U N I V E R S I T E L U M I E R E L Y O N 2

REPRESENTATIONS COLLECTIVES, LANGAGE ET SYSTEMES DE TAGGING

MÉMOIRE DE DEUXIÈME ANNÉE DE MASTER
SCIENCES HUMAINES ET SOCIALES

MENTION : SCIENCES COGNITIVES

PARCOURS : RECHERCHE, MÉMOIRE, ÉMOTION, PRISE DE DÉCISION

responsables de la formation:
Professeurs O. KOENIG, A.MAGNAN, I. TAPIERO, R.VERSACE

présenté par:

Bertrand HIGY

réalisé sous la direction des:

Professeur Salima HASSA
et Docteur Leonardo LANA DE CARVALHO

au :

Laboratoire d'Informatique pour l'Entreprise et les Systèmes de Production (LIESP)
EA 4125, Université Claude Bernard Lyon 1

Juin 2010

Table des matières

1. Contexte théorique

La technologie et en particulier le Web sont de plus en plus présents dans notre vie de tous les jours. Que ce soit pour communiquer, s'informer, travailler ou se détendre, faire ses courses même, nous nous reposons de plus en plus sur ces technologies. Au fur et à mesure que les outils évoluent, les systèmes deviennent de plus en plus complexes. Dans le cas du Web, qui voit de très nombreux utilisateurs interagir et au travers d'une structure décentralisée qui plus est, ceci est particulièrement flagrant. Cela amène un besoin grandissant de théories pour comprendre les dynamiques en jeu et guider ce développement. C'est cet objectif que tente d'atteindre la toute jeune science du web (Hendler, Shadbolt, Hall, Berners-Lee, & Weitzner, 2008). Les sciences cognitives ont un rôle particulier à jouer dans ce projet. Nous tenterons pour notre part de montrer comment la connaissance des dynamiques derrière l'émergence de représentations sociales peut être profitable au domaine. Nous présenterons notamment les modèles de l'émergence du langage qui permettent de modéliser ce genre de dynamiques. Enfin nous verrons comment ces idées peuvent être appliquées à l'étude d'un outil en particulier, les systèmes de tagging collaboratifs.

1.1. Web social, Science du Web et Sciences Cognitives

a) Le web social, la nouvelle évolution du web

Alors que les applications sociales fleurissent sur l'Internet (Facebook, Delicious, MySpace, Twitter, ...), la dimension collaborative se révèle de plus en plus importante. On a depuis longtemps dépassé le stade des pages statiques ; l'heure est au dynamique et à l'interaction. Après l'apparition des premiers outils de communication comme l'email, les forums, les wikis ou encore les blogs, la dimension sociale prend une place grandissante dans les outils actuels qui tentent d'exploiter cet aspect au mieux (réseaux sociaux, technique de filtrage collaboratif dans les systèmes de recommandation, systèmes de tagging, ...). Cela a abouti au concept d'informatique sociale ou "social computing" (voir Allen (2004) pour une discussion sur l'émergence de ce domaine et de l'ensemble des termes qui lui ont été donnés). L'idée générale est d'exploiter ce que Wang, Carley, Zeng et Mao (2007) appellent l'intelligence sociale en facilitant la collaboration et l'interaction sociale entre les utilisateurs. Notre société entière repose, et depuis longtemps, sur les interactions sociales et il est logique que cet aspect se retrouve de plus en plus dans les outils informatiques et plus particulièrement le Web.

L'idée générale est bien entendu qu'on est plus efficace à plusieurs que seul. Sans parler des

domaines ou l'expertise de plusieurs personnes est indispensable pour mener à bien une tâche, travailler en groupe présente souvent des avantages. Dans le domaine de la recherche d'information par exemple, Pirolli (2008) montre comment les avantages de la collaboration peuvent compenser les inconvénients pour des groupes de taille restreinte : si la collaboration réduit les chances de négliger une information importante, les coûts du travail en groupe viennent souvent limiter l'efficacité de groupes trop importants. Sur internet par contre, avec les systèmes de tagging par exemple, les coûts sont très limités permettant la collaboration d'un très grand nombre de personne, pour peu que l'on exploite correctement la chose. Ainsi, Chi, Pirolli et Lam (2007) montre comment on peut favoriser la découverte de ressources intéressantes en faisant le lien entre des utilisateurs du même domaine mais faisant partie de communautés différentes et ne partageant pas forcément les mêmes sources d'information.

b) La nouvelle science du Web

Dans le cadre du web, cette évolution a donné naissance à ce qu'on appelle la Science du Web. Dès les débuts du Web, celui-ci a été décrit comme un système complexe. Avec le nombre d'utilisateur croissant et les interactions de plus en plus complexes qu'il supporte, cela est d'autant plus vrai aujourd'hui. Ainsi, les approches classiques en informatiques ne sont plus suffisantes (Hendler et al., 2008) : une analyse poussée du système et de ces dynamiques est requise pour guider le développement des applications. Un exemple fourni par les auteurs est celui des wikis (on pourrait prendre d'autres systèmes de communications comme les forums) : alors que les outils utilisées pour les gérer sont généralement les mêmes, le succès d'un wiki peut-être très variable. Ce succès ne repose donc pas seulement sur des choix technologique mais également sur des dynamiques d'interaction sociale qui expliqueront qu'une communauté va perdurer et une autre non. Ces dynamiques sont beaucoup plus dures à saisir.

Jusqu'à présent, l'approche utilisée pour le développement d'applications web est plutôt la suivante : on produit un outil basé sur une réflexion à un niveau micro (l'interaction d'un utilisateur avec le système) et on observe ce qu'on obtient au niveau macro (interaction d'un grand nombre d'utilisateurs), sans garantie sur le résultat. Pour remédier à cela, Hendler et al. (2008) propose une analyse plus poussée des systèmes impliqués et surtout des dynamiques d'interaction dans le but de guider le développement logiciel. Dans ce but, la science du Web se veut une approche pluridisciplinaire comprenant bien sûr l'informatique mais également les mathématiques, l'intelligence artificielle, la biologie, l'économie et la psychologie pour ne citer qu'eux. L'étude des systèmes complexes et de l'émergence joue bien entendu un rôle central dans l'édifice, de même que

la sociologie et la psychologie sociale, en raison des dynamiques sociales impliquées.

c) Intérêt des sciences cognitives

Comme le laisse supposer le paragraphe précédent, les sciences cognitives qui sont par nature pluridisciplinaires et qui recouvrent beaucoup des disciplines cités ci-dessus, ont un rôle très important à jouer dans la science du Web. De plus en plus d'auteurs soulignent d'ailleurs cette nécessité de faire des liens entre les domaines, que ce soit dans le domaine de la Science du Web comme Hendler et al. (2008) ou celui un peu plus large de l'informatique sociale (Parameswaran & Whinston, 2007; F. Y. Wang et al., 2007).

Les sciences cognitives sont déjà bien présentes dans le domaine de l'ergonomie des logiciels et des interfaces homme-machine. Elles gagneraient à être plus largement introduites dans la conception des systèmes pour aider à comprendre la pensée et le comportement de l'utilisateur (ses représentations, ses attentes, ses buts, son raisonnement, ...). Nous verrons que cela est notamment vrai pour les outils du web social comme les systèmes de tagging collaboratif ou la compréhension de l'utilisateur est un point clé de la modélisation et de la compréhension des systèmes. Enfin, comme nous l'avons déjà dit, la psychologie sociale, qui s'intéresse à la cognition dans le cadre social, peut énormément apporter à la réflexion sur ces outils.

On remarquera que cette idée n'est pas seulement présente en informatique : Hollan, Hutchins et Kirsh (2000) ont proposé d'appliquer la théorie de la cognition distribuée aux interactions homme-machine. La cognition distribuée, qui s'intéresse à la manière dont l'homme réfléchi dans un contexte matériel et social, semble en effet particulièrement adaptée à cette problématique.

1.2. Cognition et représentations collectives

Si les applications sociales sont un des axes de l'équipe dans laquelle s'est déroulé ce stage, un autre aspect de leur travail concerne la question de l'émergence des représentation avec notamment le travail de thèse de Carvalho (2008). Une vision de la représentation comme fondamentalement sociale y est soutenue. La question des dynamiques collective et plus particulièrement des représentations socialement construites et partagées nous a donc semblé particulièrement pertinente à étudier. Voici tout d'abord un aperçu des approches de la notion de cognition collective en sciences cognitives.

a) Tour d'horizon des théories

La notion de représentation est centrale en psychologie, mais traditionnellement, on considère la personne comme un penseur isolé. Ce point de vue à été remis en question récemment et de plusieurs manières :

- la cognition **incarnée** met l'accent sur le corps de l'individu. L'acte de pensée n'est pas abstrait mais fortement influencé par l'inscription corporelle et par l'action. Les travaux de Rizollati et son équipe sur les neurones miroirs (Rizzolatti, Fadiga, Gallese, & Fogassi, 1996) ont beaucoup contribué à cette idée. La théorie de l'enaction (Varela, Thompson, & Rosch, 1992) défend également fortement ce caractère incarné, allant jusqu'à le proposer comme alternative au concept de représentation pour expliquer la cognition.
- la cognition est également **située**, dans le temps et dans l'espace. L'individu est situé dans un environnement avec lequel il interagit et cela façonne notre manière de penser et d'agir.
- la cognition est plus particulièrement **socialement située** : outre l'environnement matériel, nous sommes également entourés d'autres individus qui font partie de nos représentations et les influences. Nous sommes immergés dans des groupes sociaux, suivant des normes socialement construites. C'est ce dernier point qui nous intéresse plus particulièrement.

En psychologie cognitive, de plus en plus de travaux s'intéressent à ces questions. Les termes sont nombreux (cognition partagée, représentation collective ou encore intelligence sociale) et ce qu'ils recouvrent assez vagues. Alors que les termes représentation commune ou partagée désigne généralement une représentation individuelle (bien que partagée par plusieurs personnes), la représentation sociale ajoute une dimension sociale ou la représentation échappe à l'individu. De même, le terme d'intelligence sociale peut ou non, suivant les auteurs, supposer une forme de représentation. On peut en effet inclure sous ce terme des comportements ne supposant pas de représentation, comme ceux des insectes sociaux (voir plus loin).

Nous citerons deux théories qui nous semble particulièrement intéressantes en psychologie cognitive. Certains auteurs proposent tout d'abord la théorie de la cognition située, en mettant l'accent sur l'environnement social, pour expliquer des phénomènes touchant à la cognition sociale comme les stéréotypes (Smith & Semin, 2007). Cette idée a par exemple été appliquée à l'étude de la formation des impressions (Smith & Collins, 2009) nos impressions étant fortement influencés pas l'interaction avec la personne concernée, voire avec un tiers nous parlant de cette personne.

Un autre domaine, celui de la cognition distribuée, comme nous l'avons déjà dit, est particulièrement adapté à l'étude du travail collaboratif (Rogers & Ellis, 1994) et donc des systèmes

homme-machine (Hollan et al., 2000). Ce domaine cherche à étudier comment les processus cognitifs peuvent être répartie entre l'individu, l'environnement matériel (une calculatrice, une étagère de rangement aidant à l'organisation, ...) et l'environnement social (un groupe de travail par exemple). C'est exactement ce qu'on cherche à comprendre avec les applications sociales où les activités sont réparties entre plusieurs utilisateurs et le matériel informatique.

Un autre domaine de la psychologie, la psychologie développementale, s'est également intéressé à l'environnement social. Un précurseur, Vygotski (1978), dont les travaux sont basés sur le développement cognitif de l'enfant et l'importance de l'immersion sociale comme un guide dans ce processus, a inspiré de nombreux auteurs travaillant sur les représentations et la cognition collective.

Un domaine incontournable est bien sûr celui de la psychologie sociale. Ici aussi, alors même que l'accent est mis sur le contexte social, la cognition à longtemps été considéré comme quelque chose d'individuel (Parker, 1987). Les travaux sur l'influence sociale ont particulièrement montré combien cela est faux et que l'individu est fortement influencé par son entourage. Mais la théorie qui se rapproche le plus de l'idée d'une forme de représentation collective est celle de la représentation sociale (Moscovici, 1961), fortement influencée par les travaux de Durkheim (1898). Pour une discussion sur la notion de représentation sociale et de l'articulation entre dimension sociale et individuelle, voir l'article de Voelklein et Howarth (2005).

Les travaux en éthologie fournissent des arguments importants en faveur de la non restriction de l'intelligence sociale à l'humain. Le vol des oies ou les comportements sociaux des insectes, qui ont abouti à l'idée de stigmergie (Grassé, 1959), ont été très étudié et ont beaucoup inspiré l'informatique. Hassas (2003) s'inspire par exemple de la recherche de nourriture dans les colonies de fourmis. Cela a également abouti à ce qu'on appelle la « swarm intelligence » en informatique, où l'on cherche à obtenir un comportement de groupe intelligent avec des agents au comportement très simple.

Nous présentons dans le paragraphe qui suit les principales idées que nous avons tirés de ces travaux et qui ont construit notre vision de la représentation collective.

b) Quelques idées sur les représentations collectives

Tout cela nous a amené à une certaine vision de la représentation sociale, dont voici une rapide synthèse.

Qu'il s'agisse d'agents réactifs (fourmis, oies, neurones) ou d'agents cognitifs (humains), les individus sont plongés dans un environnement matériel et social avec lequel ils sont en interaction constante. De là émerge un certain nombre de représentations. Ainsi, au niveau du cerveau, la représentation émerge de l'interaction de nombreux neurones. De même, on peut envisager que la masse sociale soit le support d'un certain nombre de représentations. Cette représentation a un double support physique, l'un à l'interieur des agents et l'autre sur l'environnement extérieur a l'individu (support matériel, organisation sociale). A l'intérieur des individus, il s'agit d'un système émergent sous-tendu par le système cognitif.

