UNIVERSITÉ DU QUÉBEC À MONTRÉAL

RECONSTRUCTION D'UN SYSTÈME SOCIOSÉMANTIQUE PAR APPRENTISSAGE MACHINE

THÈSE

PRÉSENTÉE

COMME EXIGENCE PARTIELLE

DU DOCTORAT EN INFORMATIQUE COGNITIVE

PAR

JEAN-FRANÇOIS CHARTIER

23 JANVIER 2016

REMERCIEMENTS

Je tiens en premier lieu à remercier chaleureusement mon directeur de recherche, le professeur Jean-Guy Meunier, avec qui cette aventure intellectuelle a commencé. J'aimerais lui exprimer ma profonde gratitude pour toutes ces années d'échanges et de collaborations.

Merci également à mon codirecteur, le professeur Petko Valtchev, pour son soutien et ce, malgré mes fréquents et longs silences souvent suivis d'agitation et de précipitation.

Merci à mes collègues du LANCI pour le précieux partage intellectuel et les sincères amitiés. Ce laboratoire aura été ma deuxième maison pendant plusieurs années.

Merci à mon ami Maxime Sainte-Marie avec qui j'ai commencé ce doctorat et fais la traversée du désert.

Merci à ma chère amie Véronique Petit pour son soutient et pour m'avoir aidé à rester sain d'esprit, chose improbable que savent ceux qui sont passés par là.

Merci à tous ceux et celles, amis, professeurs et collègues qui ont croisé ma route durant ces longues années d'étude. Vous avez tous eu un impact, grand ou petit, sur mes apprentissages, mes découvertes et ma vie académique avec ses hauts et ses bas.

Je tiens également à remercier les organismes subventionnaires du CRSH et du FQRSC qui ont financé cette recherche.

TABLE DES MATIÈRES

LISTE DES FIGURES

LISTE DES TABLEAUX

RÉSUMÉ

L'objectif de recherche de cette thèse est la reconstruction par apprentissage machine du mécanisme sociocognitif à l'œuvre dans l'évolution d'un réseau sociosémantique. Cet objectif est basé sur l'hypothèse que la dynamique d'un réseau sociosémantique est déterminée par un mécanisme d'influence sociale basé sur des facteurs d'exposition sociale, de contagion sociale, de déférence et de mimétisme des semblables. Une méthode de fouille de données basée sur l'analyse des réseaux sociaux, la sémantique vectorielle et l'apprentissage machine est développée afin de reconstruire différents modèles de l'influence sociale permettant de prédire l'évolution d'un réseau sociosémantique. Ces modèles sont des arbres de décisions, des listes de règles, des forêts aléatoires et des modèles probabilistes naïfs. L'analyse de ces modèles prédictifs suggère que l'hypothèse de recherche est vraisemblable, mais que d'autres mécanismes sont également à l'œuvre dans le processus étudié.

MOTS-CLÉS : Réseau sociosémantique; apprentissage machine; réseau social; sémantique vectorielle; influence sociale.

CHAPITRE I

INTRODUCTION

1.1 Une problématique de reconstruction

Dans les sciences empiriques, une problématique de reconstruction se présente lorsque le système étudié se révèle être une boite noire, c'est-à-dire que les mécanismes à l'œuvre dans le système sont opaques et leur observation directe impossible. Par conséquent, l'étude empirique de ces systèmes est effectuée via des observations indirectes, qui, bien qu'elles ne fournissent pas d'informations sur le fonctionnement interne de la boîte noire, informent sur ses différents états possibles.

Ce type de problématique a été au cœur du programme de recherche de la cybernétique classique. Ashby la présentait de la manière suivante :

> « The engineer is given a sealed box that has terminals for input, to which he may bring any voltages, shocks, or other disturbances he pleases, and terminals for output, from which he may observe what he can. *He is to deduce what he can of its contents.* » (Ashby, 1957, p. 86) (l'italique a été ajouté)

Pour Ashby, une problématique de reconstruction est une forme de rétro-ingénierie : par l'analyse des patrons récurrents dans les transitions d'états du système, l'ingénieur doit induire un modèle plausible du mécanisme à l'œuvre entre les entrées et les sorties du processus. Le modèle induit, s'il est suffisamment vraisemblable, devrait pouvoir se substituer à la boîte noire, c'est-à-dire l'émuler, de telle sorte qu'il puisse prédire le plus fidèlement possible ses transitions d'états.

Une problématique de reconstruction peut être illustrée par le schéma général suivant :

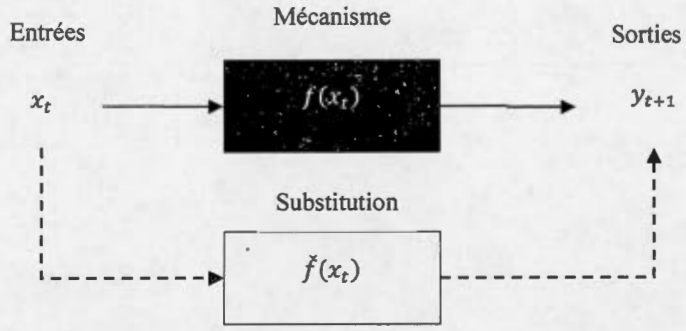

Figure 1.1: Schéma général d'une problématique de reconstruction.

Dans ce schéma, la variable indépendante x_t, aussi appelée la variable prédictive ou la variable explicative, modélise les « entrées » du système. Elle modélise l'état du système observé à un temps *t*. La variable dépendante y_{t+1}, aussi appelée la variable répondante ou la variable expliquée, modélise les « sorties » du système. Elle modélise un état du système observé à un temps *t+1*. La fonction $f(x_t)$ représente la boîte noire, c'est-à-dire un mécanisme inconnu, mais qu'on suppose être à l'œuvre dans la dynamique du système. Le symbole $\check{f}(x_t)$ est une hypothèse d'approximation de $f(x_t)$. Elle est un modèle de la boîte noire qui peut se substituer à elle et émuler son fonctionnement. L'approximation $\check{f}(x_t)$ constitue l'objet de découverte d'une problématique de reconstruction. Autrement dit, l'objectif de la reconstruction est de minimiser la quantité $\sum_{x_t \in X}\left(\check{f}(x_t) - f(x_t)\right)$, c'est-à-dire trouver l'approximation $\check{f}(x_t)$ la plus vraisemblable, celle capable de reproduire le plus fidèlement $f(x_t)$.

Une méthode de reconstruction est basée sur l'analyse des états observables de la boîte noire. Généralement, ces observations sont organisées sous la forme d'une série temporelle de données $\{(x_t, y_{t+1}) : t \in T\}$ représentant le processus empirique étudié. Dans cette série, chaque séquence (x_t, y_{t+1}) est appelée un « exemplaire » ou une « transaction » et elle correspond à une transition d'état qui a été observée empiriquement. Si un mécanisme est à l'œuvre dans le système étudié et qu'il en détermine la dynamique, alors des patrons récurrents devraient caractériser ces

transitions d'états. Une méthode de reconstruction doit permettre d'identifier ces patrons récurrents et d'en induire un modèle.

La Figure 1.2 illustre un exemple trivial d'une problématique de reconstruction. Supposons un système pour lequel nous disposons de cinq exemplaires de sa dynamique, c'est-à-dire que cinq transitions d'états d'un processus empirique donné ont été observées. Ces transitions ont été modélisées par une variable indépendante $x_t \in \mathbb{R}$ et une variable dépendante binaire $y_{t+1} \in \{0,1\}$. L'un de ces exemplaires spécifie par exemple qu'à un temps t le système a été observé être dans un état correspondant à la valeur $x_t = (1.1)$ et qu'à un temps t+1 il a été observé être dans un état correspondant à la valeur $y_{t+1} = 0$.

Supposons une fonction $f(x_t) = y_{t+1}$ qui représente un mécanisme à l'œuvre dans la dynamique du système, mais dont nous ignorons le fonctionnement.

(x_t, y_{t+1})
(1.1, 0)
(2.0, 0)
(3.0, 1)
(4.6, 1)
(5.3, 1)

Figure 1.2: Un ensemble de cinq exemplaires d'un processus empirique.

Quel modèle $\check{f}(x_t) = y_{t+1}$ peut se substituer à $f(x_t) = y_{t+1}$ et prédire le plus fidèlement possible ses transitions d'état? L'exemple de la Figure 1.2 est une problématique de reconstruction triviale car l'espace des instances de $f(x_t)$ est très restreint. Une hypothèse d'approximation $\check{f}(x_t)$ peut facilement être induite par l'analyse des patrons récurrents dans ces cinq exemplaires. Cette hypothèse pourrait, par exemple, être un ensemble de deux règles comme celles-ci :

Si $x_t < 3.0$, alors $y_{t+1} = 0$,

Si $x_t \geq 3.0$, alors $y_{t+1} = 1$.

Si nous postulons que ces cinq exemplaires représentent parfaitement l'espace des instances de la fonction $f(x_t)$, alors nous pouvons conclure que ces deux règles forment un modèle $\check{f}(x_t)$ émulant parfaitement $f(x_t)$. Évidemment, dans une problématique de reconstruction non triviale, cinq exemplaires sont rarement suffisants pour permettre l'induction d'un modèle vraisemblable. De plus, que ce soit la reconstruction d'un système électrique, d'un système cognitif ou d'un système social, la dynamique du système est généralement beaucoup plus complexe que l'exemple de la Figure 1.2. Il n'est pas rare que plusieurs centaines de milliers d'exemplaires soient nécessaires pour identifier des patrons récurrents dans les transitions d'états du système et que les variables nécessaires pour modéliser ces états soient beaucoup plus complexes qu'une variable univariée.

Par ailleurs, pour Ashby, une problématique de reconstruction ne se limite pas à l'étude des systèmes électriques. Tous les systèmes, qu'ils soient électriques, biologiques, sociaux ou autres, dont le fonctionnement est inobservable, sont potentiellement l'objet d'une problématique de reconstruction :

> « Though the problem arose in purely electrical form, its range of application is far wider. The clinician studying a patient with brain damage and aphasia may be trying, by means of tests given and speech observed, to deduce something of the mechanisms that are involved. And the psychologist who is studying a rat in a maze may act on the rat with various stimuli and may observe the rat's various behaviours; and by putting the facts together he may try to deduce something about the neuronic mechanism that he cannot observe. » (Ashby, 1957, p. 86)

Les chercheurs des sciences cognitives sont régulièrement confrontés à des problématiques de reconstruction, alors appelées des problématiques de « rétro-ingénierie cognitive » (A. Clark, 1990; Dennett, 1995; Harnad, 1994). En effet, un processus cognitif forme généralement une boîte noire. Son mécanisme n'est pas

observable directement, il doit être reconstruit par l'analyse de ses entrées et de ses sorties. En psychologie cognitive par exemple, ces données sont généralement recueillies via des protocoles expérimentaux dans lesquels on demande à des sujets d'accomplir différentes tâches cognitives de perception, de catégorisation, de mémorisation ou autres. C'est par l'analyse des régularités caractérisant les performances des sujets à ces tâches, notamment leur taux d'erreur et leur temps de réaction, qu'il devient possible de reconstruire un modèle du mécanisme cognitif impliqué dans la réalisation de ces tâches (Fortin & Rousseau, 2005, p. 14).

Dans d'autres domaines des sciences cognitives, notamment l'anthropologie, ce sont plutôt des approches ethnographiques qui sont utilisées pour la rétro-ingénierie cognitive. Ces approches permettent d'étudier des processus cognitifs dits « dans la nature » (Hutchins, 1995). Les données d'observations sont alors les traces empiriques laissées par l'activité humaine. Des archives, des œuvres d'art, des rituels en sont des exemples classiques pour l'anthropologie. Des textes publiés sur un blogue, des données de géolocalisation, des registres de transactions bancaires, des données de navigation sur le web en sont des exemples plus récents appelés « traces digitales » (Latour, 2007). C'est par l'analyse des patrons récurrents dans ces données, par exemple des associations statistiques, que les mécanismes de la boîte noire sont reconstruits.

Dans le cadre de cette thèse, la boîte noire étudiée est un mécanisme à l'œuvre dans un système sociosémantique.

1.2 Système sociosémantique

Un système sociosémantique (SSS pour la suite) correspond à une population d'agents socialement organisée au sein de laquelle des interactions sociales déterminent la production de contenus conceptuels. Un SSS a une unité d'analyse similaire à ce qui est appelé en sociolinguistique une « communauté de discours » (R. A. Blythe & Croft,

2009; H. H. Clark, 1996; Gumperz, 1969; Patrick, 2002). Un exemple canonique de SSS est une communauté de scientifiques dans laquelle les interactions sociales prendront, par exemple, la forme de collaborations entre membres et où la production de contenus conceptuels se manifestera dans la publication d'articles scientifiques (Roth, 2013). D'autres instanciations empiriques des SSS sont par exemple des communautés de blogueurs et des communautés de journalistes. Il y a également plusieurs modalités d'interactions sociales autres que la collaboration à l'œuvre dans les SSS, notamment la compétition entre membres et la coordination.

Formellement, un SSS peut être défini par un réseau bimodal $SSS = (A, C, L_c^a, L_a^a, L_c^c)$, où le paramètre $A = \{a_1 \dots a_n\}$ représente un ensemble d'agents membres du SSS, le paramètre $C = \{c_1, \dots c_m\}$ représente un ensemble de concepts, le paramètre $L_a^a = \left\{l_{a_j}^{a_i} : l_{a_j}^{a_i} \in A \times A\right\}$ représente un ensemble de liens entre agents, le paramètre $L_c^c = \left\{l_{c_j}^{c_i} : l_{c_j}^{c_i} \in C \times C\right\}$ représente un ensemble de liens entre concepts et le paramètre $L_c^a = \left\{l_{c_j}^{a_i} : l_{c_j}^{a_i} \in A \times C\right\}$ représente un ensemble de liens entre agents et concepts.[1]

La Figure 1.3 illustre le schéma général d'un SSS :

[1] Cette définition est tributaire des travaux de Roth et de sa définition des réseaux épistémiques (Roth, 2008a, 2008b, 2013; Roth & Cointet, 2010).

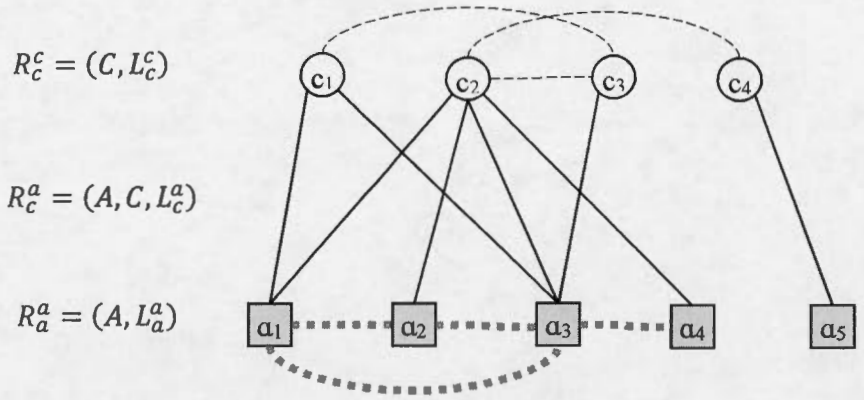

Figure 1.3: Schéma général d'un système sociosémantique.

Un SSS peut aussi être décomposé en trois différents types de réseaux. Un SSS inclut premièrement un réseau social $R_a^a = (A, L_a^a)$ composé d'agents et de liens entre agents modélisant différentes modalités d'interaction sociale dans le système. Un SSS inclut ensuite un réseau sémantique $R_c^c = (C, L_c^c)$ composé de concepts et de relations sémantiques entre concepts. Finalement, un SSS inclut un réseau sociosémantique $R_c^a = (A, C, L_c^a)$ composé d'agents et de concepts et de liens d'usages entre agents et concepts modélisant la distribution sociale des usages conceptuels dans le système.

1.2.2 Processus et mécanismes dans un SSS

Un SSS est aussi un système dynamique, c'est-à-dire évoluant avec le temps. Dénotons cette dynamique par une série temporelle $\{SSS_t : t \in T\}$ dans laquelle SSS_t correspond à l'état du système à un temps t. À chaque temps t de cette dynamique, des liens se forment et se défont entre agents, entre concepts et entre agents et concepts.

La dynamique d'un SSS se décompose en trois types de processus, soit l'évolution du réseau social, qui peut être modélisée par une série $\{R_{a_t}^a : t \in T\}$, l'évolution du réseau sémantique, modélisée par une série $\{R_{c_t}^c : t \in T\}$ et l'évolution du réseau sociosémantique modélisée par $\{R_{c\,t}^a : t \in T\}$. Chaque type de dynamique, sociale,

sémantique ou sociosémantique, peut être expliquée par des mécanismes spécifiques, soit sociaux, cognitifs et sociocognitifs.

Premièrement, on retrouve des mécanismes sociaux spécifiques à l'évolution des liens dans le réseau social d'un SSS. Ces mécanismes sont nombreux. Ce sont des mécanismes de formation des liens sociaux comme la réciprocité, l'équilibre social, l'homophilie, l'attachement préférentiel et l'assortativité (Barrat, Barthelemy, & Vespignani, 2008; Holland & Leinhardt, 1977; Monge & Contractor, 2003; Moody, 2009; Scott, 2013; T. Snijders, 2011; T. A. Snijders, Van de Bunt, & Steglich, 2010).

Deuxièmement, on retrouve des mécanismes cognitifs spécifiques à l'évolution des liens dans le réseau sémantique d'un SSS. Ce sont des mécanismes de formation des relations sémantiques entre concepts. Ces mécanismes aussi sont nombreux, ce sont notamment des mécanismes d'inférence, d'association, de cohérence ou de classification (Borge-Holthoefer & Arenas, 2010; Gärdenfors, 2014; Griffiths, Steyvers, & Tenenbaum, 2007; Sowa, 2006; Steyvers & Tenenbaum, 2010; Widdows, 2004; Zwarts, 2010).

Troisièmement, on retrouve des mécanismes sociocognitifs spécifiques à l'évolution du réseau sociosémantique d'un SSS. Ces mécanismes sont responsables de l'évolution des liens d'usage entre agents et concepts. L'évolution de ces liens correspond à des processus de propagation, de diffusion ou de partage social des contenus conceptuels parmi les membres d'un SSS. Plusieurs mécanismes sont à l'œuvre dans ces processus. En épistémologie sociale par exemple, la propagation des concepts est principalement expliquée par des mécanismes de déférence (Goldman, 1999). En anthropologie cognitive, les processus de diffusion sont expliqués notamment par des mécanismes de biais cognitifs et d'imitation (Atran, Ross, & Medin, 2005; Richerson & Boyd, 2005). Dans la linguistique évolutionniste, les processus de diffusion sociale sont expliqués par des mécanismes de coordination sociale (Croft, 2000; Tomasello, 2008). En économie, on fait appel à des mécanismes de cascade informationnelle (Bikhchandani,

Hirshleifer, & Welch, 1998). Des sociologues ont suggéré que des mécanismes d'offre et de demande cognitive puissent également expliquer la diffusion des concepts dans une population (Bronner, 2003). Plusieurs mécanismes d'influence sociale seraient également à l'œuvre dans les processus de diffusion sociale des contenus conceptuels (Dodds & Watts, 2005; Friedkin & Johnsen, 1999, 1999; Latané, 1996; Robins, Pattison, & Elliott, 2001; Watts & Dodds, 2007).

1.2.3 Les traces textuelles

Prenons par exemple le SSS formé de la communauté des contributeurs de l'encyclopédie collaborative Wikipédia. Les processus dans ce SSS se réalisent à une échelle macroscopique de plusieurs années et de manière distribuée dans une population de plusieurs milliers d'agents anonymes répartis un peu partout sur la planète. Dans ce SSS, les mécanismes d'évolution des liens sociaux entre membres, des liens sémantiques entre concepts et des liens d'usage entre agents et concepts sont totalement opaques et leur observation directe impossible. Malgré cet obstacle méthodologique, ce SSS, comme plusieurs autres, par la nature même de l'activité de ses membres laisse des traces empiriques qui elles peuvent être recueillies et constituer un corpus de données qui informeront indirectement le fonctionnement du SSS.

Ces traces empiriques sont des textes produits par les agents du système. Ce sont par exemple des articles scientifiques ou des articles de presse, ou encore des billets publiés sur un blogue. C'est par l'analyse de ces textes qu'il est possible d'identifier des régularités ou des patrons récurrents corrélés aux éléments, aux relations et aux mécanismes de fonctionnement de ces SSS (Corman, Kuhn, McPhee, & Dooley, 2002; Diesner & Carley, 2005).

À titre d'exemple, c'est par l'analyse de ces traces textuelles que Roth a reconstruit le SSS d'une communauté de scientifiques. Il a analysé des milliers d'articles publiés par les membres d'une communauté scientifique sur plus d'une dizaine d'années. C'est par

l'analyse des mots présents dans ces textes, les relations entre cosignataires et des citations qu'il est parvenu à reconstruire plusieurs structures caractérisant le SSS (Roth, 2008a; Roth & Cointet, 2010). Il a notamment montré comment des structures d'hétérogénéité caractérisent la distribution des liens sociaux et des liens d'usages dans un SSS, ce qu'il appelle la distribution du capital social et du capital sémantique des agents. Roth a aussi contribué à une meilleure compréhension de la dynamique des SSS en montrant notamment comment la formation des liens sociaux dans le SSS était déterminée par un mécanisme d'homophilie sémantique. Il a montré que la propension qu'un lien social $l_{a_j}^{a_i}$ soit créé dans le SSS dépendait de la similarité des contenus sémantiques utilisés par les agents a_i et a_j.

1.3 Objectif de recherche

Cette thèse s'intéresse à l'évolution du réseau sociosémantique d'un SSS. L'objectif de recherche est la reconstruction du mécanisme sociocognitif à l'œuvre dans l'évolution du réseau sociosémantique d'un SSS. Plus spécifiquement, le but de cette recherche est d'induire un modèle de ce mécanisme qui permttra de prédire l'évolution des liens d'usage entre agents et concepts dans le réseau sociosémantique d'un SSS.

Dénotons par $\vec{x}_{i,j,t}$ la variable indépendante de notre problématique et par $y_{i,j,t+1}$ la variable dépendante. La variable indépendante forme un vecteur $\vec{x}_{i,j,t} = (x_{1,i,j,t}, \dots x_{m,i,j,t})$ et pour l'instant, mentionnons simplement que $\vec{x}_{i,j,t}$ se situent dans un espace multidimensionnel $\mathbb{R}^m$ et modélisent différents rapports sociaux impliquant un agent a_i et un concept c_j à un temps t de la dynamique d'un SSS.

La variable dépendante $y_{i,j,t+1}$ est une variable binaire. Elle modélise l'évolution des liens d'usage dans un réseau sociosémantique, c'est-à-dire la présence ou l'absence d'un lien d'usage entre l'agent a_i et le concept c_j à un temps t+1 de la dynamique du SSS. Autrement dit, $y_{i,j,t+1} = 1$ si $l_{c_j}^{a_i} \in R_{c\,t+1}^{a}$ et $y_{i,j,t+1} = 0$ si $l_{c_j}^{a_i} \notin R_{c\,t+1}^{a}$.

Dénotons par la série temporelle $\{(\vec{x}_{i,j,t}, y_{i,j,t+1}): t \in T\}$ un ensemble d'exemplaires ou de transactions du processus empirique d'évolution d'un réseau sociosémantique d'un SSS. Finalement, dénotons par la fonction (1.1) la boîte noire que constitue le mécanisme sociocognitif à l'œuvre dans l'évolution du réseau sociosémantique :

$$f(\vec{x}_{i,j,t}) = y_{i,j,t+1} \qquad \text{Eq. (1.1)}$$

Ceci étant défini, l'objectif de recherche peut être reformulé de manière plus précise : à partir de l'analyse des patrons récurrents dans la série temporelle $\{(\vec{x}_{i,j,t}, y_{i,j,t+1}): t \in T\}$, il consiste à reconstruire le mécanisme sociocognitif à l'œuvre dans l'évolution du réseau sociosémantique d'un SSS en induisant une hypothèse d'approximation $\check{f}(\vec{x}_{i,j,t})$ capable d'émuler et de prédire le plus fidèlement possible les exemplaires de la fonction cible (1.1).

1.3.2 La reconstruction par apprentissage machine

Une hypothèse d'approximation peut être induite de plusieurs manières. On distingue généralement deux classes de solutions possibles à une problématique de reconstruction : des approches « paramétriques » et des approches « algorithmiques » (Breiman, 2001b).

Une solution paramétrique classique est un modèle de régression linéaire. C'est une solution qui consiste à approximer le fonctionnement de la boîte noire par un modèle qui a la forme générale suivante $\check{f}(x_t) = ax_t + b$. Comme d'autres solutions paramétriques, la régression linéaire impose des contraintes très fortes sur la modélisation, notamment des contraintes de linéarité. Des problématiques complexes de reconstruction sont insolubles avec ces contraintes (Breiman, 2001b).

Les modèles algorithmiques ont profondément renouvelé les solutions possibles à une problématique de reconstruction. Ils ont été développés en informatique dans des

domaines comme l'apprentissage machine (T. M. Mitchell, 1997), la reconnaissance de forme (Theodoridis & Koutroumbas, 2008) et l'apprentissage statistique (James, Witten, Hastie, & Tibshirani, 2013). Il existe de nombreux modèles algorithmiques. Une approximation sous forme de règles représente un modèle algorithmique classique. D'autres types de modèles algorithmiques sont par exemple les arbres de décisions, les réseaux de neurones artificiels, des forêts aléatoires et bien d'autres.

Dans le cadre de cette thèse, l'objectif de recherche est d'approximer la fonction (1.1) à l'aide d'apprentissage machine.

1.3.3 Communauté de journalistes

L'objectif de reconstruction est réalisé via l'analyse des traces textuelles générées par l'activité des membres d'un SSS particulier, soit celui constitué par les membres de la communauté journalistique du journal The New York Time. Ces traces sont des articles de presse produits quotidiennement par les membres du système. L'objectif de recherche consiste donc à reconstruire le mécanisme sociocognitif que représente la fonction (1.1) uniquement par l'analyse des données présentes dans ces articles de presse.

Ces données sont très partielles, ce sont les dates de publication des textes, les noms des auteurs et les mots présents dans les articles. Étant très limitées, elles imposent de très fortes contraintes méthodologiques sur l'objectif de recherche.

En effet, les articles de presse constituent des données empiriques dites « non structurées ». Les informations que ces articles contiennent (mots, dates, auteur) ne forment pas des variables qui permettraient de modéliser directement un mécanisme sociocognitif. Par conséquent, les articles de presse doivent faire l'objet de plusieurs opérations de prétraitement. À l'objectif de recherche lié à l'approximation de la fonction (1.1) s'ajoute également un objectif méthodologique complexe de fouille de données textuelles non-structurées.

1.3.4 Hypothèse de recherche

Tel que discuté précédemment, plusieurs types de mécanismes sociocognitifs sont possiblement à l'œuvre dans l'évolution d'un réseau sociosémantique. Par ailleurs, il est également possible que l'évolution de ce réseau soit un processus purement aléatoire et indéterminé. Certains ont avancé que plusieurs aspects de l'évolution linguistique pouvaient être modélisés de cette manière (Richard A. Blythe & Croft, 2012; Croft, 2000; Fitch, 2005; Keller, 1994). Si c'était le cas, son évolution serait impossible à prédire.

Cette thèse s'intéresse à une classe particulière de mécanismes possiblement à l'œuvre dans l'évolution du réseau sociosémantique d'un SSS, soit les mécanismes d'influence sociale basés sur la topologie des réseaux sociaux. L'hypothèse fondamentale de ces mécanismes est relativement simple et se résume de la manière suivante : des agents qui ont les mêmes rapports sociaux, les mêmes liens ou qui interagissent avec les mêmes personnes ont tendance à adopter les mêmes comportements (Brass, 1984; R. S. Burt, 2010b; Erickson, 1988; Friedkin, 2006; R. T. A. Leenders, 2002; Marsden & Friedkin, 1993; Rice, 1993; Robins et al., 2001).

Dans la littérature des sciences sociales, ces mécanismes d'influence sociale ont été utilisés pour expliquer différents processus de diffusion sociale. On a étudié, par exemple, les mécanismes d'influence sociale responsables de la diffusion des innovations dans un réseau social (R.S. Burt, 1987; Valente, 2005), la diffusion des opinions dans un réseau social (Watts & Dodds, 2007), la diffusion des contenus numériques entre membres d'un blogue (Gruhl, Guha, Liben-Nowell, & Tomkins, 2004), le marketing viral (Leskovec, Adamic, & Huberman, 2007), la diffusion des comportements menant à l'obésité (Christakis & Fowler, 2007), la diffusion des connaissances dans une organisation (Cowan & Jonard, 2004), la diffusion dans une entreprise des perceptions liées aux conditions de travail (Meyer, 1994), la diffusion

des défauts de paiement entre banques (Gai & Kapadia, 2010), et bien d'autres (Borgatti & Foster, 2003; Monge & Contractor, 2003).

L'hypothèse sur laquelle se base la présente recherche est que le mécanisme sociocognitif représenté par la fonction (1.1) en est un d'influence sociale entre agents d'un SSS. Cette hypothèse peut être formulée de la manière suivante :

Soit $\{R^a_{c\,t} : t \in T\}$ la dynamique sociosémantique d'un SSS, la présence ou non d'un lien d'usage $l^{a_i}_{c_j}$ dans le réseau $R^a_{c\,t+1}$ est déterminée par la magnitude au temps t de l'influence sociale du concept c_j sur l'agent a_i.

En d'autres mots, l'hypothèse de recherche conjecture que l'usage au temps t+1 d'un concept c_j par un agent a_i, est déterminé par des rapports sociaux d'influence sociale qui étaient à l'œuvre dans le SSS au temps t. Opérationnalisé au SSS étudié, une autre manière de formuler cette hypothèse consiste à dire que les rapports sociaux d'influence sociale à l'œuvre à un temps t dans la communauté de journalistes du journal The New York Time détermineront à un temps t+1 la présence ou l'absence d'un concept c_j dans les articles de presse signés par la journaliste a_i.

Si cette hypothèse est vraisemblable, l'analyse des traces textuelles produites par les journalistes devrait révéler des patrons récurrents entre la magnitude de l'influence sociale et les usages conceptuels. Autrement dit, si cette hypothèse est vraisemblable, la fonction (1.1) devrait pouvoir être approximée, un modèle algorithmique devrait pouvoir être induit par apprentissage machine et permettre de prédire l'évolution du réseau sociosémantique. Au contraire, si cette hypothèse est improbable, peu de régularités empiriques devraient caractériser la relation entre la magnitude de l'influence sociale et les usages conceptuels et aucun modèle algorithmique ne devrait pouvoir en être induit.

1.4 Plan de la thèse

La thèse se divise en sept chapitres. Les trois prochains chapitres présentent les éléments théoriques de la thèse. Ces éléments sont issus de la théorie des réseaux sociaux, de la sémantique vectorielle et de l'apprentissage machine. Le chapitre cinq présente la méthode et le chapitre six présente les résultats et les analyses. Le chapitre de conclusion souligne la contribution de la thèse et ses limites.

1.4.1 La théorie de l'influence sociale dans les réseaux sociaux

Dans le cadre de cette recherche doctorale, une théorie particulière de l'influence sociale est utilisée pour modéliser le mécanisme d'influence sociale à l'œuvre dans la dynamique sociosémantique du SSS. Il s'agit de la théorie des réseaux sociaux. Le but du deuxième chapitre est de proposer une synthèse de ce cadre théorique et de montrer comment l'influence sociale dans un réseau social peut être modélisée. Quatre types de mécanismes seront discutés : l'exposition sociale, la contagion, la déférence et le mimétisme des semblables.

1.4.2 La sémantique vectorielle

Lorsque les seules données empiriques disponibles pour l'analyse de la dynamique sociosémantique d'un SSS sont des traces textuelles, le seul moyen de savoir si un agent a_i a utilisé un concept c_j à un temps t est de vérifier si ce concept c_j est exprimé ou non dans un texte signé par a_i au temps t. Bien qu'en apparence simple, identifier les différents contenus conceptuels exprimés dans un corpus de textes est une problématique de recherche à part entière et extrêmement complexe. Est-ce que l'identification des différents contenus conceptuels exprimés dans un texte peut se réduire à l'identification des différents mots qui y sont présents? Est-ce qu'alors des mots synonymes expriment des contenus conceptuels différents? Est-ce que deux graphies comme « livre » et « livres » expriment aussi des concepts différents? Qu'est-ce qu'un mot? Voilà quelques questions inhérentes à cette problématique.

Le but du troisième chapitre est de présenter le cadre théorique qui est utilisé dans cette thèse pour la modélisation des contenus conceptuels dans les textes. Ce cadre théorique est la sémantique vectorielle. Le chapitre présentera les différents paramètres de ce cadre, notamment les notions de cooccurrence, de similarité sémantique et d'espace sémantique. Nous verrons également que dans le cadre de la sémantique vectorielle, un concept correspond à une région dans un espace sémantique et que chaque région peut être interprétée comme une classe d'équivalence regroupant des mots sémantiquement similaires.

1.4.3 L'apprentissage machine

Dans le cadre de cette thèse, l'induction d'une hypothèse d'approximation de la fonction (1.1) est réalisée à l'aide de différents d'apprenants automatiques. Le but du quatrième chapitre est de présenter le cadre théorique de l'apprentissage machine. Ce chapitre permettra de définir les principaux paramètres de ce cadre, notamment ceux d'espace des instances, d'espace des hypothèses, de bases d'apprentissage et de test, d'apprenant automatique et d'évaluateur.

1.4.4 La méthode

L'objectif du cinquième chapitre est de présenter la méthode utilisée pour reconstruire le mécanisme d'influence sociale. Cette méthode inclut les trois grandes phases d'une méthode de fouille de données, soit une phase de collecte de données, une phase de prétraitement des données et une phase analytique (Aggarwal, 2015, p. 3).

La première phase est la construction d'un corpus de textes composé des articles de presse du journal The New York Time. La deuxième phase implique plusieurs sous-étapes d'extraction des textes de différentes variables statiques et dynamiques nécessaires à la reconstruction. Ces sous-étapes impliquent notamment une extraction du corpus de texte du réseau social des relations de collaboration entre journalistes, une extraction des différents contenus conceptuels utilisés par les journalistes et une

extraction des variables modélisant les patrons d'influence sociale. Finalement, la phase de prétraitement se termine par la constitution d'un échantillon d'exemplaires du processus d'évolution du réseau sociosémantique.

La troisième phase est l'apprentissage machine de la fonction cible (1.1). Cette phase implique deux étapes. La première est l'approximation de la fonction (1.1) par apprentissage machine, c'est-à-dire l'induction à l'aide de différents apprenants automatiques de différents modèles de l'influence sociale. La deuxième étape est l'évaluation de ces modèles.

1.4.5 Les expérimentations

Le sixième chapitre porte sur les expérimentations d'apprentissage machine réalisées dans la thèse. Il se divise en deux étapes. Dans la première sont présentés les différents modèles de l'influence sociale induits par apprentissage machine. Ces modèles sont des arbres de décisions, des listes de règles, des forêts aléatoires et des modèles probabilistes. Ensuite, dans la deuxième étape, ces modèles sont analysés afin d'en évaluer la vraisemblance et pour démontrer s'ils peuvent approximer correctement la fonction (1.1) et prédire empiriquement l'évolution du réseau sociosémantique. Cette évaluation est basée sur une technique de validation croisée, l'analyse de courbes d'apprentissage, l'analyse d'une réduction dimensionnelle et l'analyse de corrélation.

La thèse conclut avec une discussion des retombées de la recherche et de ses limites.

CHAPITRE II

RÉSEAUX SOCIAUX ET MÉCANISMES D'INFLUENCE SOCIALE

2.0 Introduction

L'objectif de ce chapitre est de présenter un cadre théorique particulier pour la modélisation des mécanismes d'influence sociale, soit le cadre de la théorie des réseaux sociaux. Dans un premier temps, le texte introduit le concept d'influence sociale. Il se divise ensuite en deux sections. Dans une première section sont introduits les principaux concepts de la théorie des réseaux sociaux. Dans une deuxième section sont introduits et définis quatre types de mécanismes d'influence sociale à l'œuvre dans les réseaux sociaux.

2.1 L'influence sociale

Un mécanisme d'influence sociale est à l'œuvre lorsqu'un individu adapte son comportement en fonction du comportement des individus avec lesquels il interagit (Festinger, 1954). Un comportement doit être compris dans un sens très général qui n'est pas spécifique à l'étude des SSS. Il réfère aux décisions, aux pratiques, aux attitudes, aux cognitions ou autres variables psychologiques chez un individu.

L'influence sociale peut être directe, par exemple via des interactions face à face, ou indirecte, par exemple via des interactions stigmergiques[2], ou encore imaginées via les anticipations que génère chez un individu l'intériorisation d'une norme culturelle.

---

[2] Des interactions stigmergiques sont des interactions médiatisées par une modification de l'environnement. Il y a influence sociale stigmergique lorsqu'un agent effectue une modification de l'environnement qui affectera ensuite le comportement d'un autre agent. C'est un processus qui n'implique pas de contact social direct entre deux agents, ni même de la connaissance de l'existence d'autrui.

Compte tenu de l'importance des mécanismes d'influences sociales dans la vie sociale, ils représentent un domaine d'étude extrêmement vaste. Elle est étudiée à la fois par la psychologie sociale (Cialdini & Goldstein, 2004), la sociologie (Friedkin, 2006), la science du marketing (Watts & Dodds, 2007), l'économie (Bikhchandani et al., 1998), la communication (Mucchielli, 2009), l'étude des médias (McQuail, 1987), l'épidémiologie (Christakis & Fowler, 2007), l'épistémologie sociale (Goldman, 1999), l'intelligence artificielle collective (Kennedy & Eberhart, 2001) et bien d'autres.

Plusieurs mécanismes y sont étudiés : des mécanismes d'obéissance à une autorité légitime, de conformisme social, de conformité à une norme, de réciprocité, de conditionnement opérant, d'identification à des individus valorisés, de réduction de la dissonance cognitive, d'excitation émotive, de déférence, de re-cadrage, d'argumentation, de contagion sociale, de mimétisme, etc. (Bohner & Dickel, 2011; Cialdini & Trost, 1998; Latané, 1996; Mercier & Sperber, 2011; Moscovici, Sherrard, & Heinz, 1976; Petty & Brinol, 2010; Surowiecki, 2005; Wood, 2000).

Afin d'introduire le concept d'influence sociale, il est heuristique de rappeler une étude classique comme celle de Asch (Asch, 1951). Dans cette étude, on cherche à montrer à quel point les perceptions subjectives d'un individu peuvent être affectées par l'influence sociale. L'étude est basée sur l'expérimentation suivante. On invite dans un premier temps un groupe d'individus dans une salle et on les informe qu'ils participeront tous à un test de vision. Ce test est illustré dans la Figure 2.1. Il consiste à demander à chaque participant de mentionner lequel des traits A, B ou C est de même longueur que le trait de référence X.

En fait, dans ce groupe, tous les participants sont de connivence avec la chercheuse, sauf un, qui lui croit qu'il s'agit d'un véritable test. C'est cet individu qui représente le véritable sujet de l'expérimentation. Les complices de la chercheuse ont des directives très claires : durant les premiers essais ils doivent tous donner la bonne réponse, mais à partir du troisième essai, tous les participants de connivence doivent mentionner de

manière consensuelle une réponse erronée. On analyse alors la réaction du sujet face aux erreurs, évidentes, mais unanimes, des autres participants au test de vision. Le test est répété avec plusieurs sujets.

Or, bien que la majorité des sujets de l'expérimentation répondent correctement au test malgré un consensus différent chez les autres participants de connivence, un nombre surprenant d'entre eux, soit 36%, préférent imiter la réponse erronée, mais consensuelle. Ils reproduisent la réponse consensuelle pour différentes raisons. Certains font davantage confiance aux jugements des autres qu'à leurs propres aptitudes visuelles, mais d'autres imitent la réponse consensuelle tout en étant convaincus que ces participants sont dans l'erreur. L'expérimentation de Asch montre de manière spectaculaire l'un des principaux mécanismes d'influence sociale : le conformisme.

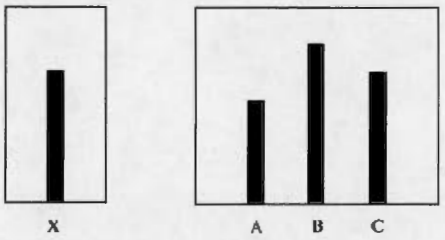

Figure 2.1: Le test de vision dans l'expérimentation de Asch.

Une synthèse de l'ensemble de la littérature sur l'influence sociale serait bien au-delà des objectifs de ce chapitre. Toutefois, afin d'introduire le concept d'influence sociale, nous pouvons rappeler une synthèse récemment proposée par Cialdini et ses collègues (Cialdini & Goldstein, 2004; Cialdini & James, 2009; Cialdini & Mortensen, 2008; Cialdini & Trost, 1998; Griskevicius et al., 2009; Sundie, Cialdini, Griskevicius, & Kenrick, 2006, 2012). Cette synthèse nous aidera dans la prochaine section à interpréter le fonctionnement des différents mécanismes d'influence sociale à l'œuvre dans les réseaux sociaux.

Cialdini et ses collègues proposent d'interpréter les différents mécanismes d'influence sociale sur la base des différentes motivations psychologiques à partir desquelles ils opèrent. Selon ces chercheurs, trois catégories de motivations psychologiques permettent de comprendre comment fonctionne l'influence sociale, soit des motivations informationnelles, relationnelles et des motivations liées à l'image de soi. Ces trois catégories de motivation peuvent être résumées de la manière suivante.

Motivation informationnelle : En situation d'incertitude, les agents sont à la recherche d'informations supplémentaires dans leur environnement social qui vont leur permettre de prendre une décision appropriée, adaptée ou pertinente. Une situation d'incertitude se présente lorsque, par exemple, un agent a peu d'expérience face à un type particulier de problème, ou que la définition du problème est ambiguë, que les conséquences sont importantes ou difficiles à anticiper. Se conformer à un comportement très répandu socialement, déférer à une personne de confiance ou d'autorité, imiter une personne similaire à soi, sont toutes des heuristiques de type « preuve sociale » très efficaces pour résoudre plusieurs problèmes rencontrés en situation d'incertitude.

Motivation relationnelle ou d'affiliation : Les individus accordent beaucoup d'importance au maintien des relations sociales significatives à leurs yeux, que ce soit avec des membres de leur famille, de leurs amis, leurs collègues de travail ou autres relations sociales. Des différences trop marquées entre les comportements (e.g. opinions, pratiques, etc.) de deux individus peuvent compromettre le lien social. Des opinions trop différentes entre deux amis peuvent mener à une rupture de l'amitié entre eux. Un individu trop différent des membres d'un groupe peut rencontrer des résistances à son affiliation à ce groupe. Pour maintenir des relations sociales, les individus sont généralement disposés à ajuster leurs comportements, en adopter de nouveaux et en abandonner d'autres. Par exemple, adopter les comportements prescrits par une norme partagée par les membres d'un groupe, notamment son lexique, ses pratiques, sera généralement indispensable à l'intégration de ce groupe.

Motivations liées à l'image de soi : Une troisième catégorie de motivations regroupe celles liées à l'importance de maintenir une image positive de soi-même. Maintenir une cohérence dans ses comportements, ses croyances ou ses attitudes, reproduire des comportements associés à des individus valorisés, obtenir une approbation sociale, sont tous des moyens par lesquels un individu construit une image positive de lui-même.

En somme, selon Cialdini et ses collègues, c'est parce qu'un individu utilise son environnement social comme une source importante d'information, parce qu'il accorde beaucoup d'importance au maintien de ses relations sociales et parce qu'il souhaite maintenir une image positive de lui-même, que des mécanismes d'influence sociale s'enclenchent et opèrent dans la vie sociale.

2.2 Les réseaux sociaux

Historiquement, les réseaux ont d'abord formé un objet d'étude purement mathématique pour la théorie des graphes, la combinatoire et l'algèbre linéaire. Ce n'est qu'à partir de la seconde moitié du 20e siècle que les réseaux sont progressivement devenus un cadre théorique et une manière de modéliser différents systèmes empiriques étudiés dans les sciences humaines et sociales, la biologie, l'informatique et la physique.

Un réseau social forme un cadre formel pour la modélisation des systèmes sociaux (Borgatti, Everett, & Johnson, 2013; Borgatti & Halgin, 2011; Carrington, Scott, & Wasserman, 2005; Monge & Contractor, 2003; Scott, 2013; Wasserman & Faust, 1994). Ce modèle est un graphe composé de nœuds qui représentent les agents d'une population et d'arcs entre ces nœuds qui représentent des liens sociaux entre ces agents. Les agents modélisés dans un réseau social peuvent être des individus, comme les membres d'une équipe sportive, les étudiants d'une école, les employés d'une entreprise, les scientifiques d'une communauté épistémique; ils peuvent aussi être des

agents collectifs, comme les entreprises d'un secteur économique, les banques d'un pays, les groupes sociaux d'une société, etc.

Plusieurs typologies des liens dans un réseau social ont été suggérées. Ce sont des typologies basées par exemple sur les différentes méthodes de collectes des données relationnelles (e.g. questionnaire sociométrique, analyse d'archive), les différents contenus des relations (e.g. relation d'amitié, relation de parenté), l'unité d'analyse (e.g. réseau égo-centré, réseau organisationnel) (Lazega, 1998). Borgatti et ses collègues suggèrent quant à eux de distinguer deux types de liens possibles dans un réseau social : des liens modélisant des états et des liens modélisant des évènements (Borgatti, Brass, & Halgin, 2014; Borgatti & Halgin, 2011). Le premier type modélise des états relationnels entre agents caractérisés par une certaine permanence dans le temps, par exemple des relations d'amitié, de parenté, d'affiliation à un groupe ou à une organisation. Le deuxième type modélise des événements interactionnels entre agents caractérisés par une discontinuité dans le temps. Ce type de lien modélise par exemple des collaborations qui ont eu lieu entre collègues, des transactions économiques entre vendeurs et acheteurs, des contacts sexuels entre partenaires ou des échanges téléphoniques entre deux individus.

Toujours selon Borgatti et ses collègues, au-delà des types de liens, le modèle des réseaux sociaux peut être séparé en deux grandes approches : une théorie des conséquences des réseaux et une théorie de causes de la formation des réseaux (Borgatti & Halgin, 2011, p. 1168). Dans la première approche, les réseaux forment la variable explicative d'un phénomène. Des mécanismes en interaction avec les structures d'un réseau social servent à expliquer un résultat, généralement un comportement individuel ou collectif. Dans cette approche, on retrouve par exemple la théorie de Granovetter sur la « force des liens faibles » (Granovetter, 1973). Cette théorie fait appel à certaines propriétés des réseaux sociaux, notamment la présence de liens faibles, pour expliquer pourquoi certains agents ont plus de succès que d'autres sur le marché de l'emploi.

Dans la deuxième approche, c'est la formation d'un réseau qui est l'objet d'une explication. Les réseaux sont la variable dépendante. L'objet d'étude est du côté des mécanismes responsables de l'apparition de structures et de propriétés spécifiques à certains réseaux. On retrouve notamment dans cette approche la théorie des réseaux complexes (Albert & Barabási, 2002; Boccaletti, Latora, Moreno, Chavez, & Hwang, 2006; Newman, 2003). Cette théorie utilise certains mécanismes comme l'attachement préférentiel et l'assortativité pour expliquer la présence de propriétés statistiques particulières aux réseaux complexes comme des invariances scalaires et une structure en petit-monde.

L'étude des mécanismes d'influence sociale à l'œuvre dans les réseaux sociaux relève de la première approche. Ce sont les structures des réseaux qui servent à expliquer un résultat particulier, soit l'adoption d'un comportement par un individu.

2.2.1 Concepts de base de la théorie des réseaux

Avant d'introduire les mécanismes d'influence sociale à l'œuvre dans les réseaux sociaux, certains concepts de base de la théorie des réseaux sociaux doivent être présentés. Cinq concepts sont introduits : le voisinage, le degré, la centralité de degré, la distance et l'équivalence structurale.

Un réseau est un graphe composé d'un ensemble de nœuds connectés entre eux par des liens. Dans sa forme la plus simple, un réseau est défini par deux éléments G = (A, L), où $A = \{a_1 \dots a_n\}$ correspond à l'ensemble des n noeuds qui composent le réseau et où $L = \{l_1 \dots l_m\} \in A \times A$ correspond à l'ensemble de m liens qui connectent différentes paires de nœuds (a_i, a_j).

Ce modèle simple peut cependant être généralisé. Un réseau peut être directionnel ou non-directionnel, dichotomique ou valué, simple ou multiplexe. Un réseau non-directionnel est caractérisé par des liens symétriques, c'est-à-dire où le lien entre les

nœuds a_i et a_j est identique au lien entre les nœuds a_j et a_i. Au contraire, dans un réseau directionnel $(a_i, a_j) \neq (a_j, a_i)$. Dans un réseau dichotomique, les liens sont simplement présents ou absents, alors que dans un réseau valué une pondération leur est associée. Enfin, dans un réseau simple il n'y a qu'un seul lien possible entre deux nœuds, alors que dans un réseau multiplexe deux nœuds peuvent être liés par plusieurs connexions parallèles.

Dans la Figure 2.2 est illustré un exemple de réseau simple, non-directionnel et dichotomique composé de 17 nœuds et 22 liens. Nous ferons référence à cet exemple à plusieurs reprises dans la suite du texte afin d'illustrer différents concepts.

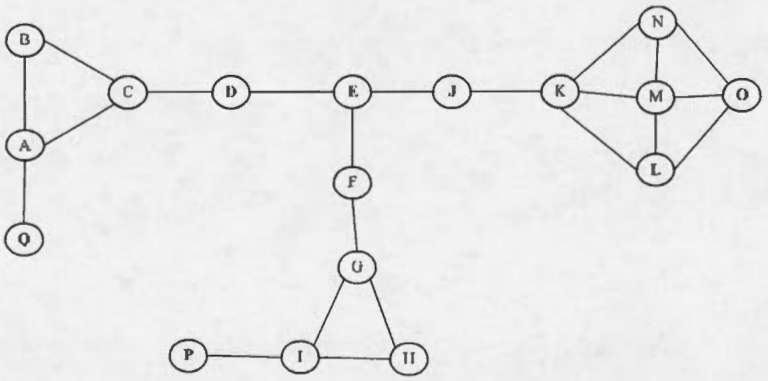

Figure 2.2: Un réseau simple, non-directionnel et dichotomique composé de 17 nœuds et 22 liens.

2.2.2 Voisinage

Un premier concept de base est le voisinage d'un nœud dans un réseau. Le voisinage d'un nœud a_i correspond à l'ensemble des nœuds $a_j \in A$ qui sont adjacents à a_i, c'est-à-dire qui sont directement connectés avec lui dans le réseau. Le voisinage d'un nœud a_i est dénoté par $V(a_i)$ et est défini ainsi :

$$V(a_i) = \{a_j : (a_i, a_j) \in L\} \qquad \text{Eq. (2.1)}$$

À titre d'illustration, dans le réseau de la Figure 2.2, le voisinage du nœud M est composé des nœuds {K, N,O,L} et celui du nœud J des nœuds {E,K}.

2.2.3 Degré

Le degré d'un nœud correspond à la taille de son voisinage, c'est-à-dire le nombre de nœuds qui lui sont adjacents. Il est dénoté par $d^{\circ}(a_i)$:

$$d^{\circ}(a_i) = |V(a_i)| \qquad \text{Eq. (2.2)}$$

Dans le réseau de la Figure 2.2, le nœud M a un degré de quatre, alors que J a un degré de deux.

