.3 Perspectives

Ces travaux ouvrent de nombreuses perspectives au sein de ce domaine en plein essor qu'est la compréhension des échanges entre utilisateurs. Parmi elles, nous pouvons citer trois développements futurs — parfois connexes — qui découlent de nos résultats.

Alignement des systèmes de communication Lors de nos expérimentations, notre compréhension des échanges entre les utilisateurs s'est trouvée limitée par notre capacité à prendre en compte l'intégralité des discussions potentiellement pertinentes. Cela est dû au fait que nos analyses ciblent un système de communication donné. Il nous est donc par nature impossible de prendre en compte les échanges situés à l'extérieur de ce système, mais qui peuvent avoir des influences sur lui. En l'état, même s'il nous est possible d'analyser plusieurs systèmes en parallèle, ces derniers seront alors vus comme autant de sources de résultats indépendants entre eux. Une perspective intéressante d'évolution serait donc de travailler à l'alignement des systèmes pour mener une analyse multi-sources. Cela passe par un alignement des utilisateurs, permettant d'identifier une même personne comme propriétaire de comptes utilisateurs sur plusieurs des systèmes analysés. Ce cas se trouve fréquemment dans les systèmes de forums de discussion, où plusieurs utilisateurs sont inscrits et échanges sur plusieurs forums indépendants.

D'autre part, nos analyses sont également contraintes par les limites de connaissances de l'ontologie choisie. Avoir une ontologie comme base de référence impose que nous ne pouvons détecter uniquement que les concepts qui sont présents dans l'ontologie. Une solution à ce problème serait de se référer à plusieurs ontologies simultanément, posant alors la problématique bien connue de l'alignement d'ontologies. Cela permettrait alors d'adjoindre plusieurs bases de connaissances distinctes pour améliorer la couverture sémantique des systèmes étudiés.

Adaptabilité des ontologies de référence Une alternative à l'alignement de plusieurs ontologies existantes peut être considérée en faisant évoluer l'ontologie de référence utilisée. Cet enrichissement, basé sur une détection des nouveaux concepts émergeant au sein du système analysé, permet de maintenir l'ontologie

de référence en pertinence par rapport aux discussions des utilisateurs. Il se pose en revanche le problème du placement dans l'ontologie des concepts nouvellement détectés, qui peut être résolu par l'intervention d'experts ou l'utilisation de bases de connaissances extérieures.

Transformation de l'information Dans nos analyses des caractéristiques des communautés, nous avons cherché à identifier les utilisateurs centraux d'un contexte sémantique donné. Cette centralité se base entre autres sur la proximité sémantique des entrées et sorties de l'utilisateur. À ce titre, l'utilisateur est alors vu comme un relais dans la communication sémantique de la communauté. Dans des travaux ultérieurs, il nous apparaît opportun de prendre en compte et de caractériser les transformations que subit un message dans un chemin de communication, d'un utilisateur initial à un utilisateur final en passant par tous ceux qui ont transféré ou altéré le message. Cette caractérisation permettrait alors de déterminer les capacités de l'utilisateur sur le message, en matière de calcul, de correction, de transformation de l'information, etc. Ces compétences ainsi explicitées pouvant alors être filtrées et détaillées en fonction des contextes sémantiques manipulés.

Annexes

Sommaire

A Principaux sites Web

Les principaux sites Web mondiaux, en terme de trafic, au 1er juin 2012 [1].

Rang	Site	Part estimée utilisateur	Description
1	**Google** http://google.com	49 %	Moteur de recherche et services
2	**Facebook** http://facebook.com	45 %	Réseau social généraliste
3	**YouTube** http://youtube.com	33 %	Hébergement de vidéo
4	**Yahoo !** http://yahoo.com	21 %	Portail web généraliste
5	**Baidu.com** http://baidu.com	11 %	Moteur de recherche chinois
6	**Wikipedia** http://wikipedia.org	14 %	Encyclopédie libre collaborative
7	**Windows Live** http://live.com	10 %	Moteur de recherche et services
8	**Twitter** http://twitter.com	9 %	Réseau social de microblogage
9	**Tencent QQ** http://qq.com	7 %	Messagerie instantanée chinoise
10	**Amazon.com** http://amazon.com	6 %	Vente en ligne
11	**Blogspot.com** http://blogspot.com	8 %	Création et hébergement de blogs
12	**LinkedIn** http://linkedin.com	6 %	Réseau social professionnel