Dans le cas d'agent cognitifs, des interactions ont lieu entre l'individu et son environnement (cf. figure 1) : l'individu externalise ses représentations à travers ses actes et internalise les représentations disponibles dans son environnement (stéréotypes, lexique, ...).

Cette vision correspond en fait assez bien à la définition de la représentation sociale par Jodelet (1989) comme «une forme de connaissance, **socialement élaborée et partagée**, ayant une **visée pratique** et concourant à la **construction d'une réalité commune à un ensemble social**.».

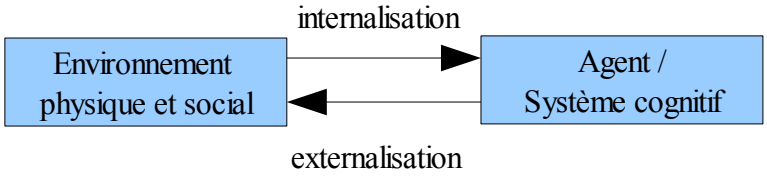


Figure 1 : Schéma des interactions entre l'individu et son environnement : internalisation et externalisation des représentations.

Comprendre les dynamiques en jeux dans la construction et le fonctionnement de ces représentations collectives est essentiel dans le cadre du web social, comme le souligne Markova (2007) : «To facilitate the development of social software, one fundamental issue is the representation of social information and social knowledge. ». Un domaine qui a été particulièrement étudié récemment est celui de l'émergence du langage et des dynamiques collectives qui le sous-tendent.

1.3. L'émergence du langage

Le langage est une forme de représentation collective un peu particulière car il est le support de communication privilégié dans les échanges humains. La construction d'une base sémantique commune pour échanger est donc essentielle.

Récemment, de nombreux travaux se sont intéressé à la modélisation informatique de l'émergence de lexique, de grammaire et plus généralement du sens, permettant d'étudier les dynamiques de construction de représentations collectives en interaction.

a) Les modèles d'émergences du langage

Le modèle le plus connu est le **naming game** de Steels. Nous le présenterons dans sa version simplifié proposée par Dall'Asta et Baronchelli (2006) ou un seul item est négocié. Le principe est simple : soit une population d'agents. A chaque itération, on sélectionne aléatoirement deux individus dans cette population, un locuteur et un auditeur. Le locuteur choisi un des noms qu'il a en mémoire pour l'item et le communique à l'auditeur. Si celui-ci le connait déjà (succès), les deux interlocuteurs se mettent d'accord sur ce nom et mettent à jour leur mémoire pour ne retenir que lui. Sinon (échec), l'auditeur l'ajoute à son lexique.

Ce jeu permet d'étudier la négociation d'un mot pour désigner un item et peu plus généralement de décrire la négociation d'une représentation quelconque et comment elle se répand dans la population. Dans une version plus complète (les Talking Heads), les agents doivent se mettre d'accord sur un lexique entier pour désigner des éléments géométriques présents sur un tableau devant eux (Steels, 2003b). Dans le cas le plus complexe (Kaplan, 2000), les agents parviennent à construire tout un système de représentation et de catégorisation du monde au travers de ce mécanisme de négociation de lexique.

Un autre modèle, celui de Swarup et Gasser (2008b), repose sur un jeu de classification (**classification game**). Les agents doivent ici tenter de classer des items qu'on leur propose. Le but là encore est de voir si les agents finissent par se mettre d'accord sur les labels à utiliser pour désigner les objets.

On citera encore les travaux de Loula (2010) où c'est cette fois-ci une population de singes qui est simulée. Les individus doivent se mettre d'accord sur un répertoire de signes pour désigner différents prédateurs.

Quelque soit le modèle, ces travaux permettent de montrer comment, à partir de mécanismes relativement simple, une population d'agents parvient à négocier des représentations langagières pour converger vers un lexique commun.

b) Lien avec les représentations collectives

Comme les auteurs eux-mêmes le précise (Barrat, Baronchelli, Dall'Asta, & V. Loreto, 2007; Dall'Asta & Baronchelli, 2006; Donetto & Cecconi, 2009; Steels, 2003a; Swarup & Gasser, 2008b) les résultats obtenus à partir de ces modèles sont généralisables à d'autres types de dynamiques sociales (négociation d'opinion, diffusion d'innovation, ...). Après tout, le lexique est une forme de représentation collective et les mécanismes de négociation mis en œuvre dans ces modèles ne sont pas spécifiques au langage. C'est dans cette optique que notre réflexion sur les représentations collectives nous a amené à nous intéresser aux modèles d'émergence du langage. Nous verrons également que ces modèles sont particulièrement proches des systèmes de tagging qui reposent sur l'utilisation de mots pour classer les ressources.

Une question qui revient souvent dans l'étude des dynamiques collectives est celle de l'impact du réseau d'interaction sur la diffusion des représentations. Cette problématique a été beaucoup étudiée dans le cadre du naming game dans sa version simplifiée (Baronchelli, C. Cattuto, V. Loreto, & Puglisi, 2007; Barrat et al., 2007; Dall'Asta & Baronchelli, 2006) mais également avec le classification game (Swarup & Gasser, 2008a) ainsi que d'autres modèles d'émergence de représentations socialement partagées (Donetto & Cecconi, 2009; Ke, Gong, & W. S. Wang, 2008).

c) Parallèle avec les systèmes de tagging

Comme nous l'avons déjà annoncé, nous allons plus particulièrement nous intéresser à un type d'applications web : les systèmes de tagging collaboratifs. Nous ne sommes pas les premiers à faire le rapprochement entre cet outil et les travaux sur l'émergence de représentations ou l'émergence de lexique (Baronchelli et al., 2007).

Swarup et Gasser (2008b), font également le lien entre leur jeu de langage et cet outil. Comme les auteurs l'expliquent eux-même, le classification game est très proche des systèmes de tagging : dans les deux cas, on voit un ensemble d'utilisateurs interagir sur différents types de ressources en leur attribuant des mots-clés. Une question soulevée dans cet article et qui nous semble très pertinente, est celle des connaissances de l'utilisateur. Contrairement aux jeux de langage vus ici, les utilisateurs réels n'arrivent pas vierge de toute connaissance. Ils ont déjà à leur disposition tout un système de représentation qui influence leur manière de taguer.

Plus généralement, on peut analyser l'activité de tagging comme l'émergence d'un système de catégorisation décrivant les différentes ressources. La représentation socialement émergente va alors

permettre d'améliorer l'activité de recherche des utilisateurs. Cet outil est donc un exemple idéal pour étudier les dynamiques de construction de telles représentations en situation réelles.

d) Autres travaux

D'autres travaux nous ont semblé intéressants même si nous n'avons pu approfondir ces différentes pistes. Un domaine très proche de l'émergence de lexique est celui de la gestion de l'hétérogénéité sémantique. On ne part plus cette fois-ci d'agents ignorants : les agents ont déjà un certain nombre de connaissance et le but pour eux est de parvenir à se mettre d'accord sur du sens pour pouvoir communiquer. Un moyen d'y parvenir est l'utilisation de protocole de négociation de sens permettant l'émergence de représentations partagées. On peut citer dans ce domaine les travaux de Sansonnet et Valencia (2005) Mazuel et Sabouret (2008) ou encore Laera et al. (2007). Cette problématique est très importante en informatique également avec des structures de traitement de plus en plus décentralisées (Aberer et al., 2004). De plus en plus de services sont disponibles en réseau, comme avec le Web, amenant le souci de la communication entre ses applications et services. Dans le cas des systèmes de tagging, comme nous l'avons vu, les utilisateurs ont également des connaissances qui peuvent diverger. Ces travaux pourraient ainsi amener des éclaircissements sur la manière dont les utilisateurs peuvent gérer cette hétérogénéité sémantique, si un tel phénomène à lieu dans les systèmes de tagging.
Dans notre équipe, Hassas (2003) s'est également intéressé à l'émergence de sens avec notamment la thèse de Stuber (2007). Le but était de permettre à un utilisateur et un agent informatique de communiquer pour collaborer sur une tâche, par des protocoles de négociation de sens similaire à ceux présentés plus haut. Enfin, dans la continuité des travaux sur l'émergence de lexique, un certain nombre d'auteurs s'intéresse au problème plus complexe de l'émergence de grammaire, tel Steels (2005).

1.4. Les systèmes de tagging

Dans l'idée d'étudier l'émergence des représentations sociales dans le cadre du web, nous nous sommes donc penchés sur les systèmes de tagging. Après une bref aperçu de ce que sont ces systèmes, nous présenterons les propriétés remarquables de ces systèmes qui nous ont amené à nous y intéresser. Enfin, nous feront le tour des modèles existants, en particulier celui de Dellschaft et Staab dont nous nous sommes inspirés pour ce mémoire.

a) Présentation des systèmes de tagging

- Qu'est-ce qu'un système de tagging ?

Les systèmes de tagging sont de plus en plus présents sur le web. Mais qu'est-ce exactement ? En fait, il ne s'agit ni plus ni moins que d'un système de classification de données par mots-clés. Ce qui est nouveau par contre, c'est la dimension sociale qu'apporte le web : chaque utilisateur peut ajouter ses propres mots-clés aux ressources tout en profitant des mots-clés des autres lorsqu'il fait une recherche. De l'interaction de nombreux utilisateurs émerge un système de classification, appelé "folksonomie", qui présente des propriétés intéressantes que nous allons développer plus loin.

Actuellement, les systèmes de tagging collaboratif sont de plus en plus largement utilisés, que ce soit pour classer de la musique (LastFM[1]), des images (Flickr[2]), des liens (Delicious[3]) ou encore des références universitaires (Connotea[4], CiteULike[5]). Les systèmes peuvent toutefois varier dans leur mode de fonctionnement. Delicious, qui est le système de tagging le plus présent dans les études, permet à chaque utilisateur de taguer les ressources. A l'inverse, Flickr, ne permet qu'au propriétaire de l'image d'ajouter des tags. Nous nous intéressons plus particulièrement aux systèmes collaboratifs, comme celui de Delicious, qui permettent aux utilisateurs d'interagir sur les ressources.

- Systèmes de tagging ou taxonomie

Au travers de cette activité collective émerge donc ce qu'on appelle une "folksonomie", en fait un système de classification. Mais un tel système offre-t-il vraiment des avantages intéressant par rapport à des systèmes plus classiques d'organisation et classification ?

Dans un système taxonomique classique, il convient de mettre au point dès le départ une hiérarchie pour la catégorisation des objets. Cette hiérarchie doit permettre de classer l'ensemble des objets de manière unique et non ambigüe. L'avantage est qu'un objet ne peut être classé qu'à un seul endroit et est donc facilement retrouvable. L'inconvénient est qu'il est souvent difficile d'établir ce type de classification a priori et en dehors d'un domaine très restreint.

Le système de tagging permet en revanche de classer les objets sans concertation préalable et sous plusieurs catégories, plusieurs tags pouvant être attribués au même objet. Si ce type de système est rapidement devenu populaire, c'est grâce à sa simplicité d'utilisation (Mathes, 2004). Cela est probablement lié au fait que ce type de classification est plus proche de la manière dont un humain

1 http://www.lastfm.fr
2 http://www.flickr.com
3 http://delicious.com
4 http://www.connotea.org
5 http://www.citeulike.org

catégorise le monde qui l'entoure : les catégories sont souvent floues et un objet peut appartenir à plusieurs catégories (voir les travaux sur la théorie du prototype de Rosch (1975) par exemple). Pour une comparaison un peu plus détaillée de ces deux types de classification, se référer aux textes de Mathes (2004), Shirky (2005) ou eGolder et Huberman (2006).

En terme de recherche, on peut évidemment craindre que le manque de concertation dans les systèms de tagging freine l'efficacité pour trouver une ressource. En cela, l'aspect collaboratif est essentiel : le fait que l'ensemble des utilisateurs tague les ressources assure que les tags reflèteront au mieux le point de vue de cet ensemble de personnes (Shirky, 2005). De plus, comme nous l'avons déjà dit, la coopération améliore la recherche d'information : un groupe à moins de chance de rater une information importante qu'un individu seul (Pirolli, 2008).

Par rapport à d'autres moyens de recherche (moteurs de recherche par exemple), plusieurs travaux ont montré les intérêts particuliers des systèmes de tagging pour rechercher de l'information (Heymann, Koutrika, & Garcia-Molina, 2008; Robu, Halpin, & Shepherd, 2009; Yanbe, Jatowt, Nakamura, & Tanaka, 2007).

- Tagging et construction de représentations collectives

Un grand nombre d'utilisateurs taguent donc chaque ressource, permettant l'émergence d'une forme de représentation collective : les tags les plus populaires pour une ressource donnée sont en fait ceux qui la décrivent pour le plus grand nombre d'utilisateurs. Cette représentation collective qui émerge de manière décentralisée, correspond à un lexique de mots permettant de décrire les ressources. Avec ce point de vue, on se rapproche beaucoup des travaux sur l'émergence de lexique comme nous l'avons déjà expliqué. L'étude de ces mécanisme de négociation de lexique se révèle très intéressante pour comprendre les systèmes de tagging et inversement.