2.2.4 Centralité de degré

La centralité d'un nœud dans un réseau peut faire l'objet de plusieurs conceptions alternatives. Les plus usités sont la centralité de degré, la centralité de proximité et la centralité d'intermédiarité (Freeman, 1979). De manière générale, la centralité d'un nœud est un concept qui rend compte de l'importance de la position occupée par un nœud dans la structure d'un réseau. Les différentes conceptions de la centralité permettent de mettre en relief différents aspects de cette position. L'une des conceptions les plus courantes est la centralité de degré. Elle correspond simplement à la quantité de connexions d'un nœud par rapport à la quantité de connexions possibles :

$$c_d(a_i) = \frac{d^{\circ}(a_i)}{n-1} \qquad \text{Eq. (2.3)}$$

En d'autres mots, plus le degré d'un nœud est élevé, plus il est central. Dans cette conception, l'aspect mis en relief de la position d'un nœud est son niveau d'intégration dans l'ensemble de la structure du réseau.

2.2.5 Distance

La distance entre deux nœuds a_i et a_j, dénotée ici par $d(a_i, a_j)$, est un autre concept central de l'analyse des réseaux. Il existe plusieurs conceptions de la distance dans un réseau. La définition la plus naturelle est le géodésique, c'est-à-dire le chemin le plus court entre deux nœuds a_i et a_j dans un réseau. C'est le nombre minimal d'intermédiaires ou le degré de séparation entre deux nœuds. Par exemple, dans la Figure 2.2, les nœuds C et D sont adjacents, ils ont donc un géodésique de un. Les nœuds B et D ne sont pas adjacents, mais ils ont un voisin commun, le nœud C, ils ont donc un géodésique de deux qui les séparent.

2.2.6 Équivalence structurale

La similitude des positions occupées par deux nœuds dans un réseau est un autre concept important de l'analyse des réseaux. Une fois de plus, il y a plusieurs conceptions possibles de cette similitude. La plus commune est appelée l'équivalence structurale. Dans cette conception, la similitude des positions occupées par deux nœuds est fonction de la taille de l'intersection de leur voisinage respectif. Autrement dit, plus deux nœuds a_i et a_j sont connectés aux mêmes nœuds, plus ils sont considérés structurellement équivalents dans le réseau. Formellement, l'équivalence structurale entre deux nœuds peut être définie de la manière suivante:

$$e_s(a_i, a_j) = \frac{|V(a_i) \cap V(a_j)|}{|V(a_i) \cup V(a_j)|} \qquad \text{Eq. (2.4)}$$

Dans cette définition, $e_s(a_i, a_j) = 1$ lorsque deux nœuds sont parfaitement équivalents structurellement, c'est-à-dire que leur voisinage est identique, comme c'est le cas pour les nœuds N et L du réseau de la Figure 2.2. Ces deux nœuds sont parfaitement équivalents structuralement et nous pouvons les substituer l'un à l'autre sans affecter

le réseau. À l'inverse, $e_s(a_i, a_j) = 0$ lorsque deux nœuds n'ont aucun voisin commun, comme c'est le cas, par exemple, pour les nœuds Q et P ou C et G.[3]

2.3 Mécanismes d'influence sociale dans les réseaux sociaux

La théorie des réseaux sociaux forme un cadre d'analyse particulier des mécanismes d'influence sociale. Dans la littérature, une très grande diversité de conceptions des mécanismes d'influence sociale a été proposée et étudiée. Toutefois, malgré cette diversité apparente, plusieurs de ces conceptions représentent en réalité différentes opérationnalisations d'un même type de mécanisme. Par conséquent, dans ce chapitre, une typologie de ces mécanismes incluant seulement quatre catégories est proposée : les mécanismes d'exposition sociale, de contagion sociale, de déférence et de mimétisme des semblables.[4]

Afin d'introduire ces quatre types de mécanismes d'influence sociale, nous supposerons la plupart du temps un réseau simple non-directionnel $G_t = (A, L, W)$ représentant l'état d'un réseau social à un temps t (lorsque ce ne sera pas le cas nous le préciserons). Dans ce réseau, $A = \{a_1 \dots a_n\}$ représente un ensemble de n agents et $L_a^a = \{l_1 \dots l_m\}$ correspond à un ensemble de m liens entre agents. Le paramètre $W = \{w_1 \dots w_n\}$ est un attribut représentant les comportements passés de chaque agent $a_i \in A$. Les valeurs de cet attribut expriment les attitudes, les opinions, les pratiques, les décisions ou autres comportements chez les membres du réseau, dont on soupçonne qu'ils ont un impact sur le processus d'influence sociale étudié. De manière à simplifier

---

[3] Dans la littérature, on fait parfois la distinction entre équivalence structurale et équivalence régulière (Borgatti & Everett, 1992). L'équivalence structurale est alors définie comme une équivalence régulière parfaite, c'est-à-dire lorsque deux nœuds ont exactement le même voisinage. Définie de cette manière, l'équivalence structurale apparaît comme un cas limite d'équivalence régulière. Empiriquement, elle est très rare. Cette distinction entre équivalence structurale et régulière n'est pas utile pour notre propos. Lorsque l'équivalence structurale est parfaite, nous préciserons simplement qu'elle est parfaite, sans lui réserver un nom particulier.

[4] Par ailleurs, d'autres typologies des mécanismes d'influence sociale dans les réseaux sociaux sont possibles, voir notamment les typologies de (P. DiMaggio & Garip, 2012; Friedkin & Johnsen, 1999; Rice, 1993; Young, 2009).

le plus possible la présentation des mécanismes d'influence sociale, nous allons restreindre cet attribut à des valeurs binaires. Ainsi, dans la présentation qui suit, $w_j = 1$ si, à un temps t, un agent a_j a déjà adopté un comportement particulier et $w_j = 0$ dans le cas inverse. Cette simplification affecte très peu la compréhension du fonctionnement des différents mécanismes.

Nous supposerons aussi une variable dépendante $y_{i,t+1}$. Cette variable est l'explicandum d'un mécanisme d'influence sociale. Elle est généralement, elle aussi, réduite à des valeurs binaires. Elle modélise l'adoption ou non d'un comportement particulier par a_i à un temps t+1. Cette variable dépendante modélise, par exemple, la décision de a_i d'adopter ou non à t+1 une innovation qui circule dans son réseau, d'utiliser ou non un concept particulier, d'adhérer ou non à une opinion qui circule dans son réseau, de voter pour ou contre un parti politique donné, etc. (Lopez-Pintado & Watts, 2008; Watts & Dodds, 2009, 2007).

2.3.1 Des mécanismes à base de seuils

Les quatre types de mécanismes d'influence sociale qui seront discutés ci-après partagent tous un même schéma général de fonctionnement. Ce sont des mécanismes dont le fonctionnement est basé sur des seuils de déclenchement (Harrison & Carroll, 2002; López-Pintado, 2008; Lopez-Pintado & Watts, 2008; Valente, 2005; Watts & Dodds, 2009).[5]

Ce schéma de fonctionnement a la forme générale suivante :

---

[5] Par ailleurs, ce schéma de fonctionnement n'est pas spécifique aux mécanismes d'influence sociale dans les réseaux. Il est commun à plusieurs mécanismes sociaux à l'œuvre dans différents processus collectifs (Granovetter, 1978; J. H. Miller & Page, 2007; Schelling, 2006).

$$y_{i,t+1} = f(x_{i,t}) = \begin{cases} 1, & x_{i,t} \geq \theta \\ 0, & x_{i,t} < \theta \end{cases} \qquad \text{Eq. (2.5)}$$

Le fonctionnement de ces mécanismes est fondé sur l'hypothèse que l'adoption d'un comportement particulier par un agent $a_i \in A$ à un temps t+1 — représentée ici par la variable dépendante $y_{i,t+1}$ — dépend de la magnitude de l'influence sociale dirigée sur a_i à un temps t. Cette magnitude est représentée dans la formule Eq. (2.5) par la variable $x_{i,t}$. Cette variable correspond à ce que Latané appelle la « quantité de forces sociales » dirigées sur a_i, c'est-à-dire le cumul des impacts exercés sur a_i par le comportement de chaque agent $a_j \in A$ (Latané, 1981, 1996).

La relation de détermination entre la variable $y_{i,t+1}$ et la variable $x_{i,t}$ prend la forme d'une fonction à seuil, dénotée dans Eq. (2.5) par le symbole θ. Cette fonction à seuil modélise les conditions de déclenchement d'un comportement chez un agent (Rolfe, 2009). Autrement dit, le fonctionnement des mécanismes d'influence sociale est fondé sur l'hypothèse que l'adoption d'un comportement par un agent à un temps t+1 est conditionnelle à ce que la magnitude de l'influence sociale dirigée sur lui dépasse un certain seuil. Par exemple : un individu acceptera de participer à une émeute que si une proportion suffisante d'individus y participent aussi; un individu acceptera de s'abonner à une plateforme web comme Facebook que si suffisamment de ses amis le font aussi; une entreprise adoptera une innovation technologique que si suffisamment de ses compétiteurs l'ont aussi adoptée; une scientifique citera les travaux d'une autre scientifique que si plusieurs scientifiques prestigieux ont déjà eux aussi cité cette dernière.

Les différents types de mécanisme d'influence sociale présentés ci-après partagent tous ce schéma général de fonctionnement. La spécificité de chacun des types de mécanismes se situe au niveau de la définition de la variable $x_{i,t}$, c'est-à-dire de la définition de la magnitude de l'influence sociale. Ces spécificités sont discutées dans les prochaines sections.

2.3.2 L'exposition sociale

Un premier type de mécanisme d'influence sociale dans les réseaux sociaux regroupe différentes opérationnalisations de l'exposition sociale. Ce type de mécanisme est basé sur l'hypothèse que, dans un réseau social, l'adoption d'un comportement par un agent dépend de l'étendue du partage social de ce comportement. Plus un agent est exposé à une quantité importante d'individus ayant déjà adopté un comportement, plus cet agent aura tendance à adopter également ce comportement.

Les différentes opérationnalisations de ce mécanisme varient en fonction de leur conception de l'exposition sociale. Dans une première opérationnalisation, l'étendue de l'exposition se fait à l'ensemble du réseau social et la magnitude de l'influence sociale est définie de la manière suivante :

$$x_{i,t} = \frac{1}{n} \sum_{a_j \in A} w_j \qquad \text{Eq. (2.6)}$$

Dans cette opérationnalisation, la magnitude de l'influence sociale d'un comportement correspond simplement à la proportion d'agents dans le réseau qui ont déjà adopté ce comportement. Plus cette proportion est grande et plus grande est cette magnitude (Bass, 1969; Granovetter, 1978).

Une variante importante de ce type de mécanisme est l'exposition sociale égo-centrée. Cette opérationnalisation est basée sur l'hypothèse que l'exposition sociale opère uniquement via des contacts sociaux directs dans un réseau. Ainsi, dans un réseau social, un agent ne serait pas influencé par une exposition aux comportements de tous les membres du réseau, mais seulement par une exposition aux comportements des agents qui font partie de son voisinage (Valente, 1996, 2005).

On justifie cette opérationnalisation alternative par l'argument selon lequel un agent n'aurait toujours qu'une perspective limitée sur un réseau social. Cette perspective

serait restreinte par la position d'un agent dans un réseau. Un agent n'aurait connaissance que des comportements des autres agents membres de son voisinage. Ce serait donc seulement ce sous-réseau égo-centré qui constituerait une source d'influence sociale.

Selon cette opérationnalisation particulière d'un mécanisme d'exposition sociale, la magnitude de l'influence sociale est définie de la manière suivante :

$$x_{i,t} = \frac{1}{|V(a_i)|} \sum_{a_j \in V(a_i)} w_j \qquad \text{Eq. (2.7)}$$

Cette opérationnalisation spécifie que la magnitude de l'influence sociale d'un comportement sur un agent a_i est liée à la proportion d'agents dans son voisinage qui a déjà adopté ce comportement. Plus un comportement est commun dans le voisinage de a_i, plus a_i aura tendance à s'y conformer.

L'exposition sociale est un mécanisme relativement simple, mais très important dans l'explication de processus de diffusion sociale. Il est au cœur de plusieurs études classiques en sociologie sur la diffusion sociale des innovations (Coleman, Katz, & Menzel, 1957; Ryan & Gross, 1943) et en économie sur la diffusion sociale des comportements d'achat dans une population (Bass, 1969). C'est un mécanisme à l'œuvre dans de nombreux et très différents contextes empiriques. Il serait par exemple à l'œuvre dans la diffusion de la consommation de la cigarette dans les réseaux d'amitié d'adolescents (Alexander, Piazza, Mekos, & Valente, 2001). Il seraiut aussi à l'œuvre dans la diffusion des opinions politiques. Des études ont montré que plus une opinion politique est répandue dans le réseau d'amis d'un individu, plus cet individu a tendance à se conformer à cette opinion (Lazer, Rubineau, Chetkovich, Katz, & Neblo, 2010).

De manière générale, l'exposition sociale est un mécanisme à l'œuvre dans la production d'externalités dans un réseau (Easley & Kleinberg, 2010, p. 449 et

suivantes). Il y a production d'externalités lorsque le bénéfice que peut retirer un agent de l'adoption d'un comportement est directement lié à la quantité de personnes qui ont déjà adopté ce comportement. Plus un comportement est répandu socialement, plus grand est le bénéfice pour un agent de l'adopter. Un exemple classique d'externalités est celui associé à l'adoption du téléphone comme technologie de communication. Ce comportement n'est bénéfique pour un individu que dans la mesure où une quantité suffisante d'individus utilisent déjà cette technologie pour communiquer, autrement celle-ci n'a que très peu de valeur pour eux (P. DiMaggio & Garip, 2012; Varian & Farrell, 2004).

Nous pouvons considérer l'exposition sociale comme une forme particulière de conformisme. Tel que mentionné en début de chapitre, selon Cialdini et Trost (Cialdini & Trost, 1998), plusieurs motivations peuvent expliquer pourquoi des agents adaptent leur comportement en conformité avec les comportements largement partagés socialement. Premièrement, en situation d'incertitude, un comportement largement partagé socialement représente généralement aux yeux d'un agent la meilleure solution disponible (c'est la motivation informationnelle). Deuxièmement, adopter un comportement largement partagé permet aussi à un agent d'obtenir l'approbation sociale des autres et ainsi maintenir une harmonie dans ses relations sociales (motivation relationnelle). Et troisièmement, adopter un comportement largement partagé permet à un individu d'éviter les stigmates sociaux (moquerie, rejet, etc.) typiquement attribués aux individus déviants et contribue ainsi à maintenir une image positive de soi (motivation liée à l'égo).

2.3.3 La contagion sociale

Un deuxième type de mécanisme d'influence sociale à l'œuvre dans les réseaux sociaux regroupe différentes variantes de contagion sociale (R.S. Burt, 1987; Dodds & Watts, 2004; Friedkin, 2006; R. Leenders, 1997; Marsden & Friedkin, 1993; Robins et al., 2001). Selon Monge et Contractor, la contagion sociale serait, de loin, le type de

mécanisme d'influence sociale qui a été le plus étudié dans la littérature sur l'analyse des réseaux sociaux (Monge & Contractor, 2003, p. 182).

On appelle ce type de mécanisme de la « contagion sociale » en référence, métaphoriquement, au processus de propagation épidémiologique. De manière analogue à une infection virale qui se propage dans une population par contacts entre individus, l'adoption d'un comportement se propagerait aussi dans une population par contacts sociaux entre individus (Easley & Kleinberg, 2010, p. 568).

La contagion sociale est un mécanisme fondé sur l'hypothèse que l'influence sociale est un processus déterminé par la cohésion sociale et la socialisation entre les agents membres d'un réseau social.[6] Plus la cohésion sociale est grande entre deux agents a_i et a_j, plus l'adoption d'un comportement par a_j génèrerait un impact important chez a_i et augmenterait la probabilité que a_i adopte aussi ce même comportement :

> « By appropriately taking into account the [...] behaviors displayed by their significant others, actors thus establish their own behavior. In the literature, this influence process has been labeled 'contagion'. » (R. T. A. Leenders, 2002, p. 21)

Contrairement à un mécanisme d'exposition sociale, dans lequel la magnitude de l'influence sociale d'un comportement dépend uniquement de la proportion d'individus dans l'ensemble du réseau ou dans le voisinage des agents ayant adopté ce comportement, le fonctionnement d'un mécanisme de contagion sociale est de type agrégatif (Grabisch & Rusinowska, 2013). La magnitude de l'influence sociale sur un agent a_i est le résultat de l'impact cumulé de chaque agent $a_j \in A$ avec lesquels a_i a des opportunités de socialisation. On considère que plus forte est la cohésion sociale

[6] On distingue parfois deux types de contagion sociale : une contagion par cohésion sociale et une contagion par équivalence structurale (R.S. Burt, 1987; R. S. Burt, 2010b). Afin d'éviter certaines ambiguïtés terminologiques, nous allons réserver l'expression « contagion sociale » aux mécanismes de contagion par cohésion sociale et nous allons appeler les mécanismes de contagion par équivalence structurale des mécanismes de « mimétisme des semblables ».

entre a_j et a_i, plus nombreuse seront les opportunités de socialisation et plus grande sera la contribution de a_j à la magnitude de l'influence sociale.

On peut retrouver dans la littérature sur l'analyse des réseaux sociaux, plusieurs opérationnalisations possibles d'un mécanisme de contagion sociale. Ces différentes opérationnalisations sont basées sur différentes conceptualisations de la cohésion sociale entre deux agents. Dans une première opérationnalisation de ce mécanisme, la magnitude de l'influence sociale est définie de la manière suivante :

$$x_{i,t} = \frac{1}{n-1} \sum_{\substack{a_j \in A \\ i \neq j}} \left(\frac{1}{d(a_i, a_j)} \cdot w_j \right) \qquad \text{Eq. (2.8)}$$

Cette opérationnalisation s'appuie sur le concept de distance sociale introduite précédemment. La cohésion sociale est ici réduite à la distance topologique dans le réseau. L'impact de l'adoption d'un comportement par a_j sur l'adoption future de ce comportement par a_i dépend de la distance sociale qui sépare a_j de a_i. L'impact de a_j est inversement proportionnelle à la longueur du géodésique qui le sépare de a_i dans le réseau social. Autrement dit, moins il y a d'intermédiaires entre eux, plus son impact est important. Au niveau agrégatif, plus il y a d'agents proches de a_i dans le réseau qui ont déjà adopté un comportement particulier, plus a_i aura tendance à également adopter ce comportement dans le futur (Latané, 1996).

Dans cette opérationnalisation, le concept de distance sociale modélise la force de la cohésion sociale entre deux agents. Cette opérationnalisation est basée sur l'hypothèse que plus deux agents sont proches socialement, plus ils ont d'opportunités de socialiser. Pour un agent a_i, c'est l'accumulation de ces opportunités de socialisation avec des agents qui ont déjà adopté un comportement particulier, qui détermine l'adoption éventuelle par a_i de ce même comportement.

Par ailleurs, la cohésion sociale entre deux agents dans un réseau peut être définie autrement. C'est le cas notamment dans les réseaux sociaux modélisés par des multigraphes. Dans ces réseaux, plusieurs liens parallèles peuvent relier deux agents a_i et a_j, indiquant par exemple la fréquence des interactions qui a eu lieu entre eux (un lien pour chaque évènement interactionnel). Dans ce type de réseau, une opérationnalisation alternative de la contagion sociale consiste simplement à remplacer la distance sociale par le nombre de liens $n_{i,j}$ entre deux agents :

$$x_{i,t} = \frac{1}{n-1} \sum_{\substack{a_j \in A \\ i \neq j}} (n_{i,j} \cdot w_j) \qquad \text{Eq. (2.9)}$$

Cette opérationnalisation alternative s'interprète de la manière suivante : plus un comportement w_j a été adopté par des agents avec qui a_i est fréquemment en contact, plus a_i aura tendance à adopter également ce comportement (R. T. A. Leenders, 2002, p. 28).

Plusieurs autres opérationnalisations sont aussi possibles, notamment via les concepts de clique ou de k-plex. Par exemple, la force de la cohésion sociale entre deux agents peut être modélisée via le nombre de cliques dans le réseau qui regroupe ces deux agents (R. T. A. Leenders, 2002, p. 28).

Des études empiriques sur les réseaux sociaux d'inventeurs permettent d'illustrer comment est à l'œuvre un mécanisme de contagion. Ces études font appel à un mécanisme de contagion sociale pour expliquer la diffusion des connaissances entre inventeurs (Agrawal, Kapur, & McHale, 2008; Sorenson, Rivkin, & Fleming, 2006). Méthodologiquement, ces études s'appuient sur l'analyse du réseau social des relations de collaboration entre inventeurs, tel que nous pouvons les identifier par la cosignature des brevets. Dans ce réseau de collaboration, la distance sociale entre deux inventeurs ayant cosigné un brevet est de zéro, la distance sociale entre deux inventeurs ayant en

commun un même collaborateur est de un, et ainsi de suite. Elles s'appuient également sur l'analyse des liens de citations entre brevets. Ainsi, on considère qu'il y a eu transfert de connaissance d'un inventeur a_j vers un inventeur a_i, lorsque a_i cite un brevet déposé par a_j.

Ces études convergent toutes vers une même conclusion : toutes choses étant égales par ailleurs (notamment en ce qui a trait au niveau de complexité des connaissances transmises), plus deux inventeurs sont socialement distants, plus grande est la probabilité que la transmission de connaissances entre eux, c'est-à-dire la contagion, soit compromise (Sorenson et al., 2006, p. 999).

D'autre part, la contagion sociale telle qu'elle a été définie précédemment est un mécanisme d'influence sociale à l'œuvre dans de nombreux et très différents processus de diffusion sociale. Elle est à l'œuvre, notamment, dans la propagation des divorces chez les couples (Åberg, 2009), la diffusion des comportements menant à l'obésité (Christakis & Fowler, 2007), la diffusion de contenus sur les blogues (Lerman & Ghosh, 2010) et la diffusion des comportements de consommation de drogues chez les adolescents (Fujimoto & Valente, 2012).

Plusieurs des motivations psychologiques identifiées par Cialdini et ses collègues permettent de comprendre le fonctionnement d'un mécanisme de contagion sociale. Tout d'abord, la cohésion sociale est un état social très valorisé par les individus. Ceux-ci sont disposés à modifier leur comportement pour préserver cet état. Ceci renvoie aux motivations relationnelles. D'autre part, la contagion sociale fonctionne aussi grâce à cette tendance chez les agents à utiliser leur environnement social comme une source d'information de confiance à partir de laquelle ils vont adapter leur comportement. La répétition des interactions sociales entre agents proches dans un réseau permettrait de construire des rapports de confiance entre agents. En situation d'incertitude, une relation de confiance permet à un agent de mieux évaluer un comportement chez autrui. Les individus font davantage confiance aux agents avec qui ils socialisent

fréquemment et parce qu'ils font davantage confiance à ces derniers, ils sont beaucoup plus enclins à adopter leurs comportements (R. S. Burt, 2010a, p. 248).

2.3.4 La déférence

La plupart des réseaux sociaux sont caractérisés par une forme de hiérarchie sociale. Dans ces réseaux, certains agents occupent des positions sociales plus importantes que d'autres, soit en termes d'accès aux ressources, de capital, de pouvoir, de contrôle ou de visibilité (P. DiMaggio & Garip, 2012). Un troisième type de mécanisme d'influence sociale est spécifique à ces réseaux. Il regroupe différentes opérationnalisations d'un mécanisme de déférence.[7] Ce type de mécanisme est basé sur l'hypothèse que l'impact généré par l'adoption d'un comportement par un agent a_j sur l'adoption future de ce même comportement par un agent a_i, est fonction de la position sociale de a_j dans le réseau. Plus importante est la position sociale d'un agent dans un réseau plus grand est son impact sur les autres.

Il y a plusieurs opérationnalisations possibles d'un mécanisme de déférence. Ces opérationnalisations varient selon la définition donnée à la hiérarchie des positions sociales dans un réseau. Une définition usuelle de cette hiérarchie est fondée sur le concept de centralité dans un réseau (Friedkin, 1991), notamment la centralité de degré définie précédemment dans Eq. (2.3). Cette opérationnalisation repose sur l'hypothèse (largement validée empiriquement) que la centralité des agents dans un réseau est corrélée à plusieurs caractéristiques spécifiques aux relations hiérarchiques entre agents. Les différences de centralité entre agents dans un réseau indiquent généralement des différences au niveau de leur statut social, des différences au niveau de leur capital social, de leur autorité, leur prestige ou de leur pouvoir (Scott, 2013, p. 122; Wasserman & Faust, 1994, Chapter 5). La centralité d'un agent dans un réseau

[7] D'autres noms peuvent être donnés à ce type de mécanisme. Par exemple, Leenders parle de « mécanisme de gravitation » (R. T. A. Leenders, 2002).

social signifie généralement qu'il occupe une position sociale importante, que son « poids social » est grand et par conséquent qu'il a un impact important sur la magnitude de l'influence sociale (Friedkin, 2006, p. 87).

L'équation suivante est une opérationnalisation de la déférence dans laquelle la position sociale d'un agent est définie par sa centralité de degré :

$$x_{i,t} = \frac{1}{n-1} \sum_{\substack{a_j \in A \\ i \neq j}} (c_d(a_j) \cdot w_j) \qquad \text{Eq. (2.10)}$$

Plus grande est la centralité de degré d'un agent a_j, plus l'adoption d'un comportement par ce dernier a un impact important sur l'adoption future de ce comportement par a_i.

De manière similaire à un mécanisme de contagion sociale, dans lequel c'est la force de la cohésion sociale qui pondère l'impact de a_j sur a_i, dans un mécanisme de déférence, c'est la centralité de la position de a_j qui pondère son impact sur a_i. La définition Eq. (2.10) montre également que le fonctionnement des mécanismes de déférence est aussi de nature agrégative. La magnitude de l'influence sociale d'un comportement sur un agent a_i est égale au cumul du « poids social » des agents qui ont déjà adopté ce comportement. Plus un comportement a été adopté par des agents importants d'un réseau social, plus ce comportement sera adopté dans le futur par d'autres agents.

Une autre opérationnalisation possible de la centralité d'un nœud dans un réseau est la centralité de proximité. Cette manière de concevoir l'importance d'un agent dans un réseau est basée sur la distance moyenne entre un agent et l'ensemble des autres agents qui composent le réseau. En d'autres mots, plus un agent est, en moyenne, proche des autres agent du réseau, plus cet agent est considéré central. Sa définition formelle est la suivante :

$$c_p(a_i) = \frac{n-1}{\sum_{\substack{a_j \in A \\ i \neq j}} d(a_i, a_j)} \qquad \text{Eq. (2.11)}$$

L'aspect mis en relief par cette conception de la centralité est le rayon d'action ou la portée sociale d'un agent sur l'ensemble de la structure d'un réseau. Plus un agent occupe une position caractérisée par une grande centralité de proximité, plus son rayon d'action et sa portée sont grands, moins il y a d'intermédiaires qui le séparent des autres agents. Les nœuds E et M dans le réseau de la Figure 2.2 en début de chapitre illustrent bien la différence entre les aspects mis en relief par la centralité de degré et ceux de la centralité de proximité. Le nœud M a une centralité de degré supérieure à celle de E, il est, en ce sens, plus intégré dans le réseau que E. Mais le nœud E est plus proche, en moyenne, de tous les autres nœuds du réseau, son rayon d'action est plus grand que M.

Une étude de Susarla et de ses collègues sur la diffusion des contenus numériques sur une plateforme web comme YouTube permet de bien illustrer le rôle des mécanismes de déférence, notamment dans des contextes d'incertitude (Susarla, Oh, & Tan, 2012). Le réseau social implémenté sur YouTube est formé des interactions entre membres abonnés à des « chaînes de diffusion ». Une chaîne de diffusion regroupe l'ensemble des vidéos publiées par un membre. Dans ce type de réseau, un agent a_i est lié à un autre agent a_j s'il est « abonné » à l'une de ses chaînes. Les agents qui occupent une position caractérisée par une grande centralité de degré sont ceux qui possèdent un très grand nombre d'abonnés.

Ce que l'étude de Susarla et de ses collègues montre est que la publication d'un contenu numérique (une vidéo) par les membres très centraux a un impact disproportionné par rapport aux autres membres plus excentrés sur la diffusion de ce contenu. Autrement dit, la publication d'une vidéo par un membre central s'ensuit généralement par une republication massive de cette vidéo sur les « chaînes » des autres membres de YouTube. Les agents centraux ont un effet dit « multiplicateur ».

Selon Friedkin, une première explication de cet « effet multiplicateur » est liée à la grande visibilité des agents centraux. Plus un agent est central, plus il est connecté ou proche socialement des autres membres du réseau, plus ses comportements sont visibles aux autres et plus les opportunités de comparaison sociale sont nombreuses :

> « In this literature on centrality, visibility is not only a precondition of interpersonal influence but also an indicator of interpersonal salience; hence, visibility implies influence. Visible opinions may be influential simply by virtue of their visibility; moreover, visibility may be a basis of power or a direct reflection of the power bases held by an actor. » (Friedkin, 2006, p. 92)

Susarla et ses collègues suggèrent une autre explication de l'impact des agents centraux. Cette explication est fondée sur ce que Cialdini et ses collègues appellent des motivations informationnelles. Sur YouTube en particulier, mais aussi dans d'autres réseaux dont l'évolution est extrêmement rapide, il est très difficile pour les membres de rester informés de tous les changements qui ont lieu, par exemple de tous les nouveaux contenus numériques publiés par ses membres. L'évolution est trop rapide et la quantité d'information trop grande:

> « The myopic nature of product discovery coupled with the range and depth of offerings and the growth of titles in YouTube substantially increase the uncertainty in searching and locating content. » (Susarla et al., 2012, p. 26)

Il est aussi très difficile pour un individu d'évaluer la valeur des innombrables nouveaux contenus numériques qui apparaissent sur YouTube. Dans ce contexte d'incertitude, que ce soit sur YouTube ou dans d'autres réseaux sociaux, les agents centraux sont souvent perçus comme des sources d'information plus fiables que les agents périphériques. Les agents centraux d'un réseau sont généralement perçus comme détenant plus d'informations sur les enjeux importants d'une situation et comme étant informés plus rapidement que les agents périphériques. Par conséquent, on leur fait davantage confiance, ils servent en quelque sorte de référence ou de repère.

Par ailleurs, la déférence n'est pas un mécanisme uniquement à l'œuvre sur YouTube. Ce type de mécanisme est, en particulier, au cœur des processus de diffusion sociale

des comportements d'achat étudiés en marketing viral (Goel & Goldstein, 2013; Iyengar, Van den Bulte, & Valente, 2011; Leskovec et al., 2007; Van den Bulte & Joshi, 2007). Il est aussi à l'œuvre dans les réseaux de médecins qui font face à beaucoup d'incertitude sur la valeur des nouveaux médicaments mis en marché par les compagnies pharmaceutiques (Strang & Tuma, 1993).

2.3.5 Le mimétisme des semblables

Un quatrième type de mécanisme regroupe différentes opérationnalisations d'un mécanisme d'influence sociale que nous appellerons le mimétisme des semblables. Leenders explique de la manière suivante l'hypothèse sur laquelle est fondé ce type de mécanisme:

> « [...] ego compares himself to those alters whom he considers similar to him in relevant respects, asking himself 'what would another person do if he were in my shoes?' Ego perceives (or assesses) alter's behavior and assumes that behavior to be the 'correct' behavior for 'a-person-like-me' or for 'a-person-in-a-position-like-mine'. » (R. T. A. Leenders, 2002, p. 27)

Dans un mécanisme d'influence sociale de type mimétisme des semblables, l'impact produit par l'adoption d'un comportement par un agent a_j sur l'adoption future de ce comportement par un agent a_i n'est pas déterminé par la force de la cohésion sociale entre eux, ni par l'importance de la position sociale occupée par a_j, l'impact de a_j sur a_i dépend de la relation d'identité entre eux. Plus a_j est similaire à a_i, plus a_j a un impact important sur a_i. Cette hypothèse suggère que, dans un réseau social, les agents ont tendance à se comparer à leurs pairs ou leurs semblables avant de se comparer avec les agents centraux ou ceux avec qui ils socialisent fréquemment (R. S. Burt, 2010b, p. 356).

Il s'agit d'une hypothèse classique en psychologie sociale, dont on peut retrouver l'une des premières formulations dans les travaux de Festinger.[8] Un agent est davantage influencé par ses semblables, car ceux-ci forment un cadre de référence à partir duquel il peut évaluer la pertinence d'un comportement :

> « When confused about an appropriate judgment or course of action, we look around to see what people "like me" are doing. Comparison to people like me provides a benchmark for my own opinion and behavior. » (R. S. Burt, 2010b, p. 15)

Ce type de mécanismes a reçu plusieurs noms dans la littérature. Certains parlent de « contagion sociale par équivalence structurale » (R.S. Burt, 1987), d'autres « d'adaptation sociale » (Borgatti & Foster, 2003; Borgatti, Mehra, Brass, & Labianca, 2009), d'autres encore de « mimétisme isomorphique » (P. J. DiMaggio & Powell, 1983). Ces variations de dénominations sont le reflet des variations au niveau de l'opérationnalisation des paramètres du mécanisme.

Ces différentes opérationnalisations sont basées sur différentes spécifications des critères de similarité qui définissent la relation d'identité entre deux agents dans un réseau. En d'autres mots, chaque opérationnalisation est une manière de spécifier qui sont les semblables avec lesquels un agent se compare et éventuellement en fonction desquels il adaptera son comportement.

L'opérationnalisation la plus commune de ce mécanisme est celle fondée sur le concept d'équivalence structurale introduit précédemment dans Eq. (2.4) :

$$x_{i,t} = \frac{1}{n-1} \sum_{\substack{a_j \in A \\ i \neq j}} \left(e_s(a_i, a_j) \cdot w_j \right) \qquad \text{Eq. (2.12)}$$

[8] La formulation de Festinger est un peu différente, il affirme qu'un individu a tendance à cesser de se comparer (et par conséquent d'être influencé) avec d'autres individus lorsqu'il perçoit une différence trop importante entre lui et eux. Plus la différence est grande, moins il se compare, moins il y a d'influence (Festinger, 1954, p. 133).

Selon cette opérationnalisation, la magnitude de l'influence sociale d'un comportement sur un agent a_i dépend de l'équivalence structurale entre la position occupée par a_i et celles occupées par les agents $a_j \in A$ qui ont déjà adopté ce comportement. Plus a_j occupe une position structurellement similaire à a_i, c'est-à-dire plus il partage les mêmes relations sociales, plus il a un impact important sur a_i.

Par ailleurs, un mécanisme de mimétisme des semblables peut être basé sur plusieurs critères alternatifs de similarité entre deux agents. On peut notamment faire une distinction entre critères endogènes et critères exogènes.

Les critères endogènes de similarité sont des critères internes à la structure d'un réseau social. L'équivalence structurale est un critère endogène. L'automorphisme est un autre critère endogène très important (Borgatti & Everett, 1992). L'automorphisme est un concept proche de l'équivalence structurale, mais il se situe à un niveau d'abstraction supérieur. Deux agents dans un réseau sont automorphiques s'ils occupent des positions qui ont des propriétés d'adjacences similaires. Burt résume la différence entre équivalence structurale et automorphisme de la manière suivante :

> « Two people are structurally equivalent when they have identical relations with the same people. Two people are [automorphic] when they have identical relations with the same kinds of people. » (R. S. Burt, 2010b, p. 8)

Dans le réseau de la Figure 2.2 en début de chapitre, les nœuds N et L sont parfaitement équivalents structurellement, car ils sont tous les deux liés aux nœuds K, M et O et à aucun autre. Par contre, les nœuds Q et P sont parfaitement automorphiques, tout comme le sont les nœuds C et G et les nœuds D et J. Les nœuds de chacune de ces paires sont caractérisés par une équivalence structurale nulle, ils sont non substituables, toutefois, ces nœuds sont liés à des nœuds qui eux ont les mêmes propriétés topologiques. Ce sont donc des nœuds dits équivalents fonctionnellement, mais non substituables dans la structure.

D'autre part, le mimétisme des semblables n'est pas un mécanisme qui opère uniquement via des critères de similarité internes aux réseaux sociaux. C'est un mécanisme qui peut aussi être fondé sur des critères exogènes de similarité, c'est-à-dire externes aux propriétés topologiques d'un réseau. La similarité entre deux agents peut être fondée sur la similitude entre les attributs caractérisant les agents d'un réseau, par exemple leur genre, leur âge, leur statut socioéconomique, etc. (Friedkin & Johnsen, 1997, p. 214).

Plusieurs études empiriques sur les processus d'influence sociale à l'œuvre dans les organisations ont montré comment opére le mimétisme des semblables (Bothner, 2003; R.S. Burt, 1987; Galaskiewicz & Burt, 1991; Galaskiewicz & Wasserman, 1989; Mizruchi, 1989, 1993). C'est un type de mécanisme particulièrement important pour les réseaux sociaux dans lesquels on retrouve de la compétition entre les membres. Burt résume l'argument de la manière suivante :

> « The argument for an equivalence criterion defining peers is competition: people engaged in relations with the same other people could replace one another in those relations. Equivalent people are expected to benchmark against one another for how to be more attractive in their relations. The more equivalent two people are, the more likely they benchmark against one another, and so the more likely they express similar opinion and display similar behavior. » (R. S. Burt, 2010a, p. 248)

Les agents occupant des positions structurellement équivalentes dans un réseau sont des agents généralement substituables l'un à l'autre. D'une part, ce sont des agents ayant souvent un même rôle social dans le réseau, c'est-à-dire ayant les mêmes obligations et les mêmes privilèges. D'autre part, ce sont aussi des agents perçus par les autres membres du réseau comme étant des agents similaires, ayant tous, par exemple, plus ou moins les mêmes compétences, les mêmes connaissances, les mêmes opinions, les mêmes comportements. Or, dans un contexte de compétition, il se développe entre agents substituables (et perçus comme substituables) des rapports sociaux de rivalité. Par exemple des employés ayant au sein d'une entreprise une même fonction entrent en compétition les uns avec les autres pour démontrer à leurs collègues

qui est le plus compétent (Galaskiewicz & Burt, 1991, p. 89). Ce sont aussi des consœurs et confrères d'une profession ayant un même statut social et qui entrent en compétition les uns avec les autres pour maintenir leur réputation (R.S. Burt, 1987).

Dans de tels rapports de rivalité, les agents similaires et substituables entre eux deviennent les uns pour les autres une source d'information cruciale afin de connaître quels sont les comportements qui confèrent un « avantage relatif » sur les autres (Bothner, 2003, p. 1180). Cet avantage relatif peut être par exemple un accès à une ressource utile, augmenter sa crédibilité aux yeux des autres, son pouvoir sur les autres, obtenir un avantage économique comme une promotion, etc. Plus deux agents a_i et a_j occupent des positions équivalentes et sont substituables l'un à l'autre — et plus le sentiment de compétition s'accentue entre eux — plus l'adoption d'un comportement avantageux par a_j sera rapidement imitée par a_i (R.S. Burt, 1987, p. 1291).

À nouveau, il est heuristique d'utiliser la typologie des motivations psychologiques de Cialdini et ses collègues pour mieux comprendre le fonctionnement d'un mécanisme de mimétisme des semblables. Les motivations psychologiques à la base du mimétisme des semblables sont avant tout informationnelles et liées à l'acquisition d'un avantage relatif, mais elles sont aussi liées à l'image de soi. Lorsqu'un comportement a été adopté par une proportion importante d'agents occupant la même position qu'un agent a_i, et que a_i est parmi les derniers à ne pas avoir encore adopté ce comportement, un sentiment d'embarras et même d'imposture peut se développer chez a_i. Ce sentiment est causé par le fait que ce comportement est reconnu par les membres du réseau comme quelque chose de typique et d'attendu pour quelqu'un occupant sa position et jouant son rôle (Bothner, 2003; Galaskiewicz & Burt, 1991, p. 90).

2.4 Conclusion

Le fonctionnement des quatre types de mécanismes d'influence sociale à l'œuvre dans les réseaux sociaux peut se résumer de la manière suivante.

L'exposition sociale : L'adoption future d'un comportement par un agent a_i est influencée par la proportion d'agents dans le réseau social qui a déjà adopté ce comportement. Plus grande est cette proportion, plus grande est la magnitude de l'influence sociale de ce comportement sur a_i.

La contagion sociale : L'adoption future d'un comportement par un agent a_i est influencée par la force de la cohésion sociale entre un a_i et les autres membres du réseau social qui ont déjà adopté ce comportement. Plus grande est cette cohésion sociale et plus grande est la magnitude de l'influence sociale de ce comportement sur a_i.

La déférence : L'adoption future d'un comportement par un agent a_i est influencée par l'importance de la position sociale des membres du réseau qui ont déjà adopté ce comportement. Plus importante est la position sociale des membres ayant adopté jusqu'à maintenant le comportement, plus grande est la magnitude de l'influence sociale de ce comportement sur a_i.

Le mimétisme des semblables : L'adoption future d'un comportement par un agent a_i est influencée par la similitude entre a_i et les membres du réseau qui ont déjà adopté ce comportement. Plus la similitude entre a_i et ces membres est grande, plus grande est la magnitude de l'influence sociale de ce comportement sur a_i.

Ces mécanismes sont hautement stylisés, c'est-à-dire qu'ils sont tous fondés sur des hypothèses extrêmes de déterminisme social. Ils supposent tous que l'adoption d'un comportement par un agent peut être entièrement expliquée par les interactions sociales de cet agent avec les autres membres du réseau social. Leenders parle en ce sens de caricature (R. T. A. Leenders, 2002, p. 25), car il serait surprenant de trouver une situation empirique dans laquelle ces mécanismes seuls peuvent prédire parfaitement l'adoption d'un comportement par les membres d'un réseau.

2.4.1 Réceptivité et résistance à l'influence sociale

Le schéma de fonctionnement le plus simple d'un mécanisme d'influence sociale à base de seuil est celui présenté dans la définition Eq. (2.5), où la valeur du paramètre θ est la même pour tous les agents du réseau. Mais ce paramètre peut aussi recevoir différentes opérationnalisations. Dans certaines variantes, les différentes valeurs de θ sont distribuées de manière uniforme parmi les membres du réseau (Watts & Dodds, 2009). Dans d'autres, elles sont distribuées en différentes catégories (Rogers, 2003; Valente, 1996). Cette dernière variante est basée sur l'hypothèse que dans un réseau social on retrouve différents groupes d'agents caractérisés par différents degrés de réceptivité à l'influence sociale. Certains agents seraient ainsi plus réceptifs à l'influence sociale, certains seraient plutôt neutres et d'autres au contraire seraient très résistants.

Friedkin suggère que la position sociale d'un agent dans un réseau déterminerait sa réceptivité ou sa résistance à l'influence sociale. Selon lui, les agents excentrés dans un réseau seraient beaucoup plus réceptif à l'influence sociale que les agents centraux. Cette hypothèse est basée sur des études empiriques qui montrent que la centralité d'un agent est généralement corrélée à l'estime de soi et à la confiance en soi d'un agent. Selon ces études, plus un agent est central, plus son estime de soi et sa confiance seraient élevées et moins il serait motivé à comparer ses comportements avec ceux des autres membres du réseau (c.f. Cialdini & Goldstein, 2004, p. 611; Friedkin, 2006, p. 87; c.f. Ibarra & Andrews, 1993).

Par ailleurs, les agents occupant une position structurellement unique dans un réseau sont généralement beaucoup moins susceptibles d'être influencés. C'est le cas par exemple des « courtiers » d'un réseau, c'est-à-dire les agents qui occupent une position d'intermédiaire unique entre différents groupes d'un réseau. Ce sont des agents non substituables dans le réseau, c'est-à-dire que personne d'autre qu'eux ne peut assurer la connexion entre deux régions d'un réseau (R. S. Burt, 2010b, p. 31). Ils ne font face,

en ce sens, à aucune compétition, ils ne ressentent pas le besoin de se comparer à d'autres et deviennent ainsi moins réceptifs à l'influence sociale. Dans le réseau fictif illustré dans la Figure 2.2, le nœud E est un exemple de nœud non substituable. Aucun autre nœud ne peut assurer son rôle d'intermédiaire entre trois régions du réseau.

2.4.2 Interactions entre différents mécanismes

Les différents types de mécanismes présentés sont tous conceptuellement distincts, mais il peut arriver qu'il soit impossible de les distinguer empiriquement (R. T. A. Leenders, 2002, p. 30; Marsden & Friedkin, 1993, p. 133; Rice, 1993, p. 53). Les agents socialement proches tendent à occuper des positions sociales similaires en terme d'équivalence structurale — ils tendent à partager, au moins en partie, leur voisinage — si bien que les deux mécanismes sont souvent corrélés et indiscernables l'un de l'autre (R. S. Burt, 2010a, p. 349; R. T. A. Leenders, 2002, p. 29).

Dans certains contextes empiriques, un mécanisme peut en subsumer un autre. C'est ce qui se produit lorsqu'un réseau est composé de plusieurs cliques.[9] Une clique a cette particularité topologique que ses membres sont à la fois tous parfaitement cohésifs et tous parfaitement équivalents structurellement. Dans ce contexte particulier, un mécanisme de mimétisme des semblables subsume complètement la contagion sociale (R. S. Burt, 2010b, p. 9).[10]

Par ailleurs, une autre classe de mécanismes à l'œuvre dans les réseaux sociaux est étroitement corrélée aux mécanismes d'influence sociale. Cette classe de mécanisme regroupe des mécanismes d'homophilie, de sélection et d'assortativité. Le fonctionnement général de ces mécanismes peut se résumer de la manière suivante :

---

[9] Une clique est un sous-ensemble d'agents tous connectés les uns avec les autres. Dans la Figure 2.2, les nœuds {A,B,C} sont une clique ainsi que les nœuds {G,H,I}.

[10] Par ailleurs, dans ce contexte, la contagion sociale est le mécanisme qui explique l'influence sociale *entre* les cliques. C'est une influence sociale réalisée par des agents que Burt appelle des « leaders d'opinion » (Ronald S. Burt, 1999, p. 46).

les agents similaires, en fonction de critères endogènes ou exogènes de similarité, tendent à entrer en interaction plus souvent que les agents différents. Or, des études ont montré qu'il est méthodologiquement très difficile de distinguer l'homophilie de la contagion sociale (Easley & Kleinberg, 2010, p. 82; Shalizi & Thomas, 2011; Steglich, Snijders, & Pearson, 2010). En effet, la contagion sociale est un mécanisme qui explique l'apparition de similarités entre agents en fonction de la socialisation entre eux, alors que l'homophilie est un mécanisme qui explique l'apparition de relations sociales entre agents en fonction des similarités entre eux. Ces deux mécanismes sont étroitement liés par une boucle de rétroaction positive : plus deux agents sont proches socialement plus ils deviennent similaires et plus ils sont similaires plus ils se rapprochent socialement, et ainsi de suite.

L'interaction entre les différents mécanismes d'influence sociale et l'interaction avec d'autres classes de mécanismes à l'œuvre dans les réseaux sociaux sont actuellement un enjeu important de recherche.

CHAPITRE III

LA SÉMANTIQUE VECTORIELLE

3.0 Introduction

Le but de ce chapitre est de présenter le modèle, appelé la sémantique vectorielle, utilisée dans cette thèse pour la modélisation et la découverte des concepts exprimés dans un corpus de texte. La sémantique vectorielle est à la fois un modèle computationnel de la représentation des contenus sémantiques exprimés par des signes, généralement linguistiques, et une méthode statistique et algorithmique de découverte de ces contenus sémantiques dans un corpus de textes. La sémantique vectorielle est fondée sur une métaphore spatiale, alors que sa méthode est basée sur une hypothèse distributionnelle (Gärdenfors, 2014; T. K. Landauer & Dumais, 1997; Lemaire & Denhière, 2006; Lund & Burgess, 1996; Sahlgren, 2006a, Chapter 2; Schütze, 1992; Turney & Pantel, 2010; Widdows, 2004).

La sémantique vectorielle (SV pour la suite) est caractérisée par une grande multidisciplinarité. C'est un modèle et une méthode que l'on retrouve dans de nombreux programmes de recherche, notamment d'analyse de textes (S. Lebart & Salem, 1994; Meunier, Forest, & Biskri, 2005), de traitement automatique des langues naturelles (Manning & Schütze, 1999; Widdows, 2004), de recherche d'information (Manning, Raghavan, & Schütze, 2008a), de fouille de textes (Feldman & Sanger, 2007), de psycholinguistique (Thomas K. Landauer, Foltz, & Laham, 1998; Lund & Burgess, 1996), de linguistique de corpus (Lenci, 2008), de sociologie (Chartier & Meunier, 2011; Larsen & Monarchi, 2004), de sémiologie (L. Lebart, Piron, & Steiner, 2003), de sciences cognitives et d'intelligence artificielle (Gärdenfors, 2014; Turney & Pantel, 2010).

Cette multidisciplinarité vient, en partie, du fait que la SV est un modèle formel qui se situe à un très haut niveau d'abstraction, à savoir celui de l'algèbre vectorielle. Les paramètres de ce modèle peuvent faire l'objet de différentes opérationnalisations adaptées à l'étude de différents objets.

Le chapitre se divise en trois sections. La première section présente les origines théoriques et méthodologiques à partir desquelles fut élaborée la SV. La deuxième section présente les différents paramètres du modèle et la troisième section présente comment sont modélisés les contenus conceptuels dans ce cadre et comment ils sont découverts dans un corpus de texte.

3.1 Origines de la sémantique vectorielle

La SV a plusieurs sources théoriques et méthodologiques. Plusieurs programmes de recherche dans les années cinquante et soixante ont eu une grande influence sur son élaboration, notamment les travaux de Salton en recherche documentaire (Salton, Wong, & Yang, 1975) et les travaux de Benzecri et l'école française d'analyse de données (Benzecri, 1973). Toutefois, dans la littérature, la SV est principalement présentée comme une synthèse entre, d'une part, la sémantique différentielle développée par Osgood et, d'autre part, la méthode distributionnelle de la linguistique structurale développée par Harris.

3.1.1 Sémantique différentielle et métaphore spatiale

Une première source de la SV, issue de la psychologie, a été le programme de recherche de la sémantique différentielle développé par Osgood et ses collègues (Charles E. Osgood, 1952, 1964; Charles Egerton Osgood, Suci, & Tannenbaum, 1957). C'est dans ces travaux que la métaphore spatiale à la base de la SV fut élaborée et formalisée.

Dans ses travaux, Osgood utilisait une technique d'amorce sémantique classique en psychologie expérimentale, dont l'objectif était de lui permettre de mesurer les

différences de connotation des mots. Il demandait à des sujets de juger, à l'aide d'une échelle ordinale, de la connotation de différents mots en fonction de plusieurs dimensions sémantiques.

La Figure 3.1 illustre un exemple dans lequel un sujet était invité à juger, sur une échelle de sept niveaux, de la connotation du mot anglais « FATHER » et ceci en fonction des trois dimensions suivantes : « happy vs sad », « hard vs soft » et « slow vs fast » (Charles Egerton Osgood et al., 1957, p. 26).

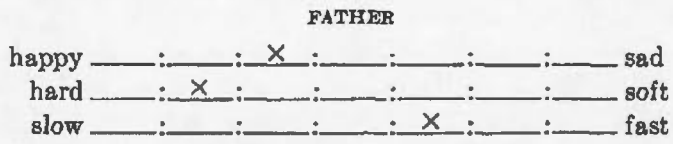

Figure 3.1: Exemple utilisé par Osgood dans ses expérimentations pour mesurer les jugements connotatifs du mot « FATHER ».

