Propagation du buzz sur Internet – Identification, analyse, modélisation et représentation dans un contexte de veille

Aurélien Lauf

Institut National des Langues et Civilisations Orientales

École doctorale N°265

Langues, littératures et sociétés du monde

Équipe de Recherche Textes, Informatique, Multilinguisme (ERTIM)

THÈSE

présentée par :

Aurélien LAUF

soutenue le 14 Octobre 2014
pour obtenir le grade de : **Docteur de l'INALCO**
Discipline : Traitement Automatique des Langues

Propagation du buzz sur Internet
Identification, analyse, modélisation et représentation dans un contexte de veille

Thèse dirigée par :

M. Mathieu VALETTE	Professeur des universités, INALCO

RAPPORTEURS :

M. Pascal MARCHAND	Professeur des universités, Université de Toulouse 3
M. Mathieu ROCHE	Chercheur HDR, CIRAD

MEMBRES DU JURY :

M. Pascal MARCHAND	Professeur des universités, Université de Toulouse 3
M. Mathieu ROCHE	Chercheur HDR, CIRAD
Mme Frédérique SEGOND	Professeure associée, INALCO
M. Julien VELCIN	Maître de conférences, Université de Lyon 2
Mme Leila KHOUAS	Docteur, AMI Software
M. Mathieu VALETTE	Professeur des universités, INALCO

Remerciements

Au début de ma thèse, je pensais que trois années suffiraient largement pour la réaliser. Quatre ans et demi plus tard, je réalise à quel point j'avais tort ! L'excitation et les rêves du début ont vite été remplacés par le doute, la frustration, et la fatigue. Ces dernières années ont néanmoins été tellement enrichissantes que je ne regrette pas de m'être lancé dans un tel projet.

Au cours de cette thèse, je me suis étrangement imaginé plusieurs fois en train de rédiger ces remerciements. Il est désormais temps de le faire pour de vrai.

Je tiens tout d'abord à remercier Mathieu Valette, d'avoir accepté de suivre cette thèse pluridisciplinaire parfois assez éloignée de la linguistique et du traitement automatique des langues. Je remercie aussi Pascal Marchand, Mathieu Roche, Frédérique Segond, et Julien Velcin de m'avoir fait l'honneur de participer à mon jury.

Merci à Eric Fourboul d'avoir cru en moi en me confiant ce passionnant sujet de thèse. Un énorme merci aussi à Leila Khouas, sans qui cette thèse ne serait pas la moitié de ce qu'elle est. Sa rigueur scientifique intransigeante et son suivi au quotidien m'ont permis de tenir le cap. Encore merci ! Je remercie aussi l'ensemble de mes collègues montpelliérains pour leur bonne humeur et, surtout, toutes ces heures endiablées de baby-foot !

Merci à mes parents, ainsi qu'à mes beaux-parents, de m'avoir permis de poursuivre mes études dans les meilleures conditions possibles.

Merci à Andréa pour ces 13 années passées ensemble et celles à venir ! Je tiens aussi à souligner son remarquable travail de relecture. Merci à ma grande Wendy, née un mois avant le début de cette thèse, de m'avoir appris le sens des priorités et m'avoir fait grandir autant. J'aurais aimé être plus présent et disponible certains soirs ! Merci aussi à ma petite Ellie, dont la naissance est prévue dans moins d'un mois, d'avoir patienté jusqu'à ce que je dépose ma thèse !

Enfin, je tiens à remercier Jean Véronis, dont les réflexions m'ont été d'une aide précieuse à un moment-clé de mon travail.

A tous, je vous souhaite une excellente lecture. J'espère avoir fait honneur à ce sujet ambitieux.

Table des matières

Introduction

Pour une entreprise, il est nécessaire de faire de la veille stratégique, c'est-à-dire surveiller et écouter ce qui se dit à son sujet, ses concurrents ainsi que sur son domaine de compétence. Collecter, organiser, et analyser l'information qui circule aide l'entreprise à devancer ses concurrents, innover, comprendre les attentes de ses clients, et réagir le plus rapidement possible.

Le Web ne cesse de grossir, atteignant des proportions incommensurables. En effet, depuis le Web 2.0, il est devenu aisé pour quiconque de diffuser du contenu. Par ailleurs, on assiste de nos jours à une perte de confiance vis-à-vis des médias classiques et à un changement majeur dans les habitudes d'accès à l'information : ces nouvelles sources numériques gagnent en importance et viennent directement concurrencer les sources classiques d'information telles que les journaux ou les revues spécialisées. Ces nouveaux intermédiaires de l'information que sont les blogs et les forums sont des mines d'informations qui doivent être exploitées. Néanmoins, du fait de la diversité des sources et de l'absence de contrôle de l'information diffusée, évaluer la pertinence ou la crédibilité d'une information ne peut plus uniquement reposer sur des critères comme le nom de l'auteur ou le contexte de publication. Par ailleurs, avec l'émergence des réseaux sociaux, l'information circule désormais en temps réel, impliquant de nouveaux modes d'extraction et d'analyse en continu des données.

S'inscrivant principalement dans un contexte de veille informationnelle et d'intelligence d'entreprise sur Internet, l'objectif de cette thèse est d'élaborer des outils et des méthodes permettant d'identifier, analyser, modéliser et représenter le cheminement des buzz sur Internet. Tout buzz a un ou plusieurs points d'origine : les sources primaires. L'information est ensuite relayée par des sources secondaires qui vont accélérer ou non la propagation en fonction de leur degré d'influence. Tout au long du cycle de vie du buzz, le contenu sémantique est amené à évoluer. Il s'agira donc de détecter des buzz en ligne, remonter jusqu'aux sources primaires et aux sources secondaires ayant joué un rôle majeur dans sa propagation, d'en dégager des sous-thématiques ainsi que des communautés de discours, et d'analyser les différences sémantiques pouvant apparaître dans le temps.

La compréhension d'un buzz sur Internet passe ainsi par l'analyse de ce qui se dit (grands thèmes abordés, sous-thématiques, etc.) et la qualification des émetteurs : qui parle, quelles sont les personnes les plus influentes, quelles interactions entretiennent-elles au sein des réseaux, appartiennent-elles à des communautés en particulier ? Cette thèse est axée autour de deux types d'analyses complémentaires : une analyse topologique des émetteurs (théorie des graphes et des réseaux) et une analyse du contenu textuel (linguistique de corpus).

De fait, de nombreuses approches complémentaires seront considérées tout au

long de cette étude. L'analyse conjointe de la topologie du Web et du contenu textuel est intéressante pour tenter de comprendre comment le sens se propage. En effet, si l'analyse du Web sous forme de graphe a permis, depuis la fin des années 90, de mettre en avant certaines régularités statistiques intéressantes et une meilleure qualification des émetteurs (appartenance communautaire, importance dans les réseaux), il est encore difficile d'évaluer la façon dont le sens se localise [Ghitalla 2004]. L'analyse du contenu du Web se fait en effet principalement par mots-clés. Les personnes "parlant de la même chose" ont-elles tendance à se connaître ? L'apparition sur le Web d'un buzz donné s'explique-t-elle par une diffusion virale au sein des réseaux ou s'agit-il d'une appropriation commune "accidentelle", à un moment donné et à divers endroits du Web ? En d'autres termes, peut-on en déduire des modèles topologico-sémantiques ? Y a-t-il corrélation entre proximité topologique et sémantique ?

L'analyse de la dimensions temps est aussi intéressante. Il s'agit donc non seulement de connaître les émetteurs et ce qu'ils disent, mais aussi de détecter ceux qui sont à l'origine d'un buzz donné. Qui en a parlé en premier ? Qui a repris l'information et à partir de quand s'est-elle propagée rapidement ? Combien de temps le buzz a-t-il perduré ? Comment l'information s'est-elle déformée au fil du temps ? A partir de l'analyse rétrospective de buzz, une modélisation et des algorithmes de prévision peuvent alors être envisagés.

Ce mémoire de thèse sera organisé de la façon suivante. Dans le chapitre 1, nous présenterons le contexte dans lequel cette étude se situe. Il s'agira alors de présenter le monde de la veille, ses acteurs et ses enjeux, ainsi que de décrire le processus de veille et les méthodes couramment mises en oeuvre. Ce sera aussi l'occasion de délimiter le périmètre de notre étude, en définissant plus précisément ce que nous entendons par "buzz".

Le chapitre 2 traitera de l'exploitation de la topologie des sous-graphes du Web dans le but de qualifier les émetteurs. L'objectif est ici de détecter les pages Web et personnes les plus influentes, les mieux connectées, mais aussi les plus réactives vis-a-vis de l'actualité. Dans un cadre de veille, ces émetteurs devront être particulièrement surveillés. Nous aborderons la problématique de la quantification de l'*autorité* d'une source sur le Web et les réseaux sociaux. A partir d'un examen des méthodes existantes et de leurs limites, nous amènerons une réflexion sur la validité des pratiques en matière de calcul d'autorité. Nous proposerons ensuite une approche permettant de recontextualiser l'information et ainsi guider le veilleur dans son parcours interprétatif des résultats. Notre approche associe calcul d'autorité, détection de communautés en ligne et outils de visualisation.