Certains travaux se sont par ailleurs intéressés à la construction d'ontologie (ensemble structuré des termes et concepts) grâce aux données disponibles à partir de systèmes de tagging (Mika, 2007), montrant ainsi la richesse des représentations construites. D'autres travaux encore tentent d'exploiter les informations disponibles à partir de ces systèmes pour améliorer l'efficacité des recommandations de tags (Sen et al., 2006; Xu, Fu, J. Mao, & Su, 2006), des recommandations de ressources lors des recherches (Rupert, 2009; Schenkel et al., 2008; Yanbe et al., 2007), mesurer le degré de similarité entre deux mots (Benz et al., 2008; C. Cattuto, Benz, Hotho, & Stumme, 2008), ou encore détecter des communautés (Rupert, 2009). On voit ainsi toute la richesse de ces représentations sociales qu’on appelle des folksonomies.

b) Dynamiques des systèmes de tagging

De nombreux articles décrivent les propriétés des folksonomies. Nous rappelons brièvement ces propriétés que notre modèle tentera de reproduire. Nous distinguerons les propriétés qui concernent les flux de ressources (ensemble des tags assignés à une ressource donnée) et les flux de co-occurrence (obtenues en prenant l'ensemble des tags co-occurrent à un tag donné, i.e. qui ont été assignés à une même ressource par un même utilisateur).

- Données rang-fréquence

On peut étudier la distribution des fréquences des tags en fonction de leur rang. En ce qui concerne les flux de co-occurrence, Cattuto et al. (2007) ont montré que cette distribution s'approche d'une loi de puissance pour les tags de rang moyen et faible, les tags de rang élevé (1 à 100 environ) présentant une pente plus faible (cf. figure 2). Ces résultats ont été retrouvés pour les flux de ressource de Delicious par Halpin et al. (2007) si ce n'est que la pente plus faible ne concerne que les tout premier tags (1 à 7 environ) suivi par une chute assez abrupt de la fréquence. Si une loi puissance a également été observée pour des corpus de textes (Montemurro & Zanette, 2002), la pente moins importante pour les tags de rang élevé semble liée à des corpus plus restreints. La chute brutale de la fréquence pour les flux de ressources aux alentours du 7e tag serait par contre liée à l'interface de Delicious (Dellschaft & Staab, 2008).

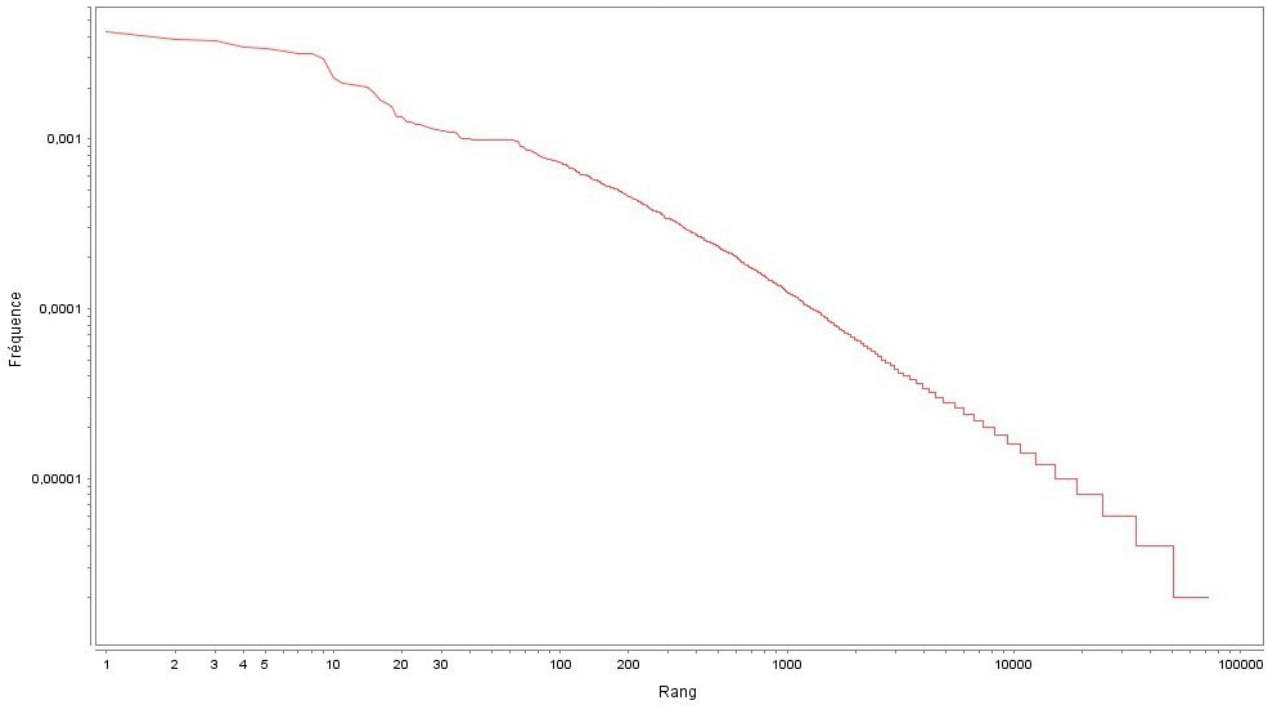


Figure 2 : Distribution des fréquences en fonction du rang pour les tags co-occurent au tag ajax sur Delicious, à partir des données de Dellschaft et Staab.

- Croissance du nombre de tag

Cattuto, Baldassari, Serviedo & Loreto (2007) ont montré que la croissance du nombre de tags différents ne se fait pas de manière linéaire mais sous-linéaire (cf. figure 3).

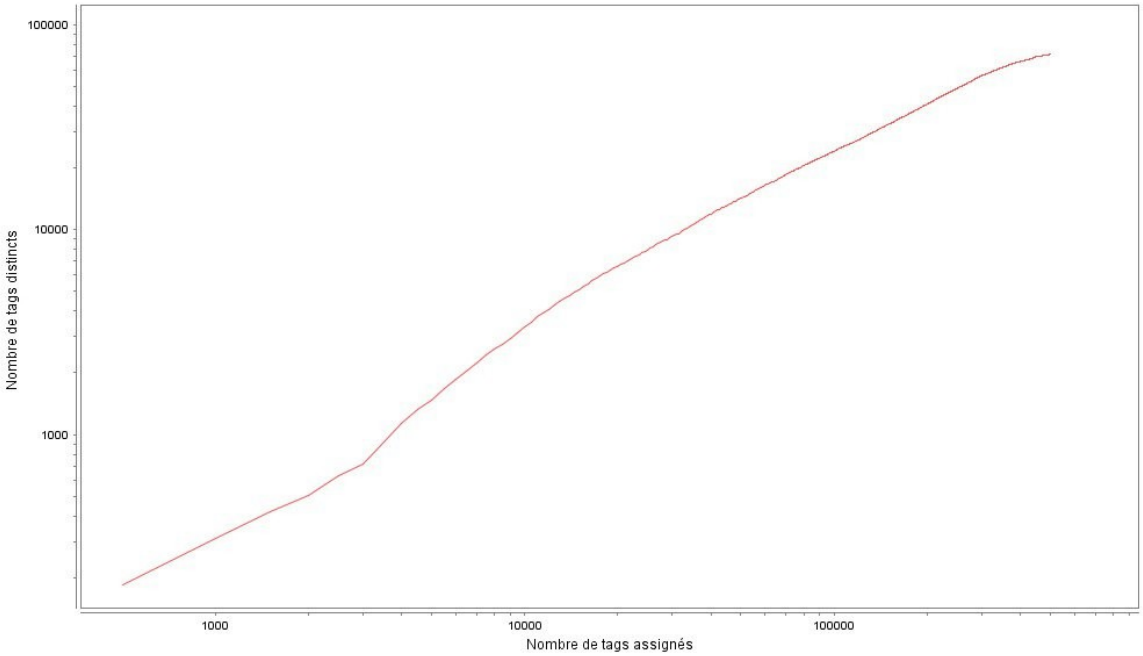


Figure 3 : Croissance du nombre de tags distincts en fonction du nombre de tags total assignés pour les tags co-occurrents au tag ajax sur Delicious, à partir des données de Dellschaft et Staab.

- Convergence des proportions de tags

Un dernier résultat qui nous semble essentiel est la convergence rapide des proportions des tags. Golder et Huberman (2006) ont mis en évidence le fait que les proportions relatives des 25 tags les plus fréquents pour une ressource se stabilisent très rapidement (dès les 100 ou 200 premiers tags). Cela montre qu'en effet un système de catégorisation émerge, les utilisateurs se mettant d'accord sur les tags permettant de décrire les ressources (comme le montre la stabilisation des proportions). Cela n'est d'ailleurs pas sans rappeler la convergence du vocabulaire dans les travaux sur l'émergence de lexique.

Il est donc regrettable que le modèle de Dellschaft et Staab (qui est tout de même le premier à vérifier les propriétés précédemment cités sur le distribution rang-fréquence et la croissance des tags) ne reproduise pas cette dynamique. Au contraire, les proportions sont très instables et continuent à évoluer même très tard au cours de la simulation. Cela est certainement dû au phénomène d'imitation et à la manière dont il est modélisé par les auteurs.

c) Aperçu des modèles existants

Avant d'en venir enfin au modèle que nous proposons pour simuler l'activité de tagging, il convient de faire un rapide tour d'horizon des modèles existants. Après une brève description de leur fonctionnement, nous essaierons de dégager leurs principaux points positifs et négatifs, et en particulier les propriétés connues des folksonomies (décrites ci-dessus) qu'ils permettent de reproduire. Nous détaillerons particulièrement le dernier modèle, celui de Dellschaft et Staab (2008) dont nous nous sommes inspirés et avec lequel nous comparerons nos résultats.

- Le modèle de l'urne – Golder et Hubermann (2006)

Golder et Huberman propose d'expliquer l'activité des utilisateurs à partir de deux facteurs : les connaissances partagées (les utilisateurs taguent en fonction de leur connaissances et du contenu du document) et l'imitation (les utilisateurs peuvent imiter les tags des autres utilisateurs). Si cette idée est intéressante, dans les faits, le modèle proposé par les auteurs rend compte de l'imitation seulement. Les auteurs reprennent en fait un modèle stochastique proposé au départ par Eggenberg et Polya (1923) et qui fonctionne de la manière suivante : soit une urne contenant n boules de couleur (une par item, en l'occurrence une par tag). A chaque tour, on pioche une boule au hasard et on la replace dans l'urne avec une autre boule de la même couleur.

L'intérêt de ce modèle est que très rapidement, la proportion de chaque type de boule dans l'urne converge vers une certaine valeur, à la manière des proportions des tags pour une ressource. Ce modèle explique donc bien le phénomène de stabilisation des proportions des tags, ce qui était son but. On remarque toutefois que le tirage des boules (le choix des tags) se fait de manière aléatoire. Avec ce modèle, les connaissances de l'utilisateur n'entre pas dans le choix tag contrairement à ce propose les auteurs (idées que l'on retrouve chez Swarup et Gasser (2008b)). Les connaissances des utilisateurs influencent pourtant très certainement le choix des tags, même s'il y a aussi une part d'imitation. Il est en revanche peu vraisemblable que les utilisateurs choisissent aléatoirement les tags qu'ils imitent et cela conduirait à un système qui n'aurait aucun intérêt dans la pratique, le but étant tout de même de pouvoir retrouver les ressources ultérieurement.

- Le modèle Yule-Simon avec mémoire à long terme – Cattuto, Loretto et Pietronero (2007)

Ce second modèle est une variante du modèle Yule-Simon (Simon, 1955; Yule, 1925) qui décrit la construction d'un texte à partir de zéro. Pour cela, à chaque tour, on choisit si l'on ajoute un nouveau mot (avec une probabilité p), ou si l'on copie l'un des mots précédents (probabilité 1 – p) en fonction

de sa fréquence d'occurrence dans le début du texte. Le même processus peut être utilisé pour construire un flux de tags, engendrant une distribution rang-fréquence qui suit une loi de puissance. Toutefois, pour s'approcher de la distribution rang-fréquence observée avec des données réelles, Cattuto et al. ajoutent une mémoire à long terme avec un déclin progressif pour l'accès aux précédents tags. Ainsi, la probabilité de copier un précédent tag décroît avec sa position dans le flux, les tags les plus récents étant ceux qui ont le plus de chance d'être sélectionnés. Cette hypothèse est supportée par des travaux en psychologie cognitive (Anderson, 2000) qui montre qu'une loi puissance pour la latence et la fréquence modélise bien la mémoire humaine.

Contrairement au modèle précédent, celui-ci se concentre sur la modélisation de la distribution rang-fréquence des tags co-occurrent avec un certain tag (les données portent sur Delicious et les tags "blog", "ajax" et "xml"). S'il reproduit relativement bien ces données, il ne permet pas par contre de modéliser la distribution rang-fréquence pour une ressource seule. Or, les utilisateurs taguent des ressources et les données sur les co-occurrences sont extraites de l'ensemble de ces ressources. Le modèle, s'il capture quelque chose de la dynamique globale, ne parvient donc pas à reproduire le comportement de l'utilisateur et n'explique pas comment ce comportement engendre la dynamique globale. Il se révèle ainsi d'une utilité limité si l'on cherche à comprendre le comportement de l'utilisateur.

- Un modèle tenant compte de la valeur informationnelle – Halpin, Robu et Shepherd (2007)

Ce modèle est le premier à introduire la valeur informationnelle dans le choix du tag : on ne tague plus dans le seul but de taguer mais les tags sont choisis en fonction de leur efficacité pour retrouver de l'information. Comme je l'ai déjà dit, un système de tagging comprend deux aspects : le tagging des ressources proprement dit et la recherche de ressource. Les deux aspects sont liés, or, la plupart des modèles se concentre sur le tagging uniquement. Comme dans le précédent modèle, l'utilisateur peut imiter un tag (probabilité p), ou ajouter un nouveau tag (probabilité 1 – p). Par contre, s'il choisit d'imiter un tag, le choix ne se fera plus seulement en fonction de la fréquence d'occurrence mais également en fonction de l'efficacité de ce mot pour retrouver la ressource désirée (cf. l'article original pour les détails du calcul).