TABLE A.1 – Principaux sites Web mondiaux

1. Source : http://www.alexa.com/topsites

Existence d'un réseau social dédié au site Web, ou présence de fonctionnalités sociales (recommander, partager avec un ami, etc.) sur le site Web.

Rang	**Site**	**Réseau social dédié**	**Fonctionnalités sociales**
1	Google	√	√
2	Facebook	√	√
3	YouTube		√
4	Yahoo !		√
5	Baidu.com		
6	Wikipedia		√
7	Windows Live	√	√
8	Twitter	√	√
9	Tencent QQ	√	√
10	Amazon.com		√
11	Blogspot.com		√
12	LinkedIn	√	√

TABLE A.2 – Présence de fonctionnalités sociales dans les principaux sites Web

B DTD de la communication extraite

```
1  <?xml version="1.0" encoding="ISO-8859-1"?>
2  <!ELEMENT article (title,date,author,body,
      numberofcomments,comments)>
3  <!ELEMENT title (#PCDATA)>
4  <!ELEMENT date (posted,updated?)>
5  <!ELEMENT posted (#PCDATA)>
6  <!ELEMENT updated (#PCDATA)>
7  <!ELEMENT author (#PCDATA)>
8  <!ELEMENT body (p* | #PCDATA)>
9  <!ELEMENT p (#PCDATA)>
10 <!ELEMENT numberofcomments (#PCDATA)>
11 <!ELEMENT comments (comment*)>
12 <!ELEMENT comment (date,author,body)>
```