Chaque jugement de ce genre était ensuite encodé sous la forme d'un vecteur dont les coordonnées exprimaient la connotation de chaque dimension sémantique. Par exemple, dans la Figure 3.2, la connotation du mot « FATHER » est encodée par un vecteur à trois dimensions dont les coordonnées sont (3, 2, 5). Ce vecteur représente la position du mot « FATHER » dans un espace vectoriel défini par les dimensions « happy vs sad », « hard vs soft » et « slow vs fast ».

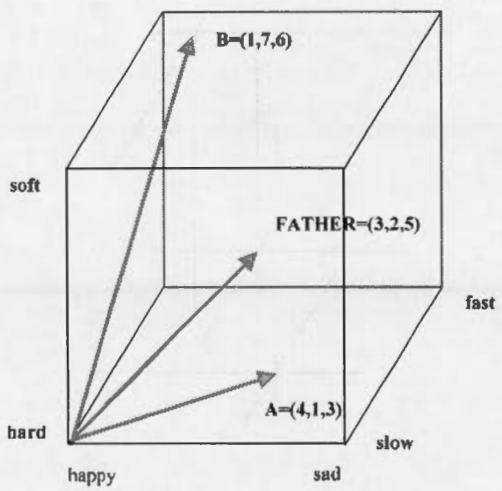

Figure 3.2: Projection dans un espace vectoriel à trois dimensions de la connotation du mot FATHER et de deux autres mots fictifs A et B.

Dans les travaux d'Osgood, ce type d'expérimentation était répété avec une centaine de sujets et pour des dizaines de mots différents. La connotation de chaque mot était ainsi modélisée par une position particulière dans l'espace. Pour Osgood, la sémantique différentielle c'était l'analyse des différences entre les positions de chaque mot dans l'espace. Ce modèle permettait, selon l'auteur, de mesurer objectivement la sémantique des mots.

L'espace vectoriel de la Figure 3.2 était appelé un « espace sémantique », lequel était considéré entretenir un isomorphisme structurel avec la sémantique des mots :

> « Since the positions checked on the scales constitute the coordinates of the [word]'s location in semantic space, we assume that the coordinates in the measurement space are functionally equivalent with the components of the representational mediation process associated with this [word]. This, then, is one rationale by which the semantic differential, as a technique of measurement, can be considered as an index of meaning. » (Charles Egerton Osgood et al., 1957, p. 30)[11]

[11] Osgood utilise de manière indifférenciée les mots « mot » et « concept ». Afin d'éviter certaines ambiguïtés théoriques, dans cette citation le mot anglais « concept » a été remplacé par le mot « word ». Ceci ne change pas le sens de la citation.

La contribution majeure des travaux d'Osgood est d'avoir montré qu'une fois modélisés dans un espace vectoriel les jugements connotatifs des mots, des propriétés mathématiques de cet espace peuvent être considérées comme fonctionnellement équivalentes à certaines propriétés des contenus sémantiques de ces mots. Dans ce type d'espace, les axes représentent des dimensions sémantiques, les vecteurs ou les coordonnées de cet espace représentent des contenus sémantiques, les relations de proximité spatiale entre vecteurs représentent des relations de similarité sémantique entre contenus et des structures algébriques, géométriques et topologiques de cet espace représentent des structures sémantiques.

En nous appuyant sur la métaphore spatiale et l'hypothèse des équivalences fonctionnelles entre propriétés d'un espace vectoriel et propriétés d'un espace sémantique, plusieurs calculs peuvent être effectués afin d'inférer les propriétés de l'espace sémantique. Cette construction abstraite, une fois appliquée à des données d'expérimentations comme celles collectées par Osgood, permet d'exploiter différentes notions de l'algèbre vectorielle afin d'inférer plusieurs propriétés des contenus sémantiques des mots modélisés. Par exemple, dans la Figure 3.2, une métrique permet de calculer que la proximité spatiale entre le mot « FATHER » et le mot fictif « A » est beaucoup plus grande que celle entre « B » et « A ». Ce calcul métrique permét ainsi d'inférer que la similarité sémantique entre « FATHER » et « B » est beaucoup plus grande que celle entre « B » et « A ».

Plusieurs des équivalences fonctionnelles conjecturées par cette métaphore spatiale seront corroborées empiriquement dans des travaux ultérieurs. Ces équivalences seront à la base de la SV et il sera démontré à plusieurs reprises que ces espaces vectoriels sont capables de reproduire de manière remarquablement fidèle plusieurs comportements cognitifs liés au traitement du langage naturel, par exemple l'acquisition du langage, sa mémorisation, certains jugements sur la synonymie des mots, la discrimination du sens des mots, la composition sémantique ou la construction

de classes sémantiques (Baroni & Lenci, 2010; Burgess, Livesay, & Lund, 1998; Charles, 2000; T. K. Landauer & Dumais, 1997; Lemaire & Dessus, 2003; J. Mitchell & Lapata, 2010; Purandare & Pedersen, 2004).

3.1.2 Méthode et hypothèse distributionnelle

Une deuxième source importante de la SV a été le programme de recherche de la linguistique structurale développé par Harris (Harris, 1951). On retrouve dans ce programme les principaux concepts méthodologiques qui seront à la base de la SV, tout particulièrement l'hypothèse distributionnelle sur laquelle elle se fondera.

Dans la linguistique structurale de Harris, un concept méthodologique central est celui d'environnement ou de contexte d'usage d'un élément linguistique. L'environnement d'un élément linguistique constitue son voisinage syntagmatique, ses cooccurrents, c'est-à-dire l'ensemble des autres éléments avec lesquels il est combiné dans un segment de texte. Chez Harris, l'ensemble des combinaisons possibles d'un élément, c'est-à-dire l'ensemble de ses environnements, forme sa distribution dans un corpus :

> « The distribution of an element is the total of all environments in which it occurs, i.e. the sum of all the (different) positions (or occurrences) of an element relative to the occurrence of other elements. » (Harris, 1951, p. 15)

L'hypothèse distributionnelle conjecture que deux éléments linguistiques caractérisés par des distributions équivalentes dans un corpus sont eux-mêmes équivalents. Harris le formulait de la manière suivante :

> « Two utterances or features will be said to be linguistically, descriptively, or distributionally equivalent if they are identical as to their linguistic elements and the distributional relations among these elements. » (Harris, 1951, p. 16)

Dans sa forme la plus générale, l'hypothèse distributionnelle de Harris s'applique de surcroît à tous les éléments du langage. L'hypothèse distributionnelle sur laquelle se fonde la SV est une opérationnalisation particulière, qui est restreinte à la sémantique des mots. Selon celle-ci, l'expression de certains contenus sémantiques dans le langage

serait concomitante à certaines régularités au niveau de la combinaison des mots, par conséquent, dans un corpus de texte, des distributions similaires exprimeraient des contenus sémantiques similaires. Harris discutait de cette relation entre distribution et contenu sémantique des mots en termes de corrélation :

> « The fact that, for example, not every adjectives occurs with every noun can be used as a measure of meaning difference. For it is not merely that different members of the one class have different selections of members of the other class with which they are actually found. More than that: if we consider words or morphemes A and B to be more different than A and C, then we will often find that the distributions of A and B are more different than the distributions of A and C. In other words, difference in meaning correlates with difference in distribution. » (Harris, 1954, p. 156)

Cette hypothèse distributionnelle représente la clé de voûte de la méthodologie de la SV. Son principe est d'une surprenante simplicité : des mots exprimant des contenus sémantiques similaires sont généralement utilisés dans des environnements ou des contextes d'usage similaires, par conséquent, il y aurait des corrélations suffisamment fortes entre des régularités statistiques caractérisant la combinaison des mots dans un corpus et l'expression du contenu sémantique de ces mots, pour que l'analyse des premières permette d'inférer les secondes.

3.2 Espace sémantique

Les principaux éléments théoriques qui forment aujourd'hui le cadre conceptuel de la SV étaient presque tous présents déjà dans les travaux d'Osgood. Ce qui va différentier le programme de la SV des travaux classiques d'Osgood est l'intégration dans sa méthodologie de l'hypothèse distributionnelle de Harris. Dans la SV, la méthode expérimentale qu'utilisait Osgood, basée sur une technique d'amorce sémantique, est remplacée par une méthode statistique d'analyse de texte qui consiste à induire, via le calcul des régularités statistiques caractérisant la combinaison des mots, différentes propriétés des contenus sémantiques exprimés dans un corpus de texte.

Les espaces sémantiques construits via cette hypothèse distributionnelle sont par conséquent différents de ceux élaborés par Osgood. Ce sont premièrement des espaces beaucoup plus complexes, composés généralement de plusieurs milliers de dimensions. Ces dimensions ne modélisent plus la connotation des mots, mais leur environnement ou leur contexte d'usage. La position des mots dans l'espace sémantique n'exprime plus la force d'une connotation, elle exprime une régularité statistique observée au niveau des combinaisons de mots dans un corpus de texte.

Formellement, un espace sémantique peut être défini à l'aide de cinq paramètres. Dénotons d'abord un mot par le symbole u_i et un segment de texte par une liste de mots $S_k = (u_1 \dots, u_t)$. Un premier paramètre, dénoté par $\mathbb{C} = \{S_1 \dots S_k\}$, représente un corpus composé d'un ensemble de segments de texte. Un deuxième paramètre, dénoté par $U = \{u_1 \dots, u_n\}$, représente un vocabulaire, c'est-à-dire un ensemble de mots présents dans les segments de texte d'un corpus et que l'on souhaite modéliser dans un espace sémantique. Un troisième paramètre, dénoté par $V = \{u_1 \dots, u_m\}$, représente l'ensemble des cooccurrents sélectionnés pour caractériser les environnements ou les contextes d'usage des mots de l'ensemble U. Le quatrième paramètre est une matrice de coordonnées $W^{n \times m} = [w_{ij}]$ dans laquelle chaque ligne de la matrice, dénotée par un vecteur $\vec{u}_i = (w_{i1} \dots w_{im})$, représente la position du mot u_i dans l'espace sémantique. Le dernier paramètre est une métrique $d: (\vec{u}_i, \vec{u}_j) \rightarrow \mathbb{R}$, qui permet de mesurer les distances ou les proximités entre mots $u_i \in U$ dans l'espace sémantique.

Ces paramètres sont discutés plus en détail dans les sections suivantes du chapitre.

3.2.1 Le corpus

Un corpus $\mathbb{C} = \{S_1 \dots S_k\}$ est composé d'un ensemble de segments de texte. Dans la SV, un segment de texte est une manière de définir l'environnement ou le contexte d'usage d'un mot. Un segment a généralement une unité grammaticale. L'unité minimale est la colocation. Les phrases, les paragraphes, les documents qui composent

un corpus constituent tous différents types de segment. Un segment peut aussi correspondre à une unité plus arbitraire comme une séquence de plusieurs mots dans un texte.

3.2.2 Les mots

Dans le cadre de la SV, un mot peut être opérationnalisé de plusieurs manières. L'opérationnalisation la plus élémentaire consiste à simplement indexer l'ensemble des chaînes de caractères séparés par deux espaces dans un corpus. C'est une opérationnalisation élémentaire, mais courante.

Il est également possible d'instancier chaque mot par sa forme normalisée, par exemple son lemme ou sa racine. Un lemme est un mot fléchi par le genre et le mode, soit l'infinitif pour les verbes et le masculin singulier pour les autres mots. La racine d'un mot est sa troncature au niveau du suffixe, par exemple les mots « LIBÉRATION », « LIBÉRAUX », « LIBÉRER » ont tous en commun la racine « LIBÉR ».

La raison pour laquelle la lemmatisation et la racinisation sont des opérationnalisations souvent préférées à simplement l'indexation des chaîne de caractères, est qu'on considère que les différentes flexions de la forme normalisée d'un mot partagent une unité sémantique suffisamment cohérente pour les regrouper sous un même dénominateur, soit leur lemme ou leur racine.

Une autre opérationnalisation possible des mots est le n-gram. Les n-grams d'un mot sont l'ensemble des séries de n-caractères qui composent un mot. Cette opérationnalisation a été développée en recherche d'information (Cavnar & Trenkle, 1994). Bien que contre intuitive en apparence, il a été démontré que pour la plupart des langues européennes, la décomposition des mots d'un corpus de texte en ses 4-grams ou 5-grams ne génère qu'une perte infime d'information (Mcnamee & Mayfield, 2004).

Afin d'illustrer la différence entre ces différentes opérationnalisations, prenons pour exemple la citation suivante de Gärdenfors :

> « [...] the meanings that we use in communication can be described as organized in abstract spatial structures that are expressed in terms of dimensions, distances, regions, and other geometric notions [...] » (Gärdenfors, 2014, p. 22)

Le Tableau suivant illustre les résultats de quatre techniques d'indexation des mots de cette citation. La première technique indexe simplement tous les termes de la citation de Gärdenfors, la seconde est une technique de racinisation basée sur l'algorithme de Porter (Porter, 1980), la troisième est une technique de lemmatisation basée sur l'algorithme MorphAdorner (Burns, 2013) et la quatrième est une indexation de tous les 5-grams de la citation.

Tableau 3.1: Exemple comparatif de différentes techniques d'indexation des mots d'un segment de texte

Technique d'indexation	Liste des mots
Termes	THE; MEANINGS; THAT; WE; USE; IN; COMMUNICATION; CAN; BE; DESCRIBED; AS; ORGANIZED; IN; ABSTRACT; SPATIAL; STRUCTURES; THAT; ARE; EXPRESSED; IN; TERMS; OF; DIMENSIONS; DISTANCES; REGIONS; AND; OTHER; GEOMETRIC; NOTIONS;
Racinisation	THE; MEAN; THAT; WE; USE; IN; COMMUN; CAN; BE; DESCRIB; AS; ORGAN; IN; ABSTRACT; SPATIAL; STRUCTUR; THAT; ARE; EXPRESS; IN; TERM; OF; DIMENS; DISTANC; REGION; AND; OTHER; GEOMETR; NOTION;
Lemmatisation	THE; MEANING; THAT; WE; USE; IN; COMMUNICATION; CAN; BE; DESCRIBE; AS; ORGANIZE; IN; ABSTRACT; SPATIAL; STRUCTURE; THAT; BE; EXPRESS; IN; TERM; OF; DIMENSION; DISTANCE; REGION; AND; OTHER; GEOMETRIC; NOTION;
5-grams	THE_M; HE_ME; E_MEA; _MEAN; MEANI; EANIN; ANING; NINGS; INGS_; NGS_T; GS_TH; S_THA; _THAT; THAT_; HAT_W; AT_WE; T_WE_; _WE_U; WE_US; E_USE; _USE_; USE_I; SE_IN; E_IN_; _IN_C; IN_CO; N_COM; _COMM; COMMU; OMMUN; MMUNI; MUNIC; UNICA; NICAT; ICATI; CATIO; ATION; TION_; ION_C; ON_CA; N_CAN; _CAN_; CAN_B; AN_BE; N_BE_; _BE_D; BE_DE; E_DES; _DESC; DESCR; ESCRI; SCRIB; CRIBE; RIBED; IBED_; BED_A; ED_AS; D_AS_; _AS_O; AS_OR; S_ORG; _ORGA; ORGAN; RGANI; GANIZ; ANIZE; NIZED; IZED_; ZED_I; ED_IN; D_IN_; _IN_A; IN_AB; N_ABS; _ABST; ABSTR; BSTRA; STRAC; TRACT; RACT_; ACT_S; CT_SP; T_SPA; _SPAT; SPATI; PATIA; ATIAL; TIAL_; IAL_S; AL_ST; L_STR; _STRU; STRUC; TRUCT; RUCTU; UCTUR; CTURE; TURES; URES_; RES_T; ES_TH; S_THA; _THAT; THAT_; HAT_A; AT_AR; T_ARE; _ARE_; ARE_E; RE_EX; E_EXP; _EXPR; EXPRE; XPRES; PRESS; RESSE; ESSED; SSED_; SED_I; ED_IN; D_IN_; _IN_T; IN_TE; N_TER; _TERM; TERMS; ERMS_; RMS_O; MS_OF; S_OF_; _OF_D; OF_DI; F_DIM; _DIME; DIMEN; IMENS; MENSI; ENSIO; NSION; SIONS; IONS_; ONS_D; NS_DI; S_DIS; _DIST; DISTA; ISTAN; STANC; TANCE; ANCES; NCES_; CES_R; ES_RE; S_REG; _REGI; REGIO; EGION; GIONS; IONS_; ONS_A; NS_AN; S_AND; _AND_; AND_O; ND_OT; D_OTH; _OTHE; OTHER; THER_; HER_G; ER_GE; R_GEO; _GEOM; GEOME; EOMET; OMETR; METRI; ETRIC; TRIC_; RIC_N; IC_NO; C_NOT; _NOTI; NOTIO; OTION; TIONS;

3.2.3 Cooccurrents

Un mot $u_j \in V$ est un cooccurrent d'un autre mot $u_i \in U$ s'ils sont coprésents au sein d'un même contexte, c'est-à-dire d'un même segment de texte $S_k \in \mathbb{C}$. L'ensemble des

cooccurrents d'un mot $u_i \in U$ forme sa distribution dans un corpus. Les mots qui composent cette distribution dépendent de la manière dont est segmenté le corpus. Cette distribution ne sera pas la même si le corpus est segmenté en colocations, en phrases ou en séquences de n-mots.

À titre d'illustration, prenons la Figure 3.3 dans laquelle il y a un corpus composé de cinq mots fictifs, soit BLA, BLE, BLI, BLO et BLU.

BLA BLI BLO.
BLE BLO.
BLI BLA.
BLO BLU BLO.

Figure 3.3: Corpus fictif composé de cinq mots différents.

Supposons que ce corpus est segmenté de manière à ce que chaque phrase représente un contexte différent. La distribution du mot BLA serait composée de deux cooccurrents, soit BLI et BLO, la distribution du mot BLE aurait un seul cooccurrent, soit le mot BLO, celle du mot BLI serait composée de deux cooccurrents, soit BLA et BLO, la distribution du mot BLO serait composée des cooccurrents BLA, BLI, BLU et BLE et la distribution du mot BLU aurait également un seul cooccurrent, soit BLO.

D'autres types de distributions de cooccurrences sont possibles et dépendent du type de contexte ou d'environnement produit par la segmentation du corpus (Riordan & Jones, 2011, p. 309). Par exemple, dans certaines opérationnalisations, l'ordre des mots est pris en compte dans la construction de la distribution des cooccurrents (e.g. Lund & Burgess, 1996). Le Tableau suivant illustre différentes distributions du mot BLO selon trois différents types de contextes de cooccurrence, soit la colocation, la phrase et une fenêtre de quatre mots.

Tableau 3.2: Opérationnalisations alternatives de segmentation et leur distribution

Contexte	Distribution du mot BLO
Colocation	BLI, BLE, BLU
Phrase	BLI, BLA, BLE, BLU
Fenêtre de 4-mots	BLI, BLA, BLE, BLA, BLU

Dans un espace sémantique, l'ensemble des cooccurrents forme la base vectorielle de l'espace. Ils ont un rôle analogue aux jugements connotatifs de la méthode qu'Osgood utilisait. Chaque cooccurrent $u_j \in V$ correspond à une dimension de l'espace sémantique. Dans l'exemple précédent, l'espace sémantique du petit corpus fictif serait par conséquent modélisé dans un espace de cinq dimensions.

3.2.4 La matrice de coordonnées

Dans une matrice de coordonnées, chaque ligne représente la position dans l'espace sémantique d'un mot $u_i \in U$ encodée sous la forme d'un vecteur $\vec{u}_i = (w_{i1} \dots w_{im})$. Chaque colonne représente un cooccurrent $u_j \in V$ et constitue une dimension de l'espace. Chaque valeur w_{ij} de la matrice représente la coordonnée du mot i sur la dimension j.

Une coordonnée peut être calculée de plusieurs manières. Généralement, w_{ij} correspond à la valeur d'un coefficient d'association entre un mot u_i et un cooccurrent u_j. À titre d'exemple, la matrice de coordonnées du corpus fictif précédent, segmenté en phrases, est représentée dans la Figure 3.4. Chaque coordonnée w_{ij} correspond à la fréquence de cooccurrence entre u_i et u_j, c'est-à-dire au nombre de phrases dans lesquelles u_i et u_j sont coprésents.

$W =$

	BLA	BLE	BLI	BLO	BLU
BLA	0	0	2	1	0
BLE	0	0	0	1	0
BLI	2	0	0	1	0
BLO	1	1	1	0	1
BLU	0	0	0	1	0

Figure 3.4: Matrice de coordonnées d'un corpus de cinq mots et de quatre segments.

Il y a plusieurs opérationnalisations alternatives au calcul des coordonnées. Le Tableau 3.3 définit quelques-unes d'entre elles. Le coefficient dénoté par le symbole fc_{ij} est la fréquence de cooccurrence entre deux mots u_i et u_j. C'est le coefficient utilisé pour construire la matrice de la Figure 3.4. Le coefficient $tfidf_{ij}$ a été élaboré dans les travaux classiques de Salton en recherche documentaire (Salton et al., 1975), mais est parfois utilisé également pour construire des espaces sémantiques. Le $tfidf_{ij}$ est la fréquence de cooccurrence fc_{ij} entre u_i et u_j pondérée par fc_i, l'étendue de u_i c'est-à-dire le nombre de segments dans lesquels apparait u_i. Le coefficient dénoté par le symbole pr_{ij} représente la probabilité conditionnelle qu'un segment contienne une occurrence de u_i sachant qu'il contient aussi une occurrence de u_j. Et le coefficient pmi_{ij}, appelé en anglais « Pointwise Mutual Information », compare la probabilité d'une cooccurrence entre deux mots u_i et u_j avec l'espérance mathématique de u_j.

Tableau 3.3: Opérationnalisations alternatives du calcul des coordonnées des mots dans un espace sémantique.

Calculs des coordonnées	Définitions	
fc_{ij}	$w_{ij} = \left\lvert \bigcup_{S_k \in \mathbb{C}} \{S_k : u_i \in S_k \ \wedge u_j \in S_k\} \right\rvert$	Eq. (3.1)
fc_i	$w_i = \left\lvert \bigcup_{S_k \in \mathbb{C}} \{S_k : u_i \in S_k\} \right\rvert$	Eq. (3.2)
$tfidf_{ij}$	$w_{ij} = fc_{ij} \times \log \frac{\lvert\mathbb{C}\rvert}{fc_i}$	Eq. (3.3)
pr_{ij}	$w_{ij} = \frac{fc_{ij}}{fc_j}$	Eq. (3.4)
pmi_{ij}	$w_{ij} = \log \frac{\mathrm{pr}_{ij}}{fc_j/\lvert\mathbb{C}\rvert}$	Eq. (3.5)

D'autres coefficients existent. On retrouve dans la littérature plusieurs dizaines d'opérationnalisations alternatives (Kiela & Clark, 2014). Le principe sous-jacent à tous ces coefficients consiste à pondérer le poids des cooccurrents qui composent l'environnement ou le contexte d'usage des mots par leur spécificité. Plus un cooccurrent $u_j \in V$ est spécifique aux environnements ou aux contextes d'un mot $u_j \in U$, plus la valeur de la coordonnée w_{ij} sera élevée.

3.2.5 Métrique

Une métrique est une fonction qui définit la distance (ou la proximité) entre deux positions dans un espace vectoriel. Dénotons par $\vec{u}_i$ et $\vec{u}_j$ deux positions (i.e. deux vecteurs) dans l'espace et par le symbole $d(\vec{u}_i, \vec{u}_j)$ la distance entre ces deux positions. Formellement, une métrique doit satisfaire les quatre conditions suivantes :

1. $d(\vec{u}_i, \vec{u}_j) \geq 0$.

2. $d(\vec{u}_i, \vec{u}_j) = 0$ si et seulement si $\vec{u}_i = \vec{u}_j$.
3. $d(\vec{u}_i, \vec{u}_j) = d(\vec{u}_j, \vec{u}_i)$.
4. $d(\vec{u}_i, \vec{u}_j) \leq d(\vec{u}_i, \vec{u}_k) + d(\vec{u}_k, \vec{u}_j)$.

La première condition stipule que la distance entre deux positions différentes doit toujours être positive. La seconde condition est l'identité des indiscernables. La troisième condition est une règle de symétrie et la quatrième condition est l'inégalité triangulaire (Widdows, 2004, pp. 99–100).

Il existe plusieurs métriques différentes qui satisfont ces quatre conditions (Ellis, Furner-Hines, & Willett, 1993; Gärdenfors, 2000; Kiela & Clark, 2014; Rajman & Lebart, 1998). Le Tableau 3.4 définit quelques-unes d'entre elles, soit la métrique euclidienne, la métrique angulaire du cosinus, une métrique basée sur l'indice de Jaccard et la métrique du khi-deux. Lorsque la valeur calculée par l'une de ces métriques se rapproche de zéro, cela signifie que les positions $\vec{u}_i$ et $\vec{u}_j$ sont proches dans l'espace, et à l'inverse, plus cette valeur est grande plus ces positions sont loin l'une de l'autre.

Tableau 3.4: Métriques et leur définition.

Métriques	Définitions	
Euclidienne	$d(\vec{u}_i, \vec{u}_j) = \sqrt{\sum_{k=1}^{m} (w_{ik} - w_{jk})^2}$	Eq. (3.6)
Cosinus	$d(\vec{u}_i, \vec{u}_j) = 1 - \frac{\vec{u}_i \cdot \vec{u}_j}{\lvert\vec{u}_i\rvert \cdot \lvert\vec{u}_j\rvert}$	Eq. (3.7)
Jaccard	$d(\vec{u}_i, \vec{u}_j) = 1 - \frac{\sum_{k=1}^{m} \min(w_{ik}, w_{jk})}{\sum_{k=1}^{m} \max(w_{ik}, w_{jk})}$	Eq. (3.8)
Khi-deux	$d(\vec{u}_i, \vec{u}_j) = \sum_{k=1}^{m} \frac{1}{w_{ik} + w_{jk}} \left(\frac{w_{ik}}{\lvert\vec{u}_i\rvert} - \frac{w_{jk}}{\lvert\vec{u}_j\rvert} \right)^2$	Eq. (3.9)

Lorsqu'un espace vectoriel modélise un espace sémantique, une métrique $d(\vec{u}_i, \vec{u}_j)$ s'interprète comme une mesure de la similarité sémantique entre deux mots u_i et u_j. Plus la distance est petite entre les positions de deux mots, plus les mots sont sémantiquement similaires, et à l'inverse, plus elle est grande plus les mots sont sémantiquement différents. Une distance égale à zéro signifie qu'il y a équivalence sémantique entre deux mots.

Dans un espace sémantique, la distance entre deux mots est une similarité sémantique d'un genre particulier. Elle est une mesure de leur degré de substituabilité ou d'interchangeabilité dans un corpus (Burgess et al., 1998; Charles, 2000; Lund & Burgess, 1996). C'est la raison pour laquelle certains appellent ce type d'espace sémantique un espace « paradigmatique », la métrique est interprétée comme une modélisation mathématique du concept de paradigme en linguistique structurale saussurienne (Sahlgren, 2006b; Schütze & Pedersen, 1993). Deux mots sont substituables l'un à l'autre lorsque nous pouvons les permuter sans changer le sens des segments de texte dans lesquels ils apparaissent.

La Figure 3.5 est une matrice des distances entre les cinq mots qui composent le corpus fictif de la Figue 3.3. Les distances sont calculées à l'aide de la métrique euclidienne.

	BLA	BLE	BLI	BLO	BLU
BLA	0	2	2,828	2,236	2
BLE		0	2	2,236	0
BLI			0	2,236	2
BLO				0	2,236
BLU					0

Figure 3.5: Matrice des distances euclidiennes entre cinq mots.

Cette matrice montre que la distance qui sépare les mots fictifs BLE et BLU est égale à zéro. La raison de cette équivalence est qu'ils ont exactement les mêmes environnements ou contextes d'usage, ils sont par conséquent parfaitement substituables l'un à l'autre dans ce petit corpus.

Ce corpus fictif est cependant trivial. Dans un corpus composé de plusieurs millions de mots, la substituabilité entre deux mots est rarement parfaite, mais la valeur de la métrique traduira leur degré de similarité sémantique.

3.3 Structures de l'espace sémantique

Un espace sémantique est caractérisé par des structures. Les mots sont organisés spatialement. Par exemple, les mots qui peuplent l'espace ne sont pas tous équidistants les uns par rapport aux autres. Certains mots sont très proches, d'autres sont très loin, certaines régions de l'espace sont très denses et d'autres au contraire sont presque vides. L'une des hypothèses de la SV est que les structures mathématiques de l'espace vectoriel, notamment des structures algébriques, géométriques et topologiques, modélisent des structures sémantiques. Il y aurait des isomorphismes entre structures mathématiques et structures sémantiques.

On ne connaît toutefois que très peu de choses concernant les structures des espaces sémantiques. Les conjectures avancées par la plupart des chercheurs sont des généralisations empiriques, c'est-à-dire des inductions fondées sur l'analyse empirique de plusieurs corpus. Une conjecture importante avancée par plusieurs chercheurs traite du phénomène d'agglutination des mots dans l'espace. La structure mathématique qui résulte de ce phénomène d'agglutination est interprétée comme étant le reflet d'une structure de classes dans l'espace sémantique.

3.3.1 Classes sémantiques

Dans un espace sémantique, des sous-ensembles de mots se regroupent dans certaines régions. Ce phénomène est schématisé dans la Figure 3.6. Le diagramme illustre un espace sémantique à trois dimensions peuplé par 27 vecteurs, chacun représentant la position d'un mot fictif. Dans cette illustration, ces 27 mots semblent être regroupés en deux régions clairement séparées l'une de l'autre. L'une d'entre elles se situe sur les dimensions x et z et l'autre principalement sur les dimensions x et y.

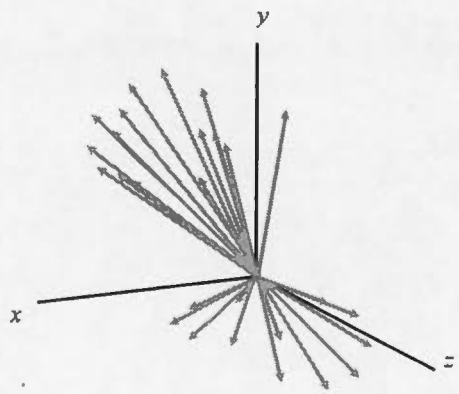

Figure 3.6: Illustration d'un espace vectoriel à trois dimensions caractérisé par deux regroupements de vecteurs.

La SV est basée sur l'hypothèse que ces régions regroupant des mots proches les uns des autres sont des structures algébriques reflétant des classes sémantiques ou des

concepts exprimés dans un corpus de texte. Une région de l'espace entretiendrait un isomorphisme structurel avec une classe sémantique, car elle regroupe des mots qui ont dans un corpus des environnements ou des contextes d'usage similaires. Dit autrement, une région peut être considérée comme une classe sémantique ou un concept, car elle regroupe des mots approximativement substituables les uns aux autres (Bullinaria & Levy, 2007; Burgess et al., 1998; Hindle, 1990; Lin & Pantel, 2002; Riordan & Jones, 2011; Widdows, 2004, p. 179).

La définition mathématique exacte de la forme de ces régions est une question de recherche ouverte. Nous pouvons toutefois retrouver quelques éléments de réponses dans un programme de recherche très similaire à la sémantique vectorielle, développé par Peter Gärdenfors et appelé les « espaces conceptuels »[12]. En nous basant sur les travaux de Gärdenfors, il est possible de préciser davantage les propriétés géométriques de cette structure de classes qui caractérise de nombreux espaces sémantiques. Ces espaces sémantiques auraient les propriétés d'une partition de Voronoi (Gärdenfors, 2014, Chapter 2).

Une partition de Voronoi est un découpage de l'espace en plusieurs régions convexes. Le diagramme de la Figure 3.7 illustre un exemple de ce type de structure géométrique. Dans cet exemple, nous avons une partition de Voronoi de 27 régions d'un espace à deux dimensions. Cette structure peut toutefois être généralisée à k régions et à des espaces de m dimensions.

[12] Peter Gärdenfors a développé une théorie des « espaces conceptuels » très similaire à la théorie de la sémantique vectorielle, mais se situant à un niveau de généralisation supérieure à cette dernière. La théorie des espaces conceptuels porte sur la cognition en général, qui inclut la sémantique des mots, mais ne s'y limitant pas.

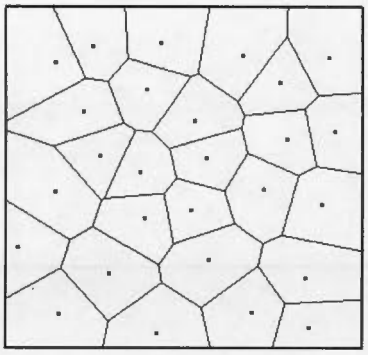

Figure 3.7: Partition de Voronoi d'un espace à deux dimensions et d'une métrique euclidienne.

Afin de définir les propriétés d'une partition de Voronoi, nous supposerons une matrice de coordonnées $W^{n\times m}$, dans laquelle le vecteur $\vec{u}_i$ représente la position du mot u_i dans un espace sémantique à m dimensions. Formellement, une partition de Voronoi en k régions forme un ensemble $VOR = \{R_1, \dots R_k\}$, dans lequel chaque région R_j forme un sous-ensemble qui satisfait la condition suivante :

$$R_j = \left\{ \vec{u}_i \in W^{n\times m} : \underset{\substack{R_h \in VOR \\ h \neq j}}{\forall} d(\vec{u}_i, \vec{c}_j) < d(\vec{u}_i, \vec{c}_h) \right\} \qquad \text{Eq. (3.10)}$$

$$\vec{c}_j = \frac{1}{|R_j|} \sum_{\vec{u}_i \in R_j} \vec{u}_i \qquad \text{Eq. (3.11)}$$

Le symbole $\vec{c}_j$ représente la position du centre[13] de la région R_j et $d(\vec{u}_i, \vec{c}_j)$ est une métrique euclidienne. La définition Eq. (3.10) précise les propriétés de convexité de chacune des régions. Interprété dans le cadre de la SV, elle stipule qu'un mot membre d'une région R_j est toujours plus proche du centre de R_j que du centre des autres régions de l'espace.

[13] Aussi appelé son « barycentre » ou son « centroïde ».

Selon Gärdenfors, cette structure géométrique permettrait d'expliquer les « effets de prototype » des classes sémantiques (Gärdenfors, 2014, p. 33). Les mots membres d'une classe sémantique n'ont pas tous le même statut, certains mots sont plus représentatifs que d'autres du prototype de la classe. Dans un espace sémantique qui a les propriétés géométriques d'une partition de Voronoi, la distance entre la position d'un mot et le centre d'une région peut s'interpréter comme une mesure de la représentativité du mot de la classe sémantique. Plus un mot est proche du centre de sa région d'appartenance, plus il est représentatif de la classe sémantique qui correspond à cette région. À l'inverse, plus un mot est loin du centre, moins il est représentatif de la classe sémantique.

3.3.2 Algorithmes de partitionnement

Dans l'exemple de la Figure 3.6, il est facile d'identifier les différentes régions de l'espace. Ces régions sont clairement séparées les unes des autres. Ceci est cependant l'exception. Dans un espace sémantique à plusieurs milliers de dimensions, peuplé de plusieurs milliers de mots, la frontière entre les régions de l'espace est généralement très difficile à identifier. Cette difficulté a plusieurs explications.

Une première explication vient du fait que dans un espace sémantique, plusieurs mots se situent à l'intersection de plusieurs régions. C'est ce qui se produit lorsqu'un mot a plusieurs sens ou acceptions, c'est-à-dire qu'il est polysémique, comme le mot SOURIS en tant qu'animal et en tant que dispositif informatique. Les mots polysémiques sont proches de plusieurs régions, car dans un corpus ils sont associés à des contextes d'usage caractérisés par plusieurs distributions de cooccurrences simultanément. Par exemple, le mot SOURIS peut être à la fois caractérisé par des cooccurrences du domaine animal et des cooccurrences qui appartiennent au domaine informatique. Or, les cooccurrents de ces domaines constituent les dimensions de régions très éloignées dans l'espace.

Pour Gärdenfors, cette idée qu'il soit possible d'identifier clairement les classes sémantiques des mots est un mythe (Gärdenfors, 2014, p. 222). Ces classes sont dynamiques et ceci se reflète au niveau des frontières des régions des espaces sémantiques. Les frontières des régions d'un espace sémantique évoluent avec le cumul des contextes d'usages. À chaque nouvel usage, le sens des mots est modifié, parfois de manière infime, parfois de manière importante. Chaque fois qu'un nouveau segment de texte s'ajoute à un corpus, la position des mots dans l'espace est ajustée.

Un autre obstacle à l'identification des classes sémantique est la quantité formidable de partitions possibles d'un espace multidimensionnel. Le nombre de partitions possible augmente de manière exponentielle avec le nombre de mots projetés dans l'espace sémantique. À titre d'exemple, il y a 115 975 partitions possibles d'un espace peuplé de seulement 10 mots, 4.76×10^{115} partitions possibles d'un espace composé de 100 mots et 2.98×10^{1927} partitions possibles d'un espace de 1000 mots.[14] À titre comparatif, c'est beaucoup plus que le nombre estimé de particules dans l'univers. Or, les espaces sémantiques ont généralement plusieurs milliers de mots. L'analyse exhaustive est impossible. Le partitionnement d'un espace sémantique n'est toujours qu'une approximation parmi plusieurs possibles.

Ce type d'approximation est réalisée à l'aide d'algorithmes de partitionnement, aussi appelés algorithmes de « regroupement automatique » ou de « classification automatique non-supervisée » (Jain, Murty, & Flynn, 1999). Dans la littérature, ces algorithmes sont présentés de la manière suivante :

> « Clustering refers to a very broad set of techniques for finding subgroups, or clustering clusters, in a data set. When we cluster the observations of a data set, we seek to partition them into distinct groups so that the observations within each group are quite similar to each other, while observations in different groups are quite different from each other. » (James et al., 2013, p. 385)

[14] Le nombre de partitions possible d'un ensemble correspond au nombre de Bell.

En d'autres mots, les algorithmes de partitionnement sont des heuristiques qui permettent de trouver une partition qui maximise un critère d'optimisation, généralement l'inertie inter-classe ou intra-classe. Il existe des centaines d'algorithmes différents de partitionnement. Il serait impossible dans le cadre de cette recherche de les présenter tous en détail. Pour une revue de la littérature voir (Berkhin, 2006; Jain, 2010; Jain et al., 1999; Steinley, 2006; Theodoridis & Koutroumbas, 2008; Xu & Wunsch, 2005).

Chaque algorithme est une heuristique différente. Le Tableau 3.5 présente sommairement quelques-uns d'entre eux couramment utilisé. KM et SOM sont deux algorithmes de partitionnement par centres mobiles très similaires. SOM peut être vu comme une extension de KM, prenant en compte également le voisinage de chaque centre mobile (Ultsch, 1995). SLINK et DBSCAN sont deux algorithmes agglomératifs (Aggarwal, 2015, p. 182). Contrairement à SLINK, DBSCAN est cependant insensible aux cas aberrants et ne produit pas de structure hiérarchique (dendogramme). Comme SLINK, BKM produit une structure hiérarchique, mais la procédure de regroupement est inversée, elle est divisive au lieu d'être agglomérative.

Dans le cinquième chapitre, l'algorithme des KM est présenté plus en détail.

Tableau 3.5: Quelques exemples d'algorithmes de partitionnement automatique.

Algorithmes	**Principaux paramètres**	**Heuristique de regroupement**	**Référence**
KM (k-means)	Fixer le nombre k de classes	Regroupement par centres mobiles	(MacQueen, 1967)
SLINK (Single-linkage clustering)	Fixer le nombre k de classes	Regroupement agglomératif hiérarchique	(Sibson, 1973)
BKM (Bisecting k-means)	Fixer le nombre k de classes	Division descendante hiérarchique	(Steinbach, Karypis, & Kumar, 2000)
DBSCAN (Density-Based Spatial Clustering of Applications with Noise)	Fixer le rayon du voisinage et la densité minimale des regroupements	Regroupement agglomératif non-hiérarchique	(Ester, Kriegel, Sander, & Xu, 1996)
SOM (Self Organizing Map)	Fixer les propriétés topologiques de la carte (dimensions et voisinage)	Regroupement par centres mobiles avec ajustement du voisinage	(Kohonen, 1982)

3.4 Conclusion

La SV est un modèle computationnel de la représentation des contenus sémantiques et une méthode algorithmique de découverte inductive de ces contenus dans un corpus de texte. La SV peut être vue comme la combinaison du cadre théorique de la sémantique différentielle d'Osgood et de la méthode distributionnelle de Harris.

On la résume de la manière suivante dans la littérature :

> « The basic idea is simply that words with similar meanings will tend to occur in similar contexts, and hence word co-occurrence statistics can provide a natural basis for semantic representations. » (Bullinaria & Levy, 2007, p. 510)

> « The terms distributional, context-theoretic, corpus-based or statistical can all be used (almost interchangeably) to qualify a rich family of approaches to semantics that share a "usage-based" perspective on meaning, and assume that the statistical distribution of words in context plays a key role in characterizing their semantic behaviour. » (Lenci 2008: 1)

> « Distributional models build semantic representations from the statistical regularities of word co-occurrences in large-scale linguistic corpora. These models are based on the distributional hypothesis: The more similar the contexts in which two words appear, the more similar their meanings. » (Riordan & Jones, 2011, pp. 303–304)

> « A semantic space is a space, often with a large number of dimensions, in which words [...] are represented by points; the position of each such point along each axis is somehow related to the meaning of the word. » (Lund & Burgess, 1996, p. 203)

Des régularités statistiques dans la distribution des mots dans un corpus permettent de construire des espaces sémantiques dont les relations et les structures sont corrélées aux contenus sémantiques exprimés dans le corpus.

La SV peut faire l'objet de nombreuses opérationnalisations autres que celles qui ont été présentées dans ce chapitre. D'une part, différents types de distributions dans un corpus permettre de construire différents types d'espaces sémantiques dans lesquels les relations de distance et de proximité peuvent correspondre à différents types de similarité sémantique. D'autre part, différentes structures mathématiques, autres que la partition, peuvent possiblement modéliser d'autres types de structures mathématiques.

CHAPITRE IV

L'APPRENTISSAGE MACHINE

4.0 Introduction

Ce chapitre est consacré à l'introduction des principaux paramètres de l'apprentissage machine. Dans le cadre de cette thèse, l'apprentissage machine est le cadre théorique et méthodologique à partir duquel est effectué la reconstruction du mécanisme à l'œuvre dans l'évolution du réseau sociosémantique du SSS.

L'apprentissage machine est un domaine à l'intersection de l'intelligence artificielle, de la statistique, de l'informatique et de la reconnaissance de forme, dont l'émergence date environ d'une trentaine d'années. Le premier colloque titré explicitement « machine learning » a eu lieu en 1980 à Carnegie-Mellon University (Sammut & Webb, 2011, p. i). C'est toutefois la publication quelques années plus tard d'un ensemble de textes fondamentaux qui allait réellement consacrer le domaine. Ces textes présentaient différents types d'apprenants automatiques qui allaient devenir les principaux paradigmes du domaine. Parmi les plus importants, on retrouve les apprenants à base de réseaux de neurones artificiels (McClelland, Rumelhart, & Group, 1986), les apprenants d'arbre de décisions (Quinlan, 1986), les apprenants à base d'exemplaires (Aha, Kibler, & Albert, 1991), les apprenants à base de règles (Michalski, 1983), les apprenants bayésiens (Langley, Iba, & Thompson, 1992) et les apprenants basés sur des techniques de séparation à vastes marges (Cortes & Vapnik, 1995).

Selon l'horizon théorique d'appartenance des auteurs, l'apprentissage machine est défini de différente manière dans la littérature. Pour certains, l'apprentissage machine

consiste à étudier comment une machine (un ordinateur) peut apprendre par expérience :

> « The field of Machine Learning seeks to answer the question "How can we build computer systems that automatically improve with experience, and what are the fundamental laws that govern all learning processes?" [...] To be more precise, we say that a machine learns with respect to a particular task T, performance metric P, and type of experience E, if the system reliably improves its performance P at task T, following experience E. » (T. M. Mitchell, 2006, p. 1)

Pour d'autres, l'apprentissage machine se définit davantage comme une méthode d'analyse de données, notamment d'analyse prédictive :

> « [...] we define machine learning as a set of methods that can automatically detect patterns in data, and then use the uncovered patterns to predict future data, or to perform other kinds of decision making under uncertainty [...]. » (Murphy, 2012, p. 1)

> « The field of pattern recognition is concerned with the automatic discovery of regularities in data through the use of computer algorithms and with the use of these regularities to take actions such as classifying the data into different categories. » (Bishop, 2006, p. 1)

Une autre perspective, plus proche de la statistique, est celle adoptée dans la présente thèse. L'apprentissage machine y est défini comme un cadre méthodologique général d'induction automatique d'une hypothèse d'approximation ou d'un modèle d'une fonction cible. Dans la littérature, cette perspective est particulièrement bien représentée par des ouvrages comme (Hastie, Tibshirani, & Friedman, 2009) et (James et al., 2013). L'apprentissage machine y est introduit de cette façon :

> « [...] suppose that we observe a quantitative response Y and p different predictors $X_1, X_2, \ldots, X_p$. We assume that there is some relationship between Y and $X = (X_1, X_2, \ldots, X_p)$, which can be written in the very general form $Y = f(X)$ [...]. Here f is some fixed but unknown function of $X_1, X_2, \ldots, X_p$ [...]. In essence, statistical learning refers to a set of approaches for estimating f. » (James et al., 2013, pp. 16–17)

Vouloir approximer une fonction cible peut être motivé par plusieurs raisons qui peuvent être regroupées en deux grandes catégories : la prédiction et l'inférence (James et al., 2013, p. 17).

Premièrement, la fonction qui est l'objet d'étude peut parfois être extrêmement complexe et posséder un grand nombre de paramètres. L'apprentissage machine peut alors être utilisé afin de substituer à cette fonction cible un modèle algorithmique beaucoup plus facile à computer, comme une forêt aléatoire de décisions. Ce modèle, bien que ne révélant pas la nature de la relation entre la variable indépendante et dépendante de la fonction cible, sera néanmoins capable de prédire approximativement ses valeurs, mais avec un coût computationnel bien moindre.

Une autre motivation est celle introduite en début de thèse. Dans une problématique de reconstruction, la fonction cible représente un mécanisme à l'œuvre dans un système empirique, mais non observable directement. L'objectif est alors de reconstruire ce mécanisme en inférant un modèle capable de l'émuler de la manière la plus transparente possible, par exemple via un arbre de décisions ou une liste de règles. Ce modèle doit pouvoir se substituer à la fonction cible, mais aussi permettre de mieux comprendre la relation entre les entrées et les sorties du mécanisme étudié.

Dans la pratique, les problématiques d'apprentissage machine sont généralement motivées à la fois par de la prédiction et de l'inférence. Même si le but est l'inférence, méthodologiquement, la vraisemblance d'une hypothèse d'approximation est généralement évaluée selon sa capacité prédictive.

L'apprentissage machine est aujourd'hui un domaine de recherche très important, qui déborde les domaines de l'informatique. Il est devenu un cadre méthodologique général utilisé dans une multitude de domaines. Il est utilisé pour la reconnaissance vocale, la traduction, la reconnaissance d'images, la fouille de textes et la fouille de données, la biométrie, la détection de fraudes bancaires, la prédiction des comportements de consommation, la construction de système de recommandation automatique, etc. (pour un aperçu des domaines d'application, voir (Kuhn & Johnson, 2013)).

Une autre contribution de l'apprentissage machine est particulièrement importante dans le cadre de cette thèse. C'est ce que Mitchell appelle « l'accélération » des découvertes dans les sciences empiriques (T. M. Mitchell, 2006). L'accélération fulgurante du séquençage du génome humain ces dix dernières années ou la multiplication des découvertes de probables exoplanètes sont étroitement liées au développement de l'apprentissage machine.

Quelque chose de similaire se produit actuellement dans les sciences humaines et sociales. Les comportements humains, individuels et collectifs, laissent de plus en plus de traces digitales (e.g. géolocalisation, navigation sur le web, achats en ligne, numérisation des textes et des images, réseaux sociaux numériques). Selon Mitchell, l'apprentissage machine sera bientôt aussi bien intégré dans les sciences humaines et sociales qu'il l'est actuellement en biologie, en météorologie ou en astronomie.

L'apprentissage machine peut se résumer par un 6-uplets $\langle \mathbb{I}, \mathbb{H}, \mathbb{A}, \mathbb{B}, \Gamma, \Omega \rangle$. Le paramètre $\mathbb{I}$ représente la définition du problème, ici appelé l'espace des instances de la fonction cible qu'on souhaite approximer. Le paramètre $\mathbb{H}$ représente l'ensemble des hypothèses d'approximation ou modèles possibles de la fonction cible. Le paramètre $\mathbb{A}$ représente une base d'apprentissage à partir de laquelle est induite une hypothèse d'approximation. Le paramètre $\mathbb{B}$ représente une base de test à partir de laquelle les hypothèses de modélisation sont évaluées. Le paramètre Γ représente l'apprenant automatique utilisé pour induire une hypothèse d'approximation. Finalement, le paramètre Ω représente l'évaluateur des hypothèses (T. M. Mitchell, 1997).

Le chapitre se divise en cinq sections, chacune introduisant l'un de ces paramètres.

4.1 L'espace des instances

La fonction cible d'un problème d'apprentissage machine peut prendre un très grand nombre de formes différentes discrètes et continues. Toutefois, dans de nombreux

domaines d'application, elle est souvent réduite à une fonction de classification binaire. Dénotons une fonction cible par $f: X \rightarrow Y$. Typiquement, la variable indépendante correspondra à un vecteur $\vec{x}_i = (x_1, \ldots x_m)$ de m attributs ou dimensions et la variable dépendante correspondra à une variable binaire $Y = \{1,0\}$. Afin de simplifier la présentation de l'apprentissage machine, et aussi parce que ce type de fonction est directement lié à la problématique de recherche de la thèse, seulement les fonctions de classifications binaires sont discutées dans ce chapitre.

Le premier paramètre est l'espace des instances de la fonction que l'on souhaite approximer. L'espace des instances est défini ainsi :

$$\mathbb{I} = \{(\vec{x}_i, y_j) \in X^m \times Y\} \qquad \text{Eq. (4.1)}$$

Cet espace correspond à l'ensemble des relations possibles, dénotées par $(\vec{x}_i, y_j)$, entre chaque valeur possible de la variable indépendante et de la variable dépendante. Cet espace correspond à la définition du problème d'approximation.

À titre d'exemple, supposons une fonction cible $f: METEO \rightarrow APPRECIATION$. Cette fonction est une classification binaire des conditions météorologiques d'une journée. Les conditions météorologiques d'une journée constituent la variable indépendante et sont décrites dans un espace de six dimensions soit les six attributs booléens <chaud, pluvieux, ensoleillé, venteux, humide, enneigé>. La variable dépendante est une classification binaire où $APPRECIATION = \{Température\ agréable,\ Température\ désagréable\}$.

Dans cet exemple, l'espace des instances de la fonction cible correspond à l'ensemble des relations entre, d'une part, les six conditions météorologiques possibles d'une journée et, d'autre part, leurs appréciations positives ou négatives. La fonction $f: METEO \rightarrow APPRECIATION$ a donc un espace de $2^6 \times 2$ instances possibles.