Ce modèle conduit à une distribution des fréquences qui suit une loi puissance stricte et une croissance du nombre de tags qui est linéaire.

- Modèle épistémique dynamique – Dellschaft et Staab (2008)

Dellschaft et Staab partent de l'idée suivante : si les autres modèles sont inaptes à reproduire un

certain nombre de propriétés des systèmes de tagging, c'est parce qu'ils n'incluent pas les connaissances des utilisateurs dans leurs modèles. Ils suivent en ce sens l'intuition initiale de Golder et Huberman (2006) ainsi que l'idée de Swarup et Les Gasser (2008b). A partir du modèle qu'ils développent autour de cette idée, ils espèrent reproduire a) les distributions de fréquence pour les flux de co-occurrence et les flux de ressources, b) la croissance sous-linéaire du nombre de tags.
Partant à nouveau des tags "ajax", "blog" et "xml", et pour modéliser les connaissances de l'utilisateur, les auteurs extraient le texte d'un grand nombre de sites sur ces trois thèmes. Ils construisent alors le lexique en extrayant les mots présents dans ces textes et en associant à chaque mot sa fréquence d'occurrence dans l'ensemble du corpus.
A partir de là, le modèle fonctionne comme suit : à chaque tour, on choisit si l'utilisateur va imiter un tag (avec une probabilité p) ou ajouter un nouveau tag (probabilité 1 – p). S'il ajoute un nouveau tag, on utilise aléatoirement un mot du lexique, la probabilité de tirer un mot étant proportionnelle à sa fréquence dans le corpus. Dans le cas de l'imitation, on établie une liste de tags recommandés (comme dans les systèmes réels) selon la méthode suivante : on choisit les n tags les plus fréquents (n = 50 ou 100 pour les flux de co-occurences qui regroupent un grand nombre de ressource, et 7 pour les flux de ressources) sur les h derniers tags attribués (h = 1000 pour les flux de co-occurrence et 300 pour les flux de ressources). On choisit alors aléatoirement un tag parmi les recommandations, en fonction de leur fréquence d'apparition dans la fenêtre h considérée.
Au niveau des résultats, le modèle reproduit très bien les flux de co-occurrences, les flux générées avec le modèle étant proches des données extraites de Delicious. Les courbes rang-fréquence obtenues suivent l'allure caractéristique des courbes obtenues à partir de données réelles et la croissance du nombre de tags est bien sous-linéaire. Du côté des flux par ressource, les résultats sont un peu moins bon : on obtient bien la chute caractéristique de la fréquence aux alentours de la 7e ressource mais les flux par ressources ayant une variabilité plus importante, le modèle est plus éloigné des données réelles. Le résultat qui n'est pas du tout reproduit par le modèle, celui qui nous intéresse le plus, est la stabilisation des proportions des tags pour une ressource.
On remarque par ailleurs que la manière dont le comportement de l'utilisateur est modélisé est assez artificielle : l'utilisateur ne va vraisemblablement pas choisir au hasard s'il ajoute un nouveau tag ou en imite un.

1.5. Synthèse

a) But de la recherche

Le but de ce travail est donc de chercher à comprendre les dynamiques des systèmes de tagging, un

nouvel outil social de plus en plus utilisé sur le Web. Pour cela, nous avons montré les besoins grandissant d'une approche pluridisciplinaire incluant en particulier les sciences cognitives, et avons proposé une approche basée sur les notions de cognition et de représentations collectives. Nous avons également fait le parallèle entre représentations collectives, modèles d'émergence du langage et systèmes de tagging. Dans les modèles actuels, les connaissances de l'utilisateur et le phénomène d'imitation sont proposés comme mécanismes explicatifs des dynamiques observées. Pourtant, suivant l'idée qu'une représentation collective est construite, on peut penser que les utilisateurs sont influencés par cette représentation et internalise le vocabulaire. Cette intuition est également présente dans la littérature (Mathes, 2004). Pour tester nos hypothèses, nous proposons à notre tour un modèle du comportement de l'utilisateur.

b) Hypothèses de travail

Voici maintenant les hypothèses que nous tenterons de vérifier avec ce modèle.

Nous pensons tout d'abord qu'un modèle plus réaliste du comportement de l'utilisateur est capable de mieux rendre compte de dynamiques observées. Comme nous l'avons souligné, le modèle de Dellschaft et ceux qui l'ont précédé, semblent un peu réductionnistes. En particulier, ils déconnectent complètement le comportement d'imitation des connaissances de l'utilisateur. Nous pensons que les sciences cognitives peuvent contribuer à l'élaboration de modèles plus réalistes et permettant de mieux rendre compte de ce qui se passe réellement. Comme une première étape dans cette direction, nous proposons un modèle où les connaissances de l'utilisateur interagissent avec les tags recommandés pour déterminer le choix des tags et tenterons de reproduire les données de Delicious avec ce modèle. En particulier, nous espérons obtenir une meilleure convergence du vocabulaire avec ce modèle.

H_1 : Notre modèle permet de reproduire les distributions de fréquences pour les flux de co-occurrence ainsi que la croissance du nombre de tags.

H_2 : Contrairement à la plupart des modèles existants, nous espérons obtenir une convergence rapide des proportions des tags

Nous pensons que cette convergence du vocabulaire peut s'expliquer par les connaissances communes des utilisateurs ainsi qu'un phénomène d'internalisation du vocabulaire employé par les utilisateurs.

Nous tenterons de montrer comment chacun des deux facteurs mentionnés, connaissances

communes et internalisation, concourt à la convergence des proportions de tags dans les flux de ressource.

H_3 : Les connaissances communes sont un facteur expliquant la convergence du vocabulaire.

H_4 : Le phénomène d'internalisation accélère cette convergence.

c) Validation des hypothèses

Il est courant dans la littérature de proposer une justification purement visuelle (sous forme de graphiques) des résultats. Pour valider nos hypothèses, nous essaierons pour notre part de proposer des tests statistiques permettant de confirmer les constatations purement visuelles. Nous analyserons donc les données obtenues avec les deux modèles (celui de Dellschaft et le notre) et celles de Delicious. Nous comparerons pour cela les courbes de rang-fréquence pour les flux de co-occurrence au moyen du test de Kruskal-Wallis. Le test de Mann-Whitney permettra d'analyser les courbes 2 à 2. Les courbes de croissance des tags, dont les données peuvent être appariées sur le nombre de tags total seront analysées au moyen du test de Friedman. Les analyses 2 à 2 seront faites au moyen du test de Wilcoxon. Enfin, nous mesurerons également la convergence des proportions de tags sur les flux de ressources au moyen de la mesure de divergence de Kullback-Leibler (décrite ci-dessous). Les données pouvant là aussi être appariées sur le nombre de tags ajoutés, nous comparerons les différentes conditions grâces aux tests de Friedman et de Wilcoxon.

2. Simulations

2.1. Modélisation

Nous allons maintenant présenter dans le détail notre modèle, qui servira à modéliser les flux de co-occurence aussi bien que les flux de ressources. Ce modèle simule l'ajout de tags par un utilisateur standard. Pour simplifier, on suppose qu'un seul tag est ajouter à chaque fois. Dans le cas de la simulation d'un flux de co-occurrence, il s'agirait de l'ajout d'un tag en même temps que le tag considéré, sur une ressource quelconque. Dans le cas d'un flux de ressource, il s'agirait de l'ajout d'un tag à la ressource considérée. Deux choses interviennent dans le choix du tag : les connaissances de l'utilisateur et les tags reccommandés. Nous allons décrire comment ces deux aspects sont rendus ici.

a) Modélisation de connaissances

Pour modéliser les connaissances de l'utilisateur, nous avons repris les données de Dellschaft et Staab disponible en ligne[6]. Pour les obtenir, les auteurs ont récupéré le contenu d'un ensemble de pages Web portant sur les thèmes *ajax*, *blog* et *xml* (termes sur lesquelles portent les flux de co-occurence qu'ils ont extraits de Delicious). A partir de ce corpus, ils ont calculé les fréquences d'occurence des mots dans le texte, ces fréquences étant sensé refléter la probabilité que les différents mots soit utilisé simultanément au tag considéré par un utilisateur. Ces données sont ainsi sensées modéliser le vocabulaire d'un utilsateur moyen, vocabulaire innaccessible en réalité. Ce procédé semblant efficace (Dellschaft & Staab, 2008), nous avons réutilisé les mêmes lexiques pour notre modèle. Cela nous permet de plus de comparer nos résultats avec ceux des auteurs sur les mêmes données.
Pour les flux de ressources (les ressources porte sur la technologie ajax), Dellschaft et Staab prenne la modélisation des connaissance sur le terme *ajax* comme approximation des connaissances sur les ressources. Nous ferons de même.

b) Recommandation des tags

Dans les systèmes de tagging, lorsque l'utilisateur veut taguer une ressource, un certain nombre de tags lui sont généralement recommandés. Il s'agit des tags les plus fréquents parmi les derniers tags ajoutés. Deux paramètres permettent de contrôler ceci :

6 http://isweb.uni-koblenz.de/Research/Tagdataset

- ***n*** : correspond au nombre de tags recommandés. Pour une ressource, 7 tags sont proposés sur Delicious. Pour les flux de co-occurrence, qui correspondent à un plus grand nombre de ressources, Dellschaft et Staab ont montré que n = 50 ou 100 (en fonction du flux) sont des valeurs adéquates. Nous reprendrons ces valeurs dans notre modèle
- ***h*** : correspond au nombre de tags sur lesquels la fréquence est calculé. En effet, seul les derniers tags ajoutés sont utilisés pour calculer la fréquence d'utilisation des différents tags. Dellschaft et Staab ont montré que les valeurs h = 300 et h = 1000 sont adaptés pour simuler, respectivement, les flux de ressources et les flux de co-occurrence. Nous utilserons ces valeurs également.

c) Influence des tags recommandés

Voulant modéliser l'influence des tags recommandés de manière un peu plus plausible, notre modèle s'éloigne ici des précédents modèles. Nous ne distinguons pas l'imitation de tags de l'ajout d'un nouveau tag. Dans tous les cas, l'utilisateur va choisir le tag à partir de ces connaissances, mais celle-ci vont être influencée par les tags recommandés (un peu comme dans un amorçage). Nous modélisons cela en augmentant la fréquence des tags reccommandés dans le vocabulaire.
Pour cela, nous introduisons un nouveau paramètre : Δ_{CT}. Ce paramètre détermine l'importance du biais en faveur des tags recommandés. Ainsi, plus Δ_{CT} est élevé, plus le phénomène d'imitation est important et plus les tags recommandés ont des chances d'être choisis préférentiellement. Concrètement, nous augmentons la valeur du nombre d'occurence du mot dans le corpus de Δ_{CT} x O_T, où O_T désigne le nombre total d'occurence des mots du corpus. Cette modification est bien sûr transitoire et n'affecte pas les itérations suivantes.

d) Internalisation de la représentation collective

Nous pensons que l'utilisateur internalise dans une certaine mesure le vocabulaire employé par les autres, adaptant son propre vocabulaire pour suivre le collectif. Pour modéliser cela, nous suivons la solution proposé pour le phénomène d'imitation, en modifiant la probabilité de sélectionner un tag donné. Nous introduisons donc un deuxième facteur : Δ_{LT} . La modification est cette fois définitive, introduisant des modification à long terme. Le biais est appliqué pour les tags des autres utilisateurs que l'utilisateur peut voir (les tags reccomandés) ainsi que les tags qu'il ajoute lui-même. On s'attend bien évidemment à des valeurs de Δ_{LT} plus faible que pour Δ_{CT}, les effets se cumulant d'une itération à l'autre.

2.2. Résultats

a) Paramétrage

- flux de co-occurence

Une première série de simulations nous a servi à paramétrer notre modèle. Si les valeurs *n* et *h* ont été reprises du modèle de Dellschaft et Staab, les valeurs adéquates pour les paramètres Δ_{CT} et Δ_{LT} nous étaient inconnues.

Le modèle a donc été testé pour les flux de co-occurence sur les tags *ajax*, *blog* et *xml*. Les données sont comparées au données extraites de Delicious par Dellschaft et Staab, données également disponibles sur internet (à la même adresse que celle citée précédemment). La figure 4 montre par exemple les résultats obtenus pour *ajax* avec différentes valeurs de Δ_{CT}, ainsi que la courbe obtenue à partir de Delicious qui correspondant au résultat attendu. On voit que les meilleurs résultats sont obtenus pour $\Delta_{CT} = 0.01$.

Le procédé a été réitéré pour les tags *blog* et *xml*, pour différentes valeurs de Δ_{CT} et Δ_{LT}. Les valeurs choisies pour simuler les trois flux sont résumées dans le tableau 1.