C Exemple de transcription de message utilisateur

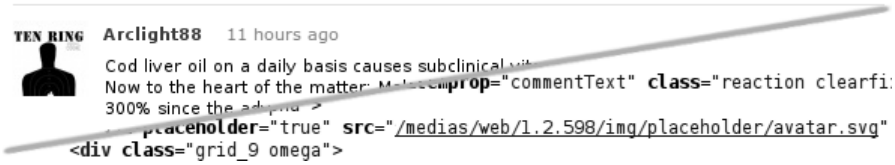


FIGURE C.1 – Message brut d'un utilisateur

FIGURE C.2 – Normalisation du message suivant la DTD

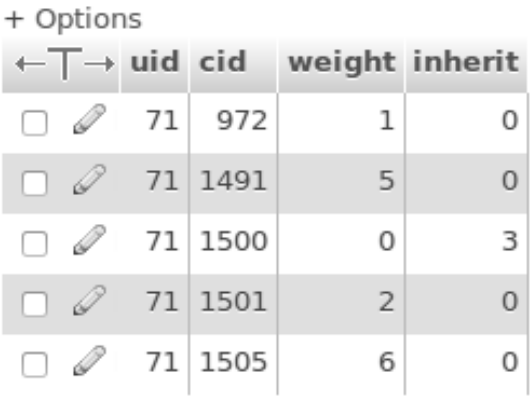


FIGURE C.3 – Stockage des concepts utilisateurs en base de données

D Captures d'écran de l'applet WebTribe