Par ailleurs, lorsque la fonction cible représente un mécanisme à l'œuvre dans un système empirique (e.g. social, cognitif, électrique), l'espace des instances est aussi associé à une distribution de probabilité $\mathbb{P}(\mathbb{I})$ pour laquelle $\sum_{(\vec{x}_i, y_j)\in\mathbb{I}} \Pr(\vec{x}_i, y_j) = 1$. Cette distribution caractérise le comportement empirique de la fonction cible que l'on souhaite approximer par apprentissage machine. Elle assigne à chaque instance de $\mathbb{I}$ une probabilité d'occurrence. Dans l'exemple de la fonction $f: METEO \rightarrow APPRECIATION$, la valeur de $\Pr(\vec{x}_i, y_j)$ indiquerait alors la probabilité qu'une journée au hasard corresponde par exemple à l'instance ($\vec{x}_i$= (chaud=oui, pluvieux=non, ensoleillé=oui, venteux=non, humide=oui, enneigé=non), yj = Température agréable).

Comme il sera discuté plus loin, la distribution de probabilité $\mathbb{P}(\mathbb{I})$ est importante, car elle doit déterminer la construction de la base d'apprentissage et de la base de test.

4.2 L'espace des hypothèses

Le paramètre $\mathbb{H}$ représente l'espace des hypothèses d'approximation possibles de la fonction cible. Formellement, l'espace des hypothèses est défini ainsi :

$$\mathbb{H} = \{\check{f}(\vec{x}_i) \in Y^X\} \qquad \text{Eq. (4.2)}$$

Chaque hypothèse d'approximation $\check{f} \in \mathbb{H}$ représente un modèle possible de la fonction cible. À titre d'exemple, l'espace des hypothèses de la fonction cible $f: METEO \rightarrow APPRECIATION$ correspond à l'ensemble des bipartitions possibles de l'espace des instances, soit 2^{64} hypothèses possibles. Certaines hypothèses d'approximation seront très similaires à la fonction cible, d'autres au contraire seront très différentes. L'objectif d'une méthode d'apprentissage machine est de sélectionner dans $\mathbb{H}$ la meilleure approximation possible.

Plus une fonction cible f possède un espace des instances complexe, c'est-à-dire plus il y a d'attributs qui décrivent sa variable indépendante et sa variable dépendante, plus l'espace des hypothèses est vaste. Même un problème d'apprentissage machine relativement simple comme l'approximation de la fonction $f: METEO \rightarrow APPRECIATION$, qui est définie par un espace de seulement 128 instances, représente un formidable problème ayant plusieurs millions de milliards de solutions possibles.

Pour tous les problèmes non triviaux d'apprentissage machine, il est impossible d'analyser de manière exhaustive l'espace des hypothèses. Dans notre exemple, si nous pouvions analyser une hypothèse par secondes, cela prendrait environ 600 milliards d'années analyser $\mathbb{H}$ dans sa totalité. Ceci montre bien toute la difficulté du problème. Il faut chercher une solution parmi des milliards d'hypothèses possibles.

Un apprenant automatique permet d'éviter cet obstacle méthodologique. À l'aide de différentes stratégies inductives et d'une base d'exemplaires, il peut fouiller de manière intelligente $\mathbb{H}$.

4.3 La base d'apprentissage et la base de test

Les paramètres $\mathbb{A}$ et $\mathbb{B}$ correspondent respectivement à la base d'apprentissage et à la base de test d'une méthode d'apprentissage machine. Une base d'apprentissage est un ensemble d'exemplaires de la fonction cible f à partir duquel sera induite une hypothèse d'approximation $\check{f}$. La base de test est quant à elle un ensemble d'exemplaires qui permettra d'évaluer la vraisemblance de l'approximation.

Formellement, une base d'apprentissage et une base de test sont deux multiensembles composés de plusieurs exemplaires des instances de f :

$$\mathbb{A} = (A, \otimes) \qquad \text{Eq. (4.3)}$$
$$A = \{(\vec{x}_i, y_j) \in \mathbb{I}\}$$

$$\mathbb{B} = (B, \otimes) \qquad \text{Eq. (4.4)}$$
$$B = \{(\vec{x}_i, y_j) \in \mathbb{I}\}$$

La fonction $\otimes: \{(\vec{x}_i, y_j) \in \mathbb{I}\} \longrightarrow \mathbb{N}_{\geq 1}$ est une multiplicité qui assigne une fréquence à une instance de $\mathbb{I}$. Cette fonction détermine le nombre d'exemplaires d'une instance particulière qui composera la base d'apprentissage ou la base de test.

Lorsque la fonction cible représente un phénomène empirique, par exemple un mécanisme à l'œuvre dans un système, les exemplaires sont issus d'un processus de collecte d'observations. Dans ce contexte, une base d'apprentissage et une base de test correspondent alors à deux échantillons aléatoires sur l'espace des instances. Ces échantillons doivent être les plus représentatifs possible du comportement empirique de f. En d'autres mots, le nombre d'exemplaires d'une instance particulière doit refléter le plus possible sa probabilité d'occurrence.

4.4 L'apprenant automatique

Le cinquième paramètre est un apprenant automatique. Un apprenant est un programme informatique responsable de fouiller $\mathbb{H}$ et d'induire à partir d'une base d'apprentissage $\mathbb{A}$ et de postulats généraux sur la forme de f une hypothèse d'approximation $\check{f}$ (Harman & Kulkarni, 2007). Autrement dit, le processus d'apprentissage machine repose sur une procédure algorithmique de fouille de $\mathbb{H}$, qui exploite d'une part les informations présentes dans une base d'apprentissage $\mathbb{A}$, et d'autre part, des connaissances a priori sur la forme de la fonction cible.

De manière générale, on peut dénoter un apprenant automatique de la manière suivante :

$$\Gamma(\mathbb{A}, \beta) = \check{f} \qquad \text{Eq. (4.5)}$$

Le paramètre β correspond à un biais inductif. C'est ce paramètre qui représente les connaissances a priori dont dispose Γ sur la forme de la fonction cible f.

Un apprenant Γ a deux objectifs. Il doit trouver dans $\mathbb{H}$ une hypothèse $\check{f}$ à la fois représentative des exemplaires d'une base d'apprentissage $\mathbb{A}$ et généralisable avec le moins d'erreurs possibles à de nouvelles instances de f. En d'autres mots, Γ doit trouver une hypothèse $\check{f}$ cohérente avec $\mathbb{A}$ et prédire $\mathbb{B}$, une base de test.

L'évaluation de ces objectifs de cohérence et de généralisation sera discutée dans la prochaine section. La présente section est consacrée au rôle de la base d'apprentissage et des biais inductifs dans le processus d'induction automatique.

Un apprenant ne fouille jamais la totalité de l'espace des hypothèses possibles d'une fonction cible. Tel que discuté précédemment, une telle procédure en force brute est inopérante pour tous problèmes d'apprentissage machine non triviaux. Un apprenant ne fouille qu'un sous-espace $H \subseteq \mathbb{H}$. Ce sous-espace H correspond à l'ensemble des hypothèses accessibles (i.e. « apprenable ») par Γ. Les frontières de ce sous-espace H sont déterminées par deux contraintes : premièrement, il est déterminé par les connaissances empiriques de Γ représentées par une base d'apprentissage et deuxièmement, il est déterminé par ses connaissances a priori représentées par un biais d'induction.

4.4.2 Les connaissances empiriques de l'apprenant

La base d'apprentissage est la première contrainte déterminant les limites du sous-espace H. À l'exception des scénarios triviaux où toutes les instances possibles de $\mathbb{I}$ sont représentées par un exemplaire dans la base d'apprentissage, généralement ceux-ci ne représentent qu'un très petit échantillon de $\mathbb{I}$. Or, plus les dimensions de la base

d'apprentissage sont restreintes, plus le nombre d'hypothèses dans le sous-espace H est limité et par conséquent plus la probabilité est grande que la meilleure approximation possible de f ne soit pas accessible à l'apprenant.

La base d'apprentissage a donc un impact majeur sur la quantité d'hypothèses accessibles à l'apprenant. Ceci peut être illustré à nouveau à l'aide de l'exemple de la fonction $f: METEO \rightarrow APPRECIATION$. Cette fonction est caractérisée par une variable indépendante de 2^6 valeurs possibles, c'est-à-dire six attributs booléens et une variable dépendante de deux valeurs possibles, soit une classification binaire. Elle possède par conséquent un espace d'hypothèses de 2^{64} bipartitions possibles. Supposons maintenant que la base d'apprentissage $\mathbb{A}$ ne contienne que des exemplaires décrits par cinq des six attributs de la variable indépendante. Par exemple, aucun exemplaire ne serait décrit par l'attribut <venteux>. En raison des effets combinatoires entre attributs sur l'espace des hypothèses, en omettant de décrire les exemplaires par un seul attribut, c'est plus de 99,99% des hypothèses de $\mathbb{H}$ qui ne se retrouveront pas dans le sous-espace H des hypothèses accessibles à l'apprenant.

4.4.3 Les connaissances a priori de l'apprenant

La deuxième contrainte déterminant le sous-espace H est le biais inductif utilisé par l'apprenant (Baxter, 2000; Gordon & Desjardins, 1995; Utgoff, 1986). Un biais inductif est un ensemble de postulats portant soit sur la forme de la fonction cible ou sur la procédure de fouille de l'espace des hypothèses. Les premiers sont appelés des biais inductifs représentationnels et les seconds des biais inductifs procéduraux.

Il existe de nombreux types de biais inductifs représentationnels. Par exemple, les inductions d'un apprenant peuvent être fondées sur le postulat que la fonction cible a la forme d'une règle conjonctive. Sur la base de ce type de postulat, seulement les hypothèses qui peuvent être représentées par une telle règle seront accessibles à l'apprenant. Pour illustrer l'impact qu'un tel biais inductif peut avoir, nous pouvons

reprendre l'exemple de la fonction cible $f: METEO \rightarrow APPRECIATION$. Tel que mentionné ci-haut, cette fonction est caractérisée par un espace de 2^{64} hypothèses de bipartitions possibles. C'est un espace de fouille immense. Toutefois, seulement quelques-unes de ces 2^{64} hypothèses peuvent être représentées sous la forme d'une règle conjonctive. En fait, il n'y a que 730 hypothèses sur 2^{64} qui soient compatibles avec ce biais représentationnel, une fraction infime.

Les postulats d'un biais inductif représentationnel ont donc un effet très restrictif sur la taille du sous-espace $H \subseteq \mathbb{H}$ (T. M. Mitchell, 1997, p. 24). Dans notre exemple, rien ne garantit que la meilleure approximation de $f: METEO \rightarrow APPRECIATION$ fasse partie de ces 730 hypothèses. Il se pourrait, peut-être, que la meilleure approximation soit plutôt une règle contenant une disjonction. Ce type de difficulté est inhérent à tout problème d'approximation, car on ne connaît généralement pas, a priori, la forme qui caractérise la fonction cible.

Il y a de nombreux autres types de biais inductifs représentationnels. Un biais utilisé par de nombreux apprenants est le postulat que la fonction cible soit de forme linéaire. Ce biais est notamment à la base de l'apprentissage machine de modèles bayésiens naïfs. Un autre exemple est le biais représentationnel utilisé l'apprenant appelé les « n-plus-proche-voisins ». Ce dernier postule que la fonction cible a les propriétés d'une partition de Voronoi (T. M. Mitchell, 1997, p. 234).

D'autres apprenants automatiques ne font aucun postulat représentationnel. Ils font cependant d'autres types de postulats appelés des biais inductifs procéduraux. Ces biais ne limitent pas a priori les dimensions du sous-espace $H \subseteq \mathbb{H}$ accessibles à l'apprenant, mais ils contraignent la procédure de fouille de $\mathbb{H}$, notamment l'ordre ou la priorité accordée à certains types d'hypothèses au détriment d'autres.

Un exemple typique de biais procéduraux est celui utilisé par un apprenant d'arbre de décisions. Ce type d'apprenant a accès, en principe, à la totalité de l'espace des

hypothèses d'approximation de la fonction cible. Les arbres que cet apprenant peut induire peuvent approximer n'importe quelles formes de fonction. Toutefois, durant la procédure de fouille de l'espace des hypothèses, les arbres courts sont systématiquement privilégiés aux arbres longs (T. M. Mitchell, 1997, p. 63). Ce biais procédural est une variante du célèbre rasoir d'Occam : toutes choses étant égales par ailleurs, les hypothèses simples sont toujours préférables aux hypothèses plus complexes. Un apprenant utilisant ce biais postulera que l'arbre de décisions le plus court cohérent avec $\mathbb{A}$ est toujours une meilleure hypothèse d'approximation qu'un autre arbre cohérent avec, $\mathbb{A}$ mais plus long. Par conséquent, bien que $\mathbb{H}$ ne soit soumis à aucune restriction a priori quant à sa taille, la procédure de fouille a pour effet que seulement un sous-espace de $\mathbb{H}$ est réellement fouillé par l'apprenant.

Pour les problèmes d'apprentissage machine non triviaux, un biais inductif est une condition de possibilité pour une fouille intelligente de l'espace des hypothèses, c'est-à-dire réalisable en temps polynomial. Sans postulats de base qui permettent de restreindre les dimensions ou la procédure de fouille de l'espace des hypothèses, un apprenant n'a aucune base rationnelle à partir de laquelle il peut induire une solution (T. M. Mitchell, 1997, p. 42).

4.5 L'évaluateur

Le dernier paramètre de l'apprentissage machine est un évaluateur. Un évaluateur a deux objectifs, celui d'évaluer la cohérence d'une hypothèse avec une base d'apprentissage et celui d'évaluer sa généralisation avec une base de test.

4.5.1 Évaluer la cohérence

L'évaluation de la cohérence d'une hypothèse d'approximation $\check{f}$ consiste à calculer le résidu produit par $\check{f}$ sur une base d'apprentissage $\mathbb{A}$. Ce résidu peut être calculé de

plusieurs manières, mais dans la majorité des cas il correspond simplement à la proportion d'erreurs que $\check{f}$ produit sur les exemplaires de la base d'apprentissage :

$$\varepsilon_{cohérence} = \frac{1}{|\mathbb{A}|} \sum_{(\vec{x}_i, y_j) \in \mathbb{A}} \begin{cases} 1, & \check{f}(\vec{x}_i) \neq y_j \\ 0, & \check{f}(\vec{x}_i) = y_j \end{cases} \quad \text{Eq. (4.6)}$$

Une hypothèse d'approximation $\check{f}$ est dite cohérente si elle est capable de décrire avec un minimum d'erreur les exemplaires qui composent $\mathbb{A}$ (T. M. Mitchell, 1997, p. 29).

4.5.2 Évaluer la généralisation

L'approximation d'une fonction doit aussi satisfaire un critère de généralisation. À partir d'une base d'apprentissage $\mathbb{A}$ ne représentant qu'un échantillon limité des instances de f, le modèle induit doit pouvoir se généraliser, c'est-à-dire prédire, de nouvelles instances de f non présentes dans $\mathbb{A}$.

Le résidu de généralisation se calcule de manière similaire au résidu de cohérence, mais le calcul est appliqué sur les exemplaires d'une base de test :

$$\varepsilon_{généralisation} = \frac{1}{|\mathbb{B}|} \sum_{(\vec{x}_i, y_j) \in \mathbb{B}} \begin{cases} 1, & \check{f}(\vec{x}_i) \neq y_j \\ 0, & \check{f}(\vec{x}_i) = y_j \end{cases} \quad \text{Eq. (4.7)}$$

Une hypothèse d'approximation $\check{f} \in \mathbb{H}$ est considérée généralisable à une base de test $\mathbb{B}$ si la probabilité que $\check{f}$ génère une erreur est faible.

4.5.3 Deux types d'erreurs

La cohérence et la généralisation d'une hypothèse dépendent étroitement du biais inductif utilisé par l'apprenant. Selon les postulats imposés par le biais inductif sur l'espace des hypothèses, deux types d'erreurs peuvent expliquer les résidus de cohérence et de généralisation. Ces types d'erreurs sont le sur-ajustement et le sous-

ajustement.[15] Ces deux types d'erreurs sont des compromis entre les objectifs de cohérence avec la base d'apprentissage et de généralisation avec la base de test (Domingos, 2012).

L'erreur de sur-ajustement est causée par un biais inductif trop faible, c'est-à-dire pas suffisamment restrictif sur l'espace des hypothèses. Par conséquent, l'hypothèse induite par l'apprenant est trop conforme à sa base d'apprentissage $\mathbb{A}$ et ne parvient pas à se généraliser à une base de test $\mathbb{B}$.

À l'inverse, l'erreur de sous-ajustement est causée par un biais inductif trop fort, c'est-à-dire trop restrictif sur l'espace des hypothèses d'approximation. Par conséquent, les hypothèses accessibles à l'apprenant sont de formes trop différentes à la fonction cible.

Ces deux types d'erreurs peuvent être illustrés graphiquement. Supposons une fonction de classification $f: X^2 \rightarrow \{+, -\}$ dont la variable indépendante est un vecteur à deux dimensions $\vec{x}_i = (x_1, x_2)$ et la variable dépendante est une classification binaire $y_j \in \{positif, négatif\}$. Dans la Figure 4.1, le trait plein en gras représente la fonction cible, la surface rectangulaire représente l'espace de ses instances, les croix représentent des exemplaires des instances de la classe <positif> et les tirets des exemplaires des instances de la classe <négatif>.

Supposons dans cet exemple que la base d'apprentissage est composée de 11 exemplaires de la classe <négatif>, de 16 exemplaires de la classe <positif> et que deux apprenants automatiques différents sont utilisés. Le premier induit une hypothèse d'approximation $\check{f}_1$ et le second une hypothèse $\check{f}_2$:

[15] On préfère parfois leur expression anglaise « overfitting » et « underfitting ».

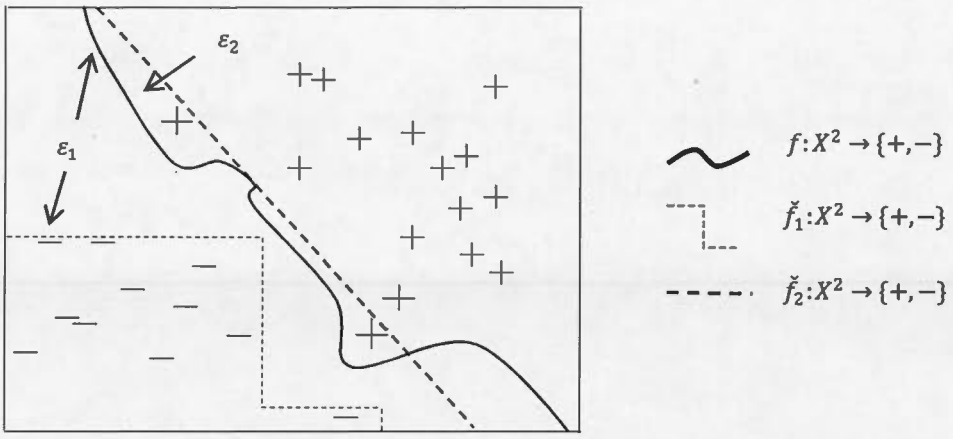

Figure 4.1: Représentation graphique d'une fonction cible, de son espace d'instances et de deux hypothèses d'approximation.

Dans ce graphique, l'hypothèse d'approximation $\check{f}_1$ illustre ce que serait une hypothèse induite par un apprenant utilisant un biais inductif trop faible. En effet, $\check{f}_1$ est caractérisée par un sur-ajustement à la base d'apprentissage. Cette hypothèse est parfaitement cohérente avec les exemplaires de la base d'apprentissage, mais ne peut se généraliser à aucune nouvelle instance de la fonction cible. Une telle hypothèse serait en effet incapable de prédire correctement la valeur des instances situées dans le graphique entre le trait pointillé de $\check{f}_1$ et le trait plein représentant la fonction cible (la zone dénotée par le résidu ε_1).

L'hypothèse d'approximation $\check{f}_2$ illustre quant à elle ce que serait une hypothèse d'approximation inférée par un apprenant utilisant un biais inductif trop fort. En effet, $\check{f}_2$ est caractérisée par un sous-ajustement. Dans ce cas, l'erreur est causée par l'utilisation du postulat faux que la fonction cible est de forme linéaire.

Dans l'illustration de la Figure 4.1, l'hypothèse $\check{f}_1$ est cohérente avec les exemplaires de la base d'apprentissage, mais se généralise très peu, alors que $\check{f}_2$ est partiellement incohérente (la zone dénotée par le résidu ε_2), mais se généralise beaucoup mieux que $\check{f}_1$.

4.5.4 Probablement approximativement correcte

Tous les biais inductifs produisent soit une erreur de sur-ajustement ou de sous-ajustement. Cette conclusion, appelée en anglais le « No free lunch theorem », suggère que sur une quantité suffisamment grande de différents problèmes d'apprentissage machine, en moyenne, tous les biais d'induction produisent la même quantité de résidu. Si un biais inductif performe bien pour l'approximation d'un type particulier de fonction, il performera moins bien pour d'autres types de fonctions, si bien qu'en moyenne, tous s'équivalent plus ou moins (Wolpert & Macready, 1997).

Compte tenu de ces limites inhérentes à un problème d'apprentissage machine, un apprenant automatique ne peut jamais trouver une hypothèse d'approximation parfaitement substituable à la fonction cible. Si nous fixons à λ le seuil minimal de résidu exigé pour un problème d'apprentissage machine, l'objectif d'un apprenant est de trouver une hypothèse $\check{f} \in \mathbb{H}$ qui maximisera la probabilité que l'erreur de cohérence et de généralisation ne dépassent pas le seuil λ, soit :

$$\underset{\check{f}\in\mathbb{H}}{arg\,max}\ \Pr[\varepsilon_{consistance} \leq \lambda \geq \varepsilon_{généralisation}|\check{f}] \qquad \text{Eq. (4.8)}$$

Sachant que des résidus de cohérence et de généralisation sont inévitables, un apprenant automatique ne peut chercher à minimiser des erreurs de cohérence de manière indépendante de sa stratégie de minimisation des résidus de généralisation. Par conséquent, l'objectif de l'apprenant est de sélectionner une hypothèse d'approximation qui génère un résidu minimal, mais très probable.[16]

[16] Il est fait implicitement référence ici au modèle PAC d'évaluation (en anglais « Probably Approximatly Correct »).

4.6 Conclusion

L'apprentissage machine est un cadre théorique et méthodologique d'approximation d'une fonction cible. Ce cadre peut se résumer par six paramètres :

1. L'espace des instances de la fonction cible, c'est-à-dire l'ensemble de ses valeurs possibles;
2. L'espace des hypothèses d'approximation possibles de la fonction cible, c'est-à-dire l'ensemble des modèles qui peuvent se substituer à la fonction cible;
3. Une base d'apprentissage, c'est-à-dire un échantillon aléatoire d'exemplaires des instances de la fonction cible;
4. Une base de test représentant un deuxième échantillon aléatoire d'exemplaire des instances de la fonction cible;
5. Un apprenant automatique, dont l'objectif est d'induire à l'aide d'une base d'apprentissage et d'un biais inductif, une hypothèse d'approximation de la fonction cible;
6. Un évaluateur, dont l'objectif est de calculer la cohérence de l'hypothèse d'approximation avec la base d'apprentissage et sa généralisation avec une base de test.

Dans ce chapitre, seulement les fonctions cibles de classification binaire ont été discutées. Ce choix est lié à la problématique de recherche de cette thèse, qui porte exclusivement sur ce type de fonction. Toutefois, l'apprentissage machine ne se limite pas à ce type de fonction. La plupart des modèles d'apprentissage machine, tels les arbres de décisions, les modèles bayésiens, les modèles à base d'exemplaires, les réseaux de neurones artificiels et autres, sont habituellement capables d'approximer des fonctions de plus de deux classes ainsi que des fonctions continues.

4.6.1 Les types d'apprenants automatiques

Un apprenant peut être opérationnalisé de plusieurs manières. Différents apprenants utiliseront différents biais d'induction et induiront différents modèles d'hypothèses d'approximation. Ces hypothèses auront la forme soit d'un arbre de décisions, une table de probabilités conditionnelles, une matrice de similitudes, un ensemble de prototypes, un réseau de neurones artificiels, un ensemble de règles, une forêt aléatoire ou un autre modèle.

Dans le chapitre cinq, différents types d'apprenants sont introduits, soit un inducteur d'un arbre de décisions, un inducteur de règles, un inducteur d'une forêt aléatoire et un inducteur bayésien naïf. Ce sont les modèles utilisés pour les expérimentations réalisées dans cette thèse, mais ce ne sont pas les seuls possibles (Bishop, 2006; Hastie et al., 2009, p. 351; T. M. Mitchell, 1997; Murphy, 2012).

4.6.2 Comment sélectionner un apprenant automatique?

Le choix d'un apprenant automatique dépend avant tout de la représentativité de la base d'apprentissage par rapport à l'espace des instances de la fonction cible. Une autre manière d'évaluer les hypothèses d'approximation induites par un apprenant automatique est par l'entremise de leur variance (Domingos, 2012). Lorsqu'un apprenant utilise un biais inductif faible, il a tendance à induire des hypothèses très cohérentes avec sa base d'apprentissage, mais il a aussi tendance à générer beaucoup de variances sur une base de test. D'un autre côté, lorsqu'un apprenant utilise un biais fort, il a tendance à induire des hypothèses moins cohérentes, mais caractérisées par peu de variance sur une base de test. Cela signifie qu'un apprenant utilisant un biais inductif faible sera très dépendant de sa base d'apprentissage. De petites variations dans les exemplaires de sa base d'apprentissage peuvent générer des hypothèses très différentes. À l'inverse, lorsqu'un apprenant utilise un biais inductif très fort, les

hypothèses d'approximation inférées vont varier très peu les unes des autres, même si elles sont induites à partir de bases d'apprentissage différentes.

La conséquence pratique de cette variance pour le choix du type d'apprenant est la suivante : lorsque la base d'apprentissage disponible est limitée et peu représentative des instances de la fonction cible, mieux vaut sélectionner un apprenant qui utilise un biais inductif fort afin de minimiser le sur-ajustement. À l'inverse, lorsque la base d'apprentissage est très représentative des instances de la fonction cible, mieux vaut sélectionner un apprenant qui utilise un biais faible de manière à minimiser les chances d'un sous-ajustement.

CHAPITRE V

MÉTHODE

5.0 Introduction

Les trois chapitres précédents ont permis de définir les éléments théoriques de la problématique de recherche définie dans l'introduction de la thèse. Le chapitre deux a permis de voir comment est modélisée l'influence sociale dans la théorie des réseaux sociaux. Quatre types de mécanisme et différentes opérationnalisations de ceux-ci ont été présentés : premièrement un mécanisme d'exposition sociale et une variante particulière appelée l'exposition sociale relative ; deuxièmement un mécanisme de contagion dont une opérationnalisation est basée sur la fréquence des interactions et une autre basée sur la proximité sociale ; troisièmement un mécanisme de déférence dont une opérationnalisation est basée sur la centralité de degré social et une sur la centralité de proximité sociale ; et quatrièmement un mécanisme de mimétisme des semblables dont une opérationnalisation est basée sur des critères endogènes de similarité (l'équivalence sociale) et une autre sur des critères exogènes de similarité (qui seront définis dans ce chapitre).

Le chapitre trois a permis de voir comment, dans le cadre de la sémantique vectorielle, sont modélisés les contenus conceptuels exprimés dans un corpus de texte. Nous avons vu que ces contenus pouvaient être modélisés sous la forme d'un espace sémantique. Des structures caractérisant cet espace, appelées des régions, forment des classes d'équivalence de mots sémantiquement similaires. Ce sont ces classes sémantiques qui sont interprétées comme étant l'expression des concepts présents dans un corpus de texte.

Le chapitre quatre a permis de voir comment, dans le cadre de l'apprentissage machine, est modélisée l'approximation d'une fonction. Nous avons vu que l'approximation d'une fonction consiste à induire à l'aide d'un apprenant automatique et d'une base d'apprentissage d'exemplaires, un modèle algorithmique de la fonction cible.

Le présent chapitre présente une méthode qui opérationnalise les éléments introduits dans les trois chapitres précédents. Le but de cette méthode est de permettre la reconstruction d'un modèle du mécanisme d'influence sociale à l'œuvre dans le processus d'évolution du réseau sociosémantique d'un SSS.

Pour ce faire, plusieurs choix d'opérationnalisations doivent être effectués. La distinction entre variables statiques et dynamiques est l'une des plus importantes (Last, Klein, & Kandel, 2001). En effet, lorsque les données empiriques forment une série temporelle de plusieurs périodes et sont caractérisées par plusieurs variables, il est généralement nécessaire de modéliser certaines variables de manière statique et d'autres de manière dynamique, ceci afin d'éviter une explosion combinatoire. Une variable statique est temporellement invariable, sa valeur est calculée une seule fois et elle est ensuite tenue constante dans la série temporelle. Au contraire, les variables dynamiques sont recalculées pour chaque période de la série temporelle.

Dans le chapitre d'introduction, un SSS a été défini comme l'union de trois réseaux, soit un réseau social $R_a^a = (A, L_a^a)$, un réseau sémantique $R_c^c = (C, L_c^c)$ et un réseau sociosémantique $R_c^a = (A, C, L_c^a)$. Dans le cadre de notre problématique de reconstruction, le réseau social et le réseau sémantique sont modélisés de manière statique, c'est-à-dire que $\underset{t \in T}{\forall}\left(R_{a_t}^a = R_{a_{t+1}}^a\right)$ et que $\underset{t \in T}{\forall}\left(R_c^c = R_{c_{t+1}}^c\right)$. Seulement les liens d'usage L_c^a dans le réseau sociosémantique et la magnitude de l'influence sociale sont modélisés de manière dynamique.[17]

[17] Techniquement, le paramètre L_c^c est omis de la problématique de reconstruction, car les liens sémantiques entre concepts ne sont pas étudiés.

La méthode utilisée suit le schéma relativement typique d'une méthode de fouille de données. Elle se déploie sur trois phases : une phase de collecte des données, une phase de prétraitement des données et une phase analytique (Aggarwal, 2015, p. 3). Chacune des phases inclut également plusieurs étapes et sous-étapes. Ce schéma est illustré par un diagramme dans la Figure 5.1.

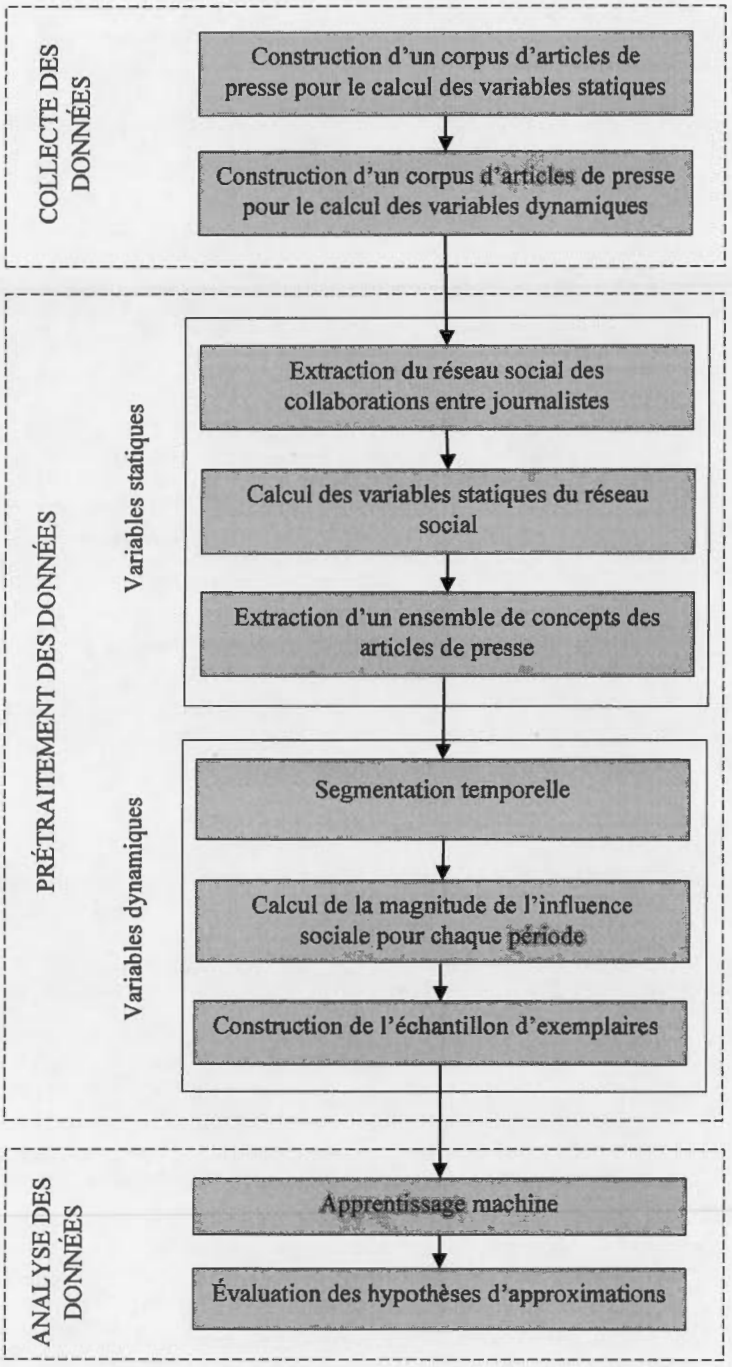

Figure 5.1: Schéma des étapes de la méthode.

La première phase est la constitution d'un corpus de texte à partir des articles de presse produits par les journalistes du journal The New York Time. Ce corpus est ensuite divisé en deux sous-corpus. Le premier sous-corpus est celui utilisé pour le calcul des variables statiques, tandis que le second est celui utilisé pour le calcul des variables dynamiques de la problématique de reconstruction.

La phase de prétraitement des données textuelles est séparée en deux grandes étapes. La première est liée aux variables statiques de la problématique et la seconde est liée aux variables dynamiques.

La première étape de prétraitement inclut trois sous-étapes. Premièrement, elle inclut une sous-étape d'extraction du réseau social à partir des articles de presse qui compose le premier sous-corpus. Deuxièmement, elle inclut une sous-étape de calcul des variables statiques liées aux propriétés du réseau social. Troisièmement, elle inclut une sous-étape d'extraction des contenus conceptuels dans les articles de presse.

La deuxième étape de la phase de prétraitement inclut également trois sous-étapes. La première est une sous-étape de segmentation temporelle du deuxième sous-corpus d'articles de presse. La deuxième sous-étape est le calcul pour chaque période de la série temporelle de la magnitude de l'influence sociale à l'œuvre dans le SSS. La dernière sous-étape est la construction d'un échantillon d'exemplaires du processus d'évolution du réseau sociosémantique du SSS étudié.

Finalement, la phase analytique inclut deux étapes. La première est l'induction à l'aide de différents apprenants automatiques et d'un échantillon d'exemplaires, d'une hypothèse d'approximation modélisant le mécanisme d'évolution du réseau sociosémantique du SSS. La deuxième étape est l'évaluation des hypothèses d'approximation à l'aide de différents indices d'adéquation entre le modèle et le processus empirique.

5.1 Collecte des données

Comme discuté dans le chapitre d'introduction de la thèse, la problématique de reconstruction est réalisée sur un SSS particulier, soit celui composé des journalistes du journal The New York Time. L'ensemble des articles de presse publiés par ces journalistes entre 1987 et 2007 est disponible sous la forme d'un corpus numérique annoté dans un format XML. Le corpus et ses annotations peuvent être manipulés à l'aide d'un parseur écrit en Java mis à la disposition de la communauté scientifique (Sandhaus, 2008).

Le corpus de texte construit pour la présente recherche inclut les articles publiés entre 2002 et 2006 inclusivement. Seulement les articles ayant une date de publication, un ou des auteurs et un corps ont été retenus. Ceci correspond à un corpus de 195 285 articles. Ces articles ont ensuite été divisés en deux sous corpus, l'un pour la modélisation des éléments statiques et l'autre pour la modélisation des éléments dynamiques.

5.1.1 Corpus lié à la modélisation des variables statiques

Le premier sous-corpus est composé des articles de presse des années 2002 à 2005. En tout, il est composé de 163 121 articles de presse. Ces traces textuelles sont utilisées pour construire le réseau social du SSS ainsi qu'identifier les concepts utilisés par les agents membres du système. C'est aussi l'analyse de ces traces textuelles qui permet d'identifier les relations de proximité sociale et d'équivalence entre agents ainsi que les différentes positions qu'ils occupent dans le réseau social du SSS.

5.1.2 Corpus lié à la modélisation des variables dynamiques

Le deuxième sous-corpus est composé des articles de presse publiés durant l'année 2006. Ce sous-corpus est composé de 32 164 articles. Ces traces textuelles sont

utilisées pour modéliser l'évolution du réseau sociosémantique du SSS et pour calculer la magnitude de l'influence sociale à chaque période de la dynamique du système.

5.2 Prétraitement des données

La phase de prétraitement des données inclut six étapes. Les trois premières sont liées aux variables statiques et les trois dernières aux variables dynamiques de la problématique de reconstruction.

Bien que l'extraction du réseau social et le calcul des variables statiques et dynamiques puissent être considérés en soi comme les résultats d'une analyse descriptive du SSS, ces résultats sont présentés dès maintenant dans le chapitre de la méthode. Deux raisons expliquent ce choix. La première raison est que, dans le cadre de la présente recherche, ces résultats ne sont pas ceux visés par la présente recherche, ils ne sont que les résultats intermédiaires des différentes opérations de prétraitement des données. Les résultats visés sont l'approximation et l'émulation de l'influence sociale par différents modèles algorithmiques. Présenter les résultats produits par chaque opération de prétraitement aide toutefois à comprendre l'opération effectuée.

5.2.1 Extraction du réseau social des collaborations entre journalistes

Comme il fut discuté dans le deuxième chapitre de cette thèse, un réseau social peut modéliser différents types d'interaction sociale entre agents. Le réseau social modélisé dans le cadre de cette recherche doctorale est celui des liens de collaboration entre journalistes, tel qu'ils peuvent être étudiés via les cosignatures d'articles de presse. Pour reprendre une typologie introduite dans le chapitre deux, ces liens de collaboration modélisent des évènements interactionnels plutôt que des états relationnels. Ces liens modélisent la fréquence des collaborations qui ont eu lieu entre journalistes du journal The New York Time.

La construction du réseau social des collaborations est effectuée à l'aide d'une méthode bibliométrique couramment utilisée dans les domaines de la scientométrie et de la bibliothéconomie (Beaver & Rosen, 1978; Katz & Martin, 1997; Laudel, 2002). Cette méthode est basée sur l'hypothèse que la cosignature d'un texte (e.g. un article, un livre, etc.) est un indicateur empirique important de la collaboration entre agents.

Cette méthode est généralement utilisée pour étudier les collaborations entre scientifiques, un territoire où la cosignature est une pratique importante (Wagner & Leydesdorff, 2003). La cosignature n'est toutefois pas une pratique unique aux collaborations scientifiques. Bien que plus rare, elle est également une pratique courante dans les collaborations journalistiques. Par conséquent, c'est par l'analyse des cosignatures d'articles de presse entre 2002 et 2005 que sont identifiés les liens de collaboration entre agents dans le SSS.

Certes, c'est une hypothèse qui a des limites méthodologiques importantes et bien connues[18], néanmoins, dans le cadre d'une problématique de reconstruction, où les données disponibles à propos de la structure interne du système sont très limitées, cette hypothèse est jugée suffisamment vraisemblable.

Le réseau social du SSS formé des journalistes du journal The New York Time est illustré dans la Figure 5.2. Chaque nœud représente une ou un journaliste et chaque lien représente une collaboration qui a eu lieu entre deux journalistes (tel qu'on peut l'observer via les cosignatures des articles de presse). La taille des nœuds est proportionnelle au nombre de collaborateurs qu'une journaliste a eus. Plus une journaliste a eu un nombre élevé de collaborateurs différents, plus le nœud est grand.

---

[18] On critique principalement deux postulats sur lesquels cette méthode s'appuie (Katz & Martin, 1997; Laudel, 2002). Premièrement, cette méthode suppose que tous les coauteurs d'une publication ont tous effectivement collaboré ensemble. Or, ce n'est pas toujours le cas, comme en témoigne le phénomène des « coauteurs d'honneur ». Deuxièmement, cette méthode suppose que tous les collaborateurs ont signé le texte publié, ce qui n'est pas toujours le cas non plus. Il y a plusieurs types de collaboration qui ne se traduisent pas toujours dans la cosignature d'un texte.

De plus, ce réseau est un multigraphe, c'est-à-dire que deux agents a_i et a_j peuvent être reliés par plusieurs liens parallèles. Le nombre de liens entre a_i et a_j est égal au nombre d'articles de presse qu'ils ont cosignés entre 2002 et 2005.

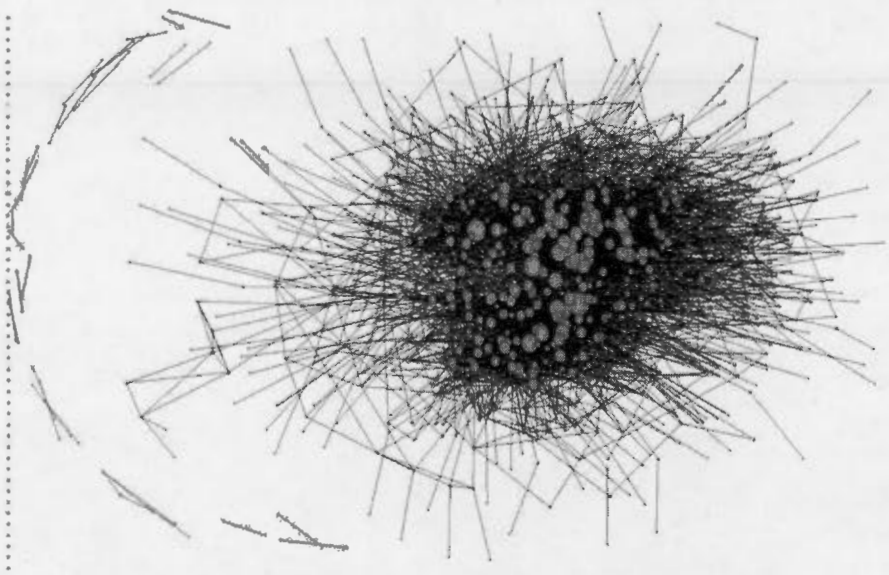

Figure 5.2: Réseau social des liens de collaborations (cosignatures) entre journalistes du The New York Time.

Les articles de presse qui composent le corpus The New York Time entre 2002 et 2005 ont été signés par 9 031 journalistes différents. Tous ces journalistes ne sont pas retenus pour la présente étude. En effet, la plupart d'entre eux n'ont signé qu'un ou deux articles durant cette période. Par conséquent, nous disposons de trop peu de données sur leurs relations avec les autres journalistes du SSS. Pour les expérimentations réalisées dans le cadre de cette recherche, seulement les 2500 journalistes les plus actifs, c'est-à-dire ayant signés le plus grand nombre d'articles, ont été sélectionnés pour la problématique de reconstruction. Ce choix d'opérationnalisation est nécessaire pour le calcul des différentes variables statistiques de la problématique.

En moyenne, chaque journaliste sélectionné a écrit 68 articles entre 2002 et 2005. Cette moyenne dissimule toutefois une grande variation. Les journalistes les moins

prolifiques ont signé seulement quatre articles, alors que le plus prolifique en a signé 1075. Le graphique de la Figure 5.3 illustre ces disparités.

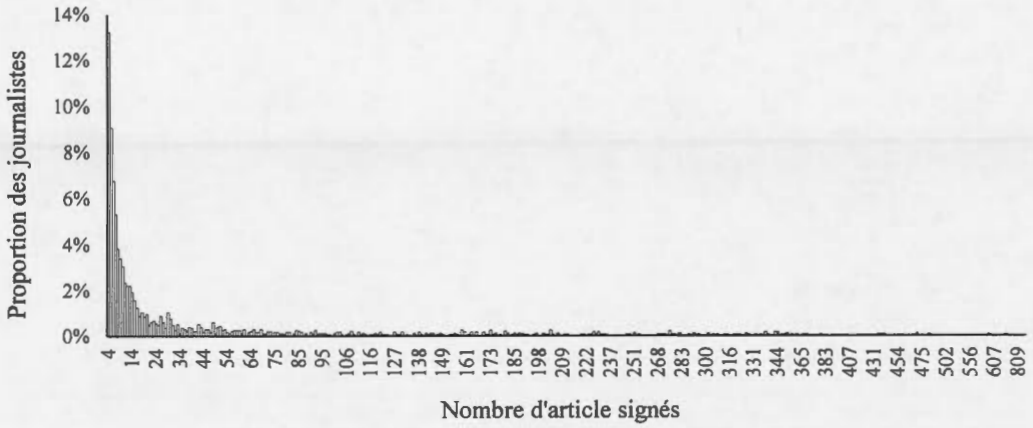

Figure 5.3: Proportion de journalistes répartie selon le nombre d'articles signés.

Dans ce graphique, nous voyons que 44,72% des journalistes n'ont signé que 10 articles et moins, 35,64% des journalistes ont signé entre 11 et 100 articles, 18,32% ont signé entre 101 et 500 articles et seulement 1,44% des journalistes ont signé plus de 500 articles.

5.2.2 Calcul des variables statiques du réseau social

En tout, six variables statiques sont calculées. Cinq de ces variables sont endogènes au réseau social et la sixième est une variable exogène. Les variables endogènes sont la fréquence des liens de collaborations, la proximité sociale entre deux journalistes, la centralité de degré, la centralité de proximité et la similitude entre journalistes basée sur l'équivalence structurale de leur position dans le réseau. La variable exogène est la similitude lexicale entre journalistes.

5.2.2.1 La fréquence des collaborations entre journalistes

Comme nous l'avons vu dans le chapitre deux, la fréquence des interactions est un facteur important de la contagion sociale. La cohésion sociale entre agents fréquemment en interaction a tendance à être plus forte. Ici, la fréquence des collaborations a été calculée pour l'ensemble des paires de journalistes du réseau social.

Les résultats de cette opération indiquent qu'il y a eu très peu de collaborations entre journalistes entre 2002 et 2005. En effet, le réseau social des collaborations est composé de seulement 4107 liens parmi 2500 nœuds. La densité du réseau est donc très faible, soit de seulement 0,001. En moyenne, chaque article a été signé par 1,05 journalistes et un peu plus de 4% des articles ont été cosignés par au moins deux personnes.

Au total, 99,8% des paires de journalistes du réseau social n'ont jamais collaboré ensemble, 0,09% ont collaboré une seule fois, 0,02% ont collaboré deux fois et 0,02% des ont collaboré ensemble entre 3 et 66 fois.

5.2.2.2 Les liens de proximité sociale entre journalistes

Un autre facteur important de la contagion sociale est la proximité sociale entre agents dans un réseau social. Cette proximité a été calculée pour l'ensemble des paires de journalistes du réseau social. Le graphique de la Figure 5.4 résume les résultats de ce calcul. Ce graphique montre la proportion des paires de journalistes dans le réseau social selon la proximité sociale qui les sépare.

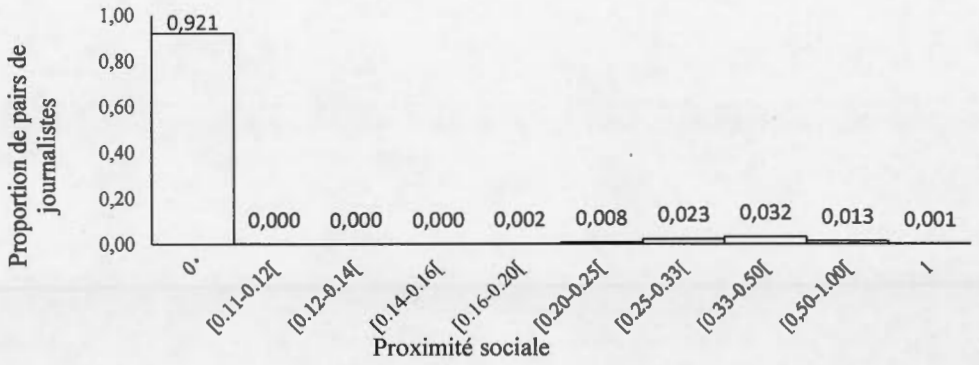

Figure 5.4: Proportion des paires de journalistes selon la proximité sociale qui les sépare dans le réseau de collaborations.

La très grande majorité des paires de journalistes dans le réseau, plus précisément 92,10% d'entre elles, ont une proximité sociale égale à zéro. Il y a 0,2% des paires de journalistes caractérisées par une proximité sociale se situant entre 0,16 et 0,20, 0,8% des paires caractérisées par une proximité sociale se situant entre 0,20 et 0,25, 2,3% des paires caractérisées par une proximité sociale se situant entre 0,25 et 0,33, 3,2% des paires caractérisées par une proximité sociale se situant entre 0,33 et 0,5, 1,3% des paires caractérisées par une proximité sociale se situant entre 0,5 et 1,00 et il y a 0,1% des paires de journalistes du réseau qui sont caractérisées par une proximité sociale de 1,00.

5.2.2.3 La centralité de degré des journalistes

Nous avons vu dans le deuxième chapitre que la position d'un agent dans le réseau, tel que modélisé par sa centralité de degré, était un facteur important de son influence sur les autres membres du réseau. Les agents centraux dans un réseau feraient l'objet d'une plus grande déférence.

La centralité de degré de chaque journaliste a été calculée. En moyenne, chaque journaliste a collaboré avec 3,29 personnes différentes. Ces collaborations sont toutefois restreintes à un petit sous-ensemble de journalistes. Le graphique de la Figure

5.5 illustre ces résultats. Ce graphique représente la proportion de journalistes répartie en fonction de la centralité de degré de leur position dans le réseau social.

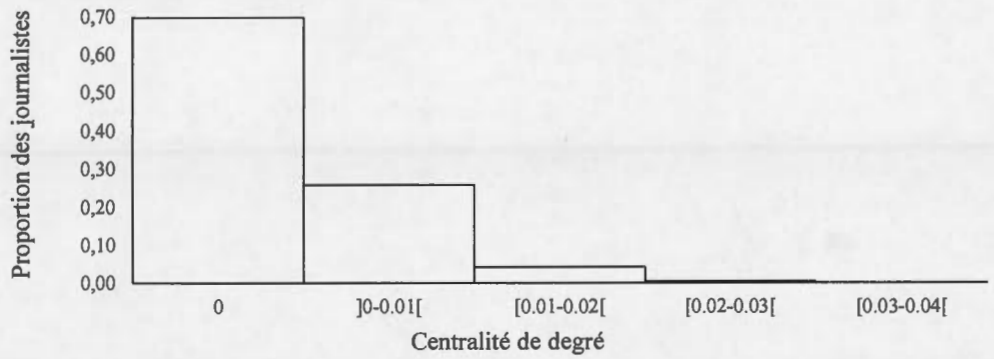

Figure 5.5: Proportion des journalistes répartie selon la centralité de leur position dans le réseau social.

On voit dans ce graphique que la majorité des journalistes, environ 70%, ont une centralité nulle, c'est-à-dire qu'ils n'ont jamais collaboré avec d'autres membres du SSS. On voit également qu'environ 26% des journalistes ont collaboré avec moins de 1% des autres membres du SSS, que 4% des journalistes ont collaboré avec de 1% à 2% des membres du SSS, que 0,44% des journalistes ont collaboré avec de 2% à 3% des membres du SSS et qu'un seul journaliste a collaboré avec plus de 3% des membres du SSS.

5.2.2.4 La centralité de proximité

Une autre manière de modéliser la centralité d'un agent dans un réseau social est la centralité de proximité. Cette centralité a également été calculée pour l'ensemble des journalistes du SSS. Le graphique de la Figure 5.6 illustre les résultats de ce calcul. Le graphique montre la proportion de journalistes répartie selon leur centralité de proximité.