- flux de ressource

Les flux de ressources extraits par Dellschaft et Staab n'étant pas disponible, nous nous contenterons d'estimer les valeurs Δ_{CT} et Δ_{LT} qui permettent d'obtenir une baisse caractéristique de

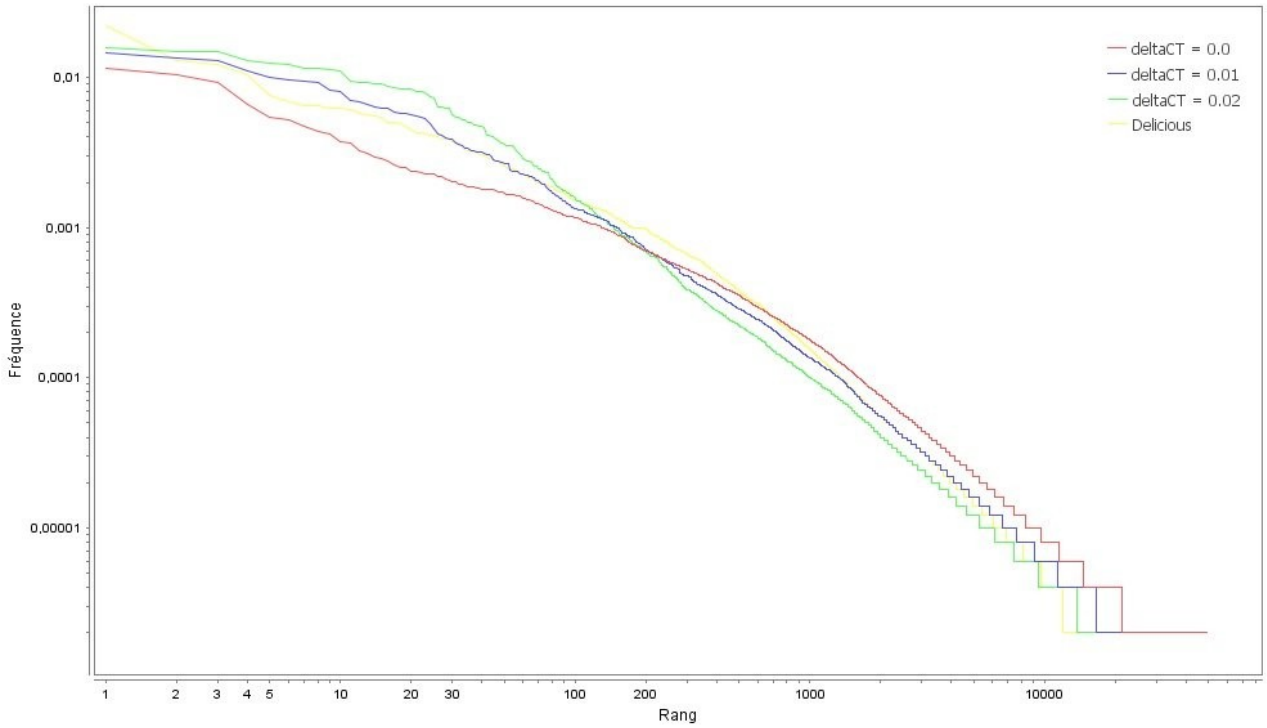


Figure 4 : Distribution rang-fréquence pour le tag ajax pour différentes valeurs de Δ_{CT} avec notre modèle et pour Delicious

Tableau 1: Valeurs des paramètre n, h, Δ_{CT}, Δ_{LT}, utilisés pour simuler les différents flux. La taille du flux simulé est également précisée

Flux	n	h	Δ_{CT}	Δ_{LT}	Taille du flux
ajax	50	1000	0.01	0.0	500000 tags
blog	100	1000	0.0	0.0	500000 tags
xml	50	1000	0.0	0.00001	500000 tags
ressource seule	7	300	0.02	0.001	17000 tags

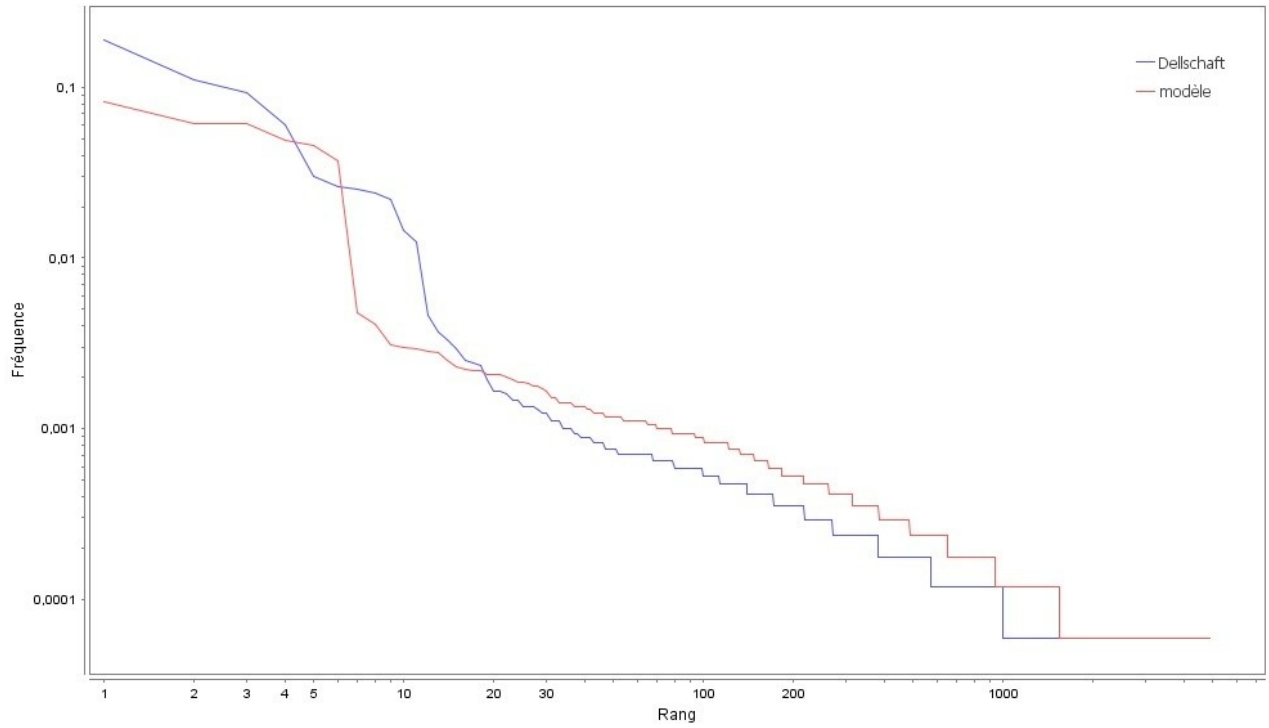


Figure 5 : Distribution rang-fréquence pour les flux de ressource simulés avec notre modèle et celui de Dellschaft.

fréquence autour de 7^{e} tag (cf. figure 5). Un flux généré par le modèle de Dellschaft et Staab nous a également servi de repère. Les valeurs que nous avons gardées sont également présenté dans le tableau 1.

b) Données rang-fréquence

Les données rang-fréquence ont été récupérés avec les deux modèles pour les flux de co-occurences. La figure 6 montre les résultats obtenus sur une simulation avec chacun des deux modèles ainsi qu'avec les données tirées de Delicious, pour les trois termes. On constate que notre modèle génère bien une distribution qui suit une loi puissance avec une pente plus faible pour les rangs élevés. Il semble même plus proche des données de Delicious que le modèle de Dellschaft.

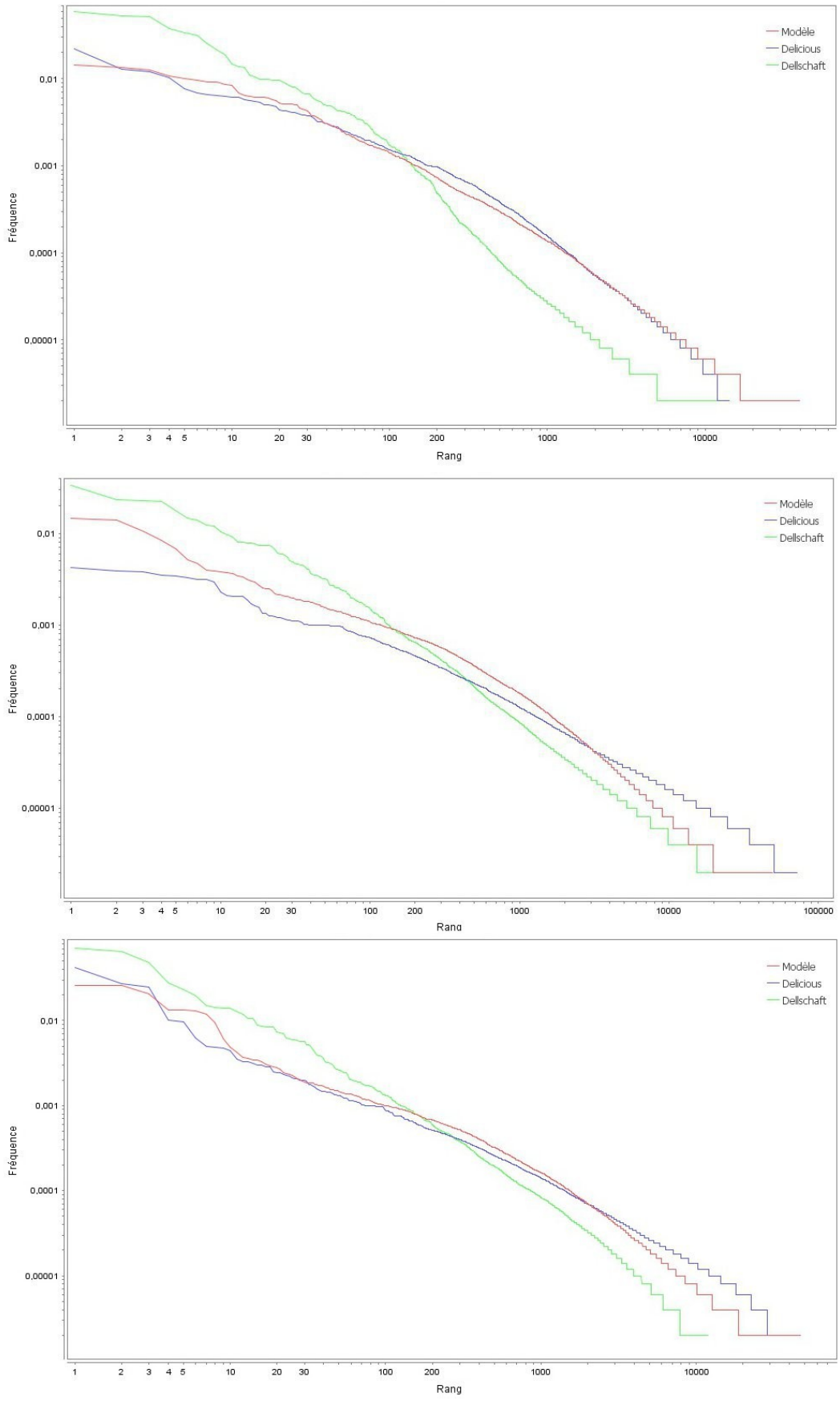
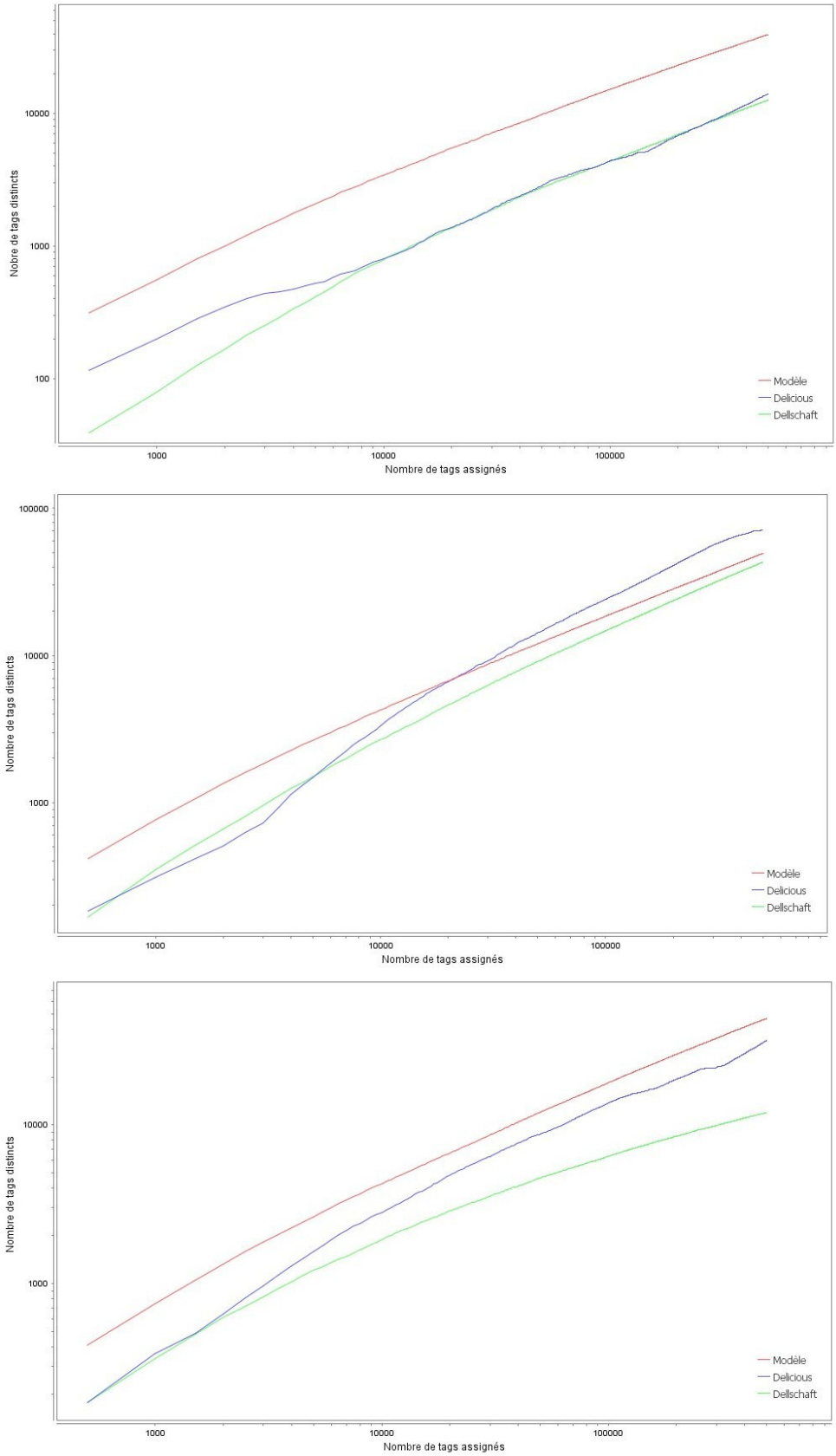


Figure 6 : Données rang-fréquence obtenues à partir de notre modèle, celuis de Dellschaft et Staab ainsi que le site Delicious, pour les tags co-occurent aux tags ajax (en haut), blog (au milieu) et xml (en bas).