Le chapitre 3 traitera de l'analyse du contenu textuel des documents extraits du Web. Il s'agit ici de chercher à modéliser et extraire ce qui se dit sur Internet à propos d'un sujet donné. Un état de l'art des travaux existants sera présenté. Deux approches seront ensuite proposées. La première consiste à ramener la tâche de détection de buzz à celle d'extraction de thématiques. Une fois les thématiques détectées, il devient possible de rapprocher les documents en fonction de leur contenu et ainsi

évaluer la volumétrie de chacune d'entre elles. L'approche que nous proposons est une alternative linguistique aux méthodes statistiques classiques de classification non supervisée de documents. Ancrée dans les traditions de la lexicométrie et de la linguistique de corpus, elle repose sur l'analyse d'un graphe de cooccurrents de deuxième ordre. Le formalisme de la théorie des graphes nous permet d'exprimer des relations sémantiques assez fines entre les mots de chaque thématique. Ces graphes peuvent par ailleurs être visualisés afin d'aider le veilleur dans son interprétation. La deuxième approche que nous proposons repose sur l'extraction de citations. Une analyse diachronique de ces citations sera par ailleurs proposée afin de mieux comprendre la façon dont le sens peut se déformer au fil du temps. Les deux approches peuvent être complémentaires. L'opération d'extraction de thématiques apporte au veilleur une vision d'ensemble macroscopique de ce qui se dit dans les documents tandis que les citations permettent une vision microscopique. Par ailleurs, les citations sont facilement détectables et permettent l'identification simple des petites histoires quotidiennes au sein de chaque thématique.

Cette thèse CIFRE a été réalisée dans un cadre de recherche applicative au sein de la section R&D de la société AMI Software, spécialisée dans l'édition de logiciels de veille stratégique et d'intelligence d'entreprise. De ce fait, le dernier chapitre sera consacré à l'intégration de nos travaux dans la plateforme de veille développée par l'entreprise. Enfin, nous illustrerons concrètement la portée de nos travaux à travers une analyse rétrospective de quelques buzz ayant eu lieu pendant l'élaboration de cette thèse. Ces derniers mettent en avant l'intérêt d'une analyse conjointe de la topologie du Web et du contenu textuel pour identifier un buzz, qualifier les émetteurs, et comprendre sa propagation au sein des réseaux.

Chapitre 1

Contexte et périmètre d'étude

Sommaire

Cette partie présentera le contexte dans lequel nos travaux se situent : la veille sur le Web. Nous décrirons la veille, son fonctionnement, ses enjeux ainsi que les différents acteurs de ce domaine. Les solutions existantes de veille seront présentées et nous décrirons AMI Software, le partenaire industriel de la collaboration CIFRE. Nous décrirons aussi en détail l'état actuel du Web.

Nous tâcherons ensuite de délimiter notre périmètre d'étude. Nous définirons de façon extensive les différents usages du mot "buzz" et préciserons ceux que nous retenons dans le cadre de nos travaux. Nous verrons par ailleurs que les mécanismes de diffusion des buzz sont assez similaires à ceux de la rumeur mais que la propagation se trouve amplifiée du fait des nouveaux outils de communication qu'offre le Web.

1.1 La veille sur le Web

1.1.1 Définitions et enjeux de la veille

La veille consiste à surveiller tout ce qui se dit ou se fait autour d'un sujet, une entité, ou un domaine donné. Il s'agit dans un second temps de diffuser de façon

ciblée les informations sélectionnées et traitées [Jakobiak 1990]. Il existe différents types de veille : commerciale, concurrentielle, juridique, technologique, etc. La veille concurrentielle surveille par exemple les entreprises concurrentes, leurs nouveaux produits ou innovations, leur situation financière, ou encore leurs recrutements ; la veille juridique prête attention aux modifications législatives ; la veille scientifique consiste à se tenir au courant de toute innovation ou nouvelle publication dans un domaine donné. Le but commun à toutes ces veilles est de pouvoir comprendre les évolutions de son environnement, anticiper certaines "situations de crise" et ainsi prendre les mesures adéquates le plus rapidement possible. Contrairement à l'espionnage, la veille a recours à des sources ouvertes. La veille est un processus itératif : de nouvelles informations sont ajoutées régulièrement et ces dernières doivent être validées et capitalisées par des experts veilleurs, ou bien filtrées.

Dans le cadre de cette thèse, nous allons nous concentrer sur l'intelligence d'entreprise (on parle parfois d'intelligence économique, ou de veille d'entreprise). L'intelligence d'entreprise est un terme vaste qui englobe veille concurrentielle, scientifique, technologique, sociétale et d'image (e-réputation). Il s'agit alors pour une entreprise ou une organisation de collecter, analyser, organiser et diffuser l'information-clé qui circule sur les réseaux. L'objectif est de devancer ses concurrents, innover, mieux comprendre les attentes de ses clients et réagir en fonction, mais aussi vérifier qu'aucune information confidentielle ne circule.

Traditionnellement effectuée par des documentalistes à partir de sources connues et contrôlées, la veille a subi une révolution majeure avec l'arrivée du Web dans les années 90. Elle s'est simplifiée grâce à l'accès aisé et rapide à l'information, mais s'est dans le même temps complexifiée du fait de la multiplication exponentielle du nombre et du type des sources. Par ailleurs, du fait de l'émergence des réseaux sociaux et des médias diffusant l'information en continu, le cycle de veille s'est récemment raccourci. Les opérations de collecte qui pouvaient avoir lieu la nuit (le veilleur a ainsi ses documents le matin) peuvent désormais devoir se faire à l'échelle du quart d'heure pour des raisons d'exhaustivité.

1.1.2 Contexte : le Web

Le Web, depuis sa création et démocratisation, a connu de nombreuses mutations. Nous décrivons brièvement ici le Web, de ses origines à maintenant [1].

1.1.2.1 Genèse du Web

Le "Web" est une abréviation de "World Wide Web" (WWW) qui désigne littéralement une toile d'araignée mondiale. Il s'agit d'un système hypertexte utilisant Internet inventé par Tim Berners-Lee en 1989, et essentiellement développé au départ par lui-même et Robert Cailliau, tout deux ingénieurs au CERN (Conseil Européen pour la Recherche Nucléaire) [2].

1. Pour plus de détails, lire cet article : http://www.vox.com/a/internet-maps.
2. http://www.webfoundation.org/about/vision/history-of-the-web/.

A l'origine, le Web était composé de pages statiques, principalement des pages personnelles et des sites institutionnels. Le degré d'interaction de l'internaute est limité, à l'exception des premiers groupes de discussion et forums permettant des échanges textuels.

Pour désigner ce chapitre du Web, on parle parfois de nos jours de "Web 1.0", par opposition au "Web 2.0" que l'on va décrire ci-dessous.

1.1.2.2 Le Web 2.0

Ce que l'on nomme "Web 2.0" et que l'on date assez arbitrairement au début des années 2000 est en réalité le fruit d'une évolution progressive des technologies, notamment Ajax, qui ont permis une nouvelle utilisation plus interactive du Web. On assiste à la naissance et à la démocratisation des blogs, des wiki [3], des flux RSS, mais aussi des premiers réseaux sociaux [4], bien que ces derniers n'aient pas encore l'ampleur qu'ils ont de nos jours [5].

Il est dès lors aisé de produire du contenu sur le Web et l'internaute passe du statut de spectateur à celui d'acteur. L'information est créée, copiée et partagée simplement, gratuitement, et instantanément n'importe où dans le monde. Des réactions spontanées sont suscitées autour d'articles de presse ou de blog. Très vite, les analystes du Web comprennent l'intérêt de se concentrer sur ces conversations à des fins notamment d'analyse d'opinion, et non plus uniquement sur les documents. L'analyse du Web devient de plus en plus complexe, avec la nécessité de faire face à des volumes colossaux et à une multitude de supports : sites institutionnels classiques, presse en ligne, blogs, agrégateurs d'actualité, réseaux sociaux, forums, wiki, etc.

1.1.2.3 Les réseaux sociaux

Le Web actuel est le prolongement logique de ce qu'avait amorcé le Web 2.0. Le degré d'interaction et de participation de l'internaute est accru. Actuellement, de nombreux supports sont connectés à Internet (téléphone portable, tablette, console de jeux vidéo, télévision, etc.). Depuis 2007-2008, on assiste par ailleurs à une réelle démocratisation des réseaux sociaux, notamment des services de microblogging (Twitter par exemple). Le commentaire et le partage de l'information sont dès lors davantage facilités (il n'est plus utile de passer du temps à rédiger un article de blog). On parle parfois de Web "temps réel".

L'ensemble des interactions effectuées sur les réseaux sociaux sont collectées et sont autant d'empreintes numériques laissées par les internautes. Ces données sont agrégées, croisées et viennent nourrir les moteurs de recherche et les systèmes de publicités ciblées. L'utilisateur devient dès lors l'unité privilégiée d'analyse sur le Web.

3. La version anglaise du Wikipédia apparaît en 2001 (http://en.wikipedia.org/wiki/Main_Page).

4. Myspace apparaît en 2003, Facebook en 2004, Twitter en 2006.

5. http://oreilly.com/web2/archive/what-is-web-20.html.

Les deux réseaux sociaux les plus connus sont Facebook et Twitter. Nous les décrivons brièvement ci-dessous.

Facebook Facebook, lancé en 2004 par Mark Zuckerberg, est le réseau social le plus utilisé actuellement : le cap du milliard d'utilisateurs a été dépassé en 2012 [6]. Chaque utilisateur crée un profil personnel dans lequel il peut partager des photos, échanger des messages avec ses amis, prendre part à des discussions dans certains groupes, etc.