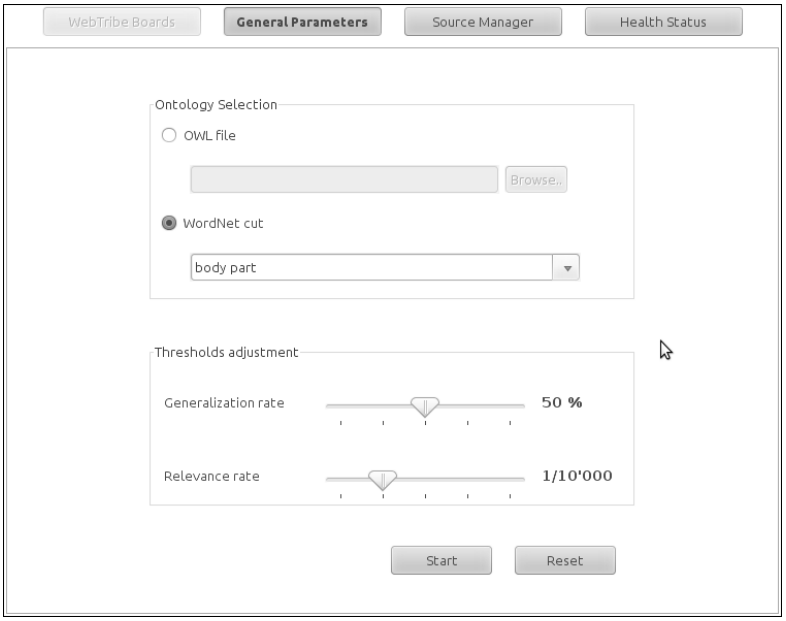


FIGURE D.4 – Applet WebTribe : paramètres généraux

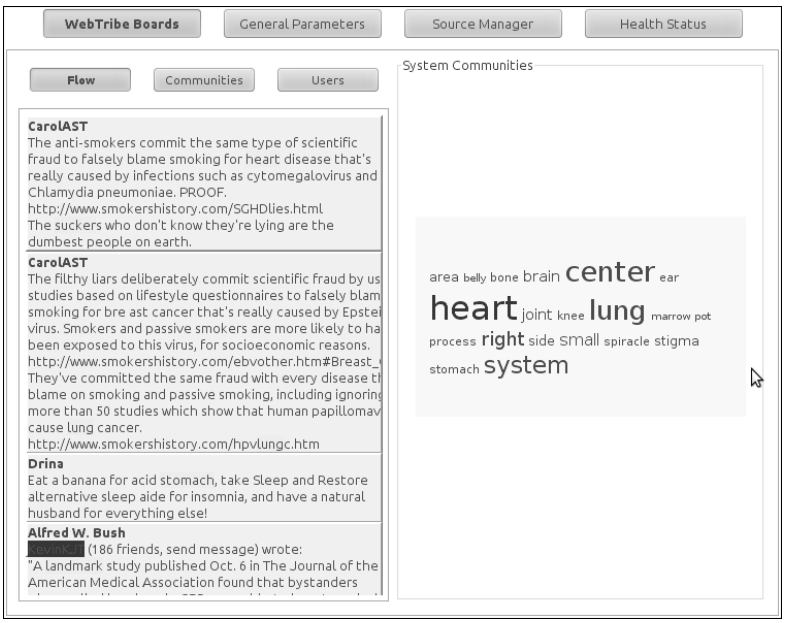


FIGURE D.5 – Applet WebTribe : flux et communautés

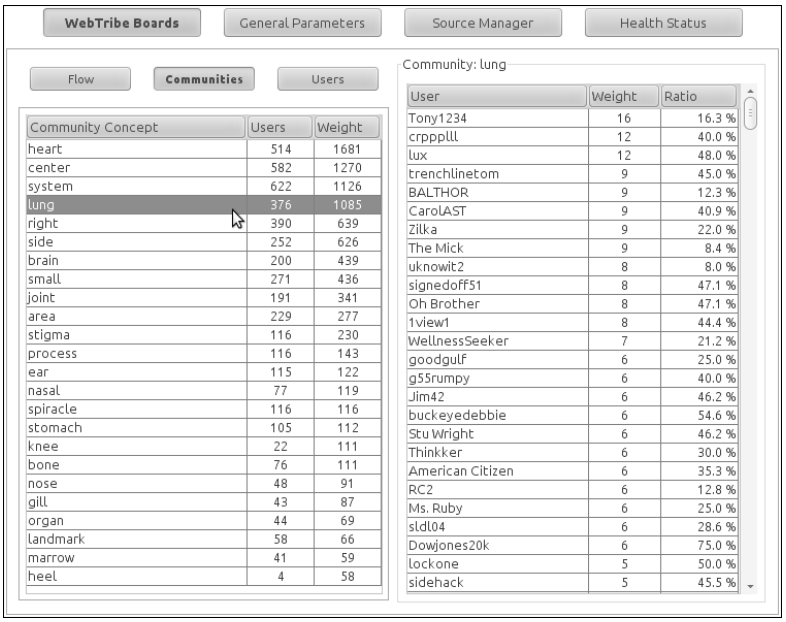


FIGURE D.6 – Applet WebTribe : communautés et utilisateurs

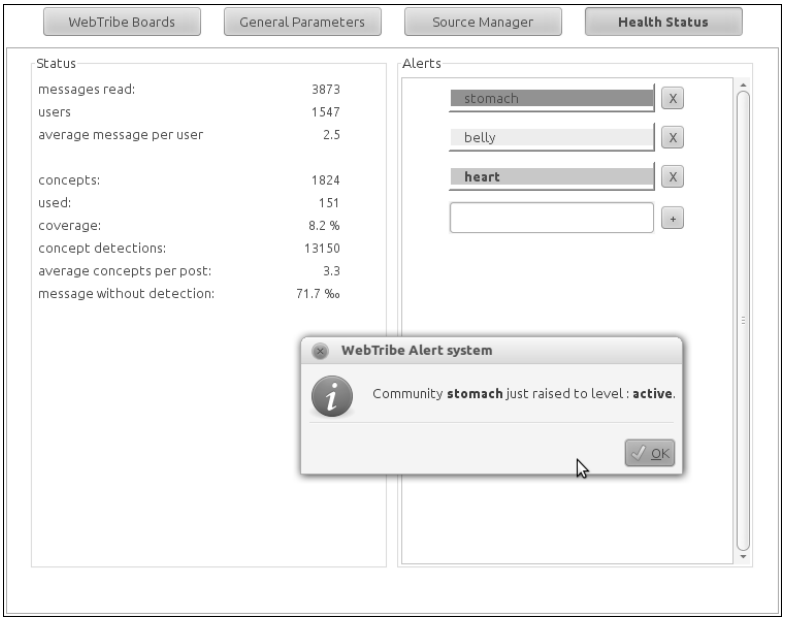
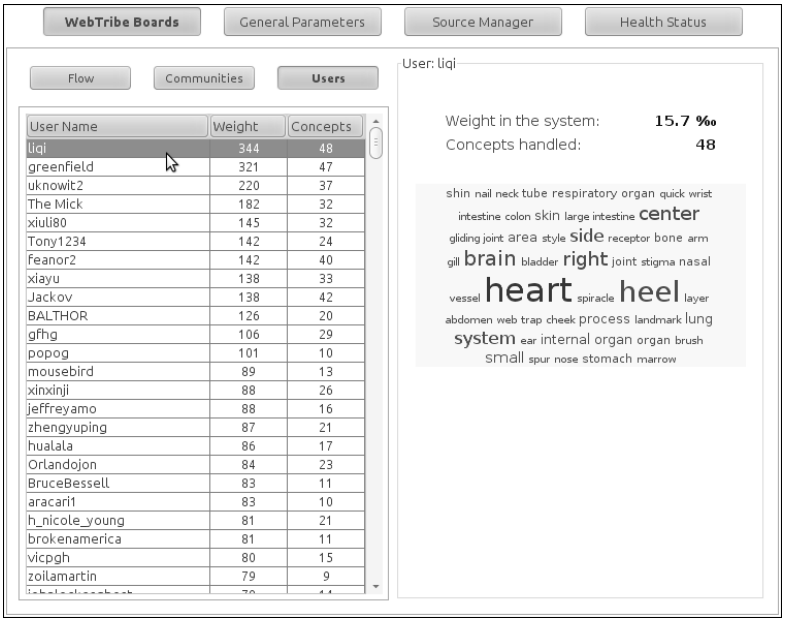