Ifigure 7 : Croissance du nombre de tags distincts en fonction du nombre total de tags assignés, pour notre modèle, celui de Dellschaft et Staab et pour les flux issues de Delicious. Les flux correspondent aux co-occurrences pour le tag ajax (en haut), blog (au milieu) et xml (en bas).

Afin de vérifier statistiquement ce résultat, nous avons moyenné les données obtenus sur 10 simulations pour chaque modèle et chacun des 3 tags. Nous avons alors appliqué le test de Kruskal-Wallis selon le plan $T*M_3$ où T désigne le facteur aléatoire Tag et M le facteur à mesure indépendantes Modèle à 3 modalités (notre modèle vs. Dellschaft vs. Delicious). La variable dépendante est la fréquence du tag dans le flux.
On note un effet significatif du modèle pour les trois termes ($p < .01$). Une analyse deux à deux avec le test de Mann-Whitney confirme une différence significative ($p<.05$) entre les 3 modèles, sauf pour *ajax* notre modèle et celui de Dellschaft ne diffère pas significativement. On peut donc conclure que notre modèle reproduit mieux les distributions rang-fréquence de Delicious pour *blog* et *xml*, et fait aussi bien que le modèle de Dellschaft pour *ajax*.

c) Croissance du nombre de tags distincts

Avec les données obtenues, nous avons également regarder la croissance du nombre de tags distincts en fonction du nombre total de tags ajoutés. Les mesures sont prélevées tout les 2500 tags ajoutés ce qui fait 200 mesure par flux. La figure 7 montre les résultats obtenus à partir d'une simulation avec chacun des deux modèles, ainsi que les données tirées de Delicious, pour les trois termes. Le résultat est un peu moins clair cette fois-ci. Pour *ajax* et *xml*, notre modèle semble s'en sortir moins bien. Pour le tag *blog* par contre, il est difficile de départager les deux modèles. On constate par contre que notre modèle génère bien une croissance sous-linéaire.

L'analyse statistique des résultats se fait toujours sur la moyenne de 10 simulations. Les résultats pouvant être appariées en fonction du temps, le test de Friedman a été appliqué aux données des 3 modèles. La variable dépendante est le nombre de tags distincts dans le flux.
On note un effet significatif du modèle pour les trois termes ($p < .00001$). Une analyse deux à deux avec le test de Wilcoxon confirme une différence significative ($p < .05$) entre les 3 modèles. A l'exception du tag blog, notre modèle semble donc s'en sortir moins bien sur la croissance du nombre de tags distincts.

d) Convergence des proportions de tags

Le principal résultat que nous cherchons à reproduire est la convergence des proportions de tags sur les flux de ressources (à titre d'exemple, la figure 8 montre l'évolution des proportions pour un flux simulé avec chacun des modèles). Nous avons donc comparé notre modèle à celui de Dellschaft et Staab au moyen de la mesure de divergence de Kullback-Leibler. Comme Halpin et al.(2007)

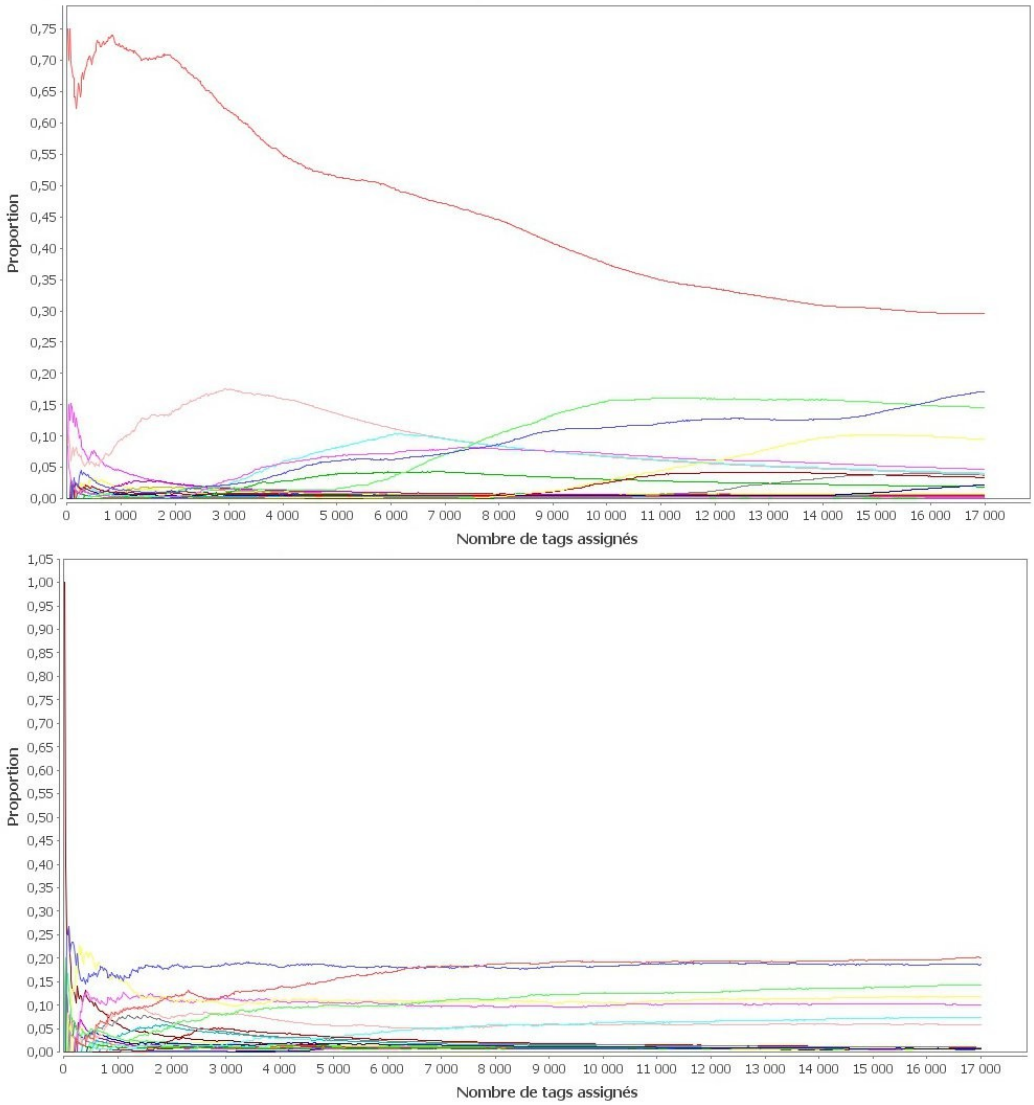


Figure 8 : Evolutions des proportions reatives pour les 25 tags les plus fréquents, pour les flux de ressources simulées avec notre modèle et celui de Dellschaft.

l'expliquent, cela permet de mesure la divergence entre deux vecteurs de données, ici entre les vecteurs de proportions à deux instincts distincts. Chaque vecteur correspond aux proportions relatives des 25 tags les plus fréquents dans le flux.

La divergence de Kullback-Leibler peut être appliquée de deux manières :

- entre deux vecteurs consécutifs (D_C). On mesure ainsi le niveau de convergence de manière temporaire.
- entre un vecteur et le vecteur final (D_F). Le vecteur final est considéré comme l'état à atteindre et on mesure la convergence vers cet état.

Les mesures de divergence ont été effectuées tout les 1700 tags (1% du flux) et les 2 premiers

vecteurs ont été ignorés car trop aléatoires.

Pour conclure que les proportions sont stables, ces deux mesures doivent être suffisament faibles. La figure 9 montre la divergence moyenne sur 10 simulation pour les deux modèles et pour les deux types de mesures. La divergence entre deux vecteurs consécutifs atteint rapidement des valeurs très faibles pour les deux modèles, indiquant que les proportions évoluent peu entre deux intervalles de temps consécutifs.

On constate en revanche que notre modèle converge bien plus rapidement vers l'état final, confirmant l'impression donnée par la figure 8. On peut d'ailleurs se demander si l'état final est stable dans le cas du modèle de Dellschaft : même si l'évolution se fait doucement, les proportions continuer à évoluer même en fin de simulation.

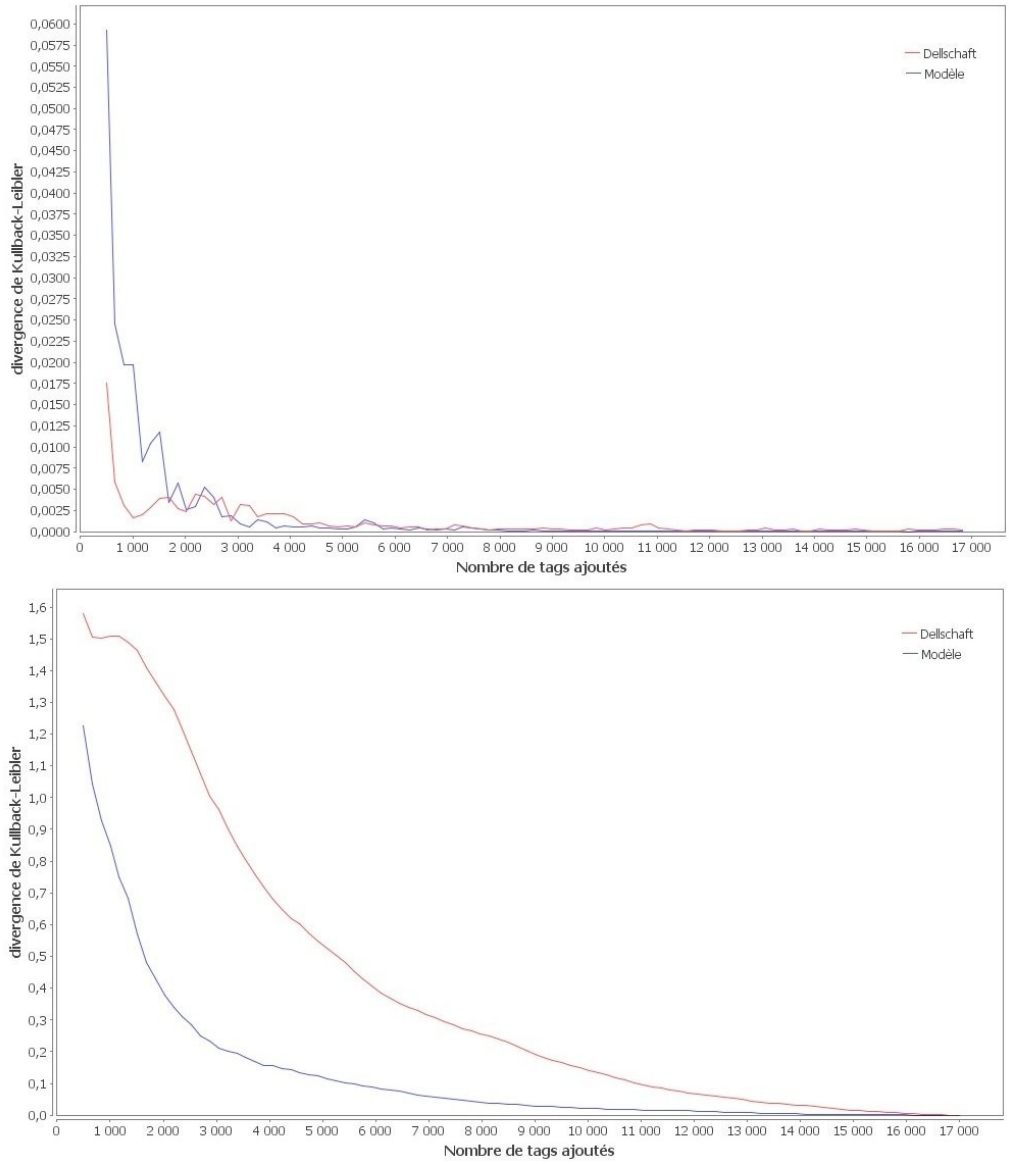


Figure 9 : Evolution de l'indice de Kullback-Leibler en fonction du nombre de tags ajoutés pour les deux modèles, entre deux instants consécutifs (en haut) et par rapport à la distribution finale (en bas).

Les données pouvant être appariés sur le nombre de tags ajoutés, elles sont testées au moyen du test de Wilcoxon. Pour D_C tout d'abord, l'effet du modèle est significatif avec p = .00001. En fait si notre modèle démarre avec des valeurs plus élevée, il finit avec un taux de divergence plus faible. Ainsi, sur la deuxième moitié de la simulation, le modèle de Dellschaft a des valeurs pour D_C comprises $5E^{-4}$ et $1E^{-4}$. Sur la même période, notre modèle à des valeurs de comprises entre $5E^{-5}$ et $1E^{-5}$ soit 10 fois plus faibles.
Concernant D_F, on note également un effet significatif du modèle avec p = $1.22E^{-7}$, confirmant ainsi notre hypothèse. Notre modèle permet d'obtenir une convergence plus rapide du vocabulaire

e) Influence des connaissances partagées et de l'internalisation

Afin de vérifier l'effet de ces deux facteurs, de nouvelles mesures ont été faites sous 3 conditions différentes :

- condition connaissances partagées seules (CP) : dans cette condition, pour vérifier l'effet des connaissances seules, Δ_{CT} et Δ_{LT} ont été mis à 0.
- condition internalisation (I) : pour voir si l'internalisation facilite la convergence des proportions, sans interférence de l'imitation, Δ_{LT} vaut 0.001 mais Δ_{CT} reste à 0.
- condition normale (N) : les valeurs normales (Δ_{LT} = 0.001 mais Δ_{CT} = 0.02) ont été utilisées pour cette série de simulation qui sert de comparaison.