Malgré sa popularité, Facebook est relativement peu exploité dans un cadre de veille étant donné qu'une grande partie des comptes sont privés, et donc inaccessibles ; seules les interactions ayant lieu sur une page publique sont exploitables. C'est la raison pour laquelle nous ne nous sommes relativement pas attardés dessus durant cette thèse.

Twitter Lancé en mars 2006 par Jack Dorsey, Twitter est un service de microblog (aussi appelé microblogue ou microblogage –*microblogging* en anglais). Bien qu'existant depuis le début du Web, le phénomène du microblog a pris de l'ampleur avec l'arrivée de Twitter. Comme son nom l'indique, le microblog est une sorte de blog miniature : il permet à tout internaute de rédiger de courts messages (moins de 200 caractères –140 pour Twitter) qui sont publiés instantanément. Ces messages, qui peuvent contenir des hyperliens, des images, ainsi que des vidéos, s'adressent le plus souvent aux personnes décidant de suivre l'utilisateur (ses *amis*).

Les microblogs sont moins coûteux en temps et en investissement qu'un blog classique et s'insèrent parfaitement dans la problématique actuelle de l'information en temps réel ; ils permettent aux utilisateurs d'être tenus au courant instantanément, mais aussi de partager une information, décrire une activité, exprimer une opinion, en se limitant au strict nécessaire.

Selon les chiffres officiels de Twitter datant de mars 2011 [7], il suffit désormais d'une semaine pour atteindre le milliard de tweets dans le monde entier. Pour comparaison, il a fallu 3 ans, 2 mois, et 1 jour pour atteindre le premier milliard de tweets. En 2011, le nombre moyen de tweets produits par jour était de 177 millions.

D'autres services de microblog ont fait leur apparition au même moment que Twitter, mais ce dernier est de loin le plus populaire. Citons notamment Jaiku, créé en 2006, racheté en 2007 par Google, puis fermé début 2012. Pownce, lancé en juin 2007 a fermé fin 2008.

Concrètement, Twitter permet à un utilisateur (*twittos* ou *twittonaute*) de suivre en temps réel les messages d'autres utilisateurs choisis. Nous décrivons ci-dessous la terminologie utilisée par Twitter.

Le ***tweet*** est l'unité fondamentale de contenu sous Twitter. Il s'agit d'un court message, de 140 caractères maximum. Hormis les rares comptes privés (8 %

6. http://money.cnn.com/2012/10/04/technology/facebook-billion-users/.
7. https://blog.twitter.com/2011/numbers.

uniquement en 2010 [Cha 2010]), ces messages sont publiquement accessibles sur la plateforme. S'abonner à un twittos donné (on devient alors *follower* ou *ami*) permet de recevoir une notification pour chaque message posté par ce dernier. Cette relation n'est pas nécessairement réciproque.

Reply : la syntaxe simple de Twitter permet aux utilisateurs d'interagir entre eux. Il est tout d'abord possible de répondre (*reply*) au tweet d'un autre utilisateur. On entame alors une discussion et apporte un contenu nouveau au tweet précédent. Pour ce faire, on précède le nom du twittos d'une arobase, par exemple :

```
@userX article très intéressant, merci !
```

A noter que le nom de l'utilisateur doit être indiqué en début de tweet. On parle sinon de *mention*, bien que le principe ne soit pas officiellement reconnu par Twitter. Dans ce cas, on ne s'adresse pas nécessairement directement à la personne :

```
En train de regarder le discours de @fhollande sur @france2tv.
```

Le ***retweet*** sert de citation ou de paraphrase d'un contenu posté par un autre utilisateur. En retweetant un message, on le diffuse à sa propre liste d'amis. La syntaxe est la suivante :

```
RT @userX Manger de la viande, une aberration écologique
http://tinyurl.com/ygnt8c9.
```

Les mots-clés "retweeting" et "reading" peuvent aussi être utilisés à la place de "RT", bien que cela soit assez rare pour des raisons de place. Dans l'exemple ci-dessus, l'utilisateur décide de partager le message de *@userX*. Il est important de préciser que le message d'origine peut être altéré, sans que cela ne soit indiqué. Il est donc, en théorie, possible de changer le propos d'autrui mais cela est le plus souvent fait pour des raisons de place (en tronquant la fin du message) ou pour ajouter un rapide commentaire au début ou à la fin.

Étant donné le nombre limité de caractères permis par tweet, les URLs sont raccourcies par des services de réduction d'URLs (on parle alors de *short URLs*) [8]. Le principe est le suivant : le service attribue une courte clé unique à l'URL que l'on souhaite réduire. La passerelle entre l'ancienne URL et l'URL courte nouvellement créée est assurée par le service de réduction qui va se charger de la redirection. La façon de générer la clé varie selon les services. Certes pratique, le principe d'URL courte est critiquable. Il est dès lors impossible de savoir vers où pointe exactement l'URL étant donné que ces dernières ne sont pas explicites [9]. Ces URLs courtes peuvent ainsi être utilisées à des fins malhonnêtes en redirigeant les utilisateurs

8. Citons notamment Bitly (`https://bitly.com/`), Goo.gl (`http://goo.gl/`), et TinyUrl (`http://tinyurl.com/`).

9. Il est parfois possible de spécifier soi-même la clé, mais cela reste marginal.

vers des sites malveillants [10].

L'attribution est assez similaire au retweet (principe de partage de l'information) mais cette dernière n'est pas, à l'instar de la mention, officiellement reconnue par Twitter. Elle est marquée par l'utilisation du mot-clé "via" suivi d'un nom d'utilisateur :

```
François Hollande remporte l'élection présidentielle
http://tinyurl.com/6ufhqog via @lemondefr
```

Une légère nuance est à souligner. Contrairement au retweet qui diffuse un tweet existant, l'attribution permet le partage d'information, sans qu'un tweet n'ait été nécessairement rédigé au préalable. Dans l'exemple ci-dessus, l'utilisateur diffuse une information publiée dans le journal *Le Monde*, bien que ce dernier n'ait jamais tweeté à ce sujet. De fait, l'attribution est souvent utilisée pour diffuser du contenu issu de la presse, ou de personnes connues (propos tenus à la télévision ou à la radio par exemple).

Enfin, le ***hashtag*** est une étiquette attribuée au tweet afin de le catégoriser, facilitant ainsi la recherche de tweets autour d'une thématique donnée. Il s'agit ainsi d'une forme de folksonomie, ou de *social bookmarking*, que l'on pourrait définir comme un système de classification collaborative. Il n'y a pas réellement de règles quant au choix du nom du tag, mais il convient de respecter certaines conventions implicites ; ils doivent être courts, explicites et sont à utiliser avec modération (ils doivent réellement apporter quelque chose autour de la thématique étiquetée). Ces hashtags prennent la forme d'une suite de caractères précédés par le symbole # :

```
Recette de buns : http://tinyurl.com/b3wa65g #miam #burger
```

1.1.3 Acteurs de la veille

Un bref panorama des principaux acteurs de la veille sera effectué et nous décrirons AMI Software, le partenaire industriel de la collaboration CIFRE.

1.1.3.1 Panorama des principaux acteurs

Il existe, pour simplifier, deux façons de faire de la veille dans un cadre professionnel [Benoist 2014]. La première est de faire appel à une agence de services. Ces agences proposent des prestations de veille et d'analyse de la e-réputation autour d'un sujet donné. Ces services sont souvent proposés parallèlement à des services d'analyse du Web, de référencement et de marketing Web. Les pays anglo-saxons ont une nette préférence pour ces prestations de service.

10. Certains services comme TinyUrl cherchent à apporter des solutions à ce problème en permettant à l'utilisateur de prévisualiser le site avant redirection, par exemple : `http://preview.tinyurl.com/awgwjsa`.

La seconde alternative est de recourir à une solution de veille. L'éventail de l'offre va d'outils gratuits peu performants aux outils professionnels plus complets. Les premiers (le plus souvent accessibles en ligne) permettent de faire de la veille sommaire, par exemple les services d'alertes mail de Google [11] et Yahoo [12] qui envoient un mail à l'utilisateur chaque fois que du nouveau contenu pertinent (pour une requête donnée) apparaît. Les seconds sont beaucoup plus rares ; on en dénombre moins d'une vingtaine, principalement d'origine française. Le point fort de ce type de solution réside dans l'agrégation d'un ensemble d'outils complémentaires. Ce qui compte avant tout, c'est la gestion totale de la chaîne de traitements, cumulée à une bonne ergonomie. L'avènement des offres SaaS [13] facturées à la consommation (ou par abonnement) permet depuis peu de rendre plus abordables et ainsi démocratiser ces solutions complètes.

1.1.3.2 AMI Software

AMI Software est, depuis juin 2007, la dénomination commerciale de Go Albert, PME française fondée en 1999. Originellement spécialisée en recherche d'information, la société a, au début des années 2000, évolué vers l'édition de logiciels de veille stratégique et d'intelligence d'entreprise. Bien que majoritairement présent en France, AMI Software intervient aussi à l'international, notamment au Maroc, au Royaume-Uni et au Canada [14].

AMI est l'acronyme de *Automatic Meaning Interpreter*, faisant référence aux outils d'analyse de contenu textuel [15] développés à l'origine dans un cadre de recherche d'information. Le produit principal de la société est AMI-EI (*AMI Enterprise Intelligence*), plateforme permettant la gestion de l'ensemble du cycle de veille et d'intelligence d'entreprise, à savoir la collecte, la recherche d'information, la capitalisation, l'analyse, et enfin la diffusion. Une description détaillée de cette plateforme est proposée dans le chapitre 4.