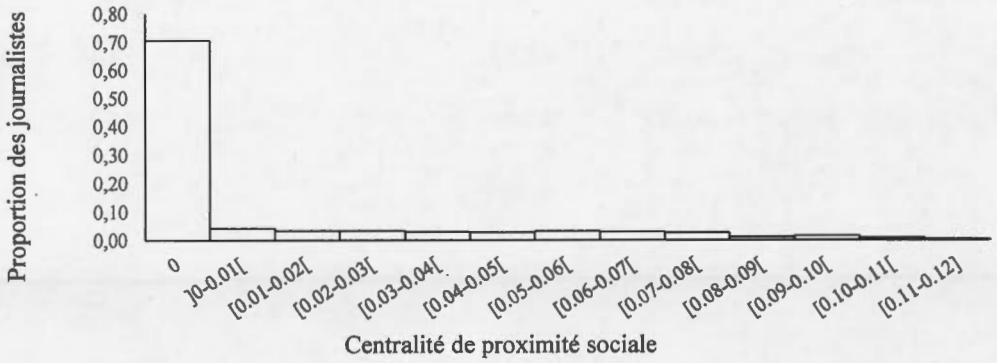

Figure 5.6: Proportion de journalistes répartie selon la centralité de proximité de leur position dans le réseau social.

Outre le fait que la plupart des journalistes (≈70%) ont une centralité de proximité nulle (car ils n'ont jamais eu de collaborateur), ce graphique montre que la proximité sociale moyenne entre journalistes est également très faible. Les journalistes les plus centraux ont en moyenne environ neuf intermédiaires qui les séparent des autres membres du SSS.

5.2.2.5 Les rapports de similarité sociale entre agents

Une autre variable endogène importante pour la problématique de recherche correspond aux rapports de similarité entre agents dans le réseau social du SSS. Deux types de similarité sont calculés.

Nous avons vu dans le chapitre deux que la similarité des position de deux agents peut être modélisée en termes d'équivalence. Cette équivalence structurale a été calculée pour l'ensemble des paires de journalistes du réseau. La Figure 5.7 illustre ces résultats. Elle montre la proportion des paires de journalistes du SSS selon l'équivalence structurale de leur position dans le réseau social.

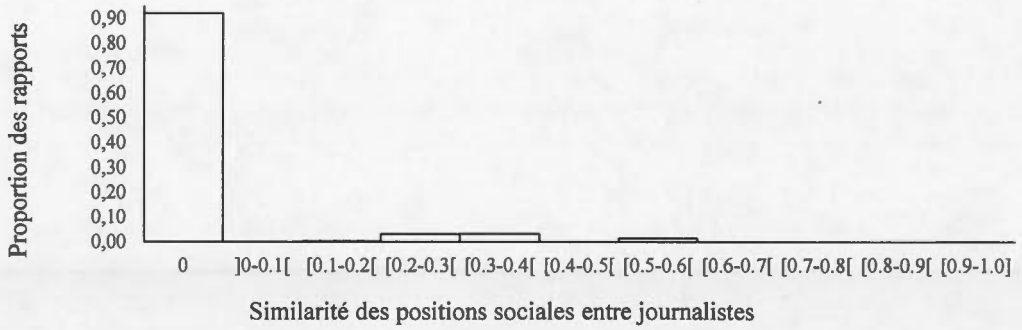

Figure 5.7: Proportion de paires de journalistes dans le SSS répartie selon l'équivalence structurale de leur position respective dans le réseau social.

Environ 92% des rapports de similarité sociale entre journalistes est égal à zéro. Le graphique de la Figure 5.7 peut donner l'impression que les positions sociales des journalistes dans le réseau sont toutes très différentes les unes des autres. Toutefois, la répartition des proportions du graphique cache des nombres très élevés. Par exemple, il y a 3% des paires de journalistes caractérisées par une équivalence structurale se situant entre 0,3 et 0,4. Ceci correspond à plus de 98 837 paires de journalistes dont l'équivalence structurelle est modérée. Il y a également 4 107 paires de journalistes qui ont un rapport de similarité sociale très fort, se situant entre 0,9 et 1,0.

5.2.2.6 Les rapports de similarité lexicale entre agents

Un autre rapport de similarité entre journalistes important pour notre problématique de reconstruction est basé sur un critère exogène de similarité. Il s'agit de la similarité entre contenus lexicaux utilisés par les journalistes. Supposons que $L(a_i) = \{u_1, \dots u_k\}$ correspond au lexique utilisé par la journaliste a_i entre 2002 et 2005, l'équivalence entre elle et une autre journaliste a_j peut être calculée de la manière suivante :

$$e_c(a_i, a_j) = \frac{|L(a_i) \cap L(a_j)|}{|L(a_i) \cup L(a_j)|} \quad \text{Eq. (5.1)}$$

Tout comme l'équivalence structurale définie dans le deuxième chapitre, la valeur de $e_c(a_i, a_j)$ varie entre zéro et un. Plus la valeur de $e_c(a_i, a_j)$ se rapproche de un et plus les lexiques utilisés par a_i et a_j sont équivalents.

Cette similarité a été calculée pour l'ensemble des paires de journalistes du SSS.

Les résultats sont illustrés dans la Figure 5.8. Ce graphique montre la proportion des paires de journalistes réparties selon la similarité entre leur lexique respectif.

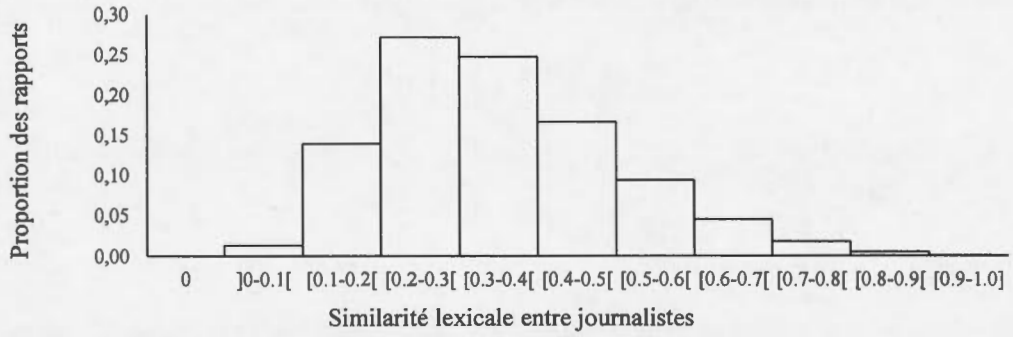

Figure 5.8: Proportion des paires de journalistes réparties selon la similarité entre leur lexique respectif.

Aucun journaliste n'a utilisé un lexique totalement différent du lexique des autres agents du SSS. La majorité des rapports de similarité se situe entre 0,2 et 0,4 et une petite proportion sont des rapports de similarité très forte, au-delà de 0,8.

5.2.3 Extraction des concepts des articles de presse

La troisième sous-étape de prétraitement liée aux variables statiques est l'extraction des concepts (aussi appelées « classes sémantiques » dans le chapitre trois) exprimés dans les articles de presse du premier sous-corpus (2002-2005). Comme il a été discuté dans le chapitre trois, plusieurs choix d'opérationnalisation sont associés à cette

opération d'extraction : la sélection des mots, la sélection des cooccurrents, le calcul de la matrice de coordonnées et l'algorithme de partitionnement de l'espace sémantique. Ces choix sont présentés dans les sections suivantes.

5.2.3.1 La sélection des mots

Ce ne sont pas tous les mots présents dans les articles de presse du premier sous-corpus qui sont utilisés pour construire l'espace sémantique de leurs contenus. Seulement les 25 000 noms les plus fréquents ont été sélectionnés. Ce choix d'opérationnalisation est motivé par trois raisons. La première raison est de nature statistique. Ce ne sont pas tous les mots qui sont pertinents pour l'analyse des contenus sémantiques exprimés dans un corpus de texte. Certains mots, notamment les articles, les pronoms et les déterminants, certains verbes et adverbes, sont si fréquents qu'ils sont considérés comme des mots dits « vides », c'est-à-dire qu'ils ne sont statistiquement spécifiques à aucun contenu particulier (Manning, Raghavan, & Schütze, 2008b, p. 27).

Une deuxième raison est de nature linguistique. Traditionnellement en linguistique, on considère que les noms sont les mots dans le langage qui expriment le mieux les contenus conceptuels (Gärdenfors, 2000, p. 60, 2014, p. 132). La troisième raison est simplement liée à un enjeu de faisabilité informatique. Au-delà de 25 000 mots, l'espace sémantique devient très difficile à manipuler sans le recours à un super ordinateur ou du calcul distribué. Or, ces technologies n'étaient pas disponibles dans le cadre de cette recherche doctorale.

De plus, c'est la forme lemmatisée des mots qui a été utilisée pour la construction de l'espace sémantique.[19]

[19] L'extraction des noms et leur lemmatisation ont été réalisées à l'aide des algorithmes de la librairie Java MorphAdorner (Burns, 2013).

5.2.3.2 La sélection des cooccurrents

Le type de contexte de cooccurrences utilisé pour construire l'espace sémantique est l'article de presse. Opérationnalisés de cette manière, deux mots sont cooccurrents s'ils sont coprésents au sein d'un même article. Pour des raisons similaires à celles mentionnées dans la section précédente, ce ne sont pas tous les cooccurrents des 25 000 mots sélectionnés précédemment qui sont utilisés pour construire l'espace sémantique. La sélection des cooccurrents est effectuée en deux étapes.

Premièrement, seulement les noms, les adjectifs et les verbes ont été conservés. Ensuite, la technique de Lund et Burgess a été utilisée (Lund & Burgess, 1996). Cette technique consiste à sélectionner les cooccurrents qui ont une valeur entropique maximale dans le corpus. Si nous dénotons par $\mathbb{C} = \{S_1 \dots S_k\}$ un corpus composé de k contextes de cooccurrence et par fc_j le nombre de contextes dans lesquels apparait le cooccurrent u_j, l'entropie de u_j se calcul de la manière suivante :

$$E(u_j) = -pr_{*j}\log_2 pr_{*j} - (1 - pr_{*j})\log_2(1 - pr_{*j}) \qquad \text{Eq. (5.2)}$$

où:

$$pr_{*j} = \frac{fc_j}{|\mathbb{C}|}$$

Sélectionner les cooccurrents ayant une valeur entropique maximale permet de ne conserver que les cooccurrents les plus informatifs dans un corpus. Ainsi, les cooccurrents très communs et très rares sont filtrés et seulement les cooccurrents les plus discriminants statistiquement sont conservés pour construire l'espace sémantique. Pour les expérimentations réalisées dans cette recherche, les 40 000 cooccurrents ayant une valeur entropique maximale ont été sélectionnés.[20]

[20] Les travaux de (Burgess, Livesay, & Lund, 1998; Lund & Burgess, 1996) suggèrent qu'au moins 10 000 cooccurrents sont nécessaires pour construire un espace sémantique.

5.2.3.3 Le calcul de la matrice de coordonnées

La matrice de coordonnées de l'espace sémantique est donc composée de 25 000 lignes et de 40 000 colonnes. Comme il a été discuté dans la section 3.2.4, différents coefficients permettent de calculer les valeurs de la matrice de coordonnées. Le coefficient utilisé dans cette recherche est la fréquence de cooccurrence. Opérationnalisée ainsi, la valeur d'une coordonnée w_{ij} correspond simplement au nombre d'articles dans le corpus dans lesquels les mots u_i et u_j sont coprésents.

5.2.3.4 Le partitionnement de l'espace sémantique avec l'algorithme des k-moyennes

L'algorithme de classification utilisé pour partitionner l'espace sémantique est l'algorithme des k-moyennes. Il s'agit d'un algorithme classique de partitionnement, probablement le plus utilisé depuis 50 ans (Jain, 2010; Steinley, 2006).

L'algorithme des k-moyennes (KM pour la suite) partitionne en k régions l'espace sémantique en minimisant la distance entre la position des mots et le centre de leur région d'appartenance. Le critère d'optimisation est la minimisation de l'inertie intra-classe. Si nous dénotons par $P = \{R_1 \dots R_k\}$ une partition en k régions, par $\vec{u}_i$ la position du mot u_i et par $\vec{c}_j$ le centre de la région j telle que définie dans la section 3.3.1 et que $d(\vec{u}_i, \vec{c}_j)$ est une métrique, la fonction objective de KM se définie ainsi :

$$\underset{P}{\operatorname{argmin}} \sum_{R_j \in P} \sum_{\vec{u}_i \in R_j} d(\vec{u}_i, \vec{c}_j) \qquad \text{Eq. (5.3)}$$

En d'autres mots, l'heuristique de KM est une procédure dite de « centres mobiles », c'est-à-dire qu'elle consiste à déplacer itérativement le centre de chacune des k régions de manière à minimiser progressivement la distance entre chaque mot et le centre le plus proche dans l'espace. La procédure s'arrête lorsqu'il y a stabilisation de la position

des centres des régions. La métrique utilisée pour les expérimentations est le cosinus défini dans la section 3.2.5

Le pseudocode de l'algorithme est présenté dans la Figure 5.9.

Soit :
$U = \{u_1 \ldots, u_n\}$, un ensemble de mots
$\vec{u}_i$, les coordonnées de la position du mot i dans l'espace vectoriel
k, le nombre de régions de la partition (1<k<n)
R_j, un ensemble de mots regroupés dans une même région (i.e. une classe)
$\vec{c}_j$, les coordonnées du centre de la région j
d, une métrique

$k\_moyennes(U, k, d)$
{
(1) Initialisation aléatoire des coordonnées des k centres;
(2) Itérer jusqu'à stabilisation des k centres
{
(3) Assigner chaque mot au centre le plus proche $R_j = \{u_i : \forall_{z=1}^{k} d(\vec{u}_i, \vec{c}_j) < d(\vec{u}_i, \vec{c}_z)\}$;
(4) Recalculer les coordonnées des centres des k régions $\vec{c}_j = \frac{1}{|R_j|}\sum_{u_i \in R_j} \vec{u}_i$;
}
(5) Retourner $\{R_1, \ldots R_k\}$;
}

Figure 5.9: Pseudocode de l'algorithme des k-moyennes.

5.2.3.5 Le paramètre k

Le principal paramètre de l'algorithme KM est le paramètre « k ». C'est le paramètre qui détermine le nombre de classes de la partition. Plusieurs heuristiques sont possibles afin d'estimer la valeur de ce paramètre (Bradley & Fayyad, 1998; Caliński & Harabasz, 1974; Milligan & Cooper, 1985; Tibshirani, Walther, & Hastie, 2001). Il n'existe toutefois aucune solution totalement satisfaisante.[21]

[21] Il existe plusieurs coefficients afin d'évaluer la qualité statistique d'une partition, notamment les coefficients Silhouette, Calinsky, Davis-Bouldin et autres. En pratique, ces coefficients sont difficilement applicables dans la présente recherche compte tenu de l'ampleur du corpus.

Comme discuté dans la section 3.3.2, identifier le bon nombre de régions dans un espace sémantique est très difficile (de surcroît impossible). Par conséquent, le paramètre k s'interprète davantage comme un paramètre de granularité de l'analyse que comme un paramètre objectif. Plus k est petit, plus l'algorithme découpe l'espace sémantique en régions vastes dans lesquelles la distance moyenne entre mots est grande. À l'inverse, plus k est grand, plus les régions sont exiguës et plus la distance moyenne entre mots est petite.

Pour les expérimentations réalisées dans le cadre de cette recherche, le paramètre k a été initialisé à 2500. C'est une valeur très élevée. Cela produit une partition ayant en moyenne seulement 10 mots différents par classe. L'avantage d'une telle partition est la granularité de l'analyse. Chaque classe sémantique regroupe un petit nombre de mots très proches les uns des autres dans l'espace sémantique.

L'inconvénient de fixer la valeur de k à 2500 est la production de nombreux cas aberrants. Ces cas aberrants sont des classes sémantiques qui ne regroupent que quelques mots atypiques généralement isolés et situés loin en périphérie de l'espace sémantique. Par exemple, parmi les 2500 classes sémantiques de la partition obtenue, il y a 1752 classes, soit 70% de la partition, qui ne regroupent qu'un seul mot et il y a 1969 classes, soit 79% de la partition, qui regroupe deux mots ou moins.

Afin de filtrer ces cas aberrants, la méthode d'extraction des contenus conceptuels utilisée dans cette thèse n'a conservé que les classes sémantiques regroupant cinq mots et plus. Par conséquent, en tout, 396 classes sémantiques ont été conservées.

Le Tableau 5.1 plus bas illustre quelques-unes des classes sémantiques obtenues. Tel que discuté dans le chapitre trois, les mots regroupés ensemble peuvent être vus comme différentes instanciations, plus ou moins représentatives, d'une même classe sémantique ou d'un même concept. Par exemple, la classe #1085 regroupe 16 différents mots tous plus ou moins liés à la classe sémantique des médicaments antidépresseurs.

La classe #1976 quant à elle regroupe 36 mots liés à la classe sémantique de la faune océanique. La classe #2482 regroupe différents mots liés à la classe sémantique de la diète. La classe #1557 regroupe différents mots liés à la classe sémantique des acteurs de la guerre en Afghanistan. La classe #1862 regroupe différents mots liés à la classe sémantique de la technologie Ethernet. La classe #2204 regroupe des mots liés à la classe sémantique de la religion. La classe #1716 regroupe différents mots liés à la physique nucléaire. La classe #484 regroupe 12 mots qui expriment tous des contenus sémantiques liés à la navigation en bateau.

Il en va ainsi pour l'ensemble des régions qui composent l'espace sémantique des articles de presse publiés entre 2002 et 2005.

Tableau 5.1: Exemples de classes sémantiques dans les données d'expérimentation.

Id classe	Mots regroupés dans la classe
#1085	antidepressant; dosage; fluoxetine; ideation; irritability; lexapro; mosholder; neurotransmitter; norepinephrine; paxil; prescriber; psychopharmacology; reducer; reuptake; serotonin; Zoloft
#1976	algae; aquaculture; bayman; bivalve; blubber; cetacean; eelgrass; fishermen; fishery; hammerhead; harpoon; hatchery; krill; leatherback; lobsterman; oceanographer; oceanography; overfishing; oysterman; phytoplankton; plankton; porpoise; salinity; sargasso; seabed; seabird; seafloor; seawater; shellfishing; shorebird; shrimper; sturgeon; submersible; toothfish; trawler; wrasse
#2482	antioxidant; bran; caloric; calorie; calory; carb; carbohydrate; carotene; dieter; dietician; dieting; dietitian; frito; glycemic; ketosis; kraft; nutritionist; omnivore; pedometer; satiety; starche; tran; whey
#1557	afghan; afghanistan; banditry; khost; mujahedeen; pashtun; peshawar; sayyaf; spinbaldak; taliban; tribesman; tribespeople; uzbek; warlord; wazir
#1862	centrino; connectivity; connexion; dsl; earthlink; gigabit; handset; kilobit; megabit; modem; palmsource; router; skype; telematics; telephony; transceiver; videophone; voip; vonage
#2204	altar; aman; apostle; atheism; atheist; auxiliary; backland; balaguer; baptism; basilica; benedict; bible; buddhist; canonization; catechism; catholic; chaplain; christian; christianity; church; churche; churchgoer; commandment; communion; confucianism; congregant; congregation; congregationalist; crucifix; crèche; denomination; divinity; ecumenism; evangelicalism; evangelism; evangelist; expiation; friar; friary; heterodox; hinduism; holiness; idolatry; immorality; intercession; interfaith; jesuit; lector; literalist; lutheran; mainline; maryknoll; megachurch; megachurche; missionary; miter; mormon; mormonism; nonbeliever; nun; obedience; orthodox; pastoral; penance; pentecostal; pew; pope; prayer; preacher; prefect; protestantism; pulpit; rector; redeemer; reformation; religion; religiosity; repentance; rev; reverend; roman; rosary; sacredness; sainthood; satanist; scripture; secularization; seminary; sermon; sinfulness; sinner; sojourner; synod; televangelist; theology; tithe; vestment; vocation; wiccan; worship; worshiper; worshipper
#1716	antimatter; cosmologist; electromagnetism; electron; neutrino; neutron; nuclei; photon; proton; quark; quasar; spectroscopy
#484	circumnavigation; conner; helmsman; mainsail; regatta; sailmaker; seamanship; skipper; spinnaker; upwind; yacht; yachtsman

5.2.4 La segmentation temporelle

La quatrième étape de la phase de prétraitement est la segmentation temporelle du deuxième sous-corpus d'articles de presse. Cette étape est liée aux variables dynamiques de la problématique de reconstruction. Son but est de séparer les articles de presse publiés durant l'année 2006 en différentes périodes de temps à partir desquelles pourra être construite la dynamique empirique du réseau sociosémnatique.

Plusieurs opérationnalisations de ce paramètre sont possibles. Le sous-corpus peut être segmenté en période d'un jour, une semaine, un mois, etc. Dans le cadre de la présente recherche, le sous-corpus des articles de l'année 2006 a été divisé en 35 périodes de 10 jours. Ce choix d'opérationnalisation est motivé par un compromis lié à la granularité de l'analyse. En effet, plus la granularité du paramètre temporel est grande, plus la probabilité d'observer un usage conceptuel est petite. Et au contraire, plus la granularité du paramètre temporel est petite, plus la probabilité d'observer le non-usage d'un concept est grande. Fixer à 10 jours le paramètre temporel permet un équilibre dans la prédiction des usages conceptuels et la prédiction des non-usages conceptuels.

À titre d'illustration, supposons que le paramètre t soit opérationnalisé en période d'une journée. Il y aurait donc 365 périodes dans le sous-corpus de 2006. L'objectif de reconstruction serait alors le suivant : si une journaliste a_i publie un article de presse demain, est-il possible de prédire, à partir de l'analyse de l'influence sociale à l'œuvre dans le SSS aujourd'hui, si elle utilisera ou non le concept c_j? Bien qu'une granularité aussi grande de l'analyse serait fascinante, méthodologiquement nous ferions face à une importante difficulté appelée « un problème déséquilibré » d'apprentissage machine (He & Garcia, 2009; Kotsiantis, Kanellopoulos, & Pintelas, 2006; Kuhn & Johnson, 2013, Chapter 16). Un tel problème se présente lorsque la probabilité d'observer un évènement est si petite, que la meilleure prédiction que l'on puisse faire consiste à prédire que l'évènement ne se produira jamais.

C'est ce qui se produit lorsque la granularité du paramètre temporelle est trop grande et c'est également ce qui se produit lorsqu'elle est trop petite. En opérationnalisant le

paramètre temporel en période, par exemple, de 100 jours, il devient impossible de reconstruire un modèle capable de prédire si un concept ne sera pas utilisé, car la meilleure prédiction possible consiste alors à prédire, trivialement, qu'un concept sera toujours utilisé par un membre du SSS dans les 100 prochains jours.

5.2.5 Le calcul de la magnitude de l'influence sociale

Pour chaque période de 10 jours, un réseau sociosémantique composé des 2500 journalistes et des 396 concepts sélectionnés précédemment est construit. En tout, 35 réseaux sociosémantiques sont donc construits formant une série $\{R^a_{c\,t}: t \in 1, \ldots, 35\}$. La construction de ces réseaux est effectuée en identifiant les concepts exprimés dans les articles de presse. Un lien d'usage $l^{a_i}_{c_j\,t}$ entre une journaliste a_i et un concept c_j est créé dans le réseau sociosémantique $R^a_{c\,t}$ si un article de presse publié durant cette période a été signé par a_i et qu'au moins l'un des mots présents dans l'article exprime le concept c_j.

La cinquième étape de la phase de prétraitement des données est le calcul des attributs modélisant la magnitude de l'influence sociale pour chacune des 35 périodes de la segmentation temporelle.

Nous avons proposé dans le chapitre deux que la magnitude de l'influence sociale pouvait être modélisée de plusieurs manières, mais que les différentes opérationnalisations dérivaient toutes de quatre principaux types de mécanismes, soit l'exposition sociale, la contagion sociale, la déférence et le mimétisme des semblables. Dans le cadre de la présente recherche, deux opérationnalisations alternatives pour chacun de ces types sont calculées. Par conséquent, en tout, huit attributs sont utilisés pour modéliser l'influence sociale.

L'influence sociale est modélisée par un vecteur composé de huit attributs $\vec{x}_{i,j,t} = (x_{1,i,j,t}, \ldots x_{8,i,j,t})$, dans lequel l'attribut $x_{z,i,j,t}$ modélise la magnitude de l'influence

sociale du concept c_j exercée sur l'agent a_i par le mécanisme z à un temps t. Chaque attribut de cette variable est défini dans le Tableau 5.2. Les définitions sont très similaires à celles introduites dans le chapitre deux, à la différence que le paramètre w_j a été remplacé par $l_{c_j\ t}^{a_k}$. Ce dernier indique si un concept c_j a été utilisé par la journaliste a_k au temps t : $l_{c_j\ t}^{a_k} = 1$ si a_k a utilisé c_j au temps t et $l_{c_j\ t}^{a_k} = 0$ dans le cas inverse.

Tableau 5.2: Définitions des huit variables modélisant la magnitude de l'influence sociale.

Attributs de l'influence sociale	Définitions de la magnitude de l'influence sociale	
Exposition sociale	$x_{1,i,j,t} = \frac{1}{n} \sum_{a_k \in A} l_{c_j}^{a_k}{}_t$	Eq. (5.4)
Exposition sociale relative	$x_{2,i,j,t} = \frac{1}{\lvert V(a_i) \rvert} \sum_{a_k \in V(a_i)} l_{c_j}^{a_k}{}_t$	Eq. (5.5)
Contagion sociale basée sur la proximité sociale	$x_{3,i,j,t} = \frac{1}{n-1} \sum_{\substack{a_k \in A \\ i \neq k}} \left(\frac{1}{d(a_i,a_k)} \cdot l_{c_j}^{a_k}{}_t \right)$	Eq. (5.6)
Contagion sociale basée sur la fréquence des collaborations	$x_{4,i,j,t} = \frac{1}{n-1} \sum_{\substack{a_k \in A \\ i \neq k}} \left(n_{i,k} \cdot l_{c_j}^{a_k}{}_t \right)$	Eq. (5.7)
Déférence basée sur la centralité sociale (de degré)	$x_{5,i,j,t} = \frac{1}{n-1} \sum_{\substack{a_k \in A \\ i \neq k}} \left(c_d(a_k) \cdot l_{c_j}^{a_k}{}_t \right)$	Eq. (5.8)
Déférence basée sur la centralité de proximité sociale	$x_{6,i,j,t} = \frac{1}{n-1} \sum_{\substack{a_k \in A \\ i \neq k}} \left(c_p(a_k) \cdot l_{c_j}^{a_k}{}_t \right)$	Eq. (5.9)
Mimétisme des semblables basé sur la similitude sociale	$x_{7,i,j,t} = \frac{1}{n-1} \sum_{\substack{a_k \in A \\ i \neq k}} \left(e_s(a_i, a_k) \cdot l_{c_j}^{a_k}{}_t \right)$	Eq. (5.10)
Mimétisme des semblables basé sur la similitude des lexiques	$x_{8,i,j,t} = \frac{1}{n-1} \sum_{\substack{a_k \in A \\ i \neq k}} \left(e_c(a_i, a_k) \cdot l_{c_j}^{a_k}{}_t \right)$	Eq. (5.11)

Le premier attribut modélise la magnitude de l'exposition sociale d'un concept. Le deuxième attribut modélise la magnitude de l'exposition sociale d'un concept relative au voisinage de chaque journaliste. Le troisième attribut modélise la magnitude de la contagion sociale basée sur la proximité sociale. Le quatrième attribut modélise la magnitude de la contagion sociale basée la fréquence des collaborations. Le cinquième attribut modélise la magnitude de la déférence basée sur la centralité de degré. Le sixième attribut modélise la magnitude de la déférence basée sur la centralité de

proximité. Le septième attribut modélise la magnitude du mimétisme des semblables basé sur un critère endogène de similarité, soit l'équivalence structurale des positions dans le réseau social. Finalement, le huitième attribut modélise la magnitude du mimétisme des semblables basé sur un critère exogène de similarité, soit la similarité entre les lexiques utilisés par les journalistes.

Les résultats de cette étape de prétraitement sont présentés sous la forme de distributions de probabilité dans le Figure 5.10. Chaque graphique de cette Figure montre la probabilité de la magnitude d'un attribut dans la segmentation temporelle. Dans l'ensemble, ce que ces distributions montrent est que la probabilité que la magnitude d'un attribut soit grande est très faible et que la probabilité que la magnitude soit très petite est très grande. Plus spécifiquement, la magnitude de l'exposition sociale varie entre 0,0 et 0,223, avec pour moyenne une valeur de 0,034. La magnitude de l'exposition sociale relative varie entre 0,0 et 1,00, avec pour moyenne une valeur de 0,023. La magnitude de la contagion basée sur la proximité sociale varie entre 0,0 et 0,063, avec pour moyenne une valeur de 0,004. La magnitude de la contagion basée sur la fréquence des collaborations varie entre 0,0 et 0,069, avec pour moyenne une valeur de 0,001. La magnitude de la déférence basée sur la centralité de degré social varie entre 0,0 et 0,591, avec pour moyenne une valeur de 0,097. La magnitude de la déférence basée sur la centralité de proximité varie entre 0,0 et 0,483, avec pour moyenne une valeur de 0,077. La magnitude du mimétisme des semblables basé sur la similitude sociale varie entre 0,0 et 0,005 avec pour moyenne une valeur de 0,000. Finalement, la magnitude du mimétisme des semblables basé sur la similitude lexicale varie entre 0,0 et 0,134, avec pour moyenne une valeur de 0,015.

Ces huit attributs ne sont pas les seules manières de modéliser l'influence sociale à l'œuvre dans le réseau social du SSS étudié. Toutefois, selon la synthèse de la littérature proposée dans le chapitre deux, elles sont parmi les opérationnalisations les plus importantes.

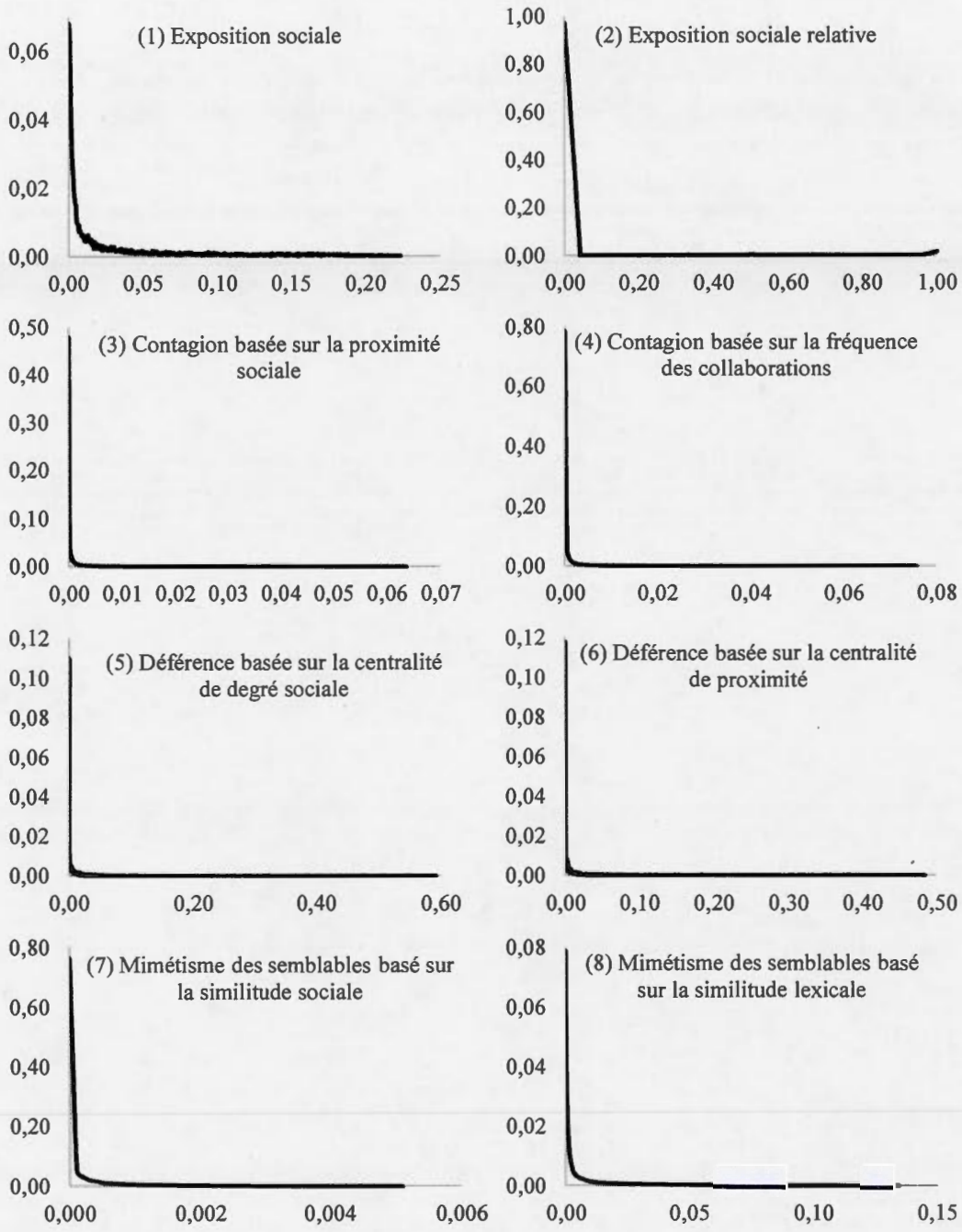

Figure 5.10: Distributions de probabilité des attributs de l'influence sociale. Dans chaque graphique, l'abscisse représente la magnitude de l'attribut et l'ordonnée représente la probabilité de cette magnitude.

5.2.6 La construction d'un échantillon d'exemplaires

La dernière étape de prétraitement est la construction d'un échantillon d'exemplaires du processus empirique d'évolution du réseau sociosémantique du SSS des journalistes du The New York Time.

Tel que discuté dans le chapitre quatre, cet échantillon doit être le plus représentatif possible du phénomène étudié. En tout, 500 000 exemplaires ont été sélectionnés de manière aléatoire (sans remise). Ce nombre représente la taille maximale de l'échantillon qui pouvait être manipulé compte tenu des ressources informatiques disponibles pour cette thèse. Dans le prochain chapitre, l'analyse des courbes d'apprentissage des apprenants automatique permettra d'évaluer si la taille de cet échantillon est suffisante pour la problématique de recherche.

L'échantillon est organisé sous la forme d'une série temporelle de données $\{(\vec{x}_{i,j,t}, y_{i,j,t+1}): t \in 1, \ldots, 35\}$ dans laquelle chaque exemplaire $(\vec{x}_{i,j,t}, y_{i,j,t+1})$ représente l'observation d'une transition d'état (ou transaction) dans le processus empirique d'évolution du réseau sociosémantique du SSS.

Le Tableau 5.3 illustre quelques-uns des exemplaires sélectionnés. Chaque ligne correspond à un exemplaire. La variable indépendante est composée de huit attributs représentant un patron particulier d'influence sociale observé dans le SSS à un temps t. La variable dépendante représente l'usage ou le non-usage du concept à un temps t+1.

Tableau 5.3: Huit exemplaires du processus d'évolution du réseau sociosémantique du SSS.

	$x_{1,i,j,t}$	$x_{2,i,j,t}$	$x_{3,i,j,t}$	$x_{4,i,j,t}$	$x_{5,i,j,t}$	$x_{6,i,j,t}$	$x_{7,i,j,t}$	$x_{8,i,j,t}$		$y_{i,j,t+1}$
#1	(0.0056,	0.0000,	0.0014,	0.0004,	0.0156,	0.0105,	0.0001,	0.0028)	→	0
#2	(0.0844,	0.0000,	0.0272,	0.0044,	0.3117,	0.2103,	0.0020,	0.0485)	→	1
#3	(0.0000,	0.0000,	0.0000,	0.0000,	0.0000,	0.0000,	0.0000,	0.0000)	→	0
#4	(0.0340,	0.0000,	0.0073,	0.0016,	0.0919,	0.0613,	0.0002,	0.0178)	→	0
#5	(0.0288,	0.0000,	0.0072,	0.0008,	0.0959,	0.0650,	0.0003,	0.0137)	→	1
#6	(0.1512,	0.0000,	0.0311,	0.0012,	0.4133,	0.3096,	0.0012,	0.0762)	→	0
#7	(0.0012,	0.0000,	0.0003,	0.0000,	0.0055,	0.0045,	0.0000,	0.0007)	→	0
#8	(0.1344,	0.0000,	0.0000,	0.0000,	0.3860,	0.2919,	0.0000,	0.0618)	→	1

Les exemplaires #2, #5 et #8 représentent des instances d'une transition d'état dans le SSS menant à l'usage du concept c_j par l'agent a_i, tandis que les exemplaires #1, #3, #4, #6 et #7 représentent des instances d'une transition d'état ne menant pas à l'usage de c_j par a_i.

Plus spécifiquement, l'exemplaire #1 correspond à l'observation dans les articles de presse publiés entre le 1er et le 10 janvier 2006 d'un patron particulier d'influence sociale suivie de l'observation dans les articles de presse publiés entre le 11 et le 21 janvier du non-usage du concept c_j par l'agent a_i. Ce patron était le suivant : la magnitude de l'exposition sociale de c_j était égale à 0,0056, la magnitude de l'exposition sociale de c_j dans le voisinage de a_i était égale à 0,0, la magnitude de la contagion basée sur la proximité sociale et sur la fréquence des collaborations étaient respectivement égales à 0,0014 et à 0,0004, la magnitude de la déférence basée sur la centralité de degré et sur la cenralité de proximité étaient respectivement égales à 0,0156 et à 0,0105 et la magnitude du mimétisme des semblables basé sur la similitude

sociale et sur la similitude des lexiques étaient respectivement égales à 0,0001 et à 0,0028.

Chaque exemplaire de l'échantillon correspond à une association entre d'une part l'observation à un temps t d'un patron particulier d'influence sociale impliquant un concept c_j et un agent a_i et d'autre part l'observation à un temps t+1 de l'usage ou non de c_j par a_j.

5.3 Analyse des données

La troisième phase de la méthode est l'analyse de l'échantillon d'exemplaires construit à l'étape précédente. La phase analytique comprend deux étapes. La première est l'induction par apprentissage machine de plusieurs hypothèse d'approximation du mécanisme d'influence sociale à l'œuvre dans le processus d'évolution du réseau sociosémantique et la deuxième étape est l'évaluation de ces hypothèses.

5.3.1 L'apprentissage automatique

Quatre apprenants automatiques sont utilisés pour l'induction des différentes hypothèses d'approximation du mécanisme d'influence sociale, soit un apprenant d'un arbre de décisions, un apprenant automatique d'une liste de règles, un méta-apprenant d'une forêt aléatoire et un apprenant bayésien naïf.

Plusieurs raisons justifient le choix de ces modèles. Premièrement, un apprenant d'un arbre de décisions et un apprenant d'une liste de règles ont été sélectionnés pour l'intelligibilité de leurs hypothèses d'approximations. Deuxièmement, un apprenant bayésien naïf a été sélectionné parce qu'il est complémentaire au modèle précédent. En effet, ces apprenants utilisent des biais inductifs très différents : alors qu'un apprenant d'arbre de décisions est basé sur un biais inductif très faible, mais générant beaucoup de variance, un apprenant bayésien naïf est basé sur un biais inductif très fort, mais générant peu de variance. Dans une problématique de reconstruction par apprentissage

machine, faire appel à des apprenants automatiques basés sur des biais inductifs différents permet de s'assurer de la robustesse des modèles induits. Finalement, un méta-apprenant d'une forêt aléatoire a été sélectionné car, bien que ses hypothèses d'approximation soient beaucoup plus complexes que celles des apprenants précédents et beaucoup plus difficiles à interpréter, il s'agit d'un modèle plus flexible que les trois précédents et reconnu pour produire moins de résidus. Un méta-apprenant de forêt aléatoire est à la fois basé sur un biais inductif très faible tout en générant peu de variance. Kuhn et Johnson suggèrent d'utiliser un apprenant de ce type comme barème pour des modèles plus simples (Kuhn & Johnson, 2013, p. 79).[22]

Les principaux concepts de chaque apprenant ainsi que leur algorithme d'induction sont introduits dans les sections suivantes. D'autres raisons justifiant le choix de ces modèles sont aussi discutées.[23]

5.3.1.1 Les arbres de décisions

Un arbre de décisions est un modèle algorithmique classique utilisé pour de l'apprentissage machine (Quinlan, 1986, 1993). Il possède plusieurs caractéristiques intéressantes. Tel que mentionné plus haut, l'une des plus importantes est l'intelligibilité de son modèle, c'est-à-dire que les hypothèses d'approximation induites par ce type d'apprenant sont très faciles à interpréter. Ce modèle s'oppose à d'autres modèles plus complexes, notamment les réseaux de neurones artificiels et les séparateurs à vaste marge qui génèrent des hypothèses d'approximation beaucoup plus opaques (Hastie et al., 2009, p. 352).

---

[22] Une autre raison justifiant ces choix est liée à la taille de l'échantillon d'exemplaires. Celui-ci étant très grand, il était nécessaire de sélectionner des apprenants extensibles. Des apprenants comme les plus-proches-voisins ou les séparateurs à vastes marges auraient nécessité des ressources informatiques qui n'étaient pas disponibles pour cette recherche.

[23] Les implémentations informatiques utilisées pour les expérimentations sont celles de WEKA 3.7 (Hall et al., 2009).

L'intelligibilité du modèle est une caractéristique essentielle lorsque l'apprentissage machine est utilisé dans le cadre d'une problématique de recherche dont l'objectif est la reconstruction d'un système empirique. En effet, des apprenants comme les réseaux de neurones artificiels ou les séparateurs à vaste marge constituent eux-mêmes des boîtes noires très difficiles à interpréter. Avec ces modèles, il peut être très difficile d'expliquer comment sont calculées les prédictions.[24] Par conséquent, utiliser ces apprenants automatiques revient, dans bien des cas, à remplacer une boîte noire (le système empirique) par une autre (l'hypothèse d'approximation).

Une autre raison justifiant le choix de ce type d'apprenant est liée aux affinités entre le modèle et la forme présumée des mécanismes d'influence sociale. En effet, un arbre de décisions modélise de manière très naturelle des fonctions à base de seuil : chaque nœud dans l'arbre représente un test sur un seuil donné. Or, comme il a été discuté dans le chapitre deux, les différents mécanismes d'influence sociale sont généralement modélisés de manière analogue. Ces mécanismes sont basés sur l'hypothèse qu'un agent adopte un comportement (e.g. une opinion, une pratique, un usage conceptuel) que si la magnitude de l'influence sociale atteint un certain seuil. De surcroît, il y aurait donc une compatibilité de forme entre la fonction cible et les arbres de décisions.

5.3.1.2 Le modèle

Ce type d'apprenant automatique cherche à approximer une fonction cible par un arbre de décisions. Un arbre de décisions est un graphe dirigé et acyclique composé de deux types de nœuds, soit des nœuds internes et des nœuds terminaux (aussi appelés les « feuilles » de l'arbre). Chaque nœud interne représente un attribut de la variable indépendante de la fonction cible et chaque nœud terminal représente une valeur possible de la variable dépendante. Chaque lien ou branche de l'arbre représente un

[24] Par ailleurs, il existe des techniques pour extraire d'un réseau de neurones artificiels entrainés des règles capables d'approximer le calcul réalisé par le réseau (Andrews, Diederich, & Tickle, 1995). Toutefois, ces techniques n'ont pas été étudiées dans cette recherche.

test conditionnel sur les valeurs d'un attribut. Dans ce modèle, la prédiction de la valeur de la variable dépendante est réalisée en parcourant de manière descendante, de la racine vers les feuilles, les différentes branches de l'arbre.

L'arbre représenté dans la Figure 5.11 est une illustration de ce à quoi pourrait ressembler un modèle du mécanisme d'influence sociale à l'œuvre dans un SSS induit par ce type d'apprenant.

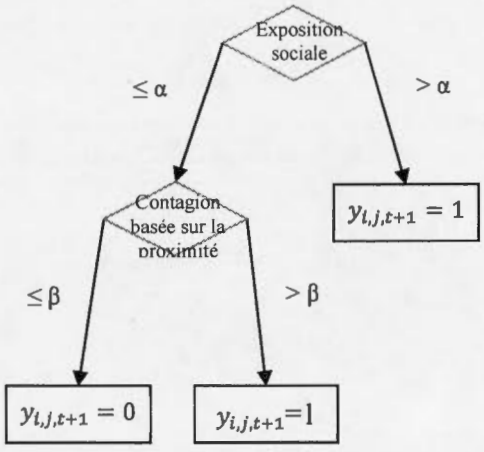

Figure 5.11: Exemple d'un arbre de décisions modélisant l'influence sociale.

Les nœuds internes sont représentés par des losanges, les nœuds terminaux par des rectangles et les branches de l'arbre par des flèches. L'arbre se lit du haut vers le bas. Dans cet exemple, ce que l'arbre prédit à propos du fonctionnement du mécanisme d'évolution des liens d'usages entre un agent a_i et un concept c_j est la chose suivante. Lorsque, à un temps t, la magnitude de l'exposition sociale d'un concept c_j est supérieure au seuil α, alors a_i utilisera c_j au temps t+1. Sinon, lorsque la magnitude de l'exposition sociale ne dépasse pas ce seuil, mais que la contagion sociale basée sur la proximité sociale dépasse le seuil β, alors c_j sera utilisé par a_i à t+1. Finalement, si aucune de ces conditions n'est satisfaite, alors le concept c_j ne sera pas utilisé par l'agent a_i à t+1.

5.3.1.3 Algorithme

L'induction automatique d'un arbre peut être réalisée via différents algorithmes (Rokach & Maimon, 2005). Dans les opérationnalisations classiques du modèle, l'induction de l'arbre est basée sur une heuristique de type « fouille gloutonne » qui consiste à séparer de manière récursive les exemplaires d'une base d'apprentissage jusqu'à ce qu'un arbre optimal soit trouvé.

L'heuristique est basée sur un coefficient statistique de discrimination dont l'objectif est d'identifier quel attribut de la variable indépendante permet de diviser de manière optimale les exemplaires de la base d'apprentissage. Ce coefficient de discrimination permet de mesurer la corrélation entre un attribut de la variable indépendante et une valeur de la variable dépendante. C'est l'élément central de l'algorithme (T. M. Mitchell, 1997, p. 55). Plusieurs coefficients peuvent jouer ce rôle, notamment le chi2 et l'index de Gini (Rokach & Maimon, 2005). Celui utilisé pour les expérimentations est le coefficient du gain d'information. Il est basé sur une mesure d'entropie définie de la manière suivante :

$$GI(x_i, \mathbb{A}) = Entropie(\mathbb{A}) - \sum_{v \in Valeur(x_i)} \frac{|\mathbb{A}_v|}{|\mathbb{A}|} Entropie(\mathbb{A}_v) \qquad \text{Eq. (5.12)}$$

Où :

$$Entropie(\mathbb{A}) = \sum_{y_j \in Y} -p_j \log_2 p_j$$

$\mathbb{A}_v$ représente le sous-ensemble d'exemplaires de la base d'apprentissage dont l'attribut x_i de la variable indépendante est égal à la valeur v. Le symbole p_j correspond à la proportion d'exemplaires de la base d'apprentissage dont la variable dépendante est égale à la valeur $y_j \in Y$.

La valeur du coefficient du gain d'information correspond à la réduction de l'entropie lorsque les exemplaires de la base d'apprentissage sont partitionnés en fonction des

valeurs de l'attribut x_i. En d'autres mots, $GI(x_i, \mathbb{A})$ permet de mesurer la quantité d'information que fournit la valeur d'un attribut x_i à propos de la valeur de la variable dépendante. Plus $GI(x_i, \mathbb{A})$ est élevé plus x_i est discriminant.

L'algorithme d'induction de l'arbre peut se résumer de la manière suivante. À la première itération, l'attribut de la variable indépendante qui possède le plus grand pouvoir de discrimination est utilisé comme racine de l'arbre. Les exemplaires sont ensuite séparés en différents sous-ensembles selon les différentes valeurs de l'attribut. À la deuxième itération, pour chaque sous-ensemble d'exemplaires est sélectionné à nouveau l'attribut le plus discriminant et les exemplaires sont à nouveau séparés en plusieurs sous-ensembles selon les différentes valeurs de l'attribut. La procédure se poursuit pour chaque nouveau sous-ensemble jusqu'à ce qu'un critère d'optimisation soit satisfait. L'arbre se construit ainsi de manière descendante et incrémentale.

Le pseudocode de l'algorithme d'induction est détaillé dans la Figure 5.12. Par ailleurs, différentes conditions d'arrêts peuvent être ajoutées à l'algorithme, notamment en fixant une profondeur maximale à l'arbre, un nombre maximal de nœuds ou un seuil entropique minimal.[25]

[25] Une étape intermédiaire dans l'algorithme n'a pas été discutée et n'est pas présentée dans le pseudocode. C'est une étape qui consiste à discrétiser les valeurs continues des attributs. La technique de discrétisation est présentée dans (Quinlan, 1996).

Soit:
$\mathbb{A}$, une base d'apprentissage
$X = \{x_1, \ldots x_m\}$, l'ensemble des attributs de la variable indépendante
$v \in x_i$, est une valeur possible de l'attribut x_i
$Y = \{1,0\}$, une classification binaire
Ψ, un arbre de décisions
$GI(x_i, \mathbb{A})$, le gain d'information de l'attribut x_i dans la base d'apprentissage $\mathbb{A}$

Induction_arbre$(\mathbb{A}, X, Y)$:
{
(1) Initialiser Ψ avec un nouveau nœud racine;
(2) Si $\underset{(\vec{x}_i, y_j) \in \mathbb{A}}{\forall} y_j = 1$, retourner Ψ avec un seul nœud étiqueté 1;
(3) Si $\underset{(\vec{x}_i, y_j) \in \mathbb{A}}{\forall} y_j = 0$, retourner Ψ avec un seul nœud étiqueté 0;
(4) Si $X = \emptyset$, retourner Ψ avec un seul nœud étiqueté selon la classe majoritaire dans $\mathbb{A}$;
(5) Sinon
{
(6) Associer au nouveau nœud racine l'attribut $\arg\max_{x_i \in X} GI(x_i, \mathbb{A})$;
(7) Pour $\underset{v \in x_i}{\forall}$
{
(8) Ajouter une branche à Ψ qui teste la condition $x_i = v$;
(9) Soit $\mathbb{A}_v$, le sous-ensemble des exemplaires pour lesquels $x_i = v$;
(10) Si $\mathbb{A}_v = \emptyset$
{
(11) Sous la nouvelle branche, ajouter un nœud terminal dont l'étiquette correspond à la classe la plus commune dans $\mathbb{A}_v$;
}
(12) Sinon
{
(13) Sous la nouvelle branche, ajouter l'arborescence suivante *Induction_arbre*$(\mathbb{A}_v, X - \{x_i\}, Y)$;
}
}
}
(14) Retourner Ψ;
}

Figure 5.12: Pseudocode d'un algorithme d'induction automatique d'un arbre de décisions d'une fonction booléenne (T. M. Mitchell, 1997, p. 56).