Visuellement, l'évolution de D_C semble à peu près équivalente entre les 3 conditions (cf. figure 10). On note pourtant un effet significatif avec le test de Friedman (p < .00001), confirmer par une différence significatif pour les conditions comparées 2 à 2 avec le test de Wilcoxon. Au vue des valeurs, il semblerait que D_C diminue un peu plus rapidement pour la condition N, suivie de la condition I puis de CP.
L'évolution de D_F est plus contrasté par contre (cf. figure 10). La condition s'en sortant le mieux est CP, suivi de I puis de N. L'effet des 3 conditions est significatif au test de Friedman (p <.00001) une différence significative est notée entre les conditions prises 2 à 2 avec le test de Wilcoxon.
On peut donc conclure que conformémant à H_3, les connaissances partagées suffisent à amener les proportions à converger. Les proportions convergent en fait même mieux que dans la condition normale. H_4 n'est pas vérifiée en revanche. Avec l'internalisation seule (i.e. sans imitation), le vocabulaire converge bien et même mieux qu'en condition normale, mais moins bien qu'avec les connaissances partagées seules.

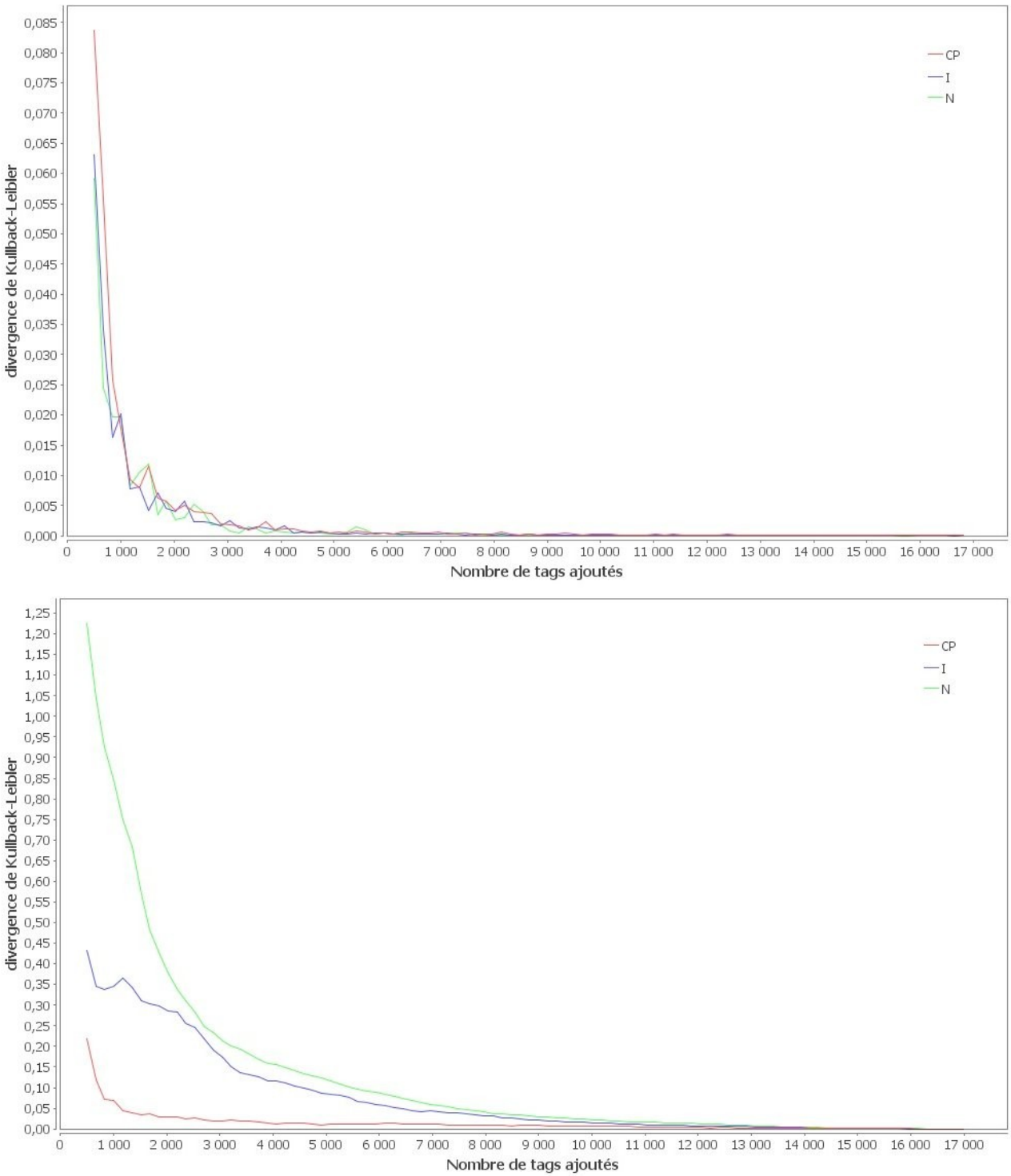


Illustration 10: Evolution de l'indice de Kullback-Leibler en fonction du nombre de tags ajoutés pour les trois condition CP, I et N, entre deux instants consécutifs (en haut) et par rapport à la distribution finale (en bas).

3. Discussion

3.1. Validité du modèle

Avant de pouvoir utiliser notre modèle pour tester certaines hypothèses sur l'émergence de représentations collectives dans les sytèmes de tagging, la première chose à faire est de montrer sa validité. Le modèle doit être capable de reproduire les propriétés connues des flux de tags.
Nos deux premières hypothèses sont justement que notre modèle est capable de reproduire les flux de co-occurence ainsi que la convergence des proportions pour les flux de ressources. Voyons ce qu'il en est.

a) Flux de co-occurrence.

D'un point de vue global, on constate que les propriétés essentielles des flux de co-occurence sont retrouvés. La distribution rang-fréquence a bien l'allure d'une loi puissance avec une pente plus faible pour les tags de rang élevé. Notre modèle est même plus proches des données de Delicious que celui de Dellschaft et Staab, pour 2 tags sur 3. Pour la croissance du nombre de tag, nous retrouvons également une croissance sous-linéaire. Sur ce point par contre, le modèle de Dellschaft et Staab semble un peu meilleur.

b) Convergence des proportions de tags pour les flux de ressources

Comme nous l'avons dit plus tôt, le résultat qui nous intéresse le plus concerne la convergence des proportions des tags. De ce côté là, notre deuxième hypothèse semble confirmé. Notre modèle converge vers une représentation stable des ressources, bien plus rapidement que le modèle de Dellschaft. On remarque toutefois que la convergence ne semble pas aussi rapide que dans la réalité, comme l'ont observé Golder et Huberman (2006).

De manière général, notre modèle est donc tout a fait adapté. Malgré sa simplicité, il parvient à reproduire les caractéristiques essentielles des flux de tags. En particulier, il reproduit la convergence des proportions de tags qui nous intéresse. De plus, notre modèle offre l'avantage de modéliser le comportement de l'utilisateur d'une manière cognitivement plus plausible. De ce point de vue, il nous semble que des progrès peuvent encore être fait. Une réflexion plus poussé sur la manière dont les utilisateurs pensent, s'appuyant sur les connaissances issues de la psychologie, permettrait certainement de modéliser encore mieux l'activité de tagging. En particulier, les

connaissances de l'utilisateur pourraient être mieux rendus.
En retour, ce type de modèle peut également permettre de tester des hypothèses sur les processus cognitifs de l'utilisateur et sur l'organiation de ces connaissances.

3.2. Mecanismes de construction de la représentation collective

On peut alors tenter de répondre à la question suivante : quels sont les facteurs qui permettent d'expliquer la convergence de cette représentation sociale ?
Nous avons formulé deux hypothèses à ce sujet : la première (H_3) est que les connaissances des sujets facilite la collaboration et ainsi la stabilisation des tags ; la deuxième (H_4) est que les utilisateurs ne restent pas passifs face au système mais adaptent leur comportement à celui des autres en internalisant la remprésentation qui se construit.

Nos résultats soutiennent fortement cette 3^e hypothèse. On constate que l'utilisation des connaissances seules, sans phénomène d'imitation ou d'internalisation, permet au proportions de converger. C'est même dans cette situation que la convergence est la plus rapide. On peut raisonablement penser que si les proportions convergent si vite dans la réalité (plus vite même que dans nos simulations), c'est parce que les individus qui taguent partagent un grand nombre de connaissances. Ces connaissances, construites lors d'interactions antérieures, sont ce qui nous permet de nous comprendre facilement et rapidement. Sans elle, les individus mettraient certainement beaucoup plus de temps pour se mettre d'accord sur les bons tags pour décrire une ressource.
En même temps, on peut penser que les connaissances des utilisateurs divergent un minimum et que ceux-ci doivent donc s'adapter. C'est le sujet de notre quatrième et dernière hypothèse. Malheureusment, les résultat ne sont pas concluant. La convergence à lieu en présence du phénomène d'internalisation mais moins rapidement que s'il n'est pas présent. Peut-être notre modèle ne rend-il pas compte correctement de la manière dont ce phénomène se passe chez un individu. Peut-être aussi que le phénomène d'internalisation ne peut être observé avec un modèle global comme le notre et sur le flux d'une seule ressource.

3.3. Vers une modélisation multi-agent

Cela nous amène au point suivant : les modélisations actuelles se contente de modéliser le flux d'une ressource, ou pire, le flux de co-occurrence pour un tag donné. L'activité de tagging sur une

ressource n'est pourtant pas détaché de l'activité de l'ensemble du système. Il est étrange également de tenter de modéliser un flux de co-occurrence, qui est en fait l'aggrégation de l'activité sur un grand nombre de ressources, par des mécanismes expliquant l'activité d'un utilisateur sur une ressource donné. On voit d'ailleurs que les valeurs des paramètres pour les flux de co-occurrence et pour les flux de ressources diffèrent. Cela laisse supposer que les paramètre permettant d'expliquer les flux de ressources ne se retrouve pas forcément au niveau des co-occurrence. Le phénomène d'imitation par exemple semble se diluer face au grand nombre de ressources sur lequel le flux de co-occurence est construit, expliquant les faibles valeurs pour l'imitation dans ce cas.
Nous pensons que la moyen le plus adéquate pour modéliser ce type de système serait d'utiliser un système multi-agent. Ce type de méthodologie est souvent proposée pour modéliser des systèmes complexes. En particulier, cette méthodologie a été proposé pour étudier les dynamiques sociales de formation des impressions (Smith & Collins, 2009), et plus largement en psychologie sociale (Smith & Conrey, 2007). L'idée est la suivante. Plutôt que de tenter de comprendre comment les différents facteur unitaire se combine pour aboutir à des dynamiques complexes et difficilement prévisibles, le plus simple est de simuler le phénomène à partir de ces facteur élémentaire et d'observer quelles dynamiques émergent.

C'est également ce que nous proposons ici. Tentons de comprendre comment un individu tag une ressource seule. Simulons un grand nombre d'utilisateur taguant un grand nombre de ressources et voyons ce que donnent ces interactions. Le système multi-agent peut alors servir à générer des hypothèses nouvelles et inattendues sur le fonctionnement du système et guider les recherches.
Un certain nombre de points nous semblent importants à prendre en compte dans ce type de simulation. Tout d'abord, il convient donc de distinguer les différents utilisateurs et les différentes ressources. Cela pose le problème de parvenir à modéliser les connaissances de chaque utilisateur (qui diffèrent) sur chacune ressource. Un enjeu important dans ce cas est la complexité de la simulation. Pour des raisons technique, les possibilités sont limités en terme de temps de calcul et de mémoire requise. Si chaque utilisateur connait ne serait-ce que 10000 mots pour 1000 ressources et qu'on considère 10000 utilisateurs différents, ont a déjà 10^{11} données à stocker en mémoire. Avec seulement un octet par mot, cela fait 100Go de données.
Ce type de modèle aurait également l'avantage de permettre de généré des flux de co-occurrence dans des conditions plus proche de la réalité. En ajoutant non plus un unique tag mais plusieurs tags à la fois (comme le font souvent les utilisateurs réels), on pourrait en effet tirer les flux de co-occurences directement des flux de ressources de l'ensemble du système.

3.4. La recherche de ressources

Un aspect souvent négligé des systèmes de tagging est la recherche de ressources. L'activité de tagging n'est qu'une partie de l'activité, les utilisateurs taguant des ressources dans le but de pouvoir les retrouver plus tard. Seul le modèle de Halpin et al. (2007) tient compte de la valeur informationnelle du tag pour le choix du tag à ajouter à une ressource. Il nous semble essentiel d'inclure l'activité de recherche dans la modélisation des systèmes de tagging afin de mieux comprendre comment et surtout pourquoi les gens tags.

3.5. L'intérêt de la science du Web pour les autres disciplines

Nous avons déjà insisté sur l'intérêt qu'a la science du Web de s'ouvrir à d'autre disciplines. Il est également vrai que cette discipline peut enrichir en retour les autres disciplines. Comme nous venons de le voir, les systèmes de tagging peuvent constituer un cas intéressant pour étudier les dynamiques sociales et l'émergence de représentations collectives. Ils peuvent également fournir des informations précieuse aux sciences cognitives sur le fonctionnement de la cognition humaine en activité. Plus généralement, le Web et les outils sociaux sont une source d'étude pour les sciences sociales et cognitives comme le soulignent Parameswaran et Whinston (2007), ou bien Salganik et Watts (2009).