Par l'automatisation des tâches "bas niveau" (indexation, recherche d'information sur Internet, collecte, tri de documents, filtrage, etc.), le but de ces logiciels est d'aider au maximum le veilleur dans son travail d'analyse. L'ensemble du processus de veille est décrit en détail dans les parties suivantes.

AMI Software fait partie des acteurs français les plus importants de ce domaine.

1.1.4 Processus de veille

Nous détaillons le processus complet de veille. La terminologie employée ci-dessous est propre à AMI Software mais le principe général est assez semblable dans toute application de veille. La figure 1.1 schématise l'ensemble de la chaîne

11. http://www.google.com/alerts.
12. http://alerts.yahoo.com/.
13. *Software as a service.*
14. http://www.amisw.com/fr/contact/nos-coordonnees.htm et http://www.amisw.com/en/who-we-are/mission/customers.htm.
15. La technologie est brevetée depuis 2001.

de traitements. Dans les paragraphes suivants, nous allons décrire plus précisément les étapes suivantes : l'acquisition de l'information (recherche et collecte de documents), la publication et le partage, et enfin l'analyse des informations. Ce sera alors l'occasion de mettre en avant les problématiques associées à chaque étape et ainsi montrer l'intérêt de nos travaux dans l'ensemble du processus.

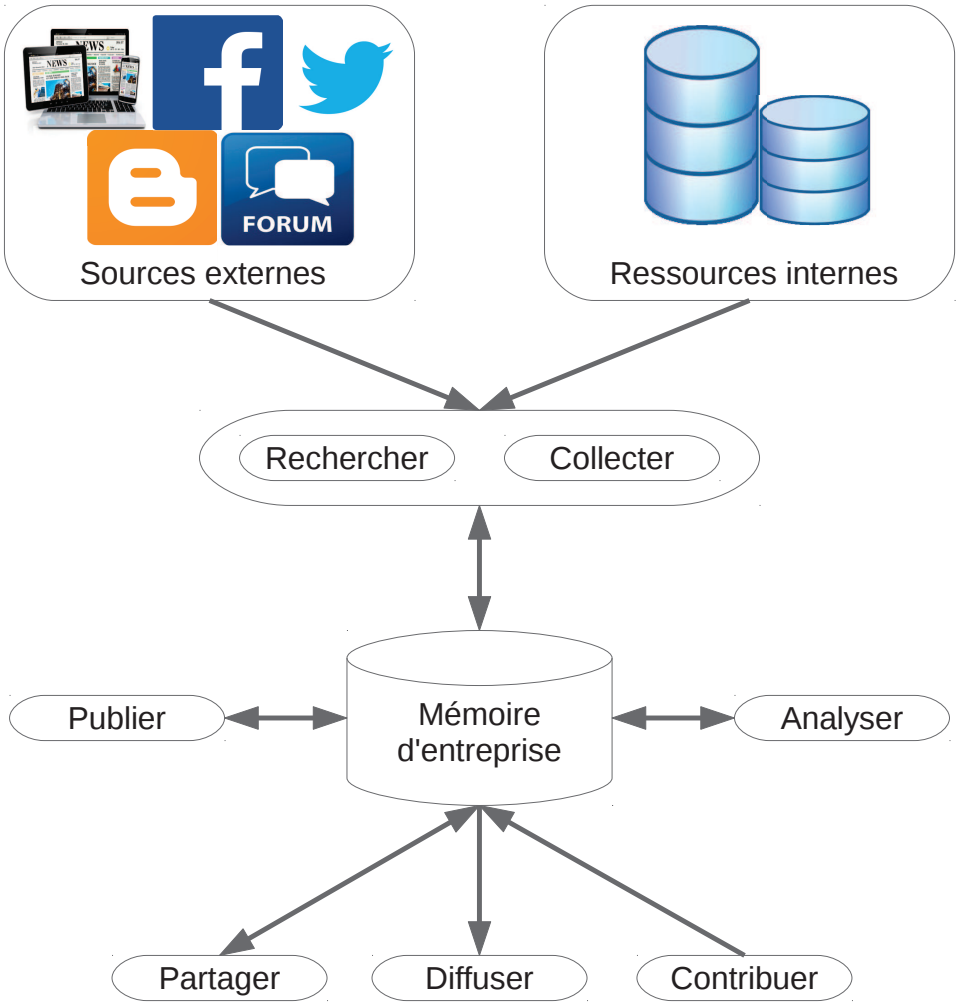

Figure 1.1 – Vision d'ensemble du processus de veille.

1.1.4.1 Acquisition de l'information : collecter et rechercher

La collecte est le point de départ de toute veille. Dans un premier temps, il convient pour le veilleur d'étudier les besoins de l'entreprise en matière de veille, en fonction de ses objectifs et de sa stratégie. Le périmètre de la veille se délimite en renseignant trois points : une ou plusieurs requêtes de recherche booléennes, un ensemble de sources d'information à surveiller, et une fréquence d'interrogation. L'automate de collecte va interroger et indexer les sources en question selon la

fréquence renseignée. La fréquence doit être d'autant plus élevée que les sources sont prolixes [16] mais le cycle de collecte s'étend le plus souvent sur 24h. Les nouveaux documents sont détectés et les anciens sont mis à jour si besoin est. L'ensemble des documents est stocké dans l'espace de publication. Il est possible de paramétrer une notification par mail à la fin de chaque collecte.

Une collecte de qualité doit répondre à trois problématiques principales. La première concerne le nettoyage des documents ; le système doit être capable d'isoler les zones intéressantes de la structure HTML des documents (contenu, titre, auteur, date, etc.) et ignorer le reste (encarts publicitaires, menus de navigation, etc.). Pour certaines sources comme les blogs ou les articles de presse, il peut par ailleurs être intéressant de séparer les commentaires et les articles, mais aussi de reconstruire les fils de discussion (le commentaire X répond au commentaire Y). Il est important de souligner que les solutions proposées à ce stade du processus de veille doivent être simples. Des techniques lourdes basées sur du text mining ne sont pas envisageables car il s'agit d'opérations à effectuer très tôt, avant même le filtrage par requêtes : la masse de documents à traiter peut être conséquente. Il faut trouver un compromis entre le temps de traitement et la qualité de la réponse.

Deux solutions sont utilisées. La première est le descripteur de site. Il s'agit d'un ensemble de règles décrites manuellement indiquant clairement au système comment chaque site Web renseigne le titre, la date ou encore l'auteur de ses articles, l'emplacement des publicités, etc. La méthode présente l'avantage d'être fiable mais le travail peut être laborieux (il faut autant de descripteurs que de sources) et la maintenance est difficile (la structure HTML des pages change assez régulièrement). C'est pourquoi on a aussi recours à une deuxième méthode qui repose sur un nettoyage complètement automatique basé sur la topologie des pages (filtrage topologique) ; il s'agit alors d'exploiter la mise en page HTML pour déterminer statistiquement les blocs susceptibles de contenir l'information utile [17]. En pratique, les deux méthodes sont utilisées conjointement : le nettoyeur automatique est utilisé par défaut et des descripteurs spécifiques sont maintenus à jour pour les sources les plus usuelles (plateformes de blogs, sites de presse en ligne, etc.).

La deuxième problématique est le filtrage des doublons. Il s'agit de détecter les documents quasi-similaires, par exemple une dépêche AFP reprise telle quelle par plusieurs médias. Le taux de doublons est à prendre en compte avant toute analyse volumétrique des données. Comme le fait remarquer [Manning 2008], le Web contient de nombreuses copies d'un même contenu, estimées à 40 % des pages (copies légales ou non). A noter néanmoins que le filtrage des doublons peut dans certains cas ne pas être souhaitable (analyse volumétrique des retombées d'un communiqué de presse par exemple).

16. Il peut être nécessaire de collecter tous les quarts d'heure pour Twitter par exemple.

17. Plus précisément : les balises *div* et *table* sont analysées. Les blocs principaux sont ceux qui contiennent le plus de texte "utile" (les blocs de texte constitués principalement d'URLs sont ignorés).

La définition de la notion de doublon est difficile : à partir de quand considère-t-on qu'il s'agit d'un doublon ? S'il est assez facile de détecter les copies exactes, la quasi-duplication est moins triviale. Il s'agit de pages dont le contenu, bien que très proche, a été modifié en partie (quelles qu'en soient les raisons). On a dans ce cas le plus souvent recours à des techniques de détection de plagiat pour détecter les doublons [Broder 1997].

La dernière problématique, moins technique, est probablement la plus critique : le choix des sources d'information à surveiller. Du fait de la perte des repères traditionnels, la détection et l'évaluation d'une source Web est loin d'être une tâche triviale. Il existe deux types de sources : les sources internes et les sources externes. Les premières renvoient aux ressources dont l'entreprise dispose dans ses propres bases de données. Les sources externes sont les sites institutionnels, la presse en ligne, les blogs, les forums, ou encore les réseaux sociaux. L'étape du choix des sources est cruciale pour une veille de qualité. Du fait de sa complexité, elle est le plus souvent laissée sous la responsabilité d'un administrateur de sources [18].