5.3.1.4 Les règles

L'une des limites des arbres de décisions est la redondance de leur structure. Les différentes sous-arborescences qui composent l'arbre sont parfois répétitives et un même attribut peut être testé plusieurs fois dans différents chemins. Ce problème est

lié à la difficulté d'exprimer les disjonctions sous forme d'arbre (Witten, Frank, & Hall, 2011, pp. 65–66). Une solution possible pour réduire cette redondance tout en conservant l'intelligibilité des hypothèses d'approximation consiste à convertir un arbre de décisions en une liste de règles.

5.3.1.5 Le modèle

Un arbre de décisions correspond à plusieurs sous-arborescences qui peuvent se lire comme une disjonction de plusieurs conjonctions mutuellement exclusives et exhaustives (T. M. Mitchell, 1997, p. 53). Par conséquent, chaque sous-arborescence peut s'interpréter de manière indépendante des autres. Toutefois, comme on peut le voir, même un arbre aussi simple que celui illustré dans la Figure 5.11 contient de la redondance. Les sous-arborescences de l'arbre sont répétitives et l'attribut <exposition sociale> est testé plusieurs fois.

Un apprenant automatique de règles peut réduire cette redondance en se basant sur deux postulats supplémentaires. Premièrement, l'apprenant postule un monde clos. De cette manière, il peut inférer que si un exemplaire n'appartient pas une classe, il appartient forcément à l'autre. Deuxièmement, l'apprenant postule un ordre parmi les règles, c'est-à-dire qu'il les ordonne sous la forme d'une liste. Il rend ainsi la validité d'une règle conditionnelle à la validité des règles la précédant dans la liste.

Dans bien des cas, ces deux postulats permettent de réduire significativement la redondance des arbres de décisions. C'est le cas de l'arbre, pourtant déjà très simple, présenté dans la Figure 5.11. Cet arbre peut en effet être réduit à une liste de deux règles comme celles présentées dans la Figure 5.13.

Si (exposition sociale $\leq \alpha$) $\wedge$ (contagion sociale $\leq \beta$) $\rightarrow$ concept non utilisé

Sinon $\rightarrow$ concept utilisé

Figure 5.13: Exemple d'un modèle de l'influence sociale à base de règles.

Dans cet exemple, on vérifie d'abord si un concept a une exposition sociale plus petite que α et si la magnitude de sa contagion sociale est plus petite que β, si c'est le cas, nous pouvons prédire que le concept ne sera pas utilisé et si ce n'est pas le cas nous pouvons prédire qu'il le sera. C'est un modèle fonctionnellement équivalent à l'arbre de décisions vu précédemment, mais sa structure est plus compacte.

5.3.1.6 Algorithme

Il existe plusieurs algorithmes d'induction automatique d'un modèle à base de règles. Frank et Witten (Frank & Witten, 1998) ont développé un algorithme d'une étonnante simplicité qui performe aussi bien que la plupart des apprenants alternatifs. L'algorithme se résume de la manière suivante. Dans une première itération, un premier arbre de décisions est induit à partir de l'ensemble de la base d'apprentissage. Cet arbre peut être induit avec l'algorithme présenté précédemment ou un autre. La sous-arborescence qui couvre le plus grand nombre d'exemplaires est convertie en une première règle. On retire ensuite de la base d'apprentissage les exemplaires couverts par cette règle. Dans une deuxième itération, on construit un nouvel arbre à partir de la base d'apprentissage réduite et on convertit en une deuxième règle la sous-arborescence qui couvre le plus grand nombre des exemplaires restants. On retire ensuite de la base d'apprentissage ces exemplaires. La procédure se poursuit ainsi jusqu'à ce que tous les exemplaires soient pris en compte par une règle. Le pseudocode de l'algorithme est présenté dans la Figure 5.14.

Soit:
$\mathbb{A}$, une base d'apprentissage
$X = \{x_1, \ldots x_m\}$, l'ensemble des attributs de la variable indépendante
$Y = \{1,0\}$, une classification binaire
Ψ, un arbre de décisions
$\mathbb{L}$, une liste de règles

Induction_règles$(\mathbb{A}, X, Y)$
{
(1) Initialiser une liste $\mathbb{L}$ vide de règles;
(2) Répéter jusqu'à ce que $\mathbb{A} = \emptyset$
{
(3) $\Psi =$ *Induction_arbre*$(\mathbb{A}, X, Y)$;
(4) Convertir en règle la feuille de Ψ qui recouvre le plus d'exemplaires dans $\mathbb{A}$ et ajouter la règle dans $\mathbb{L}$;
(5) Retirer de $\mathbb{A}$ les exemplaires pour lesquels s'applique la nouvelle règle;
}
(6) Retourne $\mathbb{L}$;
}

Figure 5.14: Pseudocode d'un algorithme d'induction de règles (Frank & Witten, 1998).

5.3.1.7 Les forêts aléatoires de décisions

La principale limite des arbres de décision est leur performance prédictive, jugée inférieure à celles de modèles plus complexes (Hastie et al., 2009, p. 352). Breiman affirme que : « While trees rate an A+ on interpretability, they are good, but not great, predictors. Give them, say, a B on prediction. » (Breiman, 2001b, p. 207). Un modèle d'arbre de décisions a habituellement plus de difficulté que d'autres modèles à approximer une fonction cible. Cette limite est liée au fait que ce type d'apprenant n'utilise aucun biais d'induction représentationnel, par conséquent, ses hypothèses d'approximation ont tendance à produire du sur-ajustement sur la base d'apprentissage et beaucoup de variance sur une base de test.

Une forêt aléatoire est un méta-apprenant qui exploite ces faiblesses typiques aux arbres de décisions. Paradoxalement, ce méta-apprenant figure parmi les plus performants en termes de performance prédictive, ce qui en fait l'un des plus utilisés dans les différents territoires de l'apprentissage machine (Breiman, 2001a; Hastie et

al., 2009, p. 587). Un modèle sous forme de forêt aléatoire est cependant très difficile à interpréter. Contrairement aux arbres de décisions ou aux règles, les forêts aléatoires ne peuvent pas être représentées visuellement de manière intelligible et les calculs du modèle sont difficilement tractables. Breiman donne la note de « A+ » aux forêts aléatoires pour leur performance prédictive, mais « F » pour leur intelligibilité (Breiman, 2001b, p. 208).

5.3.1.8 Le modèle

Ce type d'apprenant cherche à approximer une fonction cible par une collection d'arbres de décisions. C'est un méta-apprenant, c'est-à-dire un apprenant coordonnant plusieurs apprenants. Les forêts aléatoires utilisent plusieurs centaines d'arbres de décisions dé-corrélés les uns des autres et coordonnés pour approximer une fonction cible. Chaque arbre de la forêt peut être vu comme une hypothèse d'approximation partielle de la fonction cible et c'est la mise en commun de ces dernières qui constitue l'hypothèse d'approximation finale. Dans ce modèle, la prédiction de la valeur de la variable dépendante est effectuée en parcourant l'ensemble des arbres. C'est ensuite un vote majoritaire parmi l'ensemble des conjectures partielles qui permet d'inférer la prédiction.

Appliquée à l'objectif de reconstruction de la présemte recherche, une forêt va correspondre à plusieurs arbres de décisions chacun émulant partiellement le mécanisme d'influence sociale dans un SSS. Certains de ces arbres vont émuler la magnitude de l'exposition sociale et de la contagion, tandis que d'autres vont émuler la magnitude du mimétisme des semblables et ainsi de suite. Les prédictions du modèle sont ensuite calculées en faisant voter chaque arbre. Une forêt prédira qu'un concept c_j sera utilisé par un agent a_i si une majorité d'arbres prédit que ce sera le cas.

C'est ce design « d'ensachage » qui rend les forêts aléatoires si difficiles à interpréter. L'hypothèse d'approximation induite par ce méta-apprenant implique des interactions

complexes entre arbres qui sont très difficiles à analyser. C'est cependant cette mise en commun de plusieurs arbres dé-corrélés qui permet de réduire le sur-ajustement et la variance de l'hypothèse et qui explique pourquoi elle surpasse habituellement les capacités prédictives d'un seul arbre.

5.3.1.9 Algorithme

L'algorithme d'induction automatique d'une forêt aléatoire de décisions est essentiellement une boucle qui construit, à chaque itération, un nouvel arbre de décisions.

L'étape cruciale de la construction d'une forêt aléatoire est la dé-corrélation des arbres. Cette étape permet de réduire le sur-ajustement et la variance. La dé-corrélation est réalisée en introduisant de l'aléatoire dans la procédure de construction de la forêt. Cela est fait de deux manières. Premièrement, chaque arbre est induit à partir d'un ré-échantillonnage de la base d'apprentissage.[26] Deuxièmement, l'algorithme d'induction des arbres est légèrement modifié. Un paramètre supplémentaire est introduit : lors de l'ajout d'un nouveau nœud interne à un arbre, au lieu de sélectionner l'attribut le plus discriminant parmi l'ensemble des attributs de la variable indépendante, l'attribut est sélectionné parmi un sous-ensemble aléatoire de kmax attributs (dans l'algorithme de la Figure 5.12 ceci correspond à la ligne #6).

La taille de kmax est un paramètre à déterminer empiriquement. La pratique montre toutefois que prendre environ $\sqrt{|X|}$ des attributs est suffisant (Kuhn & Johnson, 2013, p. 201). Pour les expérimentations réalisées dans le cadre de cette recherche, le nombre d'arbres dans la forêt est fixé à 200 et kmax est fixée à deux.

---

[26] Appelé en anglais un « bootstrap ». C'est une technique d'échantillonnage aléatoire avec remise qui produit un échantillon de même taille que l'ensemble de référence (Hastie, Tibshirani, & Friedman, 2009, p. 249).

Le pseudocode de l'algorithme d'induction d'une forêt aléatoire est présenté dans la Figure 5.15.

```
Soit:
𝔸, une base d'apprentissage
X = {x_1, ... x_m}, l'ensemble des attributs de la variable indépendante
Y = {1,0}, une classification binaire
Ψ, un arbre de décisions
1< kmax<m, le nombre d'attributs sélectionnés de manière aléatoire dans X
𝔸_boot un ré-échantillonnage de la base d'apprentissage
B, le nombre d'arbres de la forêt
𝔽, une forêt d'arbres

Induction_forêt(𝔸, X, Y, B, kmax)
{
    (1) Initialiser une forêt 𝔽 vide;
    Pour b=1 à B
    {
        (2) Générer un ré-échantillonnage 𝔸_boot à partir de la 𝔸;
        (3) Ψ = Induction_arbre(𝔸_boot,X,Y,kmax) ;
        (4) Ajouter Ψ à 𝔽;
    }
    (5) Retourner 𝔽;
}
```

Figure 5.15: Pseudocode de l'algorithme d'induction automatique d'une forêt aléatoire de décisions.

5.3.1.10 Bayésien naïf

Le quatrième apprenant automatique utilisé dans le cadre de cette recherche est un apprenant bayésien naïf (Langley et al., 1992; Rish, 2001).

5.3.1.11 Le modèle

Un apprenant bayésien naïf approxime une fonction cible par un modèle probabiliste $\Pr(Y|X)$, c'est-à-dire via la probabilité conditionnelle que $Y = y_k$ sachant que $X = (v, \ldots v_m)$, où Y représente une variable dépendante et où v_i correspond à la valeur de l'attribut x_i de la variable indépendante. Dans le contexte de la problématique de la présente recherche, $\Pr(Y|X)$ correspondra à la probabilité qu'un concept c_j soit utilisé

par un agent a_i à un temps t+1 étant donnée la valeur de la magnitude de l'influence sociale $(v, \ldots v_m)$ au temps t.

La prédiction de la valeur de la variable dépendante est effectuée à l'aide du maximum a postériori. Supposons un exemplaire dont la valeur de la variable indépendante est égale au vecteur $(v_1, \ldots v_m)$, la prédiction de la valeur de la variable dépendante Y correspond alors simplement à la valeur de Y la plus probable :

	$\arg\max_{y_k \in Y} \Pr(Y = y_k \mid X = (v_1, \ldots v_m))$	Eq. (5.13)

En utilisant le théorème de Bayes, l'équation (5.13) est réécrite de la façon suivante :

	$\arg\max_{y_k \in Y} \frac{\Pr(y_k)\Pr(v_1, v_2, \ldots v_m \mid y_k)}{\Pr(v_1, v_2, \ldots v_m)}$	Eq. (5.14)

Dans cette équation, $\Pr(y_k)$ est la probabilité a priori que $Y = y_k$, $\Pr(v_1, v_2, \ldots v_m)$ est une probabilité jointe et $\Pr(v_1, v_2 \ldots v_m \mid y_k)$ est la probabilité conditionnelle de la valeur de la variable indépendante sachant la valeur de la variable dépendante.

5.3.1.12 Algorithme

Dans la pratique, lorsque l'espace des hypothèses est grand, il est impossible de calculer le maximum a postériori en force brute, le calcul de la probabilité jointe $\Pr(v_1, v_2 \ldots v_m \mid y_k)$ est trop complexe (T. M. Mitchell, 1997, p. 160). Afin de réduire cette complexité computationnelle, un apprenant bayésien naïf fait un postulat d'indépendance qui lui permet de simplifier grandement l'estimation du maximum a postériori. Ce postulat lui permet d'estimer la probabilité conditionnelle $\Pr(v_1, \ldots v_m \mid y_k)$ de la manière suivante :

	$\Pr(v_1, \dots v_m \mid y_k) \approx \Pr(v_1 \mid y_k) \Pr(v_2 \mid y_k) \dots \Pr(v_m \mid y_k) =$ $\Pr(v_1, \dots v_m \mid y_k) \approx \prod_{i=1}^{m} \Pr(v_i \mid y_k)$	Eq. (5.15)

Ce que montre l'équation (5.15), est que le postulat d'indépendance de l'apprenant bayésien naïf consiste à supposer que la probabilité conditionnelle $\Pr(v_1 \mid y_k)$ est indépendante de la probabilité $\Pr(v_2 \mid y_k)$ et que $\Pr(v_2 \mid y_k)$ est également indépendante de $\Pr(v_3 \mid y_k)$ et ainsi de suite pour toutes les valeurs $v_1, v_2 \dots v_m$.

La valeur de $\Pr(y_k)$ peut quant à elle être facilement estimée en comptant simplement le nombre d'exemplaires de la base d'apprentissage dont la variable dépendante est égale à y_k.

Le maximum a postériori basé sur ce postulat d'indépendance peut alors être estimé de la manière suivante :

	$\arg\max_{y_k \in Y} = \Pr(Y = y_k \mid X = (v_1, \dots v_m)) \approx \Pr(y_k) \prod_{i=1}^{m} \Pr(v_i \mid y_k)$	Eq. (5.16)

Bien qu'empiriquement faux, le postulat d'indépendance constitue un biais d'induction qui permet à l'apprenant de trouver une hypothèse d'approximation dans un temps polynomial. Dans la pratique, ce postulat aura peu d'impact sur l'estimation du maximum a postériori. Il aura tendance à générer du sous-ajustement sur la base d'apprentissage, mais en contrepartie peu de variances sur les prédictions de la base test.

Le pseudocode de l'algorithme d'induction bayésienne naïve est présenté dans la Figure 5.16.[27]

```
Soit:
𝔸, une base d'apprentissage
X = {x_1, ... x_m}, l'ensemble des attributs de la variable indépendante
V_i, l'ensemble des valeurs observées de l'attribut x_i dans 𝔸
Y = {1,0}, une classification binaire
𝕋, une table de probabilité conditionnelle entre X et Y

induction_bayésienne_naïve(𝔸, X, Y)
{
    Créer une table 𝕋;
    Pour ∀ (y_j ∈ Y)
    {
        Calculer Pr(y_k) selon les exemplaires dans 𝔸;
        Pour ∀ (x_i ∈ X)
        {
            Pour ∀ (v_i ∈ V_i)
            {
                Calculer Pr(v_i|y_k) selon les exemplaires
                dans 𝔸;
            }
        }
    }
    (5) Retourner 𝕋;
}
```

Figure 5.16: Pseudocode de l'algorithme d'induction bayésienne naïve.

5.3.2 L'évaluation

L'évaluation est la dernière étape de la phase analytique de la méthode de reconstruction.

La section 4.5.4 du chapitre quatre concluait qu'une bonne hypothèse d'approximation produit un résidu minimal et très probable, c'est-à-dire que la quantité d'erreurs d'un

[27] Il y a à nouveau un choix d'opérationnalisation qui est omis dans le pseudocode. Lorsque les valeurs des attributs de la variable indépendante sont de nature continue, comme c'est le cas dans notre problématique de recherche, celles-ci sont discrétisées en plusieurs intervalles.

modèle doit être la plus petite possible tout en étant la plus fiable possible. Une autre manière de le dire est qu'une hypothèse d'approximation doit être la plus cohérente possible avec sa base d'apprentissage et généralisable sur une base de test.

L'évaluation de la cohérence et de la généralisation d'une hypothèse d'approximation est effectuée à l'aide de différents indices quantitatifs mesurant l'adéquation entre un modèle et l'échantillon d'exemplaires du processus empirique d'évolution des liens d'usage dans le réseau sociosémantique du SSS.

Puisque ce processus est modélisé par une fonction de classification binaire (voir la fonction (1.1) du chapitre d'introduction), l'adéquation entre le processus empirique et une hypothèse d'approximation peut être résumée par la table de contingence illustrée dans le Tableau 5.4.

Tableau 5.4: Table de contingence entre la fonction cible décrivant le processus empirique d'évolution des usages conceptuels et une hypothèse d'approximation.

	$f(\vec{x}_{i,j,t}) = 1$	$f(\vec{x}_{i,j,t}) = 0$	Distorsion
$\check{f}(\vec{x}_{i,j,t}) = 1$	VP	FP	DP=(VP+FP)
$\check{f}(\vec{x}_{i,j,t}) = 0$	FN	VN	DN=(FN+VN)
Prévalence	PP=(VP+FN)	PN=(FP+VN)	N=(VP+FP+FN+VN)

Les colonnes $f(\vec{x}_{i,j,t}) = 1$ et $f(\vec{x}_{i,j,t}) = 0$ correspondent au processus empirique de l'évolution du réseau sociosémantique tel qu'observé dans l'échantillon d'exemplaires construit précédemment. Les lignes $\check{f}(\vec{x}_{i,j,t}) = 1$ et $\check{f}(\vec{x}_{i,j,t}) = 0$ correspondent à l'évolution du réseau sociosémantique prédite par un apprenant automatique. VP correspond à la quantité de vrais positifs entre $f(\vec{x}_{i,j,t})$ et $\check{f}(\vec{x}_{i,j,t})$, FP à la quantité de faux positifs, FN à la quantité de faux négatifs et VN à la quantité de vrais négatifs entre $f(\vec{x}_{i,j,t})$ et $\check{f}(\vec{x}_{i,j,t})$. En d'autres mots, VP est le nombre de fois dans l'échantillon d'exemplaires qu'un concept c_j a été utilisé par un agent a_i et que le modèle $\check{f}(\vec{x}_{i,j,t})$

l'a prédit correctement, FP est le nombre de fois que c_j n'a pas été utilisé par a_i mais que $\check{f}(\vec{x}_{i,j,t})$ l'a prédit incorrectement, FN est le nombre de fois que c_j a été utilisé par a_i mais prédit incorrectement par $\check{f}(\vec{x}_{i,j,t})$ et VN est le nombre de fois que c_j n'a pas été utilisé par a_i et que $\check{f}(\vec{x}_{i,j,t})$ l'a prédit correctement.

La prévalence est la propension d'un évènement. Plus spécifiquement, PP est la propension qu'un concept c_j soit utilisé par un agent a_i indépendamment de la magnitude de l'influence sociale, alors que PN est la propension que c_j ne soit pas utilisé par a_i indépendamment de la magnitude de l'influence sociale.

La distorsion est la propension d'une prédiction par un modèle. DP est la propension d'un modèle $\check{f}(\vec{x}_{i,j,t})$ à prédire qu'un concept c_j sera utilisé par un agent a_i, tandis que DN est la propension de $\check{f}(\vec{x}_{i,j,t})$ à prédire que c_j ne sera pas utilisé par a_i.

Basés sur ces concepts, les indices quantitatifs définis dans le Tableau 5.5 constituent différentes mesures qui permettent d'évaluer l'adéquation entre les observations empiriques et un modèle d'approximation (Kuhn & Johnson, 2013, pp. 254–260; David Martin Powers, 2011).

Tableau 5.5: Définitions de différents indices quantitatifs d'adéquation entre une fonction cible et une hypothèse d'approximation.

Indices	Définitions	
Rand	$\frac{VP+VN}{N}$	Eq. (5.17)
Chance	$\frac{(DP \times PP)+(DN \times PN)}{N}$	Eq. (5.18)
Kappa	$\frac{Rand - Chance}{1 - Chance}$	Eq. (5.19)
Matthews	$\frac{VP \times VN - VP \times FN}{\sqrt{DP \times PP \times PN \times DN}}$	Eq. (5.20)
Sensibilité	$\frac{VP}{PP}$	Eq. (5.21)
Spécificité	$\frac{VN}{PN}$	Eq. (5.22)
Précision positive	$\frac{VP}{DP}$	Eq. (5.23)
Précision négative	$\frac{VN}{DN}$	Eq. (5.24)

L'indice Rand (17) est une mesure globale de l'adéquation entre $f(\vec{x}_{i,j,t})$ et $\check{f}(\vec{x}_{i,j,t})$. Il correspond à la probabilité que l'hypothèse d'approximation $\check{f}(\vec{x}_{i,j,t})$ prédise correctement les observations empiriques de $f(\vec{x}_{i,j,t})$. Rand est égal à 1,0 lorsque l'adéquation est parfaite et qu'il n'y a aucun résidu, c'est-à-dire que le modèle prédit parfaitement le processus empirique. Rand est égal à 0,0 lorsqu'il y a absence d'adéquation. Rand est une mesure de base, mais parfois difficile à interpréter car elle doit toujours être comparée avec la prévalence et la distorsion. Sans cette comparaison il est impossible d'affirmer si une valeur élevée de Rand est due à une adéquation réelle entre $f(\vec{x}_{i,j,t})$ et $\check{f}(\vec{x}_{i,j,t})$ ou si elle est due à une forte prévalence dans les observations ou une forte distorsion dans le modèle. Pour cette raison, d'autres indices comme

Kappa et Matthews sont aussi utilisés pour mesurer l'adéquation globale entre $f(\vec{x}_{i,j,t})$ et $\check{f}(\vec{x}_{i,j,t})$ (David MW Powers, 2012).

Le coefficient Kappa (Eq. (5.19)) est une mesure d'adéquation qui permet de contrôler les effets combinés de la prévalence et de la distorsion. Kappa correspond à l'adéquation entre $f(\vec{x}_{i,j,t})$ et $\check{f}(\vec{x}_{i,j,t})$ moins l'adéquation attendue par la chance (dans ce contexte la chance signifie la probabilité que l'adéquation soit due simplement à la prévalence et la distorsion). Kappa est égal à 1,0 lorsque l'adéquation est parfaite et à 0,0 (ou moins) lorsque l'adéquation n'est pas supérieure à celle générée par la chance.

Le coefficient Matthews (Eq. (5.20)) est un autre indice qui permet de mesurer l'adéquation entre $f(\vec{x}_{i,j,t})$ et $\check{f}(\vec{x}_{i,j,t})$ en contrôlant les effets de la prévalence et de la distorsion. Matthews est une version nominale du coefficient de corrélation de Pearson. Il s'interprète de manière analogue. Lorsque le coefficient est égal à 1,0 cela signifie une corrélation parfaite entre $f(\vec{x}_{i,j,t})$ et $\check{f}(\vec{x}_{i,j,t})$, lorsqu'il est égal -1,0 cela signifie une corrélation inversée et une valeur de 0,0 signifie qu'il y a absence de corrélation.

La Sensibilité (Eq. (5.21)) et la Spécificité (Eq. (5.22)) sont deux mesures complémentaires qui permettent de décomposer l'adéquation entre $f(\vec{x}_{i,j,t})$ et $\check{f}(\vec{x}_{i,j,t})$. La Sensibilité est le taux de vrai positif par rapport à la prévalence positive. Interprétée dans le cadre de notre problématique de recherche, elle correspond à la probabilité que l'usage observé du concept c_j par l'agent a_i dans la dynamique du SSS soit prédit par le modèle $\check{f}(\vec{x}_{i,j,t})$. La Spécificité est le taux de vrai négatif par rapport à la prévalence négative. Elle correspond à la probabilité que le non-usage du concept c_j par l'agent a_i dans la dynamique du SSS soit prédit par le modèle $\check{f}(\vec{x}_{i,j,t})$.[28]

[28] Dans des domaines comme la fouille de texte ou la recherche d'information, ces deux mesures sont appelées des indices de « rappel ».

La Précision positive (Eq. (5.23)) et la Précision négative (Eq. (5.24)) permettent de mesurer la qualité des prédictions d'une hypothèse d'approximation. C'est aussi une manière d'apprécier la confiance qu'une chercheuse peut avoir en les prédictions générées par un modèle. La Précision positive est le taux de vrai positif par rapport à la distorsion positive d'un modèle. Elle correspond à la probabilité que la prédiction par le modèle $\check{f}(\vec{x}_{i,j,t})$ qu'un concept c_j sera utilisé par un agent a_i soit effectivement observée dans la dynamique du réseau sociosémantique du SSS. La Précision négative est le taux de vrai négatif par rapport à la distorsion négative d'un modèle. Elle correspond à la probabilité que la prédiction par $\check{f}(\vec{x}_{i,j,t})$ qu'un concept c_j ne sera pas utilisé par un agent a_i soit effectivement observé dans la dynamique empirique du système.

CHAPITRE VI

EXPÉRIMENTATIONS

6.0 Introduction

Ce chapitre présente et analyse les résultats des expérimentations d'apprentissage machine réalisées dans le cadre de cette recherche.

Ces expérimentations ont pour objectif de reconstruire à l'aide de différents apprenants automatiques plusieurs modèles du mécanisme d'influence sociale à l'œuvre dans le processus d'évolution du réseau sociosémantique des journalistes du The New York Time. Les quatre apprenants automatiques introduits dans le chapitre précédent sont utilisés, soit un apprenant d'arbre de décisions (qui sera dénoté par Γ_{AD}), un apprenant d'une liste de règles (Γ_{LR}), un apprenant d'une forêt aléatoire (Γ_{FA}) et un apprenant bayésien naïf (Γ_{BN}).

Le chapitre est organisé de la manière suivante. Il est divisé en deux grandes sections. Dans la première sont présentés les modèles de l'influence sociale induits par les différents apprenants automatiques et dans la deuxième section différentes analyses de ces modèles sont effectuées afin d'en évaluer la vraisemblance et les limites.

6.1 Résultats des expérimentations

L'échantillon des exemplaires a été divisé de manière aléatoire en une base d'apprentissage $\mathbb{A}$ contenant 330 000 exemplaires, soit 66% de l'échantillon total, et en une base de test $\mathbb{B}$ contenant le reste, soit 170 000 exemplaires.

Les différents modèles du mécanisme d'influence sociale sont induits à partir de la base d'apprentissage $\mathbb{A}$. Dans cette section, chaque modèle est présenté séparément à l'exception de la forêt aléatoire. Comme il a été discuté dans le chapitre précédent, la

forêt aléatoire est un modèle très complexe et difficile à interpréter. Les forêts aléatoires induites dans cette recherche sont toutes composées de 200 arbres, chacun ayant en moyenne 66 053 nœuds internes. Il est par conséquent impossible de les présenter. Les performances prédictives de ce modèle sont néanmoins présentées et comparées avec celles des autres modèles dans la section suivante.

6.1.1 La modélisation du mécanisme d'influence sociale par un arbre de décisions

Le premier modèle du mécanisme d'influence sociale à l'œuvre dans le processus d'évolution du réseau sociosémantique est un arbre de décisions induit par l'apprenant Γ_{AD}. Le modèle est illustré dans la Figure 6.1. Cet arbre est composé de 17 nœuds internes illustrés dans l'arbre par des losanges. Chacun des nœuds internes représente l'un des huit attributs de la magnitude de l'influence sociale à l'œuvre dans le SSS étudié. Chacun des nœuds internes est aussi associé à deux flèches modélisant les tests conditionnels effectués sur les valeurs des attributs. Les valeurs associées à ces flèches sont des seuils d'influence sociale. Ce sont les seuils induits par l'apprenant Γ_{AD} qui permettent d'optimiser les performances prédictives du modèle.

L'arbre contient 18 sous-arborescences. Les sous-arborescences modélisent des patrons récurrents d'influence sociale identifiés par l'apprenant Γ_{AD}. Selon Γ_{AD}, ce sont ces patrons qui déterminent le processus d'évolution du réseau sociosémantique. Chacune des sous-arborescences est associée à un nœud terminal (une feuille de l'arbre). Ils sont illustrés dans la Figure 6.1 par des rectangles. Chacun de ces nœuds contient trois valeurs. La première est une valeur binaire. Elle représente la valeur prédite par l'apprenant Γ_{AD} de la variable dépendante des exemplaires de la base d'apprentissage. Lorsque cette valeur est égale à un, cela signifie que la sous-arborescence modélise un patron récurrent d'influence sociale associé à l'usage futur d'un concept. À l'inverse, lorsque cette valeur est égale à zéro, cela signifie que la sous-arborescence modélise un patron d'influence sociale associé à la non-utilisation future d'un concept.

La deuxième et la troisième valeur permettent d'analyser l'importance des sous-arborescences de l'arbre. La deuxième valeur représente le nombre d'exemplaires dans $\mathbb{A}$ couverts par la sous-arborescence. Cette quantité indique la fréquence empirique d'un patron d'influence sociale. Plus cette quantité est grande, plus ce patron est important. Le rapport entre cette quantité et le nombre total d'exemplaires dans $\mathbb{A}$ est appelé le « support » de la sous-arborescence et s'exprime par un pourcentage de couverture.

La troisième valeur représente le nombre d'exemplaires couverts par la sous-arborescence pour lesquels Γ_{AD} a prédit incorrectement la valeur de la variable dépendante. En d'autres mots, c'est le résidu de la sous-arborescence. Le rapport entre cette quantité et la précédente est appelé le « taux d'erreurs » produit par la sous-arborescence. [29]

[29] Dans la littérature technique en fouille de données, on parle parfois de « confiance » et non de « taux d'erreur ». La confiance est l'inverse du taux d'erreur, c'est-à-dire que : confiance=(1-taux d'erreur).

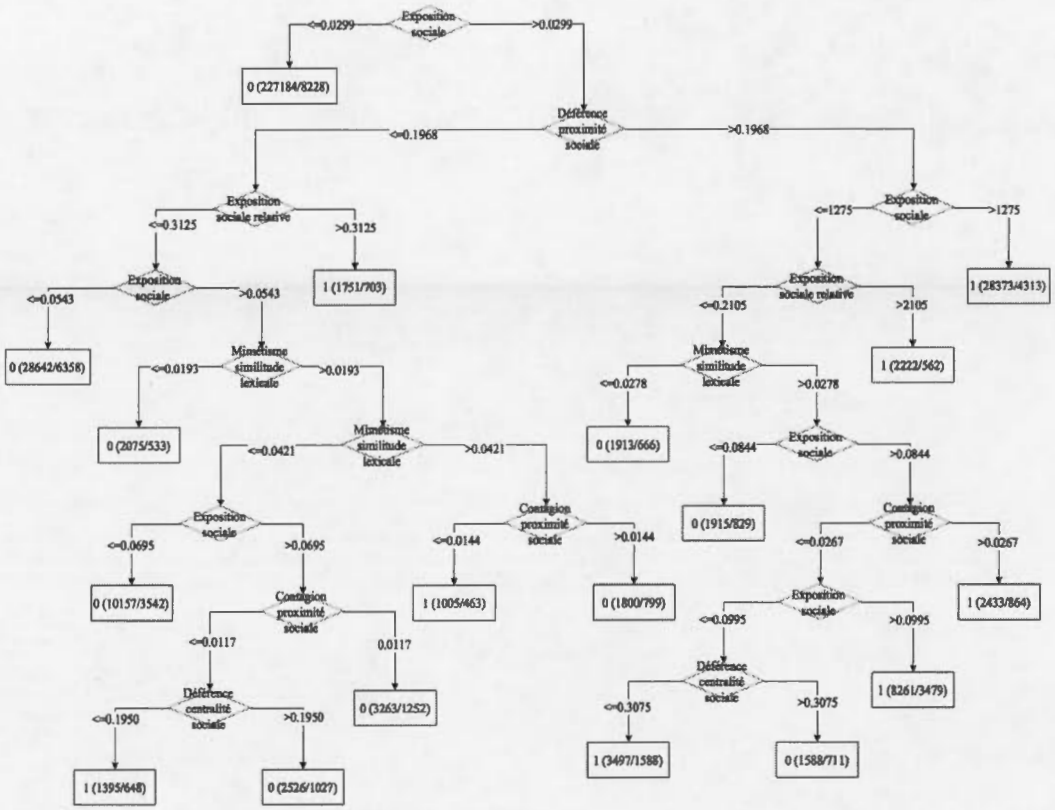

Figure 6.1: Reconstruction du mécanisme d'influence sociale par un arbre de décisions.

L'arbre de la Figure 6.1 est relativement complexe, certaines de ses sous-arborescences ont jusqu'à neuf niveaux. Cependant, ces sous-arborescences n'ont pas toutes la même importance. Certaines ont un support beaucoup plus grand que d'autres et certaines génèrent un taux d'erreur qui est aussi beaucoup plus important que d'autres. Trois groupes de sous-arborescences semblent caractériser l'arbre. Le premier groupe est le plus important. Il ne contient qu'une seule sous-arborescence, soit la suivante :

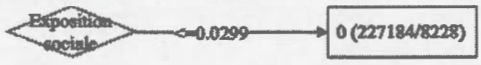

Figure 6.2: Sous-arborescence principale.

Cette sous-arborescence modélise un patron d'influence sociale associé à la non-utilisation d'un concept. Elle spécifie que, lorsqu'un concept c_j est l'objet d'une

exposition sociale inférieure ou égale au seuil 0,0299 à un temps t, alors c_j ne sera pas utilisé au temps t+1 par l'agent a_i. Bien que cette sous-arborescence soit la plus simple de l'arbre, elle est néanmoins la plus importante. Elle couvre 227 184 exemplaires dans $\mathbb{A}$, ce qui représente un support de 69%. De plus, seulement 8 228 exemplaires dans $\mathbb{A}$ sont incorrectement prédits par cette sous-arborescence, ce qui représente un taux d'erreur de seulement 4%.

Le deuxième groupe contient les deux sous-arborescences suivantes:

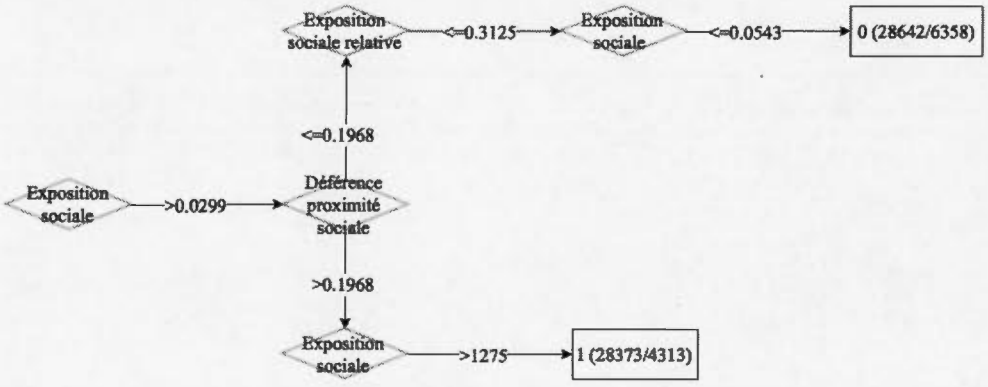

Figure 6.3: Deux sous-arborescences d'importance intermédiaire.

La sous-arborescence du haut modélise un patron complexe d'influence sociale associé à la non-utilisation d'un concept. Elle spécifie que lorsqu'un concept c_j est l'objet d'une exposition sociale plus grande que le seuil 0,0299 à un temps t, mais d'une déférence basée sur la proximité sociale plus petite que le seuil 0,1968, d'une exposition sociale relative plus petite que 0,3125 et d'une exposition sociale plus petite que 0,0543, alors ce concept c_j ne sera pas utilisé par un agent a_i à $t+1$. Cette sous-arborescence couvre 28 642 exemplaires. Comparativement à la sous-arborescence précédente, elle est donc d'une importance beaucoup moindre et son support est de seulement 9%. Elle génère également beaucoup plus de résidus que la précédente. Son taux d'erreur est de 22%.

La deuxième sous-arborescence de ce groupe modélise un patron complexe d'influence sociale associé à l'usage futur d'un concept. Elle spécifie que lorsqu'un concept c_j est l'objet à un temps t d'une exposition sociale plus grande que le seuil 0,0299, que la déférence basée sur la proximité sociale est plus grande que le seuil 0,1968 et que l'exposition sociale est également plus grande que le seuil 0,1275, alors ce concept c_j sera utilisé à t+1 par l'agent a_i. Cette sous-arborescence est d'une importance similaire à la précédente. Elle couvre 28 373 exemplaires dans $\mathbb{A}$, ce qui correspond aussi à un support de 9%. Elle est toutefois caractérisée par un résidu beaucoup moindre, son taux d'erreur est de seulement 15%.

Ensemble, les 15 sous-arborescences qui forment le troisième groupe ne couvrent en tout que 45 801 exemplaires de la base d'apprentissage, soit un support total de seulement 13%. Ce groupe de sous-arborescences génère beaucoup plus de résidu que les sous-arborescences précédentes. Ensemble, leur taux d'erreur est de 39%. Ces sous-arborescences modélisent des patrons d'influence sociale beaucoup plus rares que les précédents et beaucoup plus contingents ou incertains. Ce sont également des patrons très complexes, comme en témoigne celui modélisé par la sous-arborescence suivante :

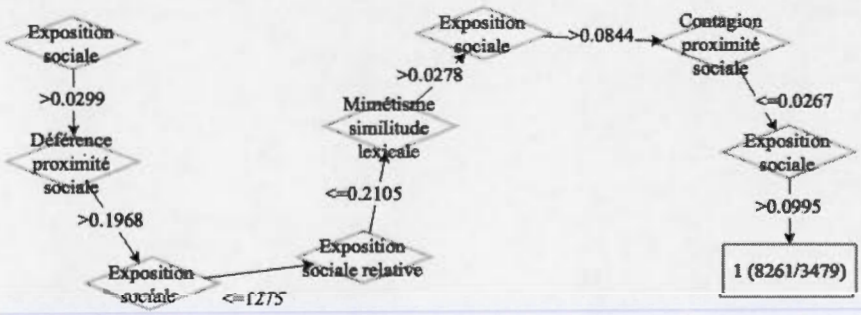

Figure 6.4: Une sous-arborescence complexe, mais mineure et incertaine.

Cette sous-arborescence est la plus importante du troisième groupe. Elle modélise un patron d'influence sociale associé à l'usage futur d'un concept. Toutefois, son support est de seulement 2,5% et son taux d'erreur est très élevé, soit de 42%.

Certains patrons d'influence sociale sont donc beaucoup plus récurrents que d'autres dans le processus d'évolution du réseau sociosémantique du SSS et ils déterminent de manière beaucoup plus certaine l'usage et le non-usage des concepts. C'est le cas des sous-arborescences des deux premiers groupes. Ensemble, elles ont un support de 87% et leur taux d'erreur est de seulement 7%. Autrement dit, ces arborescences permettent, avec un taux d'erreur de seulement 7%, de prédire 87% des transitions d'état du réseau sociosémantique étudié.

6.1.2 La modélisation du mécanisme d'influence sociale par une liste de règles

Le deuxième modèle du mécanisme d'influence sociale est une liste de règles induites par l'apprenant Γ_{LR}. Le modèle est illustré dans le Tableau 6.1.

La première colonne du Tableau 6.1 montre l'ordre de déclenchement des règles, la deuxième et la troisième colonne montrent les deux composantes d'une règle, soit l'antécédent et le conséquent. Le conséquent est l'équivalent du nœud terminal d'un arbre de décisions, c'est la valeur de la variable dépendante prédite par l'apprenant. Une valeur égale à un signifie que l'antécédent de la règle modélise un patron récurrent d'influence sociale associé à l'usage futur d'un concept. À l'inverse, lorsque cette valeur est égale à zéro, cela signifie que l'antécédent de la règle modélise un patron récurrent d'influence sociale associé à la non-utilisation d'un concept.

De plus, comme pour les sous-arborescences vues précédemment, chaque règle est associée à un support et un taux d'erreur.

Tableau 6.1: Reconstruction du mécanisme d'influence sociale par une liste de règles.

Ordre	Antécédent	Conséquent	Support	Taux d'erreur
#1	(Exposition_sociale <= 0,0299) ∧ (Mimetisme_similitude_lexicale) <= 0,0599)	0	69%	4%
#2	(Exposition_sociale > 0,112)	1	11%	19%
#3	(Mimetisme_similitude_lexicale <= 0,0448) ∧ (Exposition_sociale <= 0,0543) ∧ (Exposition_sociale_relative <= 0,1499)	0	9%	22%
#4	(Exposition_sociale_relative <= 0,3125) ∧ (Mimetisme_similitude_lexicale <= 0,0448) ∧ (Mimetisme_similitude_lexicale > 0,0187) ∧ (Exposition_sociale <= 0,0708)	0	4%	35%
#5	(Mimetisme_similitude_lexicale > 0,0222) ∧ (Exposition_sociale_relative <= 0,2222) ∧ (Mimetisme_similitude_lexicale <= 0,0437) ∧ (Exposition_sociale <= 0,0923)	0	3%	43%
#6	(Mimetisme_similitude_lexicale > 0,0222) ∧ (Exposition_sociale_relative <= 0,4615) ∧ (Mimetisme_similitude_lexicale <= 0,0589) ∧ (Contagion_proximite_sociale <= 0,0252) ∧ (Deference_proximite_sociale > 0,2062) ∧ (Deference_centralite_sociale <= 0,3066)	1	1%	44%
#7	(Mimetisme_equivalence_lexicale <= 0,0222)	0	1%	29%
#8	(Exposition_sociale_relative <= 0,4615) ∧ (Mimetisme_similitude_lexicale <= 0,0587) ∧ (Contagion_proximite_sociale <= 0,0252) ∧ (Exposition_sociale > 0,0851) ∧ (Exposition_sociale <= 0,1011) ∧ (Mimetisme_similitude_lexicale <= 0,0492)	0	1%	44%
#9	()	1	3%	41%

Ce modèle est similaire au précédent, ce qui ne devrait pas surprendre compte tenu de la similitude des algorithmes d'inductions. Ce modèle à base de règles est toutefois beaucoup plus compact que le précédent. Il n'est composé que de neuf règles. Il contient moins de redondance. Non seulement ne contient-il que neuf règles, mais en

plus celles-ci sont beaucoup moins complexes que les sous-arborescences qui composent l'arbre de la Figure 6.1.

Par ailleurs, comme c'était le cas pour les sous-arborescences de l'arbre de décisions, les règles qui composent cette liste n'ont pas toutes la même importance. Certaines règles ont un support beaucoup plus élevé que d'autres et un taux d'erreur beaucoup plus faible. Il est, à nouveau, possible de distinguer trois groupes de règles.

Le premier groupe contient une seule règle. Il s'agit de la règle #1 du Tableau 6.1. Cette règle modélise le principal patron récurrent d'influence sociale associé à la non-utilisation d'un concept. Cette règle spécifie que lorsqu'à un temps t la magnitude de l'exposition sociale d'un concept c_j est inférieure ou égale au seuil 0,0299 et que la magnitude du mimétisme des semblables basé sur la similitude lexicale est inférieure ou égale au seuil 0,0599, alors c_j ne sera pas utilisé au temps t+1 par un agent. C'est la règle la plus importante de la liste. Son support est de 69% et son taux d'erreur de seulement 4%.

Le deuxième groupe contient les règles #2 et #3 du Tableau 6.1. La règle #2 modélise le principal patron d'influence sociale associé à l'usage futur d'un concept. Cette règle spécifie que lorsqu'à un temps t la magnitude de l'exposition sociale d'un concept est plus grande que le seuil 0,112, alors ce concept sera utilisé à un temps t+1. Le support de cette règle est de 9% et son taux d'erreur de 19%.

La règle #3 modélise un patron récurrent d'influence sociale associé à la non-utilisation d'un concept. Cette règle spécifie que lorsqu'à un temps t la magnitude du mimétisme des semblables basé sur la similitude lexicale est inférieure ou égale au seuil 0,0448, que la magnitude de l'exposition sociale est inférieure ou égale au seuil 0,0543 et que la magnitude de l'exposition sociale relative est inférieure ou égale au seuil 0,1499, alors ce concept ne sera pas utilisé au temps t+1. Le support de cette règle est 9% et son taux d'erreur est de 22%.

Dans le troisième groupe, on retrouve les règles #4 à #9. Ces règles modélisent des patrons d'influence sociale beaucoup plus rares et beaucoup plus contingents et incertains dans le processus d'évolution du réseau sociosémantique. Le support de ces règles varie entre 1% et 4% et le taux d'erreur entre 29% et 44%. Au total, le support de ces six règles est de seulement 12%, mais leur taux d'erreur de 39%.

La modélisation du mécanisme d'influence sociale par un arbre de décisions et une liste de règles est similaire à bien des égards. Ces deux apprenants automatiques ont identifié en commun l'exposition sociale comme le principal facteur de l'influence sociale. Dans les deux modèles, cet attribut fait partie de tous les principaux patrons récurrents d'influence sociale. Autre caractéristique commune aux deux modèles, la contagion basée sur la fréquence des collaborations est absente tout comme le mimétisme basé sur l'équivalence des positions sociales. Ces attributs ne seraient pas selon les deux apprenants suffisamment associés à l'évolution des usages et des non-usages conceptuels pour les intégrer dans les modèles. Enfin, bien que présente, la déférence basée sur la centralité de degré sociale joue également un rôle mineur dans les deux modèles.

6.1.3 La modélisation de l'influence sociale par un modèle bayésien naïf

Le modèle induit par l'apprenant Γ_{BN} est un modèle probabiliste du mécanisme d'influence sociale. Ce modèle est présenté dans le Tableau 6.2. La présentation de ce modèle est plus complexe. Les paramètres du modèle sont plus nombreux. Ils sont présentés en six colonnes.

Les valeurs des colonnes Pr(y=0|x) et Pr(y=1|x) peuvent être interprétées de manière similaire au taux d'erreur des règles ou des sous-arborescences des modèles précédents. Les valeurs des colonnes Pr(x|y=0) et Pr(x|y=1) peuvent être interprétées de manière similaire à leurs supports.

Dans la première colonne sont présentées les différentes valeurs de la magnitude de chaque attribut de l'influence sociale. Ces valeurs ont été discrétisées en quatre intervalles.

La deuxième colonne, qui a pour entête Pr(x), représente l'approximation par l'apprenant Γ_{BN} de la probabilité qu'un concept soit caractérisé par une magnitude donnée d'influence sociale. Par exemple, selon Γ_{BN}, la probabilité qu'un concept soit caractérisé par une exposition sociale se situant dans l'intervalle [0,0-0,0558] est de 0,78.

Les valeurs de cette colonne sont des approximations de la distribution de probabilité des attributs de l'influence sociale calculée dans la section 5.2.5 du chapitre cinq. La forme générale de cette distribution semble être préservée. En effet, on peut voir qu'une magnitude très basse de l'influence sociale est beaucoup plus probable qu'une magnitude élevée et ceci s'observe pour tous les attributs. Pour certains attributs, notamment la contagion basée sur la fréquence des collaborations, c'est 99% des valeurs de l'attribut qui se situent dans le premier intervalle.

La troisième colonne, qui a pour entête Pr(x|y=0), représente l'approximation par l'apprenant Γ_{BN} de la probabilité conditionnelle qu'un concept soit caractérisé par une magnitude donnée d'influence sociale, sachant que ce concept ne sera pas utilisé à un temps t+1. Par exemple, selon Γ_{BN}, la probabilité qu'un concept soit caractérisé à un temps t par une magnitude d'exposition sociale se situant dans l'intervalle [0,0-0,0558], étant donné qu'il ne sera pas utilisé à t+1, est de 0,90. Lorsque la magnitude d'un attribut est associée à une probabilité Pr(x|y=0) élevée, cela signifie qu'une grande proportion de non-usages conceptuels sont caractérisés par cette magnitude.

La colonne ayant pour entête Pr(x|y=1) représente l'approximation par Γ_{BN} de la probabilité conditionnelle qu'un concept soit caractérisé à un temps t par une magnitude donnée d'influence sociale, sachant que ce concept sera utilisé par un agent

à un temps t+1. Par exemple, selon Γ_{BN}, la probabilité qu'un concept soit caractérisé par une magnitude d'exposition sociale se situant dans l'intervalle [0,0-0,0558], sachant qu'un agent l'utilisera à t+1 est de 0,25. Lorsque la magnitude d'un attribut est associée à une probabilité Pr(x|y=1) élevée, cela signifie qu'une grande proportion d'usages conceptuels sont caractérisés par cette magnitude.

La colonne ayant pour entête Pr(y=0|x) représente l'approximation par Γ_{BN} de la probabilité conditionnelle qu'un concept ne soit pas utilisé par un agent à un temps t+1, étant donné la magnitude d'un attribut d'influence sociale au temps t. Par exemple, selon Γ_{BN}, la probabilité qu'un concept ne soit pas utilisé par un agent à un temps t+1 sachant que son exposition sociale au temps t se situait dans l'intervalle [0,0-0,0558] est de 0,94. Lorsque la magnitude d'un attribut est associée à une probabilité Pr(y=0|x) élevée, cela signifie que cette magnitude conditionne fortement le non-usage d'un concept.

Finalement, la colonne ayant pour entête Pr(y=1|x) représente l'approximation par Γ_{BN} de la probabilité conditionnelle qu'un concept soit utilisé par un agent à un temps t+1, sachant la magnitude d'un attribut d'influence sociale au temps t. Par exemple, selon l'apprenant Γ_{BN}, la probabilité qu'un concept soit utilisé par un agent à t+1 étant donnée que la magnitude de l'exposition sociale au temps t se situe dans l'intervalle [0,0-0,0558] est de 0,06. Lorsque la magnitude d'un attribut est associée à une probabilité Pr(y=1|x) élevée, cela signifie que cette magnitude conditionne fortement l'usage futur d'un concept.

Tableau 6.2: Modèle probabiliste du mécanisme d'influence sociale.