3.6. Conclusion

Le Web social n'en est qu'à ses débuts. On est encore loin de la vision du web proposé par Godin (2007), où nous serions interconnectés en permanence, profitant ainsi du travail et de la connaissance des autre, le tout facilité par l'outil informatique. Comme le disent Wang et al (2007), le modèle actuel de l'informatique personnel sera bientôt dépassé. Les nouveaux outils doivent faciliter la collaboration et l'interaction sociale, pour peut-être un jour aboutir à la machine sociale dont parle Hendler et al. (2008). Derrière ce terme se cache l'idée de systèmes informatique qui ne travaillerait plus de manière indépendante. A travers le réseau, l'ensemble des machines seraient interconnectées, travaillant ainsi de concert. Mais pour permettre cela, une meilleure compréhension des dynamiques sociales est nécessaire. Les connaissances sur la cognition humaine doivent également guider la conception de tels systèmes. A travers ce travail, nous avons tenté d'apporter notre pierre à l'édifice. Nous n'en sommes actuellement qu'au prémisses de tels systèmes.

Références

Aberer, K., Cudré-Mauroux, P., Ouksel, A. M., Catarci, T., Hacid, M. S., Illarramendi, A., Kashyap, V., et al. (2004). Emergent semantics principles and issues. *Lecture notes in computer science*, 25–38.

Allen, C. (2004). Tracing the Evolution of Social Software - Life With Alacrity. Retrouvé Mars 29, 2010, de http://www.lifewithalacrity.com/2004/10/tracing_the_evo.html

Anderson, J. (2000). *Cognitive psychology and its implications 5th Ed*. New York : Worth.

Baronchelli, A., Cattuto, C., Loreto, V., & Puglisi, A. (2007). Complex systems approach to the emergence of language. Dans *Language, Evolution and the Brain*. J. W. Minett & W. S-Y. Wang.

Barrat, A., Baronchelli, A., Dall'Asta, L., & Loreto, V. (2007). Agreement dynamics on interaction networks with diverse topologies. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, *17*, 026111.

Benz, D., Grobelnik, M., Hotho, A., J\äschke, R., Mladenic, D., Servedio, V. D., Sizov, S., et al. (2008). Analyzing Tag Semantics Across Collaborative Tagging Systems. Dans *Dagstuhl Seminar 08391–Working Group Summary.*

Carvalho, L. L. D. (2008). *Représentations Emergentes - Une Approche Multi-Agents des Systèmes Complexes Adaptatifs en Psychologie Cognitive* (Thèse de doctorat en Psychologie, mention dimension cognitive et modélisation). Université Lumière, Lyon.

Cattuto, C., Benz, D., Hotho, A., & Stumme, G. (2008). Semantic analysis of tag similarity measures in collaborative tagging systems. Dans *Proceedings of the 3rd Workshop on Ontology Learning and Population (OLP3)* (p. 39-43). Patras, Greece.

Cattuto, C., Loreto, V., & Pietronero, L. (2007). Semiotic dynamics and collaborative tagging. *Proceedings of the National Academy of Sciences*, *104*(5), 1461-1464.

Cattuto, C., Baldassarri, A., Servedio, V. D. P., & Loreto, V. (2007). Vocabulary growth in

collaborative tagging systems. Arxiv e-print. Retrouvé de http://arxiv.org/abs/0704.3316

Chi, E. H., Pirolli, P., & Lam, S. K. (2007). Aspects of augmented social cognition: Social information foraging and social search. *Online Communities and Social Computing*, 60-69.

Dall'Asta, L., & Baronchelli, A. (2006). Microscopic activity patterns in the naming game. *Journal of Physics A: Mathematical and General*, *39*, 14851–14867.

Dellschaft, K., & Staab, S. (2008). An epistemic dynamic model for tagging systems. Dans *Proceedings of the 19th ACM Conference on Hypertext and Hypermedia HT'08* (p. 71-80).

Donetto, D., & Cecconi, F. (2009). The emergence of shared social representations in complex networks. Dans *Social Networks and Multi-Agent Systems Symposium (SNAMAS-09)*.

Durkheim, É. (1898). *Représentations individuelles et représentations collectives*. Paris: Presses universitaire de France.

Eggenberger, F., & Polya, G. (1923). Über die statistik verketteter vorgänge. *Zeitschrift für Angewandte Mathematik und Mechanik*, *3*(4), 279-289.

Godin, S. (2007). Seth's Blog: Web4. Retrouvé Mars 11, 2010, de http://sethgodin.typepad.com/seths_blog/2007/01/web4.html

Golder, S., & Huberman, B. A. (2006). The structure of collaborative tagging systems. *Journal of Information Science*, *32*(2), 198-208.

Grassé, P. P. (1959). La reconstruction du nid et les coordinations interindividuelles chez les bellicositermes natalensis et cubitermes, la théorie de la stigmergie : essai d'interprétation du comportement des termites constructeurs. *Insectes sociaux*, *6*(1), 41-81.

Halpin, H., Robu, V., & Shepherd, H. (2007). The complex dynamics of collaborative tagging. Dans *Proceedings of the 16th international conference on World Wide Web* (p. 211-220).

Hassas, S. (2003). *Systèmes complexes à base de multi-agents situés* (Habilitation à Diriger les Recherches). Université Claude Bernard, Lyon.

Hendler, J., Shadbolt, N., Hall, W., Berners-Lee, T., & Weitzner, D. (2008). Web science: an

interdisciplinary approach to understanding the web. *Communications of the ACM*, *51*(7), 60-69.

Heymann, P., Koutrika, G., & Garcia-Molina, H. (2008). Can social bookmarking improve web search? Dans *Proceedings of the international conference on Web search and web data mining* (p. 195–206).

Hollan, J., Hutchins, E., & Kirsh, D. (2000). Distributed cognition: toward a new foundation for human-computer interaction research. *ACM Transactions on Computer-Human Interaction (TOCHI)*, *7*(2), 174-196.

Jodelet, D. (1989). *Folies et représentations sociales*. Paris: Presses univ. de France.

Kaplan, F. (2000). *L'émergence d'un lexique dans une population d'agents autonomes* (Thèse de doctorat spécialité informatique). Université Paris 6.

Ke, J. Y., Gong, T., & Wang, W. S. (2008). Language change and social networks. *Communications in Computational Physics*, *3*(4), 935–949.

Laera, L., Blacoe, I., Tamma, V., Payne, T., Euzenat, J., & Bench-Capon, T. (2007). Argumentation over ontology correspondences in mas. Dans *Proceedings of the 6th international joint conference on Autonomous agents and multiagent systems* (p. 228).

Loula, A., Gudwin, R., El-Hani, C. N., & Queiroz, J. (2010). Emergence of self-organized symbol-based communication in artificial creatures. *Cognitive Systems Research*, *11*(2), 131-147.

Mathes, A. (2004). Folksonomies - Cooperative Classification and Communication Through Shared Metadata. Retrouvé Avril 6, 2010, de http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html

Mazuel, L., & Sabouret, N. (2008). Un modèle d'interaction pour des agents sémantiquement hétérogènes. *Proc. 16th Journées Francophones sur les Systèmes Multi-Agents (JFSMA)*, 233–242.

Mika, P. (2007). Ontologies are us: A unified model of social networks and semantics. *Web*

Semantics: Science, Services and Agents on the World Wide Web, *5*(1), 5–15.

Montemurro, M., & Zanette, D. (2002). Frequency-rank distribution of words in large text samples: Phenomenology and models. *Glottometrics*, *4*(87-98).

Moscovici, S. (1961). *La psychanalyse, son image et son public*. Paris: Presses Universitaires de France.

Parameswaran, M., & Whinston, A. B. (2007). Research issues in social computing. *Journal of the Association for Information Systems*, *8*(6), 336–350.

Parker, I. (1987). 'Social representations': Social psychology's (mis) use of sociology. *Journal for the theory of social behaviour*, *17*(4), 447–69.

Pirolli, P. (2008). Social Information Foraging and Sensemaking. Dans *Sensemaking Workshop*.

Rizzolatti, G., Fadiga, L., Gallese, V., & Fogassi, L. (1996). Premotor cortex and the recognition of motor actions. *Cognitive brain research*, *3*(2), 131-141.

Robu, V., Halpin, H., & Shepherd, H. (2009). Emergence of consensus and shared vocabularies in collaborative tagging systems. *ACM Transactions on the Web (TWEB)*, *3*(4), 14.

Rogers, Y., & Ellis, J. (1994). Distributed cognition: an alternative framework for analysing and explaining collaborative working. *Journal of information technology*, *9*, 119–119.

Rosch, E. (1975). Cognitive reference points. *Cognitive Psychology*, *7*(4), 532-547.

Rupert, M. (2009). *Coévolution d'organisations sociales et spatiales dans les systèmes multi-agents: application aux systèmes de tagging collaboratifs* (Thèse de doctorat spécialité Informatique). Université Claude Bernard, Lyon.

Salganik, M. J., & Watts, D. J. (2009). Web-Based Experiments for the Study of Collective Social Dynamics in Cultural Markets. *Topics in Cognitive Science*, *1*(3), 439-468.

Sansonnet, J. P., & Valencia, E. (2005). Terminological heterogeneity between agents using a generalized simplicial representation. Présenté au European Workshop on Multi-Agent Systems (EUMAS'05).

Schenkel, R., Crecelius, T., Kacimi, M., Neumann, T., Parreira, J., Spaniol, M., Weikum, G., et al. (2008). Social wisdom for search and recommendation. *IEEE Data Engineering Bulletin*, *31*(2), 40–49.

Sen, S., Lam, S. K., Rashid, A. M., Cosley, D., Frankowski, D., Osterhouse, J., Harper, F. M., et al. (2006). Tagging, communities, vocabulary, evolution. Dans *Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work* (p. 181-190).

Shirky, C. (2005). Ontology is Overrated : Categories, Links, and Tags. Retrouvé Avril 1, 2010, de http://www.shirky.com/writings/ontology_overrated.html

Simon, H. (1955). On a class of skew distribution functions. *Biometrika*, *42*(3/4), 425-440.

Smith, E. R., & Collins, E. C. (2009). Contextualizing person perception: Distributed social cognition. *Psychological Review*, *116*(2), 343–364.

Smith, E. R., & Semin, G. R. (2007). Situated social cognition. *Current Directions in Psychological Science*, *16*(3), 132.

Smith, E. R., & Conrey, F. R. (2007). Agent-Based Modeling: A New Approach for Theory Building in Social Psychology. *Personality and Social Psychology Review*, *11*(1), 87-104.

Steels, L. (2003a). Intelligence with representation. *Philosophical Transactions: Mathematical, Physical and Engineering Sciences*, *361*(1811), 2381–2395.

Steels, L. (2003b). Evolving grounded communication for robots. *Trends in Cognitive Sciences*, *7*(7), 308-312.

Steels, L. (2005). The role of construction grammar in fluid language grounding. *Artificial Intelligence*, *164*.

Stuber, A. (2007). *Co-construction de sens par négociation pour la réutilisation en situation de l'expérience tracée-Vers le partage et l'échange d'expérience collective* (Thèse de doctorat spécialité informatique). Université Claude Bernard, Lyon.

Swarup, S., & Gasser, L. (2008a). Language Evolution on a Dynamic Social Network. Présenté au

MORS Workshop on "Analyzing the Impact of Emerging Societies on National Security", Argonne National Laboratory, Argonne.

Swarup, S., & Gasser, L. (2008b). *Collaborative Tagging as Information Foraging*. Tech Report UIUCLIS--2008/1+ID. Graduate School of Library and Information Science University of Illinois at Urbana-Champaign.

Varela, F. J., Thompson, E., & Rosch, E. (1992). *The embodied mind: Cognitive science and human experience*. The MIT Press.

Voelklein, C., & Howarth, C. (2005). A Review of Controversies about Social Representations Theory: A British Debate. *Culture & Psychology*, *11*(4), 431-454.

Vygotski, L. S. (1978). *Mind in society.* Cambridge, MA: Harvard University Press.

Wang, F. Y., Carley, K. M., Zeng, D., & Mao, W. (2007). Social computing: From social informatics to social intelligence. *IEEE Intelligent Systems*, 79–83.

Xu, Z., Fu, Y., Mao, J., & Su, D. (2006). Towards the semantic web: Collaborative tag suggestions. Dans *Collaborative Web Tagging Workshop at WWW2006*. Edinburgh, Scotland.

Yanbe, Y., Jatowt, A., Nakamura, S., & Tanaka, K. (2007). Can social bookmarking enhance search in the web? Dans *Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries* (p. 116).

Yule, G. (1925). A mathematical theory of evolution, based on the conclusions of Dr. JC Willis Philos. *Trans. R. Soc. Lond. B Biol. Sci*, *213*, 21-87.

Résumé :

Avec le développement du web social, le besoin se fait sentir de mieux comprendre les dynamiques en jeu sur internet. C'est ce que propose la toute jeune science du web. Les sciences cognitives et sociales ont un rôle prépondérant à jouer dans lédifice. Nous proposons dans ce mémoire une analyse des systèmes de tagging (un outil collaboratif de classification de ressources) du point de vue de l'émergence des représentations collectives et de l'émergence du langage. Nous proposons également un modèle capable de rendre compte des propriétés connues de ces systèmes et en particulier la convergence des proportions des tags. Cette convergence s'expliquerait essentiellement par les connaissances partagées par les utilisateurs.

Mots clés : représentations collective, émergence du langage, modélisation informatique, web social, science du web, systèmes de tagging.