Un bon veilleur expert de son domaine de compétence a souvent déjà un catalogue de sources de qualité sur le sujet. Si la qualité de son jugement n'est pas remise en cause, le problème de l'exhaustivité est néanmoins à considérer. En effet, le bouquet des sources est en perpétuelle évolution et il n'existe pas de source (non) intéressante a priori et indéfiniment ; certaines peuvent devenir obsolètes et de nouvelles peuvent apparaître. Le problème de la qualification des sources sur le Web (et des émetteurs sur les réseaux sociaux) est un des piliers de cette thèse. Nous abordons ce point dans le chapitre 2.

1.1.4.2 Capitalisation, publication et partage

Les documents collectés viennent nourrir une base de données centrale, appelée mémoire d'entreprise. Cette dernière favorise la corrélation d'informations entre elles, permet de retrouver des connaissances enregistrées précédemment, et devient un véritable lieu de partage de connaissances et d'informations. L'utilisateur doit pouvoir naviguer parmi l'ensemble, organiser les documents par rubriques, valider, trier, enrichir, ou encore modifier. Cet espace est collaboratif et permet la diffusion interne de l'information en vue d'analyses futures, sous forme de synthèses, commentaires ou newsletters automatiques.

Outre les problématiques d'indexation des documents (comment indexer et traiter rapidement des millions de documents ?), l'important est la fluidité et la convivialité de l'interface. Il convient donc de proposer des outils d'aide à la lecture rapide tels que la surbrillance des concepts-clés ou des entités nommées, un résumé pour chaque document, des outils de recherche et de tri avancés, etc. Des fonctionnalités de classification automatique des documents selon leur contenu sont aussi à

18. En général, le panel de sources des clients est entièrement supervisé et maintenu par AMI Software.

envisager.

1.1.4.3 Analyse de l'information

Cette phase consiste à dégager du sens à partir de la masse brute collectée, c'est-à-dire l'information stratégique exploitable par les décideurs. Il s'agit d'une phase exploratoire et itérative (interactions entre l'expert et l'outil). Les outils d'analyse ont pour fonction de donner, au mieux, des pistes et des angles d'analyse. Ces informations peuvent être recoupées afin de valider des hypothèses ou d'en émettre de nouvelles, détecter les tendances fortes ou évènements marquants, mais aussi déceler les phénomènes marginaux annonciateurs de changement à venir (signaux faibles). En d'autres termes, les buts de navigation sont flous au début et vont s'affiner au fur et à mesure de l'exploration ; une information n'est exploitable que si elle est contextualisée, croisée et qualifiée.

Deux grands types d'outils sont alors proposés. Les premiers sont principalement statistiques et permettent de délimiter le périmètre observé : analyse par domaine, rubrique, mot-clé, date, ou encore métadonnée (par exemple : l'auteur du document). Les seconds consistent en l'analyse du contenu : détection d'entités nommées et de termes-clés, nuage de tags, analyse de tonalité, etc. Les travaux que nous décrivons dans le chapitre 3 viennent nourrir ces outils d'analyse.

1.2 Le buzz

1.2.1 Les différents usages du mot "buzz"

Le mot "buzz"[19] a de nombreux usages, aussi proposons-nous dans cette partie de le définir par extension. La méthodologie employée est la suivante. Tout d'abord, nous avons lancé une collecte à l'aide de AMI-EI sur les principaux moteurs de recherche en français avec la requête "buzz". L'ensemble des 650 documents obtenus au bout des premières 24h a été analysé et trié manuellement afin d'isoler l'ensemble des usages du mot. Outre les rares homonymes (le bruit sourd dans l'électroménager par exemple) ainsi que les noms propres (Buzz l'Éclair, Buzz Aldrin, etc.), nous avons constaté les usages principaux suivants.

Le mot "buzz" sert en premier lieu à désigner quelque chose dont "tout le monde parle", en bien comme en mal, pendant une période donnée, généralement assez courte.

Par extension, "buzz" désigne aussi la technique marketing visant à faire habilement du bruit[20] autour d'un produit ou un message (politique par exemple). On parle alors de marketing viral (ou marketing de propagation) étant donné que

19. L'Académie Française préconise l'usage de "ramdam" (lui-même emprunté de l'arabe...) à la place de "buzz". Parmi les candidats évoqués, citons "actuphène", "échoweb", "ibang", ou encore "réseaunance".

20. Notons les expressions "faire le buzz" ou encore "surfer sur le buzz".

la communication est faite par les utilisateurs par bouche à oreille en se diffusant à travers les réseaux sociaux et les blogs de façon épidémique. A noter que cette forme de marketing n'est pas nécessairement numérique étant donné qu'elle peut avoir pour point de départ, par exemple, une campagne de publicité à la télévision ou une affiche dans la rue. Il existe des agences de communication spécialisées dans la génération de buzz.

Cette idée de communication virale à grande échelle est reprise à travers certains noms de services comme Google Buzz et Yahoo Buzz. Google Buzz était un réseau social intégré à Gmail. Yahoo Buzz était un agrégateur communautaire d'articles d'actualité, au principe assez proche de BuzzFeed [21].

Ce lien entre actualité et buzz est intéressant. Dans certains cas, "buzz" désigne toute l'actualité sur un sujet donné. Par exemple, le site `www.buzzecolo.com` parle *de tout et de rien, mais toujours de développement durable.* Le site `www.le-buzz-immobilier.com` traite de *tout sur l'immobilier*, tandis que `www.buzz-litteraire.com` concerne *la littérature nouvelle génération, de bouche-à-oreille.* Notons toutefois que l'accent est mis, le plus souvent, sur l'actualité concernant des personnes célèbres, les informations insolites, marquantes, ou drôles.

1.2.2 Buzz et rumeur

Comme le signale [Froissart 2007], la propagation d'un buzz sur Internet est comparable à un effet de rumeur.

[Stern 1902] met en place un protocole expérimental pour étudier la rumeur. Il constate que les rumeurs s'appauvrissent et se déforment au fil des transmissions successives. L'idée est approfondie par [Allport 1951] qui définit trois comportements : la réduction, l'accentuation, et enfin l'assimilation. Il y a réduction quand le message d'origine est simplifié. A l'instar de Stern, les auteurs quantifient l'appauvrissement de la rumeur par le décompte du nombre de détails associés à cette dernière. Il y a accentuation quand certains détails sont ajoutés ou exagérés. Enfin, l'assimilation a lieu quand certains détails sont déformés suite à une mauvaise interprétation en fonction des croyances ou opinions de la personne. Ces travaux considèrent que la transmission de l'information suit un schéma linéaire et descendant. Or, [Froissart 2004] signale à juste titre que cette idée de linéarité est simplificatrice car elle implique que l'information a une origine unique et claire.

Les travaux sur la rumeur s'accordent à dire qu'il n'existe pas de règles figées sur la façon dont la rumeur se propage. Sur Internet, la diffusion d'une information peut être multi-format (vidéo, blogs, réseaux sociaux, etc.).

Peut-on ainsi définir le buzz comme une rumeur numérique à grande échelle ? A l'instar de la rumeur, le schéma de diffusion du buzz est le bouche à oreille et il est très souvent difficile de déterminer d'où provient l'information et qui en sont les acteurs principaux. Ce phénomène est largement amplifié par les outils de communication actuels. En effet, la diffusion sur le Web est aisée, rapide et l'information

21. `http://www.buzzfeed.com/`.

peut être partagée via les réseaux sociaux ou encore les flux RSS, anonymement, sans aucun contrôle, et dans le monde entier. Ajoutons à cela l'avènement de l'actualité en temps réel où l'instantanéité de l'information peut primer sur la vérification des sources [Kovach 1999].

Néanmoins, et c'est ce que traduisent des expressions telles que "surfer sur le buzz", [Froissart 2007] signale qu'*il ne s'agit pas uniquement d'une appropriation collective d'une icône, soudainement, aléatoirement. C'est aussi un effet largement utilisé dans l'industrie pour constituer de l'audience. Il faut donc prendre en compte d'autres phénomènes collectifs instrumentalisés, dans le but de générer des bouffées d'audience.*

1.2.3 Notre définition

Retenons de l'ensemble de ces usages les aspects viraux et massifs du buzz. Dans notre contexte de veille professionnelle, nous définissons le buzz comme tout sujet largement propagé sur le Web pendant une période donnée –le buzz est toujours ponctuel. Notons aussi que cette propagation doit être soudaine, inattendue ; quelques minutes suffisent parfois à déclencher un buzz.

Chapitre 2

Analyse topologique du buzz

Sommaire

Dans le chapitre précédent, nous avons abordé les enjeux et objectifs de la veille ainsi que de nos travaux.

Ce chapitre abordera en détail la problématique de la qualification des émetteurs, qui passe par l'exploitation de la topologie des sous-graphes du Web. Dans un cadre de veille, il est en effet crucial d'être en mesure de qualifier les émetteurs et de se constituer très rapidement un bouquet de sources pertinentes sur un sujet donné. Il s'agit de sources influentes (pages très consultées par exemple), mais aussi de sources réactives par rapport à l'actualité (celles qui ont l'information en premier). Un commentaire positif ou négatif ne présentera pas le même intérêt pour le veilleur selon que l'émetteur est influent ou non.