	Magnitude	Pr(x)	Pr(x|y=0)	Pr(x|y=1)	Pr(y=0|x)	Pr(y=1|x)
Exposition sociale	[0,0-0,0558]	0,78	0,90	0,25	0,94	0,06
	]0,0558-0,1116]	0,11	0,08	0,28	0,55	0,45
	]0,1116-0,1674]	0,07	0,02	0,30	0,24	0,76
	]0,1674-1,0]	0,03	0,00	0,17	0,08	0,92
Exposition sociale relative	[0,0-0,25]	0,97	0,99	0,88	0,84	0,17
	]0,25-0,5]	0,01	0,00	0,02	0,37	0,63
	]0,5-0,75]	0,00	0,00	0,01	0,27	0,73
	]0,75-1,0]	0,02	0,00	0,09	0,14	0,86
Contagion basée sur la proximité sociale	[0,0-0,015571]	0,89	0,96	0,59	0,88	0,12
	]0,015571-0,031142]	0,07	0,04	0,22	0,42	0,59
	]0,031142-0,046713]	0,04	0,01	0,16	0,14	0,86
	]0,046713-1,0]	0,01	0,00	0,03	0,04	0,96
Contagion basée sur la fréquence des collaborations	[0,0-0,017307]	0,99	1,00	0,97	0,82	0,18
	]0,017307-0,034614]	0,01	0,00	0,02	0,20	0,80
	]0,034614-0,05192]	0,00	0,00	0,00	0,17	0,84
	]0,05192-1,0]	0,00	0,00	0,00	0,06	0,94
Déférence basée la centralité sociale	[0,0-0,147827]	0,76	0,88	0,24	0,94	0,06
	]0,147827-0,295654]	0,11	0,08	0,24	0,61	0,40
	]0,295654-0,44348]	0,07	0,03	0,26	0,34	0,66
	]0,44348-1,0]	0,06	0,01	0,27	0,11	0,89
Déférence basée sur la centralité de proximité	[0,0-0,120748]	0,77	0,89	0,24	0,94	0,06
	]0,120748-0,241495]	0,11	0,08	0,25	0,59	0,41
	]0,241495-0,362242]	0,08	0,03	0,29	0,30	0,71
	]0,362242-1,0]	0,05	0,01	0,22	0,09	0,91
Mimétisme basé sur la similitude sociale	[0,0-0,001268]	0,95	0,99	0,79	0,85	0,15
	]0,001268-0,002535]	0,04	0,01	0,14	0,30	0,70
	]0,002535-0,003803]	0,01	0,00	0,06	0,13	0,87
	]0,003803-1,0]	0,00	0,00	0,01	0,04	0,96
Mimétisme basé sur la similitude lexicale	[0,0-0,033103]	0,83	0,94	0,36	0,92	0,08
	]0,033103-0,066206]	0,11	0,06	0,37	0,41	0,59
	]0,066206-0,099308]	0,05	0,01	0,24	0,14	0,86
	]0,099308-1,0]	0,01	0,00	0,04	0,04	0,96

Deux caractéristiques sont particulièrement importantes pour comprendre le modèle probabiliste du Tableau 6.2. La première est la distribution de Pr(y=1|x). Le Tableau 6.2 montre que, pour tous les attributs de l'influence sociale, plus leur magnitude est grande, plus la probabilité que ce concept soit utilisé à t+1 est grande également. Prenons à titre d'illustration la magnitude de la contagion basée sur la proximité

sociale. La probabilité Pr(y=1|x) pour chaque intervalle de sa magnitude est illustrée dans l'histogramme de la Figure 6.5.

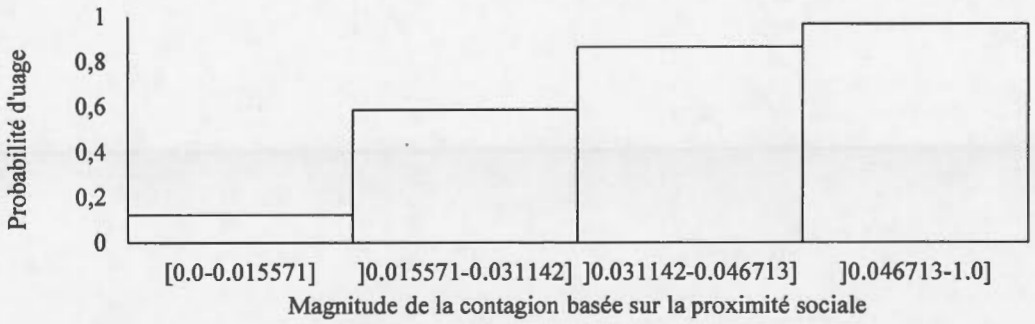

Figure 6.5: Histogramme montrant la probabilité qu'un concept soit utilisé par un agent à un temps t+1 étant donnée la magnitude de la contagion basée sur la proximité sociale à un temps t.

La probabilité qu'un concept soit utilisé par l'agent a_i à un temps t+1 lorsque la magnitude de la contagion se situe dans l'intervalle [0,0-0,015571] est très petite, soit 0,12. Mais, dès que cette magnitude dépasse le seuil 0,015571, la probabilité que le concept soit utilisé devient plus grande que celle qu'il ne soit pas utilisé. Finalement, lorsque cette magnitude dépasse le seuil 0,046713, la probabilité que le concept soit utilisé devient presque certaine, soit de 0,96. Tous les attributs de l'influence sociale sont caractérisés par une distribution de cette forme croissante.

Une deuxième caractéristique saillante dans le modèle probabiliste du Tableau 6.2 est la distribution de Pr(x|y=1). Si nous reprenons, à titre d'illustration, la magnitude de la contagion basée sur la proximité sociale, nous pouvons voir dans le Tableau 6.2 que plus la magnitude de cet attribut est grande, plus elle est rare. Par exemple, la probabilité qu'un usage conceptuel ait une magnitude se situant dans l'intervalle [0,0-0,015571] est de 0,59, ce qui est plutôt élevé, alors que la probabilité qu'un usage conceptuel ait une magnitude se situant dans l'intervalle]0,046713-1,0] est très faible, soit de seulement 0,03.

Les attributs de l'influence sociale les plus importants du modèle sont ceux qui ont une magnitude associée à la fois à des probabilités Pr(x|y=1) et Pr(y=1|x) élevées. Par exemple, bien qu'une forte magnitude de la contagion permette de prédire avec une probabilité très élevée de 0,96 qu'un concept sera utilisé à t+1, l'évènement est si rare, d'une probabilité de 0,03, qu'il est sans effet dans le modèle probabiliste.

Or, ce que montre le Tableau 6.2 montre est que ces deux conditions sont rarement réunies. Pour plusieurs attributs de l'influence sociale, plus la probabilité Pr(x|y=1) est grande, plus la probabilité Pr(y=1|x) est petite. Au mieux, pour certains attributs, les valeurs de Pr(x|y=1) sont relativement constantes lorsque Pr(y=1|x) augmente. C'est le cas de l'exposition sociale, de la déférence basée la centralité sociale, de la déférence basée sur la centralité de proximité et du mimétisme basé sur la similitude lexicale. Selon l'apprenant Γ_{BN}, ces quatre attributs sont par conséquent les plus importants dans la modélisation de l'influence sociale.

6.2 Analyse des résultats

La deuxième section de ce chapitre est consacrée à l'analyse des modèles induits par les différents apprenants automatiques. En tout, sept analyses sont menées sur les modèles. La première analyse évalue la cohérence des modèles envers leur base d'apprentissage. La deuxième analyse évalue la capacité des modèles à se généraliser sur les exemplaires d'une base de test. La troisième analyse est une validation croisée des modèles et une analyse comparative de leur performance prédictive. La quatrième analyse évalue les courbes d'apprentissage des différents apprenants automatiques. La cinquième analyse évalue les effets de la réduction dimensionnelle sur les performances prédictives des modèles. La sixième analyse évalue la colinéarité des attributs de l'influence sociale et la septième analyse cherche à expliquer le résidu des modèles.

6.2.1 Analyse de la cohérence des modèles de l'influence sociale

Le Tableau 6.3 suivant présente les résultats d'analyse de la cohérence des différents modèles avec leur base d'apprentissage. Les résidus de la cohérence sont mesurés à l'aide des différents indices d'adéquation introduits à la fin du chapitre précédent, soit les indices Rand, Kappa, Matthews, la Sensibilité, la Spécificité, la Précision positive et négative.

Tableau 6.3: Évaluation de la cohérence avec la base d'apprentissage des modèles de l'influence sociale induits par quatre apprenants automatiques, soit un inducteur d'arbre de décisions (Γ_{AD}), un inducteur de règles (Γ_{LR}), un inducteur de forêt aléatoire (Γ_{FA}) et un inducteur bayésien naïf (Γ_{BN}).

	Apprenants automatiques					
	$\Gamma_{AD}(\mathbb{A})$	$\Gamma_{LR}(\mathbb{A})$	$\Gamma_{FA}(\mathbb{A})$	$\Gamma_{BN}(\mathbb{A})$	$\bar{x}$	σ
Rand	0,89	0,89	1,00	0,87	0,92	0,06
Kappa	0,60	0,60	1,00	0,60	0,70	0,17
Matthews	0,60	0,60	1,00	0,60	0,70	0,17
Sensibilité	0,60	0,59	1,00	0,74	0,73	0,17
Spécificité	0,95	0,96	1,00	0,90	0,95	0,04
Précision +	0,74	0,75	1,00	0,63	0,78	0,14
Précision -	0,92	0,91	1,00	0,94	0,94	0,03
$\lvert\mathbb{A}\rvert$	330000	330000	330000	330000		

L'indice de Rand est identique pour les apprenants Γ_{AD} et Γ_{LR}, soit d'une valeur de 0,89. Rand a une valeur de 1,00 pour Γ_{FR}, soit la valeur maximale pour cet indice, alors que le Rand de Γ_{BN} est de 0,87. En moyenne, la valeur de Rand est de 0,92 avec un écart-type de 0,06.

Ces valeurs sont très élevées. Selon Rand, les modèles induits par les apprenants Γ_{AD}, Γ_{LR} et Γ_{BN} sont capables de prédire correctement entre 87% et 89% des transitions

d'états dans l'évolution des liens d'usages dans le réseau sociosémantique des journalistes du The New York Time. De plus, selon Rand, la modélisation du mécanisme d'influence sociale sous forme d'une forêt aléatoire prédit correctement 100% des transitions d'états du processus.

Ces résultats doivent toutefois être interprétés avec prudence. Premièrement, Rand doit être mis en contexte avec la prévalence présente dans la base d'apprentissage. La prévalence négative dans la base d'apprentissage est de 81,74%, ce qui suggère que Rand surestime l'adéquation réelle entre les hypothèses d'approximation et le processus empirique d'évolution des usages conceptuels dans le réseau sociosémantique. De plus, ces résultats sont possiblement le fruit d'un sur-ajustement sur la base d'apprentissage. Ce sur-ajustement sera analysé dans la prochaine section via l'analyse des résidus de généralisations des modèles.

Les indices Kappa et Matthews permettent quant à eux d'évaluer les effets de la prévalence dans l'évaluation de la cohérence des modèles. Les valeurs de Kappa et de Matthews sont identiques pour les trois apprenants Γ_{AD}, Γ_{LR} et Γ_{BN}, soit d'une valeur de 0,60. Les valeurs de Kappa et de Matthews sont de 1,00 pour l'apprenant Γ_{FR}. En moyenne, Kappa et Matthews ont une valeur de 0,70 et sont caractérisés par un écart-type de 0,174.

À nouveau, le modèle induit par Γ_{FR} se démarque clairement des autres apprenants. Sa cohérence avec la base d'apprentissage est parfaite et n'est nullement affectée par la prévalence. Le Kappa et le Matthews des trois autres apprenants indiquent pour leur part que la cohérence de leur hypothèse d'approximation était en partie due à la prévalence. En effet, leur cohérence avec la base d'apprentissage, qui variaient entre 0,87 et 0,89 lorsqu'elles étaient mesurées avec Rand, est maintenant de 0,60 lorsque mesurées avec Kappa et Matthews. Ceci indique qu'une fois contrôlés les effets de la prévalence, les modèles de l'influence sociale induits par les apprenants Γ_{AD}, Γ_{LR} et Γ_{BN} ne prédisent correctement en réalité que 60% des transitions d'états du processus

d'évolution du réseau sociosémantique. Malgré cette différence, les valeurs de Kappa et de Matthews peuvent être interprétées comme moyennement élevées (Kuhn & Johnson, 2013, p. 256).

La Sensibilité des modèles est de 0,60 pour l'apprenant Γ_{AD} et de 0,59 pour l'apprenant Γ_{LR}. Celle de l'apprenant Γ_{FA} est à nouveau parfaite avec une valeur de 1,00. Ayant comme valeur 0,74, la Sensibilité de Γ_{BN} est plus petite que celle de Γ_{FA}, mais plus grande que celles des apprenants Γ_{AD} et Γ_{LR}. En moyenne, la Sensibilité des apprenants est de 0,73 avec un écart-type de 0,17. Ces résultats indiquent que les modèles induits par les différents apprenants prédisent correctement entre 59% et 100% de l'évolution des usages conceptuels représentés dans la base d'apprentissage.

La Spécificité des modèles est de 0.95 pour l'apprenant Γ_{AD}, de 0.96 pour Γ_{LR} et de 1.00 pour Γ_{FA}. Alors que sa Sensibilité était plus élevée, la Spécificité du modèle induit par Γ_{BN} est légèrement plus faible que celles des apprenants Γ_{AD} et Γ_{LR}, avec une valeur de 0,90, En moyenne, la Spécificité des modèles est de 0,95 avec un écart-type de 0,04. Ces résultats indiquent que les modèles de l'influence sociale induits par les différents apprenants prédisent correctement entre 90% et 100% de l'évolution des non-usages conceptuels représentés dans la base d'apprentissage.

La Précision positive des apprenants Γ_{AD} et Γ_{LR} est similaire. Elle est de 0,74 pour Γ_{AD} et de 0,75 pour Γ_{LR}. La Précision positive de l'apprenant Γ_{FA} est parfaite, alors que celle de Γ_{BN} est plus faible que les autres avec une valeur de 0,63. En moyenne, la Précision positive des modèles est de 0,78 avec un écart-type de 0,14. Ces résultats indiquent que les prédictions que permettent de faire les différents modèles de l'influence sociale à propos de l'évolution des usages conceptuels dans le réseau sociosémantique sont valides dans une proportion variant entre 63% et 100%.

La Précision négative du modèle induit par Γ_{AD} est de 0,92, celle de Γ_{LR} est de 0,91, celle de Γ_{FA} est à nouveau de 1,00 et celle de Γ_{BN} est de 0,94. La moyenne est de 0,94

et l'écart-type de 0,03. Ces résultats indiquent que les prédictions que permettent de faire les différents modèles de l'influence sociale à propos de l'évolution des non-usages conceptuels dans le réseau sociosémantique sont valides dans une proportion variant entre 91% et 100%.

En somme, ces résultats d'analyse montrent que la modélisation de l'influence sociale sous la forme d'un arbre de décisions, d'une liste de règles ou sous la forme d'un modèle probabiliste, a une cohérence très forte. La différence entre ces trois modèles se situe au niveau de la prédiction des vrais positifs et des vrais négatifs.

Comme le montre leur Sensibilité, les apprenants Γ_{AD} et Γ_{LR} ont davantage de difficulté que les autres à identifier les patrons récurrents d'influence sociale qui permettent de prédire quels concepts seront utilisés par un agent. Toutefois, comme le montre leur Précision positive, ces prédictions sont de meilleure qualité que celles de Γ_{BN}. Ces deux modèles identifient mieux les patrons récurrents d'influence sociale qui permettent de prédire quels concepts ne seront pas utilisés par un agent.

Le modèle probabiliste de l'influence sociale a une Sensibilité plus grande, mais une Spécificité plus petite. C'est un modèle qui prédit mieux que les précédents quels concepts seront utilisés par un agent, mais prédit moins bien quels concepts ne seront pas utilisés dans l'évolution du réseau sociosémantique.

Enfin, ces résultats montrent que la modélisation du mécanisme d'influence sociale par une forêt aléatoire a une Sensibilité, une Spécificté et une Précision parfaite. Par contre, pour nous assurer de la vraisemblance de ces résultats, il faut également évaluer les résidus de généralisation sur une base de test.

6.2.2 Analyse de la généralisation des modèles de l'influence sociale

La deuxième analyse consiste à évaluer la capacité des modèles à se généraliser sur les exemplaires d'une base de test. En d'autres mots, le but est de vérifier si les modèles

induits précédemment sont le résultat d'un sur-ajustement à la base d'apprentissage ou si au contraire ils sont des modèles suffisamment vraisemblables pour être généraliser à d'autres observations empiriques. La base de test est composée de 170 000 exemplaires.

Le Tableau suivant présente les résultats de ces évaluations. Les mêmes indices d'adéquation sont utilisés, soit l'indice Rand, Kappa, Matthews, la Sensibilité, la Spécificité, la Précision positive et négative.

Tableau 6.4: Évaluation de la généralisation avec la base de test des modèles de l'influence sociale induits par quatre apprenants automatiques, soit un inducteur d'arbre de décisions (Γ_{AD}), un inducteur de règles (Γ_{LR}), un inducteur de forêt aléatoire (Γ_{FA}) et un inducteur bayésien naïf (Γ_{BN}).

	Apprenants automatiques					
	$\Gamma_{AD}(\mathbb{B})$	$\Gamma_{LR}(\mathbb{B})$	$\Gamma_{FA}(\mathbb{B})$	$\Gamma_{BN}(\mathbb{B})$	$\bar{x}$	σ
Rand	0,89	0,89	0,88	0,87	0,88	0,01
Kappa	0,59	0,59	0,57	0,60	0,59	0,01
Matthews	0,60	0,60	0,58	0,60	0,60	0,01
Sensibilité	0,60	0,59	0,59	0,73	0,63	0,06
Spécificité	0,95	0,95	0,95	0,90	0,94	0,02
Précision +	0,74	0,74	0,71	0,63	0,71	0,05
Précision -	0,91	0,91	0,91	0,94	0,92	0,01
$\lvert\mathbb{B}\rvert$	170000	170000	170000	170000		

Le Tableau 6.4 montre que, pour les modèles induits par les apprenants Γ_{AD}, Γ_{LR} et Γ_{BN}, les valeurs de Rand, de Kappa, de Matthews, la Sensibilité, la Spécificité, la Précision positives et négatives sont restées relativement inchangées. Ceci indique que les modèles induits par ces trois apprenants sont généralisables empiriquement sans compromettre les performances prédictives des modèles.

Il en va autrement pour la modélisation du mécanisme d'influence sociale sous la forme d'une forêt aléatoire. Alors que les différents indices d'adéquation suggéraient précédemment que ce modèle formait une approximation parfaite du mécanisme d'influence sociale, lorsque ce dernier est appliqué sur un nouvel échantillon d'exemplaires, ses performances prédictives chutent de manière importante. Les valeurs des différents indices pour ce modèle sont dorénavant très similaires à celles des autres modèles. Selon les indices Kappa et Matthews, elles sont même un peu plus faibles.

L'ensemble de ces résultats convergent vers une même conclusion : indépendamment des apprenants automatiques utilisés pour reconstruire le mécanisme d'influence sociale, l'adéquation selon l'indice de Rand entre les modèles et le processus empirique d'évolution des liens d'usage entre agent et concept est de 0,88. De plus, si nous tenons compte des effets de la prévalence, cette adéquation est de 0,59 selon Kappa et de 0,60 selon Matthews. Ces analyses montrent également que la prédiction de l'évolution des usages conceptuels est plus difficile que la modélisation de l'évolution des non-usages conceptuels.

6.2.3 Analyse des moyennes et validation croisée des modèles

La troisième analyse est une technique de validation croisée des modèles de l'influence sociale induits par les différents apprenants automatiques. Une validation croisée consiste répéter plusieurs fois l'induction d'un modèle, mais à partir de différentes bases d'apprentissage et à évaluer la généralisation de chaque version d'un même modèle sur plusieurs bases de test. Le but d'une telle analyse est double. Premièrement, il s'agit de s'assurer que les résidus de généralisation obtenus précédemment n'étaient pas dus au hasard ou à un biais d'échantillonnage de la base d'apprentissage et de la base de test. Deuxièmement, en comparant la moyenne des performances prédictives des différents modèles (à l'aide d'un test statistique), la validation croisée permet de

déterminer si un modèle est significativement supérieur aux autres (Hothorn, Leisch, Zeileis, & Hornik, 2005; Witten et al., 2011, p. 154).

Pour effectuer cette analyse, les 500 000 exemplaires de l'échantillon sont d'abord divisés de manière aléatoire en 100 sous-échantillons de 5000 exemplaires. La technique d'analyse est ensuite répétée 100 fois pour chaque apprenant automatique. À chaque itération, 99 sous-échantillons d'exemplaires sont regroupés pour former une base d'apprentissage sur laquelle chaque apprenant induit un modèle. En tout, chaque apprenant induit donc 100 modèles du mécanisme d'influence sociale, mais chaque fois à partir d'une base d'apprentissage différente. À chaque itération, le sous-échantillon de 5000 exemplaires restant est utilisé pour former une base de test et évaluer la généralisation des modèles. À chaque itération, un nouveau sous-échantillon de 5000 exemplaires forme donc la base de test si bien qu'à la fin de l'analyse tous les exemplaires sont utilisés une fois pour le test.

Le Tableau 6.5 illustre les résultats de cette analyse. Les valeurs dans le Tableau correspondent aux performances prédictives moyennes des modèles induits par chaque apprenant. Les résidus de généralisation sont mesurés par les différents indices d'adéquation utilisés précédemment. Un test statistique de Student est calculé afin de déterminer si les performances prédictives moyennes de certains apprenants sont supérieures autres autres. Le seuil de signification utilisé pour le test est $p < 0{,}00001$.

Tableau 6.5: Validation croisée des modèles de l'influence sociale induits par les apprenants automatiques. L'échantillon d'exemplaires a été divisé en 100 sous-ensembles égaux. Les valeurs sur fond noir correspondent aux performances significativement supérieures aux valeurs sur fond blanc (selon un test de Student, au seuil p.<0,00001).

	Apprenants automatiques					
	$\bar{\Gamma}_{AD}(\mathbb{B})$	$\bar{\Gamma}_{LR}(\mathbb{B})$	$\bar{\Gamma}_{FA}(\mathbb{B})$	$\bar{\Gamma}_{BN}(\mathbb{B})$	$\bar{x}$	σ
Rand	0,89	0,89	0,88	0,87	0,88	0,01
Kappa	0,59	0,59	0,57	0,60	0,59	0,01
Matthews	0,60	0,60	0,58	0,60	0,60	0,01
Sensibilité	0,58	0,58	0,59	0,73	0,62	0,06
Spécificité	0,96	0,96	0,95	0,90	0,94	0,02
Précision +	0,75	0,75	0,71	0,63	0,71	0,05
Précision -	0,91	0,91	0,91	0,94	0,92	0,01

L'ensemble des valeurs du Tableau 6.5 sont approximativement identiques aux résultats obtenus précédemment. Ces moyennes représentent de fortes évidences que les modèles induits par les apprenants automatiques sont robustes et ne sont pas dus au hasard ou à un biais d'échantillonnage.

Les valeurs sur fond noir sont les moyennes significativement supérieures à celles sur fond blanc. Selon l'indice Rand, les modèles induits par les apprenants Γ_{AD} et Γ_{LR} produisent des modèles de l'influence sociale en moyenne significativement supérieurs aux modèles induits par Γ_{FA} et Γ_{BN}. Cependant, lorsque les effets de la prévalence sont contrôlés, ce sont plutôt les modèles probabilistes de Γ_{BN} qui sont en moyenne significativement supérieurs aux autres. La moyenne de Kappa indique que les modèles induits par Γ_{BN} sont supérieurs à ceux des trois autres alors que la moyenne de Matthews indique qu'il n'y a pas de différence significative entre Γ_{AD}, Γ_{LR} et Γ_{BN}.

Lorsque nous analysons les autres indices, nous arrivons à des conclusions similaires à celles discutées précédemment. Les modèles de l'influence sociale sous forme de forêt

aléatoire se généralisent systématiquement moins bien que les autres modèles. Les modèles probabilistes de Γ_{BN} ont une Sensibilité et une Précision négative moyennes supérieures aux autres modèles, alors que les modèles sous forme d'arbre de décisions et de règles ont une Spécificité et une Précision positive en moyennes supérieures aux autres modèles.

6.2.4 Analyse des courbes d'apprentissage

La quatrième analyse porte sur l'évaluation des courbes d'apprentissage des différents apprenants automatiques. Le but de cette analyse est d'évaluer si l'échantillon d'exemplaires utilisé pour l'apprentissage machine était représentatif du processus empirique d'évolution du réseau sociosémantique ou si au contraire des exemplaires supplémentaires auraient été nécessaires pour assurer la vraisemblance des modèles (Witten et al., 2011). Les courbes d'apprentissage sont une technique d'analyse de la puissance statistique des modèles (Figueroa, Zeng-Treitler, Kandula, & Ngo, 2012).

La technique d'analyse des courbes d'apprentissage consiste à ajouter de nouveaux exemplaires à l'échantillon et à évaluer si cela augmente les performances prédictives des modèles induits. Si ce n'est pas le cas, ceci constitue alors une évidence supplémentaire que l'échantillon utilisé était représentatif du processus empirique et que les performances prédictives des modèles sont optimales.

Pour effectuer cette analyse, on extrait de l'échantillon original de 500 000 d'exemplaires plusieurs sous-échantillons, soit un premier contenant seulement 1% des exemplaires, un deuxième contenant 2% des exemplaires, une troisième contenant 3% et ainsi de suite jusqu'à 100%. Ensuite, pour chaque sous-échantillon on applique la technique de validation croisée utilisée précédemment. Les résultats sont affichés dans les deux graphiques de la Figure 6.6. Seulement les indices Kappa et Matthews sont comparés.

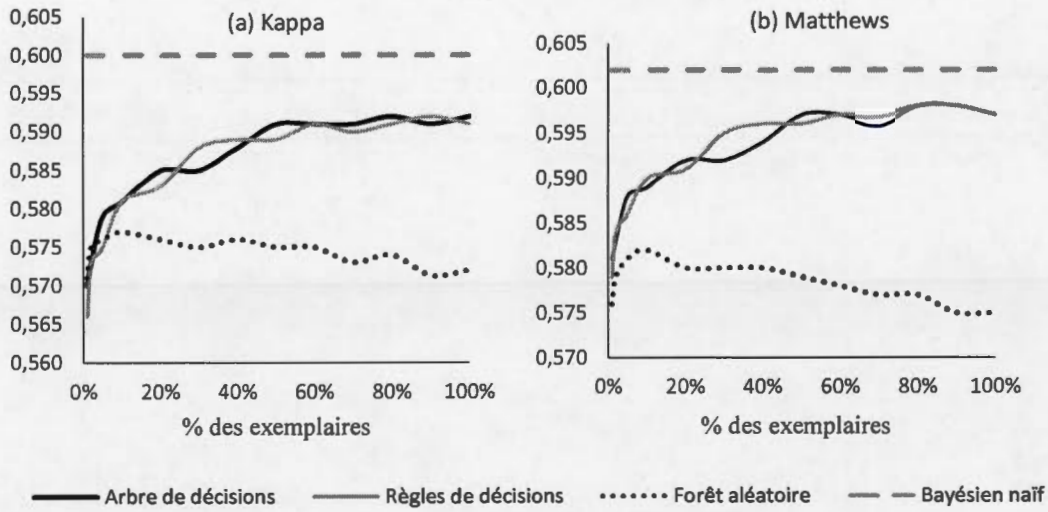

Figure 6.6: Courbes d'apprentissage des apprenants automatiques. L'abscisse des graphiques représente la proportion d'exemplaires dans les sous-échantillons. L'ordonnée des graphiques représente à gauche le coefficient Kappa et à droite le coefficient Matthews.

Les graphiques de la Figure 6.6 montrent que les performances prédictives moyennes des modèles probabilistes induits par l'apprenant Γ_{BN} ne sont aucunement affectées par la taille des sous-échantillons. En utilisant seulement 1% des exemplaires, l'adéquation (mesurée via Kappa et Matthews) entre les modèles induits et le processus empirique reste constante.

Pour les modèles d'arbre de décisions et de liste de règles, des différences significatives (au seuil $p<0,01$) sont observées lorsque 10% et moins des exemplaires sont utilisés pour induire les modèles. En d'autres termes, les performances prédictives des modèles diminuent faiblement (mais significativement) lorsque seulement 50 000 exemplaires et moins sont utilisés pour induire les modèles.

Les modèles sous la forme de forêt aléatoire ont des performances prédictives qui varient très légèrement selon la quantité d'exemplaires utilisés, mais ces différences ne sont pas statistiquement significatives.

En somme, les courbes d'apprentissage des apprenants automatiques indiquent que l'échantillon d'exemplaires utilisé était suffisamment représentatif du processus empirique d'évolution du réseau sociosémantique. Ajouter de nouveaux exemplaires n'ajoute pas d'informations supplémentaires qui permettraient aux apprenants d'identifier des patrons inédits d'influence sociale. En fait, cette analyse indique qu'un échantillon d'un peu plus de 50 000 exemplaires aurait été suffisant pour les expérimentations.

6.2.5 Analyse des effets de la réduction dimensionnelle

La cinquième analyse porte sur les effets de la réduction dimensionnelle des modèles. Elle consiste à vérifier si des modèles plus simples que ceux présentés précédemment permettraient de prédire avec autant d'exactitude le processus empirique d'évolution du réseau socisoématique étudié.

Cette technique d'analyse est inspirée des travaux de Holte qui a montré que les solutions trouvées à de nombreuses problématiques de reconstruction par apprentissage machine étaient bien souvent des modèles beaucoup plus complexes que nécessaire. En effet, dans ses travaux Holte a montré que dans bien des cas un modèle réduit à une seule dimension ou un seul attribut permettait des prédictions empiriques d'une surprenante exactitude (Holte, 1993).

Dans la présente recherche, la reconstruction par apprentissage machine du mécanisme d'influence sociale est réalisée dans un espace d'instances à huit dimensions. En d'autres termes, les patrons d'influence sociale sont modélisés par huit attributs, qui vont de la magnitude de l'exposition sociale à la magnitude du mimétisme des semblables basé sur la similitude lexicale. Or, ces différents attributs de l'influence sociale ne sont pas tous de même importance. Comme nous l'avons vu précédemment, certains sont beaucoup plus déterminants que d'autres dans le processus d'évolution du réseau sociosémantique.

L'importance de ces attributs ou dimensions de l'influence sociale peut être calculée à l'aide d'un coefficient d'association statistique. Elle peut par exemple être calculée à l'aide du coefficient du gain d'information défini précédemment dans la section 5.3.1.3 du chapitre cinq. Ce coefficient permet de quantifier la force de l'association entre d'une part la magnitude d'un attribut observé à un temps t de la dynamique du SSS et, d'autre part, l'évolution des usages conceptuels observés à un temps t+1.

La Figure 6.7 montre le gain d'information pour chacun des huit attributs de l'influence sociale.

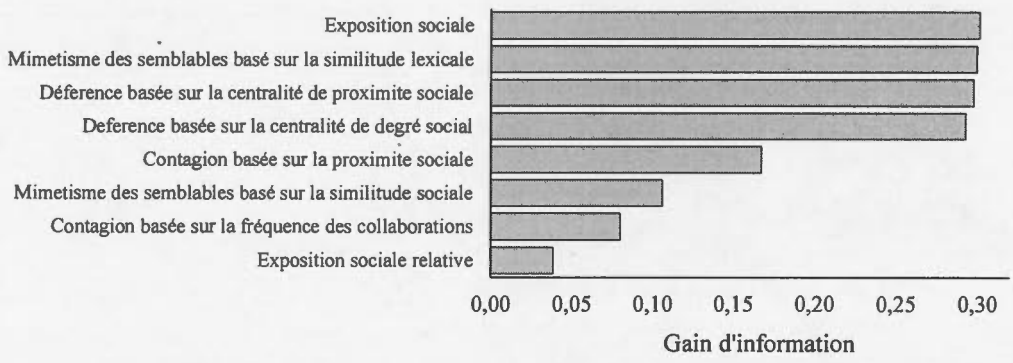

Figure 6.7: Gain d'information pour chaque attribut de l'influence sociale.

Le coefficient du gain d'information varie entre 0,0 et 1,0 : une valeur égale à 0,0 indique que l'attribut n'est pas associé à l'évolution du réseau sociosémantique et une valeur égale à 1,0 indique que l'attribut est parfaitement corrélé à celle-ci.

Selon ce coefficient, l'attribut le plus important est l'exposition sociale. Son gain d'information est cependant presque identique à trois autres attributs, soit le mimétisme des semblables basé sur les similitudes lexicales, la déférence basée sur la centralité de proximité sociale et la déférence basée sur la centralité de degré social. L'attribut le plus important ensuite a un gain d'information de près de 50% moindre. Il s'agit de la contagion basée sur la proximité sociale. Ensuite vient le mimétisme des semblables

basé sur la similitude des positions sociales, la contagion basée sur la fréquence des collaborations et finalement l'exposition sociale relative.

Dans l'analyse qui suit, afin d'évaluer l'impact des différents attributs, des modèles de l'influence sociale sont induits, mais seulement à partir d'un sous-ensemble d'attributs. Cette technique d'analyse procède de la manière suivante. Dans un premier temps, chaque apprenant induit un modèle en utilisant seulement les sept attributs les plus corrélés à l'évolution du réseau sociosémantique. Dans un deuxième temps, chaque apprenant induit un modèle, mais en utilisant cette fois-ci seulement les six attributs les plus corrélés à l'évolution du réseau sociosémantique, ensuite les cinq attributs les plus importants et ainsi de suite pour ne finalement utiliser à la fin qu'un seul attribut, le plus important, soit l'exposition sociale.

Les performances prédictives des modèles réduits dimensionnellement sont ensuite comparées avec celles des modèles complets. La comparaison utilise à nouveau la technique de validation croisée utilisée précédemment.

Le Tableau 6.6 montre les résultats d'analyses obtenus pour les différents apprenants Γ_{AD}, Γ_{LR}, Γ_{FA} et Γ_{BN}. Les valeurs sur fond gris indiquent que les modèles induits avec ce nombre réduit d'attributs prédisent moins fidèlement l'évolution du réseau sociosémantique que les modèles complets induits sur l'ensemble des huit attributs de l'influence sociale. Les valeurs sur fond noir indiquent que les modèles induits avec ce nombre réduit d'attributs prédisent plus fidèlement le processus d'évolution du réseau sociosémantique que les modèles complets. Le seuil de signification statistique utilisé pour la comparaison est $p<0,001$.

Tableau 6.6: Comparaisons des effets de la réduction dimensionnelle sur les modèles induits par chaque apprenant. Les valeurs affichées sont les moyennes obtenues à partir d'une validation croisée sur 100 sous-échantillons. Les valeurs sur fonds gris sont significativement inférieures aux modèles complets et les valeurs sur fond noir sont significativement supérieures ($p<0,001$).

		Rand	Kappa	Matthews	Sensibilité	Spécificité	Précision +	Précision -
Arbre de décisions (Γ_{AD})	1	0,886	0,576	0,584	0,566	0,957	0,745	0,908
	2	0,887	0,584	0,590	0,579	0,955	0,742	0,910
	3	0,887	0,584	0,591	0,578	0,9555	0,744	0,910
	4	0,887	0,583	0,590	0,576	0,956	0,745	0,910
	5	0,887	0,586	0,5924	0,578	0,956	0,746	0,910
	6	0,887	0,585	0,5921	0,576	0,957	0,748	0,910
	7	0,887	0,584	0,591	0,574	0,957	0,749	0,910
	8	0,889	0,589	0,596	0,579	0,958	0,753	0,911
Liste de règles (Γ_{LR})	1	0,886	0,576	0,584	0,566	0,957	0,745	0,908
	2	0,887	0,585	0,591	0,580	0,955	0,742	0,911
	3	0,887	0,582	0,589	0,571	0,957	0,749	0,909
	4	0,887	0,581	0,590	0,567	**0,958**	**0,753**	0,909
	5	0,887	0,584	0,591	0,576	0,957	0,747	0,910
	6	0,887	0,585	0,592	0,577	0,956	0,747	0,910
	7	0,887	0,587	0,594	0,582	0,955	0,745	0,911
	8	0,888	0,591	0,597	0,586	0,956	0,747	0,912
Forêt aléatoire (Γ_{FA})	1	**0,885**	0,576	**0,583**	0,565	**0,957**	**0,744**	0,908
	2	0,862	0,517	0,518	0,571	0,926	0,634	0,906
	3	0,873	0,546	0,549	0,570	0,941	0,683	0,907
	4	0,867	0,530	0,532	0,569	0,934	0,656	0,907
	5	0,879	0,566	0,569	0,583	0,945	0,703	0,910
	6	0,880	0,569	0,572	0,583	0,946	0,708	0,911
	7	0,880	0,570	0,574	0,585	0,946	0,709	0,911
	8	0,880	0,572	0,575	0,588	0,946	0,708	0,911
Bayésien naïf (Γ_{BN})	1	**0,883**	0,531	0,557	0,470	**0,975**	**0,806**	0,892
	2	**0,883**	0,598	0,599	0,649	**0,935**	**0,691**	0,923
	3	**0,883**	0,600	0,600	0,657	**0,933**	**0,686**	0,924
	4	0,873	0,598	0,601	0,729	**0,905**	**0,632**	0,937
	5	0,873	0,599	0,602	0,733	0,904	0,630	0,938
	6	0,873	0,599	0,602	0,734	0,904	0,630	0,938
	7	0,873	0,599	0,602	0,734	0,904	0,629	0,938
	8	0,873	0,600	0,602	0,734	0,904	0,631	0,938

Les modèles induits par les apprenants Γ_{AD} et Γ_{LR} sont les plus affectés par la réduction dimensionnelle. Selon les indices Rand, Kappa et Matthews, réduire les patrons

d'influence sociale d'une seule dimension, c'est-à-dire en excluant l'exposition sociale relative, génère en moyenne des arbres de décisions et des listes de règles aux capacités prédictives significativement inférieures. La Spécificité et la Précision positive restent toutefois très peu affectées par la réduction dimensionnelle. Des modèles très simples, même composés d'un seul attribut, permettent de prédire avec autant de précision (et parfois de manière significativement supérieure) le non-usage des concepts dans le processus d'évolution du réseau sociosémantique. La Sensibilité des modèles diminue légèrement, mais le taux de précision de la prédiction des usages conceptuels reste constant (Prédiction positive).

Des résultats d'analyse similaires sont obtenus avec les modèles induits par l'apprenant Γ_{FA}. Toutefois, l'analyse montre que, selon les indices Rand, Kappa et Matthews, les modèles de forêt aléatoires construits seulement à partir de l'attribut le plus important, soit l'exposition sociale, permettent de prédire l'évolution du réseau sociosémantique de manière équivalente ou supérieure aux modèles plus complexes.

Finalement, les modèles probabilistes induits par l'apprenant Γ_{BN} sont les modèles les moins affectés par la réduction dimensionnelle. Selon les indices Kappa et Matthews, il n'y a pas de différence significative entre les performances prédictives des modèles induits à partir de la totalité des attributs caractérisant l'influence sociale et les performances prédictives des modèles induits en utilisant entre deux et sept attributs. Selon Rand, les performances prédictives des modèles à un seul attribut sont significativement supérieures aux modèles complets. Toutefois ces résultats ne sont pas corroborés par Kappa et Matthews. C'est surtout la Sensibilité et la Précision négative qui sont affectées par une réduction dimensionnelle aussi importante.

Ces résultats d'analyses montrent essentiellement deux choses. La réduction dimensionnelle affecte effectivement, bien que d'une manière différente, les performances prédictives des modèles induits par les apprenants automatiques. Toutefois, comme le conjecturait Holte, des modèles extrêmement simples, composés

de deux et même d'un seul attribut, ont des performances prédictives presque équivalentes aux modèles complets. Ceci constitue un résultat d'analyse important pour la présente recherche. Prenons par exemple le modèle suivant qui a la forme d'un arbre de décisions :

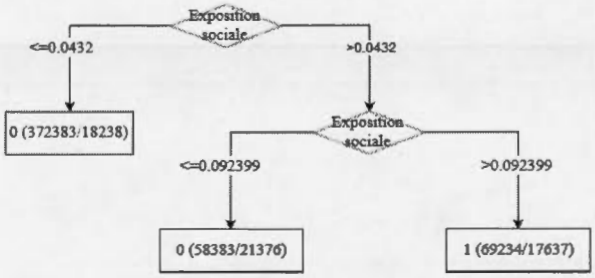

Figure 6.8: Arbre de décisions basé seulement sur la magnitude de l'exposition sociale.

Cet arbre est extrêmement simple. Il est basé sur un seul attribut de l'influence sociale, soit l'exposition sociale. Pourtant, ce modèle permet de prédire 88,55% des transitions d'états dans le processus d'évolution du réseau sociosémantique. Les valeurs des indices Kappa et Matthews sont toutes les deux de 0,58, leur Sensibilité de 0,57, leur Spécificité de 0,96, leur Précision positive de 0,75 et leur Précision négative de 0,91. C'est un modèle robuste (peu affecté par des variations d'échantillonnage) et plausible statistiquement (supérieur à la prévalence). Ses performances prédictives sont significativement inférieures à des modèles plus complexes, mais la différence est très petite. L'écart n'est que de 0,003 entre le Rand de cet arbre très simple et le Rand d'un arbre complet construit à partir de tous les attributs de l'influence sociale.

6.2.6 Analyse de corrélation

Comment expliquer que la réduction à un seul attribut des modèles du mécanisme d'influence sociale puisse malgré tout prédire avec presque autant d'exactitude le processus d'évolution du réseau sociosémantique? Une analyse de la corrélation entre les différents attributs de l'influence sociale apporte quelques éléments de réponse.

Le Tableau 6.7 présente les résultats d'une analyse de corrélation avec le r de Pearson entre les différents attributs caractérisant l'influence sociale. Toutes les valeurs de r sont significatives au seuil p<0,00001.

Tableau 6.7: Matrice de corrélation entre les différents attributs de l'influence sociale (p<0,00001).

	Exposition sociale	Exposition sociale relative	Contagion sociale basée sur la fréquence des collaborations	Contagion sociale basée sur la proximité sociale	Déférence basée sur la centralité de degré social	Déférence basée sur la centralité de proximité sociale	Mimétisme des semblables basé sur la similitude lexicale	Mimétisme des semblables basé sur la similitude sociale
Exposition sociale		0,322	0,379	0,685	**0,985**	**0,992**	**0,960**	0,548
Exposition sociale relative	0,322		0,510	0,495	0,325	0,325	0,360	0,552
Contagion sociale basée sur la fréquence des collaborations	0,379	0,510		0,647	0,385	0,384	0,436	0,793
Contagion sociale basée sur la proximité sociale	0,685	0,495	0,647		0,687	0,689	0,753	0,885
Déférence basée sur la centralité de degré social	**0,985**	0,325	0,385	0,687		**0,997**	**0,949**	0,555
Déférence basée sur la centralité de proximité	**0,992**	0,325	0,384	0,689	**0,997**		**0,954**	0,554
Mimétisme des semblables basé sur similitude lexicale	**0,960**	0,360	0,436	0,753	**0,949**	**0,954**		0,636
Mimétisme des semblables basé sur la similitude sociale	0,548	0,552	0,793	0,885	0,555	0,554	0,636	

La matrice de corrélation montre que les attributs de l'influence sociale sont tous positivement corrélés entre eux. Ces corrélations varient de moyennement faible à très forte. La corrélation la plus faible est celle entre l'exposition sociale et l'exposition sociale relative, avec r = 0,322. La corrélation la plus forte est celle entre la déférence

basée sur la centralité de degré sociale et la déférence basée sur la centralité de proximité, avec r = 0,997.

Par ailleurs, les corrélations les plus pertinentes pour la présente recherche sont celles qui concernent les quatre attributs de l'influence sociale identifiés précédemment comme étant les plus importants en termes de gain d'information, soit l'exposition sociale, la déférence basée sur la centralité de degré social, la déférence basée sur la centralité de proximité et le mimétisme des semblables basé sur la similitude lexicale. Le Tableau 6.7 montre que ces quatre attributs sont tous très fortement corrélés entre eux, avec un coefficient r variant de 0,954 à 0,997.

Ces corrélations très fortes expliquent pourquoi l'écart entre un modèle de l'influence sociale très simple réduit seulement à de l'exposition sociale est presque équivalent aux modèles plus complexes. Les attributs de l'influence sociale relativement indépendants de l'exposition sociale sont peu importants dans le processus d'évolution du réseau sociosémantique. En fait, comme le montre la Figure 6.9, il y a une relation linéaire très forte entre d'une part l'importance d'un attribut tel que mesuré par son gain d'information et d'autre part la corrélation de ce dernier avec l'exposition sociale. Plus un attribut est caractérisé par un gain d'information important, plus il est corrélé à l'exposition sociale.

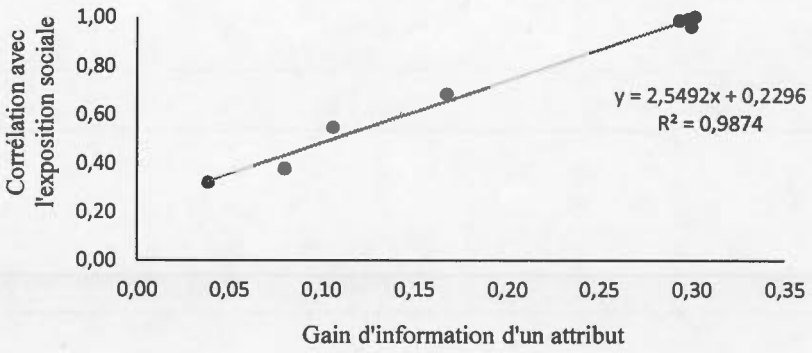

Figure 6.9: Relation entre le gain d'information d'un attribut de l'influence sociale et la corrélation de cet attribut avec l'exposition sociale.

6.2.7 Analyse du résidu de l'influence sociale

La septième analyse porte sur le résidu des modèles d'influence sociale. Nous avons vu jusqu'à maintenant que grâce à la récurrence de plusieurs patrons d'influence sociale dans la dynamique d'un SSS, différents apprenants automatiques sont capables de reconstruire, soit sous la forme d'un arbre de décisions, d'une liste de règles, d'une forêt aléatoire ou d'un modèle probabiliste, un mécanisme vraisemblablement à l'œuvre dans le processus d'évolution des liens d'usage dans un réseau sociosémantique. Selon les modèles, nous avons vu qu'il est possible de prédire en moyenne 88% des transitions d'état du processus, mais qu'un résidu en moyenne de 12% résistait à la reconstruction. Plus spécifiquement, c'est la prédiction de l'usage futur d'un concept (plutôt que la prédiction de la non-utilisation d'un concept) qui résiste à la reconstruction. En d'autres mots, la faiblesse des modèles se situe au niveau de leur Sensibilité (le taux de vrais positifs).

L'objectif de l'analyse du résidu de l'influence sociale est d'apporter des éléments de réponse à cette limite des modèles. L'analyse montrera qu'il y a vraisemblablement un deuxième mécanisme à l'œuvre dans le processus d'évolution du réseau sociosémantique.

Pour le démontrer, nous allons analyser le résidu produit par l'arbre de décisions induit par Γ_{AD} (présenté dans la Figure 6.1). Ce résidu est illustré dans la Figure 6.10. Il est composé de 55 632 exemplaires pour lesquels l'arbre de décisions ne parvient pas à prédire la valeur de la variable dépendante. Il y a 19 305 exemplaires qui représentent des patrons d'influence sociale associés à la non-utilisation d'un concept et il y a 36 327 exemplaires qui représentent des patrons d'influence sociale associés à l'usage futur d'un concept par un agent. En d'autres mots, ce résidu est composé de 19 305 faux négatifs et de presque deux fois plus de faux positifs.

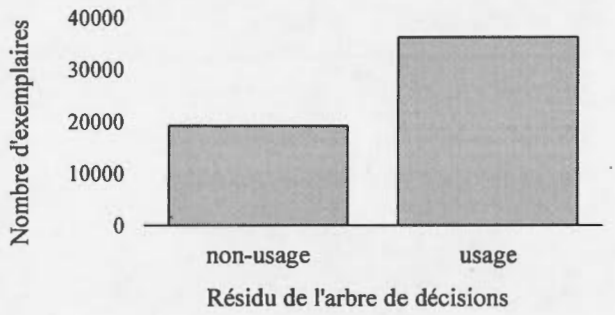

Figure 6.10: Résidu d'un arbre de décisions modélisant l'influence sociale.

Afin de comprendre pourquoi ces exemplaires résistent à la reconstruction, un nouvel arbre de décisions est induit avec Γ_{AD}, mais à partir de ces exemplaires uniquement. L'objectif est d'induire un modèle des exemplaires non expliqués par un modèle de l'influence sociale.

La technique de validation croisée est à nouveau utilisée, mais puisque l'échantillon d'exemplaires est plus petit celui-ci a été divisé en seulement 10 sous-échantillons au lieu de 100. L'arbre est présenté dans la Figure 6.11.

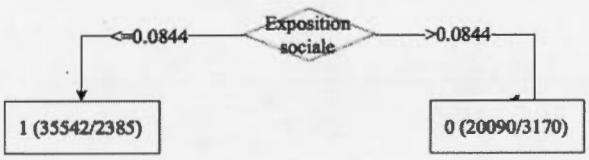

Figure 6.11: Arbre de décisions modélisant le résidu de l'influence sociale.

L'arbre de la Figure 6.11 est extrêmement simple et, à nouveau, c'est l'exposition sociale qui constitue l'attribut central (et le seul) du modèle.

Toutefois, contrairement aux modèles d'approximation du mécanisme d'influence sociale, dans lesquels la relation entre la magnitude de l'exposition sociale et les usages conceptuels était positive, la relation modélisée dans l'arbre de la Figure 6.11 est négative. Comme le montre la sous-arborescence de gauche, un concept sera utilisé à un temps t+1 si la magnitude de l'exposition sociale à un temps t est plus petite ou égale au seuil 0,0844. Selon la deuxième sous-arborescence, un concept ne sera pas utilisé à t+1 si la magnitude de l'exposition sociale dépasse le seuil 0,0844.

Selon ce modèle, l'évolution d'une partie des liens d'usage dans le réseau sociosémantique, plus spécifiquement 12% des transitions, est déterminée par un modèle complètement opposé au modèle de l'influence sociale. Un concept est utilisé si l'exposition sociale est faible et un concept n'est pas utilisé si l'exposition est forte. Le Tableau suivant montre l'adéquation entre cet arbre de décisions et les exemplaires qui forment le résidu de l'influence sociale.

Tableau 6.8: Évaluation de l'adéquation entre l'arbre de décisions de la Figure 6.11 et le résidu de l'influence sociale.

	$\bar{\Gamma}_{AD}(résidu)$
Rand	0,900
Kappa	0,782
Matthews	0,782
Sensibilité	0,923
Spécificité	0,876
Précision +	0,932
Précision -	0,843

Comme le montre la valeur de Rand dans le Tableau 6.8, cet arbre de décisions est capable de prédire 90% des exemplaires formant le résidu de l'influence sociale. La Sensibilité du modèle est de 0,923, sa Spécificité est de 0,876, sa Précision positive est de 0,876 et sa Précision négative est de 0,843. C'est aussi un modèle largement supérieur à la prévalence, comme en témoignent les indices Kappa et Matthews d'une valeur de 0,782.

En somme, cette analyse montre que le résidu de l'influence sociale est hautement structuré. Il représente des transitions d'états dans le processus d'évolution du réseau sociosémantique fortement déterminées par l'exposition sociale, mais par une relation inverse à celle de l'influence sociale. Deux mécanismes contraires, mais complémentaires semblent donc être à l'œuvre dans le processus d'évolution du réseau sociosémantique. Le premier, celui de l'influence sociale, est largement dominant, mais sa reconstruction par apprentissage machine ne permet de prédire que 88% des transitions d'états dans le processus. Le deuxième, moins important, peut aussi être reconstruit par apprentissage machine. Sa modélisation permet de prédire 90% des transitions d'états prédites incorrectement par un modèle de l'influence sociale.

6.3 Retour sur l'objectif de reconstruction

L'objectif de recherche de cette thèse était la reconstruction par apprentissage machine du mécanisme à l'œuvre dans l'évolution du réseau sociosémantique d'un SSS.