FIGURE D.7 – Applet WebTribe : statistiques et alertes



User Name	Weight	Concepts
liqi	344	48
greenfield	321	47
uknowit2	220	37
The Mick	182	32
xiuli80	145	32
Tony1234	142	24
feanor2	142	40
xiayu	138	33
Jackov	138	42
BALTHOR	126	20
gfhg	106	29
popog	101	10
mousebird	89	13
xinxinji	88	26
jeffreyamo	88	16
zhengyuping	87	21
hualala	86	17
Orlandojon	84	23
BruceBessell	83	11
aracari1	83	10
h_nicole_young	81	21
brokenamerica	81	11
vicpgh	80	15
zoilamartin	79	9

FIGURE D.8 – Applet WebTribe : détails utilisateur

Table des figures

Liste des tableaux

Bibliographie

[1] Lylia Abrouk, David Gross-Amblard, and Damien Leprovost. Découverte de communautés par analyse des usages. In *Proceedings of the EGC 2010 Workshops (Workshop Web social)*, pages A5–5–A5–16, 2010.

[2] R. Albert, H. Jeong, and A. L. Barabasi. The diameter of the world wide web. *Nature*, 401 :130–131, 1999.

[3] L. A. N. Amaral, A. Scala, M. Barthélémy, and H. E. Stanley. Classes of small-world networks. *Proceedings of the National Academy of Sciences*, 97(21) :11149–11152, October 2000.

[4] José Javier Astrain, Francisco Echarte, Alberto Córdoba, and Jesús Villadangos. A tag clustering method to deal with syntactic variations on collaborative social networks. In *Proceedings of the 9th International Conference on Web Engineering*, ICWE '09, pages 434–441, Berlin, Heidelberg, 2009. Springer-Verlag.

[5] D. Auber. Tulip : A huge graph visualisation framework. In P. Mutzel and M. Jünger, editors, *Graph Drawing Softwares*, Mathematics and Visualization, pages 105–126. Springer-Verlag, 2003.

[6] A. L. Barabasi and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439) :509–512, october 1999.

[7] Albert-Laszlo Barabasi, Reka Albert, and Hawoong Jeong. Scale-free characteristics of random networks : The topology of the world wide web. *Physica A*, 281 :69–77, 2000.

[8] Marin Bertier, Rachid Guerraoui, Vincent Leroy, and Anne-Marie Kermarrec. Toward personalized query expansion. In *SNS '09 : Proceedings of the Second*

ACM EuroSys Workshop on Social Network Systems, pages 7–12, New York, NY, USA, 2009. ACM.

[9] Mikhail Bilenko and Matthew Richardson. Predictive client-side profiles for personalized advertising. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '11, pages 413–421, New York, NY, USA, 2011. ACM.

[10] Christian Bizer, Jens Lehmann, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak, and Sebastian Hellmann. Dbpedia - a crystallization point for the web of data. *J. Web Sem.*, 7(3) :154–165, 2009.

[11] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. *Comput. Netw. ISDN Syst.*, 30 :107–117, April 1998.

[12] Ciro Cattuto, Andrea Baldassarri, Vito D. P. Servedio, and Vittorio Loreto. Emergent community structure in social tagging systems. *Advances in Complex Systems (ACS)*, 11(04) :597–608, 2008.

[13] Rudi Cilibrasi and Paul M.B. Vitanyi. Automatic meaning discovery using google. In Marcus Hutter, Wolfgang Merkle, and Paul M.B. Vitanyi, editors, *Kolmogorov Complexity and Applications*, number 06051 in Dagstuhl Seminar Proceedings, Dagstuhl, Germany, 2006. Internationales Begegnungs- und Forschungszentrum für Informatik (IBFI), Schloss Dagstuhl, Germany.

[14] Thomas H. Cormen, Clifford Stein, Ronald L. Rivest, and Charles E. Leiserson. *Introduction to Algorithms*. McGraw-Hill Higher Education, 2nd edition, 2001.

[15] Sanjiv R. Das and Mike Y. Chen. Yahoo ! for amazon : Sentiment extraction from small talk on the web. *Management Science*, 53(9) :1375–1388, 2007.