Nous définissons une source de la façon suivante : il s'agit d'un vecteur potentiel de production et de diffusion de contenu informatif sur le Web. Concrètement, cela se traduit par des pages Web définies par une URL. A noter néanmoins qu'avec l'émergence du Web social, une source peut aussi désormais désigner un utilisateur diffusant du contenu sur un réseau social. C'est la raison pour laquelle nous parlons aussi parfois d'émetteur. Pour des raisons de clarté, nous utilisons indifféremment dans ce mémoire "sources" et "émetteurs" pour désigner tout producteur de contenu informatif sur le Web.

Nous nous intéresserons dans un premier temps à cette notion d'autorité et décrirons les approches existantes pour l'évaluer. Dans un second temps, nous poserons les notions essentielles qui nous serviront tout au long de ce mémoire. Nous décrirons ensuite l'approche que nous proposons. Cette dernière associe calcul d'autorité, détection de communautés en ligne et outils de visualisation dans le but de fournir au veilleur des outils de contextualisation de l'information, essentiels pour une bonne interprétation des résultats.

2.1 L'autorité des sources sur le Web

Le passage du document traditionnel au document numérisé sur internet a entraîné une perte des repères traditionnels. Du fait de la diversité des sources, évaluer la pertinence ou encore la crédibilité d'un document sur le Web ne peut désormais plus reposer uniquement sur des critères comme le nom de l'auteur ou le contexte de publication. Dès lors, la qualification des émetteurs et la détection de pages dites "d'autorité" est un enjeu crucial pour la veille sur Internet.

Dans un cadre de veille, les sources sont généralement sélectionnées et qualifiées par un expert du domaine. Si la qualité de son jugement n'est pas remise en cause, le problème de l'exhaustivité est néanmoins à considérer. En effet, le bouquet des sources est en perpétuelle évolution et il n'existe pas de source (non) intéressante a priori et indéfiniment ; certaines peuvent devenir obsolètes et de nouvelles peuvent apparaître. Il est ainsi primordial de proposer des outils aidant le veilleur expert dans sa tâche de qualification et de découverte des émetteurs.

Le terme "autorité" est imprécis et mêle dans la littérature popularité, influence, voir même pertinence. Avant de dresser un panorama critique des approches existantes, nous apportons quelques axes de réflexion afin de tenter de délimiter plus précisément cette notion.

2.1.1 Définitions

[Arendt 1989] définit l'autorité comme la capacité de pouvoir agir sur autrui *sans recourir à la contrainte par la force ou à la persuasion par arguments.* En ce sens, l'autorité ne doit pas être confondue avec l'autoritarisme qui implique une certaine forme de contrainte physique et/ou morale. Précisons néanmoins que l'autorité classique n'est pas pour autant nécessairement égalitaire et juste ; l'auteur rappelle que les hommes dotés d'autorité étaient les anciens ou le Sénat, et que cette autorité était obtenue par héritage. Cette autorité traditionnelle repose ainsi exclusivement sur la croyance et le respect de l'ordre des choses.

La notion d'autorité a évolué avec la société, qui a au fur et à mesure pris certaines distances par rapport à la religion et aux traditions. L'autorité moderne ne découle désormais plus de la transcendance mais de l'influence ou de la popularité [Le Deuff 2006]. Citons ici les travaux de [Wilson 1983] sur l'autorité cognitive [1]. Cette forme d'autorité intègre ainsi des rapports de confiance, d'influence, de crédibilité et de réputation. Tout comme l'autorité traditionnelle, l'autorité cognitive est volontaire, mais cette dernière est un choix conscient selon des critères particuliers. Lorsque l'on choisit un journal contre un autre, que l'on tend vers les idées d'un politicien donné ou que l'on suit l'exemple d'un musicien en particulier, il s'agit d'autant de marques d'autorité cognitive. L'autorité n'est pas absolue ; il s'agit d'une forme de reconnaissance mesurable : une entité fait plus ou moins autorité sur un sujet donné.

Le psychologue Thomas Gordon distingue quatre sortes d'autorités [Gordon 2003]. La première découle de l'expérience de la personne, de son savoir ou encore de ses compétences. Elle est ainsi à rapprocher de l'autorité cognitive. La deuxième repose sur la position, les fonctions et les responsabilités, par exemple l'autorité d'un patron sur ses employés. Une troisième forme d'autorité découle d'ententes informelles, accords ou contrats. La dernière repose sur le pouvoir qu'a une personne sur une autre, impliquant des relations de domination et de force. Elle est ainsi à rapprocher de l'autoritarisme. L'auteur considère que les trois premières sont des moyens constructifs et puissants d'influencer autrui tandis que la dernière vise principalement à dominer [2].

Ces concepts à l'esprit, comment définir une autorité au sens documentaire ? [Broudoux 2007] introduit le concept d'autorité informationnelle : *contrairement à l'autorité cognitive, l'autorité informationnelle, susceptible d'être portée par un individu ou un groupe, un objet ou un outil cognitif ou encore un média, n'a pas pour fonction principale l'influence mais celle d'in-former (donner une forme).* L'autorité informationnelle intègre aussi les notions de confiance, d'influence, de crédibilité et de réputation. Plus précisément, elle mêle plusieurs notions interdépendantes [3] :

1. *Cognitive authority is influence on one's thoughts that one would consciously recognize as proper* (p.15).

2. Précisons tout de même que l'ouvrage cité concerne l'éducation des enfants. Néanmoins, les principes de la communication non violente peuvent aussi parfaitement s'appliquer aux adultes.

3. *L'institution qui publie un document prend la décision de le faire paraître en fonction de sa connaissance de l'auteur, sa connaissance du sujet traité, son programme de publication, ses*

- *autorité énonciative* : identité et statut du ou des auteur(s) ;
- *autorité institutionnelle* : groupe régi par des règles ;
- *autorité de contenu* qui inclut le genre du document, sa qualité, ses sources et toute autre métadonnée (maison d'édition, contexte de publication, etc.) ;
- *autorité du support de publication* : le type ainsi que la fréquence de publication.

Le contexte du numérique implique un changement de paradigme. Avec de moins en moins de contrôle des sources, qu'est-ce qu'une "source d'autorité" sur le Web ?

[Broudoux 2007] rappelle ces changements de paradigmes. Premièrement, l'espace de publication est ouvert à tous, ce qui est d'autant plus vrai depuis l'arrivée du Web 2.0 et désormais des réseaux sociaux ; il est en effet désormais aisé d'écrire sur le Web, que ce soit en rédigeant un article de blog, en postant un commentaire ou une critique sur un site d'information, ou encore en envoyant un message sur un réseau social. Cette ouverture de l'espace de publication implique des genres et des contenus hétérogènes. On assiste par ailleurs à un changement considérable d'échelle ; tout contenu posté sur le Web est instantanément visible partout dans le monde. On constate aussi une baisse (voire dans certains cas la disparition) du filtrage institutionnel, en faveur parfois d'un filtrage de groupe institué [4], qui découle naturellement de l'ouverture de l'espace de publication, mais aussi de l'instantanéité de l'information. Enfin, on assiste à une réutilisabilité et un partage des contenus [5]. Dès lors, il devient difficile d'identifier l'auteur (autorité énonciative) d'un document (parfois anonyme, parfois collectif [6]), et les autorités institutionnelles se trouvent concurrencées par des groupes formés spontanément autour d'intérêts ou objectifs communs ("communautés").

2.1.2 Méthodes d'évaluation de l'autorité d'une source

Une mesure intuitive de l'importance d'une page sur le Web, tout du moins de sa popularité, pourrait être le nombre d'usagers qui la consultent quotidiennement. Il convient néanmoins de ne pas confondre l'*audience* et le simple *trafic* généré par la page. L'audience est le nombre de personnes qui se connectent volontairement à la page et la lisent avec intérêt. Le trafic correspond au nombre de visites du site via un moteur de recherche [7]. Une page peut être bien référencée dans un moteur et n'intéresser paradoxalement que peu d'internautes. Nous abordons plus en détail ce point dans la partie 2.1.5. Le trafic est donc à corréler systématiquement au taux

possibilités commerciales. La notoriété de l'auteur est susceptible d'augmenter selon les actions publicitaires des groupes ou institutions auxquels il s'est affilié. L'auteur dans ses processus créatifs doit prendre en compte les contraintes du support de diffusion dont l'éditeur a habituellement la responsabilité (p.6).

4. Cas du Wikipédia (`http://fr.wikipedia.org/`) avec un filtrage collaboratif a posteriori.

5. Avec notamment l'apparition de licences comme creative commons (`http://fr.creativecommons.org/`).

6. Certains documents n'ont d'ailleurs pas d'auteur. C'est le cas notamment des bulletins météo générés automatiquement.

7. On parle aussi de *daily reach*.

de rebond (*bounce rate*) qui correspond au pourcentage d'internautes ayant quitté le site sans naviguer à l'intérieur de ce dernier [8]. Un site avec une bonne audience aura un trafic élevé et un taux de rebond faible. Il pourrait aussi être intéressant de considérer à la fois le nombre absolu de visites et le nombre de visites uniques, permettant ainsi d'inférer un pourcentage d'usagers habitués.

Comment obtenir le nombre de connexions uniques à une page donnée ? Ce dernier est certes proposé par des services du type Alexa [9] mais les données sont collectées uniquement à travers les personnes ayant installé l'application. Similairement, Google Insight [10] se limitait aux personnes ayant accédé à la page en passant par le moteur Google. Notons cependant que la taille du lectorat d'une page peut être appréhendée par le nombre (et le type –certains utilisateurs sont des habitués) de commentaires ainsi que, dans certains cas [11], par les interactions sur les réseaux sociaux (partages, *likes*, etc.).