Un SSS formant une boîte noire, seulement des données indirectes étaient disponibles sur le fonctionnement de ce mécanisme. Le SSS étudié était la communauté de journalistes du journal The New York Time et les données utilisées pour la reconstruction furent les articles de presse publiés quotidiennement par ces journalistes entre 2002 et 2006. Une autre manière de formuler l'objectif de reconstruction consiste à dire qu'il s'agissait, à partir par l'analyse de ces données textuelles, d'induire un modèle capable de prédire si une journaliste va utiliser ou non un contenu conceptuel particulier dans un article de presse. L'objet de découverte visé par cette recherche était un modèle permettant de calculer des prévisions sur quels concepts allaient être utilisés par quels agents dans la dynamique du système.

Cet objectif était basé sur l'hypothèse que le mécanisme à l'œuvre dans l'évolution du réseau sociosémantique est un mécanisme d'influence sociale entre agents et que l'utilisation ou la non-utilisation d'un concept par un agent est déterminée par des patrons d'influence sociale particuliers. En d'autres mots, l'hypothèse de recherche conjecturait que l'usage au temps t+1 d'un concept c_j par un agent a_i, est déterminé par des patrons d'influence sociale à l'œuvre dans le SSS au temps t.

Quatre types de modèle de l'influence sociale sociale ont été reconstruits via quatre types d'apprenants automatiques : un modèle sous forme d'un arbre de décisions, d'une liste de règles, d'une forêt aléatoire et un modèle probabiliste. L'analyse de ces modèles a montré qu'ils permettaient de prédire en moyenne 88% des exemplaires de l'échantillon. Plus spéciquement, c'était en moyenne 62% des usages conceptuels dans la dynamique du réseau sociosémantique qui pouvaient être prédits par ces modèles (avec une précision de 71%) et 94% des non-usages (avec une précision de 92%). Une

autre manière de résumer ces résultats consiste à dire qu'en moyenne les modèles induits avaient un taux de vrais positifs de 62% et un taux de faux positif de 94%.

L'analyse des modèles a aussi montré qu'ils étaient cohérents avec les données d'apprentissages et que la plupart d'entre eux (trois sur quatre) ne souffraient pas de sur-ajustement et pouvaient par conséquent se généraliser à une base de test. La validation croisée des modèles et l'analyse des courbes d'apprentissage ont aussi montré que ces résultats étaient très robustes et significativement supérieurs à la prévalence (voir les valeurs de Kappa et Matthews).

Par ailleurs, un résidu a persisté à la reconstruction. En moyenne, 12% des exemplaires ne pouvaient pas être expliqués par les différents modèles de l'influence sociale. L'analyse de ce résidu a suggéré qu'un autre mécanisme est vraisemblablement à l'œuvre dans le processus d'évolution du réseau sociosémantique étudié. Ce mécanisme semble fonctionner de manière opposée à l'influence sociale. En effet, pour les exemplaires composant ce résidu d'influence sociale, plus la magnitude de l'exposition sociale d'un concept est grande, plus petite est la probabilité que le concept soit utilisé, et plus l'exposition sociale est petite, plus grande est la probabilité que le concept soit utilisé.

L'objectif de reconstruction a donc été partiellement atteint et l'hypothèse de recherche partiellement corroborée : la reconstruction par apprentissage machine d'un modèle d'influence sociale permet effectivement de prédire l'évolution du réseau sociosémantique d'un SSS, mais pas totalement. Pour reprendre une expression consacrée dans le jargon de l'apprentissage machine, les modèles reconstruits sont approximativement corrects, mais très probables.

Par ailleurs, au-delà de l'objectif de reconstruction, l'analyse des effets de la réduction dimensionnelle et de la colinéarité des attributs de l'influence sociale permettent aussi

de mieux comprendre certains éléments du fonctionnement de l'influence sociale et de la dynamique sociosémantique d'un SSS.

Dans la littérature sur la théorie des réseaux sociaux, la colinéarité dans l'influence sociale est un aspect de la théorie qui a été peu étudié empiriquement. Comme il a été mentionné en conclusion du chapitre deux, traditionnellement dans la théorie des réseaux sociaux, l'influence sociale est modélisée via un seul attribut à la fois : on étudie par exemple l'impact de la contagion ou l'impact du mimétisme, mais rarement l'impact combiné de la contagion et du mimétisme, ou de la contagion et de la déférence. L'interaction entre les différentes dimensions de l'influence sociale est peu comprise. Toutefois, plusieurs chercheurs soupçonnent que ces différents attributs ne sont pas indépendants les uns des autres. Par exemple, plusieurs études ont observé que des agents socialement proches dans un réseau social tendent aussi à occuper des positions sociales similaires en termes d'équivalence structurale (i.e. ils tendent à partager leur voisinage). Par conséquent, plusieurs conjecturent que, bien que les différentes dimensions de l'influence sociale soient toutes conceptuellement distinctes, elles peuvent être empiriquement corrélées et indiscernables (R. S. Burt, 2010a, p. 349; R. T. A. Leenders, 2002, p. 30; Marsden & Friedkin, 1993, p. 133; Rice, 1993, p. 53). Pour certains, notamment Burt (R. S. Burt, 2010b, p. 9), certaines dimensions de l'influence sociale sont aussi plus importantes que d'autres et peuvent les subsumer.

Les résultats d'analyse de la présente recherche apportent des éléments de réponses sur ces enjeux théoriques. Les résultats obtenus montrent en effet que tous les attributs de l'influence sociale sont corrélés positivement et certaines de manière presque parfaite. De plus, les résultats d'analyse montrent que, bien que tous les attributs de l'influence sociale semblent déterminer, mais à des degrés différents, l'évolution du réseau sociosémantique, certains sont beaucoup plus importants que d'autres.

Les résultats d'analyses montrent que l'exposition sociale domine presque entièrement le processus d'évolution du réseau sociosémantique. En effet, il est possible d'induire

un modèle extrêmement simple de l'influence sociale, basé uniquement sur quelques patrons d'exposition sociale. Malgré cette réduction dimensionnelle radicale, ce modèle produit des performances prédictives presque équivalentes à des modèles multidimensionnels beaucoup plus complexes. Ces résultats suggèrent que l'influence sociale, du moins celle à l'œuvre dans le SSS étudié, peut être presque entièrement réduite à quelques patrons élémentaires d'exposition sociale.

Par exemple, l'un de ces patrons est basé sur le seuil 0,0432 de l'exposition sociale (voir l'arbre de la Figure 6.8). Il est probablement le seuil d'exposition sociale le plus important pour la prédiction de la non-utilisation d'un concept par un agent. Ce patron, exprimé sous la forme d'une règle ou d'une sous-arborescence, permet de prédire que tant qu'un concept ne sera pas utilisé par plus de 4% des agents du SSS, il est très peu probable qu'il soit utilisé par un agent à la période suivante. Bien qu'élémentaire, ce patron d'exposition sociale permet pourtant de prédire 74% des exemplaires de l'échantillon avec un taux d'erreur de seulement 5%.

Il y a plusieurs autres patrons d'influence sociale qui ne sont pas basés sur l'exposition sociale et qui permettent de prédire l'évolution du réseau sociosémantique. En fait, tous les attributs de l'influence sociale permettent de prédire les usages et les non-usages des concepts, mais leur pouvoir prédictif est bien moindre. Les patrons basés sur ces autres attributs semblent être beaucoup plus ponctuels et contingents. Peut-être sont-ils spécifiques à certains agents du système ou à certains concepts. Quoi qu'il en soit, les résultats d'analyse suggèrent une forme de hiérarchie entre les patrons d'influence sociale. L'influence sociale fonctionnerait avant tout par exposition sociale, ensuite, plusieurs patrons secondaires, caractérisés par des supports très petits et des taux d'erreur très élevés, semblent permettre de prédire des cas atypiques ou très rares dans l'évolution du réseau sociosémantique.

CHAPITRE VII

CONCLUSION

7.1 Contributions

La principale contribution scientifique de cette recherche consiste en la démonstration qu'il est possible de faire la rétro-ingénierie par apprentissage machine d'un mécanisme sociocognitif à l'œuvre dans l'évolution du réseau sociosémantique d'un SSS.

Il s'agit, au mieux de nos connaissances, de la première fois qu'une telle problématique est étudiée dans le contexte des réseaux sociosémantiques. Ceci rend par conséquent la comparaison des résultats obtenus avec d'autres travaux plus difficile. Il est néanmoins possible de trouver certains travaux pertinents pour la comparaison. Les plus intéressants sont ceux issus d'une problématique de recherche se situant à l'intersection de l'informatique et des sciences sociales et appelée en anglais « the link prediction problem » (Aggarwal, 2015, p. 653; Al Hasan & Zaki, 2011; Han & Kamber, 2006, p. 565; Liben-Nowell & Kleinberg, 2007; Lichtenwalter, Lussier, & Chawla, 2010; Lü & Zhou, 2011; Rattigan & Jensen, 2005).

Cette problématique de recherche est introduite de la manière suivante dans la littérature :

> « Link prediction is the problem of predicting the existence of a link between two entities, based on attributes of the objects and other observed links. Examples include predicting links among actors in social networks, such as predicting friendships; predicting the participation of actors in events [...], such as email, telephone calls and co-authorship; and predicting semantic relationships such as "advisor-of" based on web page links and content [...]. Most often, some links are observed, and one is attempting to predict unobserved links, or there is a temporal aspect: a snapshot of the set of links at time t is given and the goal is to predict the links at time t + 1. This problem is often

> viewed as a simple binary classication problem: for any two potentially linked objects oi and oj, predict whether lij isl or 0. » (Getoor & Diehl, 2005, p. 6)

Prédire l'évolution des liens d'usage entre agents et concepts dans un réseau sociosémantique peut s'inscrire dans le cadre de ce programme de recherche. En comparant les résultats obtenus dans la présente recherche avec ceux issus des travaux dans ce programme, la contribution de cette thèse peut être résumée en trois principaux points.

Premièrement, la contribution de cette thèse se situe au niveau de l'identification d'un objet de recherche original, c'est-à-dire les réseaux sociosémantiques. Ce type de réseaux forme un objet empirique de recherche très récent et actuellement étudié dans les sciences sociales que par une petite communauté de chercheurs (Diesner & Carley, 2010; Mongeau & Saint-Charles, 2014; Monge & Contractor, 2003; Phelps, Heidl, & Wadhwa, 2012; Roth, 2013; Roth & Cointet, 2010; Sieck, Rasmussen, & Smart, 2010). La définition même de l'objet est actuellement un enjeu de recherche important. Cette thèse apporte non seulement des éléments de réponse à ce problème de définition, c'est également la première fois que ce type de réseau est étudié dans le cadre d'une problématique de prédiction de l'évolution des liens.

L'originalité de cette recherche doctorale est liée ensuite à l'hypothèse émise sur la nature du mécanisme sociocognitif à l'œuvre dans la dynamique du réseau sociosémantique. L'évolution d'un réseau sociosémantique est modélisée en termes d'influence sociale. Nous avons vu dans le chapitre deux que la théorie de l'influence sociale est largement utilisée pour expliquer la diffusion ou la propagation de comportement dans un réseau social. Toutefois, en comparaison avec les travaux existants sur la prédiction des liens dans un réseau, l'utilisation de cette théorie est inédite et par conséquent une contribution importante de la présente recherche est d'avoir développée une nouvelle classe de patrons qui peut être utilisés comme variables prédictives.

Finalement, une autre contribution importante de cette recherche est d'avoir modélisé le problème de la prédiction de l'évolution des liens dans un réseau dans un cadre d'apprentissage machine supervisé. La plupart des cadres d'analyse utilisés dans les travaux liés à cette problématique sont basés sur des méthodes non-supervisées (pour une synthèse de ces méthodes, voir (Liben-Nowell & Kleinberg, 2007)). L'utilisation d'apprenants automatiques dans ce domaine relève encore de la preuve de concept (Aggarwal, 2015, p. 653; Al Hasan & Zaki, 2011; Davis, Lichtenwalter, & Chawla, 2011; Lichtenwalter et al., 2010). La recherche réalisée dans cette thèse apporte des évidences supplémentaires que l'apprentissage machine est une perspective très prometteuse pour le domaine.

7.2 Limites et perspectives

La recherche présentée dans cette thèse est une recherche en cours de développement. Les travaux réalisés jusqu'à maintenant furent fondés sur plusieurs choix théoriques et méthodologiques qui ont imposé inévitablement des limites aux résultats obtenus.

7.2.1 D'autres variables prédictives

Un choix déterminant dans cette recherche fut évidemment l'hypothèse de l'influence sociale. Cette hypothèse a déterminé comment fut modélisée la variable indépendante de la problématique de recherche (i.e. les huit attributs de l'influence sociale). Or, comme l'a montré l'analyse des résultats du chapitre six, d'autres mécanismes sont sans doute à l'œuvre dans le processus d'évolution du réseau sociosémantique et des variables supplémentaires permettraient probablement d'augmenter les performances prédictives des modèles.

À cet égard, une première piste de recherche se situe du côté de l'analyse du réseau sémantique d'un SSS. Dans la recherche menée dans cette thèse, le réseau sémantique du SSS n'a jamais été pris en compte dans la reconstruction. Pourtant, il est probable que des variables liées aux relations et aux structures de ce réseau puissent jouer un

rôle important dans la dynamique du système. Prenons par exemple la centralité d'un concept dans le réseau sémantique d'un SSS. Est-ce que cette variable pourrait permettre de prédire les usages conceptuels des agents du système? Par exemple, est-ce que plus un concept est central dans le réseau sémantique d'un SSS plus grande est la probabilité qu'il soit utilisé par un agent dans le futur? Des expérimentations encore au stade exploratoire laisse croire que c'est le cas et suggèrent également que d'autres variables liées au réseau sémantique d'un SSS — notamment la transitivité des concepts — sont aussi associées à l'évolution des liens d'usage entre agents et concepts dans un SSS.

7.2.2 Les apprenants automatiques

Étroitement lié au choix de la variable indépendante de la problématique de reconstruction, le choix des apprenants automatiques fut aussi une décision importante dans cette recherche. Quatre apprenants automatiques, par ailleurs très classiques en informatique, ont été utilisés. Toutefois, des centaines d'autres auraient été possibles. Est-ce que d'autres apprenants automatiques auraient permis d'induire de meilleurs modèles ayant des performances prédictives supérieures? C'est une question empirique qui demande davantage d'expérimentations. À nouveau, des expérimentations exploratoires avec d'autres apprenants, notamment un séparateur à vastes marges et un perceptron multicouches, suggèrent jusqu'à maintenant que ce n'est pas le cas. Par conséquent, l'avancement de la recherche se situe probablement davantage du côté de la sélection des variables prédictives que du côté de la sélection des apprenants automatiques.

7.2.3 La modélisation des contenus conceptuels

Dans cette recherche, les contenus conceptuels ont été modélisés dans un cadre théorique particulier, soit celui de la sémantique vectorielle. Les contenus conceptuels ont été modélisés sous la forme de régions dans un espace sémantique et le

partitionnement en régions de cet espace été réalisé avec un algorithme particulier, c'est-à-dire l'algorithme des k-moyennes.

D'autres choix de modélisation auraient été possibles. Par exemple, les contenus conceptuels auraient pu être modélisés dans le cadre de la sémantique latente (Thomas K. Landauer et al., 1998), de l'analyse conceptuelle formelle (Cimiano, Hotho, & Staab, 2005) ou en utilisant un thésaurus comme WordNet (G. A. Miller, 1995). L'impact de la modélisation des contenus conceptuels reste une question de recherche ouverte.

7.2.4 La dérive du modèle

Un dernier point important de souligner avant de conclure concerne le problème de la « dérive du modèle » (appelé en anglais « concept drift »). Dans le domaine de l'apprentissage machine appliqué sur des séries temporelles, ce problème fait référence à l'évolution inattendue dans le temps de la relation entre une variable indépendante et une variable dépendante (Aggarwal, 2015, p. 390). Un modèle « dérive » lorsqu'il reste inchangé alors que de nouvelles données empiriques ont de nouvelles propriétés que ce dernier ne peut pas prédire.

Dans le cadre de la recherche effectuée dans cette thèse, ce problème renvoie à la possibilité que la relation entre la magnitude des différents attributs de l'influence sociale et l'évolution du réseau sociosémantique puisse changer avec le temps. Par exemple, il se pourrait que pour une période donnée, des seuils très bas d'exposition sociale soient déterminants dans l'évolution des liens d'usages, mais qu'à la période suivante, ces seuils changent soudainement et qu'ils soient beaucoup plus élevés. Il se pourrait par exemple que le résidu de l'influence sociale soit en réalité causé par une « dérive » du modèle. Il se pourrait aussi qu'à différentes périodes de l'évolution d'un réseau sociosémantique différents patrons d'influence sociale soient déterminants. La

« dérive » possible des modèles d'influence sociale analysés dans cette recherche n'a pas été contrôlée. Elle reste donc elle aussi une question de recherche ouverte.

BIBLIOGRAPHIE

Åberg, Y. (2009). The contagiousness of divorce. In P. Hedström & P. Bearman, *The Oxford Handbook of analytical sociology* (pp. 342–364). The Oxford handbook of analytical sociology.

Aggarwal, C. C. (2015). *Data Mining: The Textbook*. Springer.

Agrawal, A., Kapur, D., & McHale, J. (2008). How do spatial and social proximity influence knowledge flows? Evidence from patent data. *Journal of Urban Economics*, *64*(2), 258–269.

Aha, D. W., Kibler, D., & Albert, M. K. (1991). Instance-based learning algorithms. *Machine Learning*, *6*(1), 37–66.

Albert, R., & Barabási, A.-L. (2002). Statistical mechanics of complex networks. *Reviews of Modern Physics*, *74*(1), 47.

Alexander, C., Piazza, M., Mekos, D., & Valente, T. (2001). Peers, schools, and adolescent cigarette smoking. *Journal of Adolescent Health*, *29*(1), 22–30.

Al Hasan, M., & Zaki, M. J. (2011). A survey of link prediction in social networks. In *Social network data analytics* (pp. 243–275). Springer.

Andrews, R., Diederich, J., & Tickle, A. B. (1995). Survey and critique of techniques for extracting rules from trained artificial neural networks. *Knowledge-Based Systems*, *8*(6), 373–389.

Asch, S. E. (1951). Effects of group pressure upon the modification and distortion of judgments. *Groups, Leadership, and Men. S*, 222–236.

Ashby, W. R. (1957). An introduction to cybernetics.

Atran, S., Ross, N. O., & Medin, D. L. (2005). The Cultural Mind: Environmental Decision Making and Cultural Modeling Within and Across Populations. *Psychological Review*, *112*(4), 744–776.

Baroni, M., & Lenci, A. (2010). Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, *36*(4), 673–721.

Barrat, A., Barthelemy, M., & Vespignani, A. (2008). *Dynamical processes on complex networks* (Vol. 1). Cambridge University Press Cambridge.

Bass, F. M. (1969). A new product growth for model consumer durables. *Management Science*, *15*(5).

Baxter, J. (2000). A model of inductive bias learning. *J. Artif. Intell. Res.(JAIR)*, *12*, 149–198.

Beaver, D., & Rosen, R. (1978). Studies in scientific collaboration: Part I. The professional origins of scientific co-authorship. *Scientometrics*, *1*(1), 65–84.

Benzecri, J.-P. (1973). *L'analyse des données, tome II: l'analyse des correspondances*. Paris: Dunod.

Berkhin, P. (2006). A survey of clustering data mining techniques. In J. Kogan, C. Nicholas, & M. Teboulle (Eds.), *Grouping multidimensional data* (pp. 25–71). Springer.

Bikhchandani, S., Hirshleifer, D., & Welch, I. (1998). Learning from the behavior of others: Conformity, fads, and informational cascades. *The Journal of Economic Perspectives*, 151–170.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.

Blythe, R. A., & Croft, W. (2012). S-curves and the mechanisms of propagation in language change. *Language*, *88*(2), 269–304.

Blythe, R. A., & Croft, W. A. (2009). The speech community in evolutionary language dynamics. *Language Learning*, *59*(s1), 47–63.

Boccaletti, S., Latora, V., Moreno, Y., Chavez, M., & Hwang, D.-U. (2006). Complex networks: Structure and dynamics. *Physics Reports*, *424*(4), 175–308.

Bohner, G., & Dickel, N. (2011). Attitudes and Attitude Change. *Annual Review of Psychology*, *62*, 391–417.

Borgatti, S. P., Brass, D. J., & Halgin, D. S. (2014). Social network research: Confusions, criticisms, and controversies. *Research in the Sociology of Organizations*, *40*, 1–29.

Borgatti, S. P., & Everett, M. G. (1992). Notions of position in social network analysis. *Sociological Methodology*, *22*(1), 1–35.

Borgatti, S. P., Everett, M. G., & Johnson, J. C. (2013). *Analyzing social networks*. SAGE Publications Limited.

Borgatti, S. P., & Foster, P. C. (2003). The network paradigm in organizational research: A review and typology. *Journal of Management*, *29*(6), 991–1013.

Borgatti, S. P., & Halgin, D. S. (2011). On network theory. *Organization Science*, *22*(5), 1168–1181.

Borgatti, S. P., Mehra, A., Brass, D. J., & Labianca, G. (2009). Network analysis in the social sciences. *Science*, *323*(5916), 892–895.

Borge-Holthoefer, J., & Arenas, A. (2010). Semantic networks: Structure and dynamics. *Entropy*, *12*(5), 1264–1302.

Bothner, M. S. (2003). Competition and Social Influence: The Diffusion of the Sixth-Generation Processor in the Global Computer Industry1. *American Journal of Sociology*, *108*(6), 1175–1210.

Bradley, P. S., & Fayyad, U. M. (1998). Refining Initial Points for K-Means Clustering. In *ICML* (Vol. 98, pp. 91–99). Citeseer.

Brass, D. J. (1984). Being in the right place: A structural analysis of individual influence in an organization. *Administrative Science Quarterly*, 518–539.

Breiman, L. (2001a). Random forests. *Machine Learning*, *45*(1), 5–32.

Breiman, L. (2001b). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, *16*(3), 199–231.

Bronner, G. (2003). *L'empire des croyances*. Paris: PUF.

Bullinaria, J. A., & Levy, J. P. (2007). Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior Research Methods*, *39*(3), 510–526.

Burgess, C., Livesay, K., & Lund, K. (1998). Explorations in context space: Words, sentences, discourse. *Discourse Processes*, *25*(2-3), 211–257.

Burns, P. R. (2013). MorphAdorner v2: A Java Library for the Morphological Adornment of English Language Texts. Northwestern University. Retrieved from https://morphadorner.northwestern.edu/morphadorner/download/morphadorner.pdf

Burt, R. S. (1987). Social contagion and innovation: Cohesion versus structural equivalence. *American Journal of Sociology*, 1287–1335.

Burt, R. S. (1999). The social capital of opinion leaders. *The Annals of the American Academy of Political and Social Science*, *566*(1), 37–54.

Burt, R. S. (2010a). *Neighbor networks: Competitive advantage local and personal*. Oxford University Press.

Burt, R. S. (2010b). The shadow of other people: Socialization and social comparison in marketing. *The Connected Customer: The Changing Nature of Consumer and Business Markets*, 217–256.

Caliński, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics-Theory and Methods*, *3*(1), 1–27.

Carrington, P. J., Scott, J., & Wasserman, S. (2005). *Models and methods in social network analysis* (Vol. 28). Cambridge university press.

Cavnar, W. B., & Trenkle, J. M. (1994). N-gram-based text categorization. *Ann Arbor MI*, *48113*(2), 161–175.

Charles, W. G. (2000). Contextual correlates of meaning. *Applied Psycholinguistics*, *21*(04), 505–524.

Chartier, J.-F., & Meunier, J.-G. (2011). Text Mining Methods for Social Representation Analysis in large corpora. *Papers on Social Representations*, *20*, 37.1–37.47.

Christakis, N. A., & Fowler, J. H. (2007). The spread of obesity in a large social network over 32 years. *New England Journal of Medicine*, *357*(4), 370–379.

Cialdini, R. B., & Goldstein, N. J. (2004). Social influence: Compliance and conformity. *Annu. Rev. Psychol.*, *55*, 591–621.

Cialdini, R. B., & James, L. (2009). *Influence: Science and practice* (Vol. 4). Pearson education Boston, MA.

Cialdini, R. B., & Mortensen, C. R. (2008). Social influence. In S. F. Davis & W. Buskist (Eds.), *21st century psychology: a reference handbook* (Vol. 2, pp. 123–133). Sage.

Cialdini, R. B., & Trost, M. R. (1998). Social influence: Social norms, conformity and compliance.

Cimiano, P., Hotho, A., & Staab, S. (2005). Learning Concept Hierarchies from Text Corpora using Formal Concept Analysis. *J. Artif. Intell. Res.(JAIR)*, *24*, 305–339.

Clark, A. (1990). Connectionism, competence, and explanation. *The British Journal for the Philosophy of Science*, *41*(2), 195–222.

Clark, H. H. (1996). *Using language*. Cambridge: Cambridge University Press.

Coleman, J., Katz, E., & Menzel, H. (1957). The diffusion of an innovation among physicians. *Sociometry*, 253–270.

Corman, S. R., Kuhn, T., McPhee, R. D., & Dooley, K. J. (2002). Studying Complex Discursive Systems. *Human Communication Research*, *28*(2), 157–206.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297.

Cowan, R., & Jonard, N. (2004). Network structure and the diffusion of knowledge. *Journal of Economic Dynamics and Control*, *28*(8), 1557–1575.

Croft, W. (2000). *Explaining language change: An evolutionary approach*. Pearson Education.

Davis, D., Lichtenwalter, R., & Chawla, N. V. (2011). Multi-relational link prediction in heterogeneous information networks. In *Advances in Social Networks Analysis and Mining (ASONAM), 2011 International Conference on* (pp. 281–288). IEEE.

Dennett, D. C. (1995). Cognitive science as reverse engineering several meanings of "Top-down" and "Bottom-up." *Studies in Logic and the Foundations of Mathematics*, *134*, 679–689.

Diesner, J., & Carley, K. M. (2005). Revealing social structure from texts: meta-matrix text analysis as a novel method for network text analysis. *Causal Mapping for Information Systems and Technology Research: Approaches, Advances, and Illustrations*, 81–108.

Diesner, J., & Carley, K. M. (2010). A methodology for integrating network theory and topic modeling and its application to innovation diffusion. In *Social Computing (SocialCom), 2010 IEEE Second International Conference on* (pp. 687–692). IEEE.

DiMaggio, P., & Garip, F. (2012). Network effects and social inequality. *Annual Review of Sociology*, *38*, 93–118.

DiMaggio, P. J., & Powell, W. W. (1983). The Iron Cage Revisited: Institutional Isomorphism and Collective Rationality in Organizational Fields. *American Sociological Review*, *48*(2), 147–160.

Dodds, P. S., & Watts, D. J. (2004). Universal behavior in a generalized model of contagion. *Physical Review Letters*, *92*(21), 218701.

Dodds, P. S., & Watts, D. J. (2005). A generalized model of social and biological contagion. *Journal of Theoretical Biology*, *232*(4), 587–604.

Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM*, *55*(10), 78–87.

Easley, D., & Kleinberg, J. (2010). *Networks, crowds, and markets* (Vol. 8). Cambridge Univ Press.

Ellis, D., Furner-Hines, J., & Willett, P. (1993). Measuring the degree of similarity between objects in text retrieval systems. *Perspectives in Information Management*, *3*(2), 128–149.

Erickson, B. H. (1988). The relational basis of attitudes. *Social Structures: A Network Approach*, *99*, 121.

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd* (Vol. 96, pp. 226–231).

Feldman, R., & Sanger, J. (2007). *The Text Mining Handbook*. New York: Cambridge University Press.

Festinger, L. (1954). A theory of social comparison processes. *Human Relations*, *7*(2), 117–140.

Figueroa, R. L., Zeng-Treitler, Q., Kandula, S., & Ngo, L. H. (2012). Predicting sample size required for classification performance. *BMC Medical Informatics and Decision Making*, *12*(1), 8.

Fitch, W. T. (2005). The evolution of language: a comparative review. *Biology and Philosophy*, *20*(2), 193–203.

Fortin, C., & Rousseau, R. (2005). *Psychologie cognitive: une approche de traitement de l'information*. Télé-université.

Frank, E., & Witten, I. H. (1998). Generating Accurate Rule Sets Without Global Optimization. In *Proceedings of the Fifteenth International Conference on Machine Learning* (pp. 144–151). Morgan Kaufmann Publishers Inc.

Freeman, L. C. (1979). Centrality in social networks conceptual clarification. *Social Networks*, *1*(3), 215–239.

Friedkin, N. E. (1991). Theoretical foundations for centrality measures. *American Journal of Sociology*, 1478–1504.

Friedkin, N. E. (2006). *A structural theory of social influence* (Vol. 13). Cambridge University Press.

Friedkin, N. E., & Johnsen, E. C. (1997). Social positions in influence networks. *Social Networks*, *19*(3), 209–222.

Friedkin, N. E., & Johnsen, E. C. (1999). Social influence networks and opinion change. *Advances in Group Processes*, *16*(1), 1–29.

Fujimoto, K., & Valente, T. W. (2012). Social network influences on adolescent substance use: disentangling structural equivalence from cohesion. *Social Science & Medicine*, *74*(12), 1952–1960.

Gai, P., & Kapadia, S. (2010). Contagion in financial networks. In *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences* (p. rspa20090410). The Royal Society.

Galaskiewicz, J., & Burt, R. S. (1991). Interorganization contagion in corporate philanthropy. *Administrative Science Quarterly*, 88–105.

Galaskiewicz, J., & Wasserman, S. (1989). Mimetic processes within an interorganizational field: An empirical test. *Administrative Science Quarterly*, 454–479.

Gärdenfors, P. (2000). *Conceptual Spaces: The Geometry of Thought*. Cambridge: MIT Press.

Gärdenfors, P. (2014). *The Geometry of Meaning: Semantics Based on Conceptual Spaces*. MIT Press.

Getoor, L., & Diehl, C. P. (2005). Link mining: a survey. *ACM SIGKDD Explorations Newsletter*, *7*(2), 3–12.

Goel, S., & Goldstein, D. G. (2013). Predicting individual behavior with social networks. *Marketing Science*, *33*(1), 82–93.

Goldman, A. I. (1999). *Knowledge in a Social World*. Clarendon Press.

Gordon, D. F., & Desjardins, M. (1995). Evaluation and selection of biases in machine learning. *Machine Learning*, *20*(1-2), 5–22.

Grabisch, M., & Rusinowska, A. (2013). A model of influence based on aggregation functions. *Mathematical Social Sciences*, *66*(3), 316–330.

Granovetter, M. (1973). The Strength of Weak Ties. *American Journal of Sociology*, *78*(6), 1360–13.

Granovetter, M. (1978). Threshold models of collective behavior. *American Journal of Sociology*, *83*(6), 1420.

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review*, *114*(2), 211.

Griskevicius, V., Goldstein, N. J., Mortensen, C. R., Sundie, J. M., Cialdini, R. B., & Kenrick, D. T. (2009). Fear and loving in Las Vegas: Evolution, emotion, and persuasion. *Journal of Marketing Research*, *46*(3), 384–395.

Gruhl, D., Guha, R., Liben-Nowell, D., & Tomkins, A. (2004). Information diffusion through blogspace. In *Proceedings of the 13th international conference on World Wide Web* (pp. 491–501). ACM.

Gumperz, J. J. (1969). The speech community (pp. 381–386). New York: Macmillan.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: an update. *ACM SIGKDD Explorations Newsletter*, *11*(1), 10–18.

Han, J., & Kamber, M. (2006). *Data Mining: Concepts and Techniques* (2e ed.). San Francisco: Morgan Kaufmann.

Harman, G., & Kulkarni, S. (2007). *Reliable Reasoning: Induction and Statistical Learning Theory*. Cambridge: A Bradford Book.

Harnad, S. (1994). Levels of functional equivalence in reverse bioengineering. *Artificial Life*, *1*(3), 293–301.

Harrison, J. R., & Carroll, G. R. (2002). The dynamics of cultural influence networks. *Computational & Mathematical Organization Theory*, *8*(1), 5–30.

Harris, Z. S. (1951). *Methods in structural linguistics.* Chicago & London: The University of Chicago Press.

Harris, Z. S. (1954). Distributional structure. *Word*, *10*(23), 146–162.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning*. New York, NY: Springer New York. Retrieved from http://link.springer.com/10.1007/978-0-387-84858-7

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *Knowledge and Data Engineering, IEEE Transactions on*, *21*(9), 1263–1284.

Hindle, D. (1990). Noun classification from predicate-argument structures. In *Proceedings of the 28th annual meeting on Association for Computational Linguistics* (pp. 268–275). Association for Computational Linguistics.

Holland, P. W., & Leinhardt, S. (1977). A dynamic model for social networks†. *Journal of Mathematical Sociology*, *5*(1), 5–20.

Holte, R. C. (1993). Very simple classification rules perform well on most commonly used datasets. *Machine Learning*, *11*(1), 63–90.

Hothorn, T., Leisch, F., Zeileis, A., & Hornik, K. (2005). The design and analysis of benchmark experiments. *Journal of Computational and Graphical Statistics*, *14*(3), 675–699.

Hutchins, E. (1995). *Cognition in the Wild.* Cambridge: The MIT Press.

Ibarra, H., & Andrews, S. B. (1993). Power, social influence, and sense making: Effects of network centrality and proximity on employee perceptions. *Administrative Science Quarterly*, 277–303.

Iyengar, R., Van den Bulte, C., & Valente, T. W. (2011). Opinion leadership and social contagion in new product diffusion. *Marketing Science*, *30*(2), 195–212.

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, *31*(8), 651–666.

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: a review. *ACM Computing Surveys (CSUR)*, *31*(3), 264–323.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. Springer.

Katz, J. S., & Martin, B. R. (1997). What is research collaboration? *Research Policy*, *26*(1), 1–18.

Keller, R. (1994). *On language change: The invisible hand in language*. New York: Routledge.

Kennedy, J., & Eberhart, R. C. (2001). *Swarm Intelligence*. San Francisco: Morgan Kaufmann Publisher.

Kiela, D., & Clark, S. (2014). A systematic study of semantic vector space model parameters. In *Proceedings of the 2nd Workshop on Continuous Vector Space Models and their Compositionality (CVSC) at EACL* (pp. 21–30).

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, *43*(1), 59–69.

Kotsiantis, S., Kanellopoulos, D., & Pintelas, P. (2006). Handling imbalanced datasets: A review. *GESTS International Transactions on Computer Science and Engineering*, *30*(1), 25–36.

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*. Springer.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review; Psychological Review*, *104*(2), 211.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, *25*(2-3), 259–284.

Langley, P., Iba, W., & Thompson, K. (1992). An analysis of Bayesian classifiers. In *AAAI* (Vol. 90, pp. 223–228).

Larsen, K. R., & Monarchi, D. E. (2004). A mathematical approach to categorization and labeling of qualitative data: The latent categorization method. *Sociological Methodology*, *34*(1), 349–392.

Last, M., Klein, Y., & Kandel, A. (2001). Knowledge discovery in time series databases. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, *31*(1), 160–169.

Latané, B. (1981). The psychology of social impact. *American Psychologist*, *36*(4), 343.

Latané, B. (1996). Dynamic social impact: The creation of culture by communication. *Journal of Communication*, *46*(4), 13–25.

Latour, B. (2007). Beware, your imagination leaves digital traces. *Times Higher Literary Supplement*, *6*(4), 2007.

Laudel, G. (2002). What do we measure by co-authorships? *Research Evaluation*, *11*(1), 3–15.

Lazega, E. (1998). *Réseaux sociaux et structures relationnelles*. Presses universitaires de France.

Lazer, D., Rubineau, B., Chetkovich, C., Katz, N., & Neblo, M. (2010). The coevolution of networks and political attitudes. *Political Communication*, *27*(3), 248–274.

Lebart, L., Piron, M., & Steiner, J.-F. (2003). La sémiométrie. *Dunod, Paris*.

Lebart, S., & Salem, A. (1994). *Statistique textuelle*. Paris: Dunod.

Leenders, R. (1997). Longitudinal behavior of network structure and actor attributes: modeling interdependence of contagion and selection. *Evolution of Social Networks*, *1*.

Leenders, R. T. A. (2002). Modeling social influence through network autocorrelation: constructing the weight matrix. *Social Networks*, *24*(1), 21–47.

Lemaire, B., & Denhière, G. (2006). Effects of High-Order Co-occurrences on Word Semantic Similarity. *Current Psychology Letters. Behaviour, Brain & Cognition*, (18, Vol. 1, 2006).

Lemaire, B., & Dessus, P. (2003). Modèles cognitifs issus de l'Analyse de la sémantique latente. *Cahiers Romans de Sciences Cognitives*, *1*(1), 55–74.

Lenci, A. (2008). Distributional semantics in linguistic and cognitive research. *Italian Journal of Linguistics*, *20*(1), 1–31.

Lerman, K., & Ghosh, R. (2010). Information Contagion: An Empirical Study of the Spread of News on Digg and Twitter Social Networks. *ICWSM*, *10*, 90–97.

Leskovec, J., Adamic, L. A., & Huberman, B. A. (2007). The dynamics of viral marketing. *ACM Transactions on the Web (TWEB)*, *1*(1), 5.

Liben-Nowell, D., & Kleinberg, J. (2007). The link-prediction problem for social networks. *Journal of the American Society for Information Science and Technology*, *58*(7), 1019–1031.

Lichtenwalter, R. N., Lussier, J. T., & Chawla, N. V. (2010). New perspectives and methods in link prediction. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 243–252). ACM.

Lin, D., & Pantel, P. (2002). Concept discovery from text. In *Proceedings of the 19th international conference on Computational linguistics-Volume 1* (pp. 1–7). Association for Computational Linguistics.

López-Pintado, D. (2008). Diffusion in complex social networks. *Games and Economic Behavior*, *62*(2), 573–590.

Lopez-Pintado, D., & Watts, D. J. (2008). Social influence, binary decisions and collective dynamics. *Rationality and Society*, *20*(4), 399–443.

Lü, L., & Zhou, T. (2011). Link prediction in complex networks: A survey. *Physica A: Statistical Mechanics and Its Applications*, *390*(6), 1150–1170.

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers*, *28*(2), 203–208.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability* (Vol. 1, pp. 281–297). Oakland, CA, USA.

Manning, C., Raghavan, P., & Schütze, H. (2008a). *Introduction to Information Retrieval*. Cambridge: Cambridge University Press.

Manning, C., Raghavan, P., & Schütze, H. (2008b). *Introduction to Information Retrieval*. Cambridge: Cambridge University Press.

Manning, C., & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. Cambridge: MIT Press.

Marsden, P. V., & Friedkin, N. E. (1993). Network studies of social influence. *Sociological Methods & Research*, *22*(1), 127–151.

McClelland, J. L., Rumelhart, D. E., & Group, P. R. (1986). *Parallel distributed processing. Explorations in the Microstructure of Cognition: Psychological and Biological Models* (Vol. 2). Cambridge: The MIT Press.

Mcnamee, P., & Mayfield, J. (2004). Character n-gram tokenization for European language text retrieval. *Information Retrieval*, *7*(1-2), 73–97.

McQuail, D. (1987). *Mass communication theory: An introduction* . Sage Publications, Inc.

Mercier, H., & Sperber, D. (2011). Why do humans reason? Arguments for an argumentative theory. *Behavioral and Brain Sciences*, *34*(2), 57.

Meunier, J.-G., Forest, D., & Biskri, I. (2005). Classification and categorization in computer assisted reading and analysis of texts. In H. Cohen & C. Lefebvre, *Handbook of Categorization in Cognitive Science* (pp. 955–978). Elsevier.

Meyer, G. W. (1994). Social information processing and social networks: A test of social influence mechanisms. *Human Relations*, *47*(9), 1013–1047.

Michalski, R. S. (1983). *A theory and methodology of inductive learning*. Springer.

Miller, G. A. (1995). WordNet: a lexical database for English. *Communications of the ACM, 38*(11), 39–41.

Miller, J. H., & Page, S. E. (2007). *Complex adaptive systems. Introduction to computational models of social life*. Princeton: Princeton University Press.

Milligan, G. W., & Cooper, M. C. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika, 50*(2), 159–179.

Mitchell, J., & Lapata, M. (2010). Composition in distributional models of semantics. *Cognitive Science, 34*(8), 1388–1429.

Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill Science/Engineering/Math.

Mitchell, T. M. (2006). *The discipline of machine learning* (Rapport technique No. 17). Carnegie Mellon University, School of Computer Science, Machine Learning Department.

Mizruchi, M. S. (1989). Similarity of political behavior among large American corporations. *American Journal of Sociology*, 401–424.

Mizruchi, M. S. (1993). Cohesion, equivalence, and similarity of behavior: a theoretical and empirical assessment. *Social Networks, 15*(3), 275–307.

Mongeau, P., & Saint-Charles, J. (2014). Réseaux sociaux et réseaux sociosémantiques et phénomènes de communication. *Communiquer. Revue de Communication Sociale et Publique*, (12), 1–7.

Monge, P. R., & Contractor, N. S. (2003). *Theories of Communication Networks*. Oxford: Oxford University Press.

Moody, J. (2009). Network Dynamics. In P. Hedström & P. Bearman (Eds.), *Oxford Handbook of Analytical Sociology*. Oxford University Press.

Moscovici, S., Sherrard, C., & Heinz, G. (1976). *Social influence and social change* (Vol. 10). Academic Press London.

Mucchielli, A. (2009). *L'art d'influencer*. Armand Colin.

Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective*. The MIT Press.

Newman, M. E. (2003). The structure and function of complex networks. *SIAM Review*, *45*(2), 167–256.

Osgood, C. E. (1952). The nature and measurement of meaning. *Psychological Bulletin*, *49*(3), 197.

Osgood, C. E. (1964). Semantic differential technique in the comparative study of Cultures1. *American Anthropologist*, *66*(3), 171–200.

Osgood, C. E., Suci, G. J., & Tannenbaum, P. H. (1957). *The measurement of meaning*. University of Illinois Press.

Patrick, P. L. (2002). The speech community. In J. K. Chambers, P. Trudgill, & Schilling-Estes (Eds.), *Handbook of language variation and change*. Oxford: Blackwell.

Petty, R. E., & Brinol, P. (2010). Attitude change. *Advanced Social Psychology: The State of the Science*, 217–259.

Phelps, C., Heidl, R., & Wadhwa, A. (2012). Knowledge, networks, and knowledge networks a review and research agenda. *Journal of Management*, *38*(4), 1115–1166.

Porter, M. F. (1980). An algorithm for suffix stripping. *Program: Electronic Library and Information Systems*, *14*(3), 130–137.

Powers, D. M. (2011). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *Journal of Machine Learning Technologies*, *2*(1), 37–63.

Powers, D. M. (2012). The problem with kappa. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 345–355). Association for Computational Linguistics.

Purandare, A., & Pedersen, T. (2004). Word sense discrimination by clustering contexts in vector and similarity spaces. In *Proceedings of the Conference on Computational Natural Language Learning* (Vol. 72). Boston.

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, *1*(1), 81–106.

Quinlan, J. R. (1993). *C 4.5: Programs for machine learning*. Morgan Kaufmann.

Quinlan, J. R. (1996). Improved use of continuous attributes in C4. 5. *Journal of Artificial Intelligence Research*, 77–90.

Rajman, M., & Lebart, L. (1998). Similarités pour données textuelles. In *4th International Conference on Statistical Analysis of Textual Data (JADT'98), Nice, France* (pp. 545–555).

Rattigan, M. J., & Jensen, D. (2005). The case for anomalous link discovery. *ACM SIGKDD Explorations Newsletter*, *7*(2), 41–47.

Rice, R. E. (1993). Using network concepts to clarify sources and mechanisms of social influence. *Progress in Communication Sciences*, *12*, 43–62.

Richerson, P. J., & Boyd, R. (2005). *Not By Genes Alone: How Culture Transformed Human Evolution*. Chicago: The University of Chicago Press.

Riordan, B., & Jones, M. N. (2011). Redundancy in perceptual and linguistic experience: Comparing feature-based and distributional models of semantic representation. *Topics in Cognitive Science*, *3*(2), 303–345.

Rish, I. (2001). An empirical study of the naive Bayes classifier. In *IJCAI 2001 workshop on empirical methods in artificial intelligence* (Vol. 3, pp. 41–46). IBM New York.

Robins, G., Pattison, P., & Elliott, P. (2001). Network models for social influence processes. *Psychometrika*, *66*(2), 161–189.

Rogers, E. M. (2003). *Diffusion of Innovations* (5th edition). Free Press.

Rokach, L., & Maimon, O. (2005). Decision trees. In *Data Mining and Knowledge Discovery Handbook* (pp. 165–192). Springer.

Rolfe, M. (2009). Conditional choice. In P. Hedström & P. Bearman (Eds.), *The Oxford Handbook of analytical sociology* (pp. 419–446). Oxford University Press.

Roth, C. (2008a). Coévolution des auteurs et des concepts dans les réseaux épistémiques: le cas de la communauté «zebrafish». *Revue Française de Sociologie*, *49*(3), 523–558.

Roth, C. (2008b). Réseaux épistémiques: formaliser la cognition distribuée. *Sociologie Du Travail*, *50*(3), 353–371.

Roth, C. (2013). Socio-Semantic Frameworks. *Advances in Complex Systems*, *16*(04n05).

Roth, C., & Cointet, J. P. (2010). Social and semantic coevolution in knowledge networks. *Social Networks*, *32*(1), 16–29.

Ryan, B., & Gross, N. C. (1943). The Diffusion of Hybrid Seed Corn in Two Iowa Communities. *Rural Sociology*, (8), 15–24.

Sahlgren, M. (2006a). *The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces*. Stockholm, Stockholm.

Sahlgren, M. (2006b). *The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces*. Stockholm, Stockholm.

Salton, G., Wong, A., & Yang, C.-S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, *18*(11), 613–620.

Sammut, C., & Webb, G. I. (2011). *Encyclopedia of machine learning*. Springer Science & Business Media.

Sandhaus, E. (2008). The new york times annotated corpus. *Linguistic Data Consortium, Philadelphia*, *6*(12), e26752.

Schelling, T. C. (2006). *Micromotives and macrobehavior*. WW Norton & Company.

Schütze, H. (1992). Dimensions of meaning. In *Proceedings of the 1992 ACM/IEEE Conference on Supercomputing* (pp. 787–796). IEEE Computer Society Press.

Schütze, H., & Pedersen, J. (1993). A vector model for syntagmatic and paradigmatic relatedness. In *Proceedings of the 9th Annual Conference of the UW Centre for the New OED and Text Research* (pp. 104–113). Citeseer.

Scott, J. (2013). *SOCIAL NETWORK ANALYSIS*.

Shalizi, C. R., & Thomas, A. C. (2011). Homophily and contagion are generically confounded in observational social network studies. *Sociological Methods & Research*, *40*(2), 211–239.

Sibson, R. (1973). SLINK: an optimally efficient algorithm for the single-link cluster method. *The Computer Journal*, *16*(1), 30–34.

Sieck, W. R., Rasmussen, L. J., & Smart, P. (2010). Cultural network analysis: A cognitive approach to cultural modeling. *Network Science for Military Coalition Operations: Information Extraction and Interaction*, 237–255.

Snijders, T. (2011). Network Dynamics. *The SAGE Handbook of Social Network Analysis*, 501–513.

Snijders, T. A., Van de Bunt, G. G., & Steglich, C. E. (2010). Introduction to stochastic actor-based models for network dynamics. *Social Networks*, *32*(1), 44–60.

Sorenson, O., Rivkin, J. W., & Fleming, L. (2006). Complexity, networks and knowledge flow. *Research Policy*, *35*(7), 994–1017.

Sowa, J. F. (2006). Semantic networks. *Encyclopedia of Cognitive Science*.

Steglich, C., Snijders, T. A., & Pearson, M. (2010). Dynamic networks and behavior: Separating selection from influence. *Sociological Methodology*, *40*(1), 329–393.

Steinbach, M., Karypis, G., & Kumar, V. (2000). A comparison of document clustering techniques. In *KDD workshop on text mining* (Vol. 400, pp. 525–526). Boston.

Steinley, D. (2006). K-means clustering: a half-century synthesis. *British Journal of Mathematical and Statistical Psychology*, *59*(1), 1–34.

Steyvers, M., & Tenenbaum, J. B. (2010). The Large-Scale Structure of Semantic Networks: Statistical Analyses and a Model of Semantic Growth. *Cognitive Science*, *29*(1), 41–78.

Sundie, J. M., Cialdini, R. B., Griskevicius, V., & Kenrick, D. T. (2006). Evolutionary social influence.

Sundie, J. M., Cialdini, R. B., Griskevicius, V., & Kenrick, D. T. (2012). The world's (truly) oldest profession: Social influence in evolutionary perspective. *Social Influence*, *7*(3), 134–153.

Surowiecki, J. (2005). *The wisdom of crowds*. Anchor.

Susarla, A., Oh, J.-H., & Tan, Y. (2012). Social networks and the diffusion of user-generated content: Evidence from YouTube. *Information Systems Research*, *23*(1), 23–41.

Theodoridis, S., & Koutroumbas, K. (2008). *Pattern Recognition* (4th ed.). London: Academic Press.

Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *63*(2), 411–423.

Tomasello, M. (2008). *Origins of human communication* (Vol. 2008). Cambridge: MIT Press.

Turney, P. D., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, *37*(1), 141–188.

Ultsch, A. (1995). Self organizing neural networks perform different from statistical k-means clustering. In *Proc. Conf. Soc. for Information and Classification, Basel* (Vol. 1995).

Utgoff, P. E. (1986). Shift of bias for inductive concept learning. *Machine Learning: An Artificial Intelligence Approach*, *2*, 107–148.

Valente, T. W. (1996). Social network thresholds in the diffusion of innovations. *Social Networks*, *18*(1), 69–89.

Valente, T. W. (2005). Network models and methods for studying the diffusion of innovations. In *Models and methods in social network analysis* (pp. 98–116). Cambridge University Press.

Van den Bulte, C., & Joshi, Y. V. (2007). New product diffusion with influentials and imitators. *Marketing Science*, *26*(3), 400–421.

Varian, H. R., & Farrell, J. V. (2004). *The economics of information technology: An introduction*. Cambridge University Press.

Wagner, C. S., & Leydesdorff, L. (2003). Mapping global science using international co-authorships: a comparison of 1990 and 2000. In *Proceedings of ninth international conference on scientometrics and informetrics, Beijing*.

Wasserman, S., & Faust, K. (1994). *Social network analysis: Methods and applications*. Cambridge: Cambridge University Press.

Watts, D. J., & Dodds, P. (2009). Threshold Models of Social Influence. In P. Hedström & P. Bearman (Eds.), *The Oxford Handbook of analytical sociology* (pp. 475–497). Oxford University Press.

Watts, D. J., & Dodds, P. S. (2007). Influentials, networks, and public opinion formation. *Journal of Consumer Research*, *34*(4), 441–458.

Widdows, D. (2004). *Geometry and Meaning*. Standford: CSLI Publications.

Witten, I. H., Frank, E., & Hall, M. (2011). *Data Mining: Practical machine learning tools and techniques* (3e ed.). Morgan Kaufmann.

Wolpert, D. H., & Macready, W. G. (1997). No free lunch theorems for optimization. *Evolutionary Computation, IEEE Transactions on*, *1*(1), 67–82.

Wood, W. (2000). Attitude change: Persuasion and social influence. *Annual Review of Psychology*, *51*(1), 539–570.

Xu, R., & Wunsch, D. (2005). Survey of clustering algorithms. *Neural Networks, IEEE Transactions on*, *16*(3), 645–678.

Young, H. P. (2009). Innovation diffusion in heterogeneous populations: Contagion, social influence, and social learning. *The American Economic Review*, 1899–1924.

Zwarts, J. (2010). Semantic map geometry: Two approaches. *Linguistic Discovery*, *8*(1), 377–395.