[16] Munmun De Choudhury, Winter A. Mason, Jake M. Hofman, and Duncan J. Watts. Inferring relevant social networks from interpersonal communication. In *International conference on World wide web (WWW)*, pages 301–310, New York, NY, USA, 2010. ACM.

[17] E. Desmontils and C. Jacquin. Indexing a web site with a terminology oriented ontology. In *SWWS'01 : International Semantic Web Working Symposium*, pages 181–198. IOS Press, 2002.

[18] M. C. Díaz-Galiano, M. Á. García-Cumbreras, M. T. Martín-Valdivia, A. Montejo-Ráez, and L. A. Ure na-López. Integrating mesh ontology to improve medical information retrieval. In *Advances in Multilingual and Multimodal Information Retrieval : 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007, Budapest, Hungary, September 19-21, 2007, Revised Selected Papers*, pages 601–606, Berlin, Heidelberg, 2007. Springer-Verlag.

[19] Yon Dourisboure, Filippo Geraci, and Marco Pellegrini. Extraction and classification of dense communities in the web. In *WWW'07 : Proceedings of the 16th international conference on World Wide Web*, pages 461–470, New York, NY, USA, 2007. ACM.

[20] Jack Edmonds and Richard M. Karp. Theoretical improvements in algorithmic efficiency for network flow problems. *Journal of the ACM*, 19(2) :248–264, April 1972.

[21] M. Eirinaki, M. Vazirgiannis, and I. Varlamis. Sewep : using site semantics and a taxonomy to enhance the web personalization process. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '03, pages 99–108, New York, NY, USA, 2003. ACM.

[22] B. Falissard. *Comprendre et utiliser les statistiques dans les sciences de la vie.* Abrégés (Paris. 1971). Masson, 2005.

[23] Gary William Flake, Steve Lawrence, and C. Lee Giles. Efficient identification of web communities. In *KDD'00 : Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 150–160, New York, NY, USA, 2000. ACM.

[24] L. R. Ford and D. R. Fulkerson. Maximal flow through a network. *Canadian Journal of Mathematics*, 8 :399–404, 1956.

[25] C. Linton Freeman. A set of measures of centrality based on betweenness. *Sociometry*, 40 :35–41, Mars 1977.

[26] C. Linton Freeman. Centrality in social networks : Conceptual clarification. *Social Networks*, 1(3) :215–239, 1979.

[27] H. Fuehres, K. Fischbach, P. Gloor, J. Krauss, and S. Nann. Adding taxonomies obtained by content clustering to semantic social network analysis. *On Collective Intelligence, Advances in Intelligent and Soft Computing*, 76, 2010.

[28] Susan Gauch, Jason Chaffee, and Alaxander Pretschner. Ontology-based personalized search and browsing. *Web Intelli. and Agent Sys.*, 1 :219–234, December 2003.

[29] Eirini Giannakidou, Vassiliki Koutsonikola, Athena Vakali, and Yiannis Kompatsiaris. Co-clustering tags and social data sources. In *Proceedings of the 2008 The Ninth International Conference on Web-Age Information Management*, WAIM '08, pages 317–324, Washington, DC, USA, 2008. IEEE Computer Society.

[30] David Gibson, Jon Kleinberg, and Prabhakar Raghavan. Inferring web communities from link topology. In *HYPERTEXT'98 : Proceedings of the ninth ACM conference on Hypertext and hypermedia : links, objects, time and space—structure in hypermedia systems*, pages 225–234, New York, NY, USA, 1998. ACM.

[31] M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *PNAS*, 99(12) :7821–7826, June 2002.

[32] Peter A. Gloor and Yan Zhao. Analyzing actors and their discussion topics by semantic social network analysis. In *Conference on Information Visualization*, pages 130–135, 2006.

[33] A V Goldberg and R E Tarjan. A new approach to the maximum flow problem. In *Proceedings of the eighteenth annual ACM symposium on Theory of computing*, STOC '86, pages 136–146, 1986.