D'autres critères pourraient aussi être envisagés, mais ces pistes restent à démontrer.

La qualité de l'orthographe : une page doit-elle être nécessairement bien orthographiée ? La question demeure aussi ouverte pour le respect des conventions structurelles du Web (normes W3C, etc.). Ces considérations de forme ont-elles un impact sur le fond ?

Le nombre de thématiques abordées : certains sites sont spécialisés sur certains sujets et sont de ce fait plus susceptibles de fournir une information de qualité. Les pages tenues par des journalistes indépendants proposant des analyses détaillées de plusieurs sujets distincts sont néanmoins de bons contre-exemples.

La réactivité vis-a-vis de l'actualité : la fraicheur de l'information [Dai 2010] est indéniablement un point fort. Lui donner la priorité pénaliserait cependant les articles de fond écrits avec plus de recul. Dans le même ordre d'idée, que penser de la fréquence, de la densité et de la taille des publications ?

L'ancienneté de la page : le Web est en évolution constante et il apparaît sans cesse de nouvelles sources d'information. La prise d'importance des médias alternatifs (blogueurs influents, agences de conseil, etc.) en est la preuve. Certaines pages peuvent devenir obsolètes et ne plus être tenues à jour.

L'identité du ou des auteur(s) : ce dernier est-il connu en dehors du Web ? S'agit-il d'un professionnel ou d'un particulier ? Garde-t-il son anonymat ?

L'originalité : les articles sont-ils personnels ou s'agit-il d'une agrégation d'autres articles pré-existants ? Gardons à l'esprit que certaines sources, de par leur popularité, servent de grands propagateurs. Il peut par exemple être plus pertinent, et moins chronophage, de surveiller ce que dit une personne particulièrement bien informée et qui diffuse l'information. Ces personnes jouent le rôle de hubs, une

8. Le taux de rebond se calcule simplement par le nombre d'internautes ne visitant qu'une seule page, divisé par le nombre total de visites.

9. http://www.alexa.com/.

10. Le service a été fusionné en 2012 avec Google Trends (http://www.google.com/trends/).

11. Certains blogueurs utilisent par exemple leurs comptes Facebook et Twitter pour publier automatiquement un message chaque fois qu'ils écrivent un nouvel article.

notion que nous décrirons plus en détail dans la partie 2.1.3.6.

Afin d'évaluer l'autorité d'une source, on distingue principalement trois grandes approches : l'évaluation humaine, le classement par l'usage, et l'utilisation de la citation. Ces approches ne sont pas nécessairement exclusives. Nous les présentons dans les parties qui suivent.

2.1.2.1 L'évaluation humaine

L'évaluation humaine fut historiquement la première méthode à voir le jour ; on se souvient notamment de l'annuaire de Yahoo [12] apparu au milieu des années 90. La pertinence des résultats proposés par les moteurs de recherche étant alors discutable, il était primordial de proposer autrement un bouquet de pages de qualité. Cette dernière était donc évaluée en amont par des experts humains, qui parcouraient le Web et qui classaient les pages par (sous-)catégories et/ou mots-clés. L'exemple le plus connu de classification humaine aujourd'hui est probablement l'Open Directory Project [13]. Ces annuaires peuvent être généralistes ou, au contraire, spécialisés à un domaine donné.

Les inconvénients sont nombreux. Tout d'abord, le temps de création et de mise à jour est inadapté à la croissance fulgurante du Web, entraînant un réel problème d'exhaustivité. Par ailleurs, la grande force de la méthode, à savoir une évaluation humaine de qualité (par opposition à une évaluation automatique "aveugle"), est problématique : les mots-clés et les catégories choisis sont parfois incomplets, ou en inadéquation avec les catégories qu'aurait choisi l'utilisateur (il s'agit avant tout de choix humains figés). Par ailleurs, l'objectivité de certains de ces annuaires peut être mise en doute dès lors qu'ils peuvent faire figurer un site contre financement, ou tout autre intérêt comme l'ajout de liens réciproques en échange de la présence du site dans l'annuaire. Il existe en effet plusieurs modèles. Yahoo et About.com [14], par exemple, sont gratuits et sont ainsi financés uniquement par la publicité. D'autres, comme l'Open Directory Project, sont collaboratifs : tout internaute peut proposer des sites mais le choix définitif reste tout de même à la seule discrétion des experts du système.

Bien que l'impartialité de ces annuaires ne soit pas nécessairement discutable, les modalités de sélection et validation restent floues si bien que ces annuaires sont souvent critiqués.

2.1.2.2 Classement par l'usage

Le classement par l'usage consiste à collecter et utiliser les informations que fournit l'utilisateur à un moteur de recherche. Par exemple, pour une requête donnée, si la majorité des utilisateurs préfère cliquer sur le troisième résultat renvoyé par

12. http://dir.yahoo.com/.
13. http://www.dmoz.org/.
14. http://www.about.com/#!/browse-categories/.

le système, c'est qu'il est d'une façon où d'une autre plus "pertinent" que les deux premiers. Le classement va donc être mis à jour au fur et à mesure des contributions de chacun. Cette approche est le plus souvent employée en complément de l'approche décrite dans le point suivant.

Le principal problème de cette approche est la nécessité de collecter une quantité de données conséquente, et ce pour chaque requête[15]. Par ailleurs, se pose le problème de la prétendue qualité universelle d'une page, indépendamment des besoins de l'utilisateur, et de la logique du *rich get richer* : plus une page est choisie, plus elle devient populaire et visible, ce qui va la rendre encore plus susceptible d'être choisie à l'avenir par les autres utilisateurs. Ici, on favorise clairement la popularité, le choix de la majorité[16].

L'approche est utilisée depuis plusieurs années par Google pour affiner les résultats, en complément de leurs algorithmes principaux. A noter que cette logique a récemment été poussée encore plus loin : la situation géographique de l'utilisateur ou son historique personnel sont désormais autant de critères pouvant influer sur les résultats renvoyés par le moteur. Les critères retenus sont encore plus nombreux si l'on s'est identifié sur Gmail au moment de la requête : le contenu des mails envoyés ou discussions échangées, qui sont analysés en permanence par mots-clés, peut aider le moteur à aiguiller l'utilisateur selon ce qu'il pense être ses centres d'intérêt. Deux personnes faisant la même requête sur deux machines différentes auront des résultats différents. Comme le signale [Pariser 2012], les indicateurs propres à l'utilisateur sont au nombre de 57 ; il n'y a plus de Google générique.

2.1.2.3 Utilisation de la citation

La dernière méthode que nous décrirons ici, et la plus utilisée depuis une dizaine d'années, repose sur l'analyse des citations. Il s'agit alors de considérer un hyperlien d'une page A vers une page B comme un vote de A en faveur de B. En d'autres termes, A confère une certaine autorité à B. Pour simplifier, plus une page possède de liens entrants, plus son rang dans la hiérarchie est élevé. Si les liens entrants proviennent de pages elles-mêmes considérées "importantes", le score de la page cible est d'autant plus élevé. L'approche est récursive et les pages se renforcent donc mutuellement. Cette relation de renforcement mutuel est cruciale car elle permet, entre autres, de lutter contre des contournements du système. En effet, si l'on compte uniquement le nombre de liens entrants d'une page, il suffirait de créer des milliers de pages Web factices pointant vers un site donné.

Il est important de préciser qu'un hyperlien n'est pas nécessairement un vote po-

15. Seuls les grands moteurs de recherche en sont capable. Notons néanmoins que ce point n'est pas spécifique au classement par l'usage.

16. Ce classement par l'usage est omniprésent sur le Web, par exemple sur les sites de partage de vidéo : "vidéos les plus vues", "vidéos les plus commentées". Sur les réseaux sociaux, citons notamment les *trending topics* (TT) de Twitter ; il s'agit de hashtags employés massivement sur une période donnée. Les TT sont indiqués sur la page de présentation de Twitter. Le but est de montrer aux utilisateurs ce qui "se passe dans le monde". Souvent, par émulation, ces mots ou expressions deviennent encore plus utilisées dès lors qu'ils apparaissent parmi les TT.

sitif ; ce dernier peut critiquer et non soutenir, et il peut s'agir de liens publicitaires. Cependant, le postulat est que ces cas particuliers sont minoritaires.

Parmi les méthodes utilisant de cette façon la structure des hyperliens entre les pages, le représentant le plus connu est le PageRank (ou PR) de Google [17] [Page 1999], largement inspiré de l'algorithme HITS [Kleinberg 1999a]. Ces deux approches sont décrites plus en détail respectivement dans les parties 2.1.3.5 et 2.1.3.6. Ce type d'approches a été proposé en réponse aux résultats discutables des moteurs de recherche jusqu'à la fin des années 90. Citons notamment Altavista [18] fondé en 1995, un des moteurs les plus populaires jusqu'à l'arrivée de Google en septembre 1998, qui proposait de classifier automatiquement les résultats pour faciliter la navigation. Ces moteurs reposaient principalement sur le nombre d'occurrences dans les pages des termes de la requête, ainsi que de leur proximité. Le balisage HTML des pages Web permettait aussi de pondérer les termes selon la place qu'ils occupaient : un terme présent dans le titre a, par exemple, plus de poids qu'un mot au milieu d'un paragraphe. L'approche présente de nombreux inconvénients. Tout d'abord, le système n'a aucun moyen pour gérer efficacement la synonymie et la polysémie. Ce problème, bien connu de la recherche par mot-clé, peut être contourné en partie en élargissant la requête avec d'autres termes proches, de façon à restreindre à un usage donné du mot. Néanmoins, les pages les plus pertinentes ne contiennent pas nécessairement ces mots-clés. Par exemple, il n'est jamais question de "constructeur automobile japonais" sur les sites de Honda ou Toyota ; il y a donc souvent inadéquation entre la façon de formuler de l'utilisateur et les mots utilisés par les pages Web en question [Kleinberg 1999b] [Chakrabarti 1999a] [19]. Enfin, l'approche présente une très faible résistance au spam. Il est en effet très simple pour un internaute d'améliorer le poids de sa page en répétant certains mots importants à des endroits stratégiques de la page (titres, entêtes, etc.), par exemple écrits en blanc sur fond blanc.