[34] Michelle Gumbrecht. Blogs as 'protected space'. In *WWW Workshop on the Weblogging Ecosystem : Aggregation, Analysis and Dynamics*, pages 5+, 2004.

[35] Peter M. Haas. Introduction : Epistemic communities and international policy coordination. *International Organization*, 46(1) :1–35, 1992.

[36] Noriko Imafuji and Masaru Kitsuregawa. Effects of maximum flow algorithm on identifying web community. In *WIDM'02 : Proceedings of the 4th international workshop on Web information and data management*, pages 43–48, New York, NY, USA, 2002. ACM.

[37] Hidehiko Ino, Mineichi Kudo, and Atsuyoshi Nakamura. Partitioning of web graphs by community topology. In *WWW '05 : Proceedings of the 14th in-*

ternational conference on World Wide Web, pages 661–669, New York, NY, USA, 2005. ACM.

[38] Hideaki Ishii and Roberto Tempo. Distributed randomized algorithms for the pagerank computation. *IEEE Trans. Automat. Contr.*, 55(9) :1987–2002, 2010.

[39] Jay J. Jiang and David W. Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings on International Conference on Research in Computational Linguistics*, pages 19–33, 1997.

[40] Jason J. Jung. Query transformation based on semantic centrality in semantic social network. *Journal of Universal Computer Science*, 14(7) :1031–1047, 2008.

[41] Joon Hee Kim, Brian Tomasik, and Douglas Turnbull. Using artist similarity to propagate semantic information. In *ISMIR'09 : 10th International Conference on Music Information Retrieval*, pages 375–380, 2009.

[42] Jon M. Kleinberg. Authoritative sources in a hyperlinked environment. In *SODA'98 : Proceedings of the ninth annual ACM-SIAM symposium on Discrete algorithms*, pages 668–677, Philadelphia, PA, USA, 1998. Society for Industrial and Applied Mathematics.

[43] Jean Lave and Etienne Wenger. *Situated Learning : Legitimate Peripheral Participation (Learning in Doing : Social, Cognitive and Computational Perspectives)*. Cambridge University Press, 1 edition, September 1991.

[44] Damien Leprovost. Webtribe : Implicit community clustering by semantic analysis. In *Proceedings of the Fourth International Conference on Advances in Semantic Processing (SEMAPRO)*, pages 186–190, 2010.

[45] Damien Leprovost, Lylia Abrouk, Nadine Cullot, and David Gross-Amblard. Temporal semantic centrality for the analysis of communication networks. In *ICWE*, pages 177–184, 2012.

[46] Damien Leprovost, Lylia Abrouk, and David Gross-Amblard. Discovering implicit communities in web forums through ontologies. *Web Intelligence and Agent Systems*, 10(1) :93–103, 2012.

[47] J. C. R. Licklider and Robert W. Taylor. The computer as a communication device. *Science and Technology*, 76 :21–31, 1968.

[48] Dekang Lin. An information-theoretic definition of similarity. In *Proceedings of the Fifteenth International Conference on Machine Learning*, ICML '98, pages 296–304, San Francisco, CA, USA, 1998. Morgan Kaufmann Publishers Inc.

[49] S. Lozano, J. Duch, and A. Arenas. Community detection in a large social dataset of european projects. In *Workshop on Link Analysis, Counterterrorism and Security (SIAM on Data mining 2006)*, 2006.

[50] Ericka Menchen-Trevino. Blogger motivations : Power, pull, and positive feedback. *Internet Research 6.0*, 2005.

[51] George A. Miller. Wordnet : A lexical database for english. *Commun. ACM*, 38(11) :39–41, 1995.

[52] Gilad Mishne and Natalie Glance. Leave a reply : An analysis of weblog comments. In *WWW06 Workshop on the Weblogging Ecosystem*, 2006.

[53] Marija Mitrović, Georgios Paltoglou, and Bosiljka Tadić. Quantitative analysis of bloggers' collective behavior powered by emotions. *Journal of Statistical Mechanics : Theory and Experiment*, 2011(02) :P02005, 2011.

[54] Mark E. J. Newman, Albert L. Barabási, and Duncan J. Watts, editors. *The Structure and Dynamics of Networks*. Princeton University Press, 2006.

[55] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking : Bringing order to the web. In *Proceedings of the 7th International World Wide Web Conference*, pages 161–172. Elsevier Science Publishers B. V., 1998.

[56] C.A.R. Pinheiro. *Social Network Analysis in Telecommunications*. Wiley and SAS Business Series. John Wiley & Sons, 2011.

[57] D.E. Poplin. *Communities : a survey of theories and methods of research*. Macmillan, 1979.

[58] Erzsébet Ravasz and Albert L. Barabási. Hierarchical organization in complex networks. *Physical Review E*, 67(2) :026112+, February 2003.

[59] Philip Resnik. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th international joint conference on*

Artificial intelligence - Volume 1, IJCAI'95, pages 448–453, San Francisco, CA, USA, 1995. Morgan Kaufmann Publishers Inc.

[60] Matthew Richardson and Pedro Domingos. Mining knowledge-sharing sites for viral marketing. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '02, pages 61–70, New York, NY, USA, 2002. ACM.

[61] Cornelius Rosse and José L. Mejino. A reference ontology for biomedical informatics : the foundational model of anatomy. *Journal of biomedical informatics*, 36(6) :478–500, December 2003.

[62] Maayan Roth, Assaf B. David, David Deutscher, Guy Flysher, Ilan Horn, Ari Leichtberg, Naty Leiser, Yossi Matias, and Ron Merom. Suggesting friends using the implicit social graph. In *KDD '10 : Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 233–242, New York, NY, USA, 2010. ACM.

[63] Alessandra Sala, Haitao Zheng, Ben Y. Zhao, Sabrina Gaito, and Gian Paolo Rossi. Brief announcement : revisiting the power-law degree distribution for social graph analysis. In *Proceedings of the 29th ACM SIGACT-SIGOPS symposium on Principles of distributed computing*, PODC '10, pages 400–401, New York, NY, USA, 2010. ACM.

[64] Gerard Salton and Michael J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc., 1986.

[65] Thomas Schoberth, Jenny Preece, and Armin Heinzl. Online communities : A longitudinal analysis of communication activities. In *Hawaii International Conference on System Sciences*, volume 7, Los Alamitos, CA, USA, 2003. IEEE Computer Society.

[66] Ahu Sieg, Bamshad Mobasher, and Robin Burke. Web search personalization with ontological user profiles. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, CIKM '07, pages 525–534, New York, NY, USA, 2007. ACM.

[67] Edwin Simpson. Clustering Tags in Enterprise and Web Folksonomies. Technical report, Hewlett Packard, 2007.

[68] Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. Yago : A core of semantic knowledge. In *16th international World Wide Web conference (WWW 2007)*, New York, NY, USA, 2007. ACM Press.

[69] John Tang, Mirco Musolesi, Cecilia Mascolo, Vito Latora, and Vincenzo Nicosia. Analysing information flows and key mediators through temporal centrality metrics. In *Proceedings of the 3rd Workshop on Social Network Systems*, SNS '10, pages 3 :1–3 :6, New York, NY, USA, 2010. ACM.

[70] Jeffrey Travers and Stanley Milgram. An experimental study of the small world problem. *Sociometry*, 32(4) :425–443, 1969.

[71] Johan Ugander, Brian Karrer, Lars Backstrom, and Cameron Marlow. The anatomy of the facebook social graph. *CoRR*, abs/1111.4503, 2011.

[72] Etienne Wenger. Communities of practice : Learning as a social system. *Systems Thinker*, 9(5), June 1998.

[73] Zhibiao Wu and Martha Palmer. Verbs semantics and lexical selection. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, ACL '94, pages 133–138, Stroudsburg, PA, USA, 1994. Association for Computational Linguistics.

[74] Valentina Zanardi and Licia Capra. Social ranking : uncovering relevant content using tag-based recommender systems. In *RecSys '08 : Proceedings of the 2008 ACM conference on Recommender systems*, pages 51–58, New York, NY, USA, 2008. ACM.