Tous les moteurs de recherche de nos jours ont recours à l'analyse de la structure des hyperliens pour donner un poids aux documents. La recherche par mots-clés demeure la même, mais sont alors renvoyées en priorité les pages les plus citées. C'est ainsi que le site `www.lemonde.fr/` devra être renvoyé en priorité par rapport aux millions d'autres pages pour la requête "le monde". Comme le signale [Le Deuff 2006], on assiste à un glissement de concept de l'autorité vers la popularité. Le processus de légitimation émane du peuple, et non plus d'institutions. On passe par ailleurs de la pertinence à l'influence ; la pertinence caractérise l'information fiable, provenant de sources sûres, tandis que l'influence est la capacité à être lu et écouté, en d'autres termes, sa visibilité sur le Web.

Cette utilisation des hyperliens pour déterminer le poids des pages hérite des travaux en bibliométrie [20] qui se servent du nombre et du type de citations, afin de

17. `http://www.google.fr/`.
18. `http://fr.altavista.com/`.
19. *Authorities are often not particularly self-descriptive.*
20. Évoquons aussi les travaux en analyse des réseaux sociaux [Katz 1953] [Wasserman 1994] [Rogers 2003] [Gladwell 2002] qui s'intéressent à mesurer la centralité des acteurs au sein des ré-

produire des estimations quantitatives de l'importance de certains articles ou journaux scientifiques. Le principe est introduit par [Garfield 1955] (le *facteur d'impact*). Si deux articles citent un troisième, cela signifie que ce dernier a une certaine importance mais aussi qu'il existe un lien de similarité thématique entre ces deux travaux (co-citation). Le lien est pondéré par le nombre de citations qu'ils ont en commun. En théorie, les travaux les plus importants sont les plus cités et les journaux importants ont de nombreux papiers influents [Kessler 1963], [Garfield 1972], [Small 1973], [White 1989]. La bibliothèque scientifique en ligne Citeseer [21] base par exemple une partie de ses classements sur ces principes statistiques. Dans [Pinski 1976], les auteurs étendent le principe en conférant un poids plus important aux citations provenant de journaux qui ont eux-mêmes un score élevé (relation de renforcement mutuel mentionnée précédemment –ici nommé *facteur d'amortissement*), ce qui permet de mieux rendre compte des relations de hiérarchies. Ce principe d'évaluation de la recherche par le nombre de citations est parfois critiqué car cela a tendance à favoriser notamment les grandes équipes de recherche où les chercheurs peuvent se citer mutuellement (communautés denses). Par ailleurs, les scores peuvent fortement varier selon le type d'indicateur utilisé [Coutrot 2008] et certains travaux fortement précurseurs doivent parfois attendre plusieurs années avant d'être cités, le temps que le domaine gagne en maturité [Van Raan 2004]. Néanmoins, ces problèmes se posent moins sur le Web, du moins en théorie, puisque les communautés sont plus larges et surtout parce que les modalités de publication sont différentes : plus qu'un vote positif, l'hyperlien est avant tout le principal élément de navigation sur Internet.

2.1.3 Le Web vu comme un graphe

Depuis quelques travaux pionniers de la fin des années 90, on pense le Web comme un graphe dont les noeuds sont les pages Web et les arêtes entre ces noeuds les liens hypertextes [Bharat 1998a] [Brin 1998b] [Chakrabarti 1998c] [Kleinberg 1999d] [Chakrabarti 1999a] [Kumar 2000]. Il devient alors possible, à l'aide de la théorie des graphes, de mieux modéliser la complexité du Web et de calculer des régularités statistiques [Davison 2000a] [Henzinger 2005a] [Kritikopoulos 2006] [Nie 2006] [Wu 2006a] [Lee 2008] [Qi 2009].

Modéliser le Web sous forme de graphe revient à le considérer comme un système social stratifié, et non plus comme une grande bibliothèque de documents. On pense l'espace comme quelque chose de relationnel.

2.1.3.1 Définitions

La théorie des graphes est un formalisme mathématique permettant de modéliser des réseaux. On attribue son origine au mathématicien suisse Leonhard Euler qui est le premier à proposer un traitement mathématique sous forme de

seaux, et à comprendre les mécanismes de diffusion au sein de ces derniers.

21. citeseerx.ist.psu.edu/.

graphe au problème des sept ponts de Königsberg [22]. Le formalisme de la théorie des graphes permet d'appréhender plus aisément certains phénomènes complexes, ce qui explique son utilisation dans de nombreuses disciplines : linguistique [23], bibliométrie, ou encore biologie [24]. Nous définissons brièvement ci-dessous quelques notions-clés qui nous serviront tout au long de ce mémoire.

Un graphe permet de spécifier des relations entre objets. Un graphe est donc composé d'un ensemble d'objets appelés *noeuds* [25], et d'un ensemble d'*arêtes* formant des liens entre ces noeuds [26]. Une arête forme une boucle si elle relie un point vers lui-même. On dit que deux noeuds sont *voisins* s'ils sont reliés par une arête. Un graphe peut être *orienté* [27] (A pointe vers B mais l'inverse n'est pas vrai) ou non (relations symétriques). La figure 2.1 illustre un graphe simple, orienté ou non.

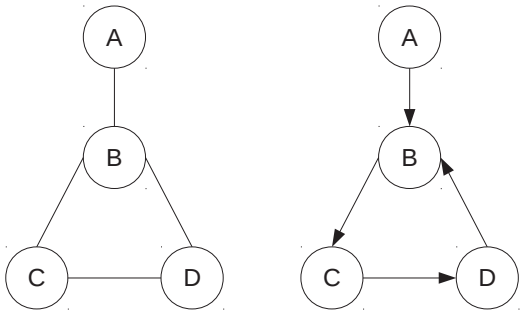

Figure 2.1 – Illustration d'un graphe simple composé de 4 noeuds (orienté à droite, non orienté à gauche). La position des noeuds dans l'espace n'est pas pertinente.

Les arêtes partant d'un noeud donné sont appelées *liens sortants*. A l'inverse, les *liens entrants* sont les arêtes pointant vers ce noeud. Les *degrés* entrant et sortant désignent respectivement le nombre de liens entrants et sortants.

On appelle *chemin* toute séquence de noeuds reliés par une arête [28]. La *longueur* du chemin correspond au nombre d'arêtes à parcourir pour aller du début à la fin du chemin. La *distance* entre deux noeuds correspond à la longueur du plus court chemin entre eux [29]. Pour reprendre la figure 2.1, la distance entre les noeuds A et

22. L'objectif était de trouver un chemin permettant un aller-retour à partir d'un point donné, sachant qu'il n'est possible de passer qu'une fois par chacun des sept ponts de la ville.

23. Citons notamment les analyses syntaxiques sous forme d'arbres. Un arbre n'est jamais qu'un graphe non orienté, acyclique et connexe.

24. Modélisation des schémas de diffusion de maladie, des flux migratoires, etc.

25. On parle aussi parfois de *sommets* ou de *points*.

26. On définit donc simplement un graphe G par un couple $G = (V, E)$ où V (*vertices*) désigne l'ensemble des noeuds et E (*edges*) désigne l'ensemble des arêtes.

27. Les arêtes sont alors appelées *arcs* ; on parle de *réciprocité* ou *mutualité* si le lien est dans les deux sens. Pour des raisons de simplicité, nous parlerons d'arêtes indépendamment de l'orientation ou non du graphe.

28. A noter que les arêtes se parcourent dans un seul sens si le graphe est orienté.

29. Parmi les algorithmes permettant de calculer le plus court chemin entre deux noeuds, citons

B est de 1, et de 2 entre A et C. A l'inverse, le *diamètre* correspond à la distance maximale entre toute paire de noeuds d'un graphe.

Un graphe est dit *connexe*[30] (ou *connecté*) s'il existe un chemin entre toute paire de noeuds. Dans le cas d'un graphe non connexe, il existe au moins deux noeuds entre lesquels il n'existe aucun chemin : on a alors plusieurs *agrégats*, ou *composantes connexes* (sous-graphe connexe maximal). Par exemple, la figure 2.2 illustre un graphe composé de 3 composantes connexes. En pratique, les réseaux complexes ont très souvent un *agrégat géant* qui va regrouper un nombre considérable des noeuds du graphe. Il est en général unique. Cela nous permet d'introduire la notion de *noeud central*[31]. Il s'agit de noeuds qui, s'ils sont retirés, briseraient le graphe en plusieurs composantes connexes. On peut voir ces noeuds comme des passages obligés au sein des réseaux.

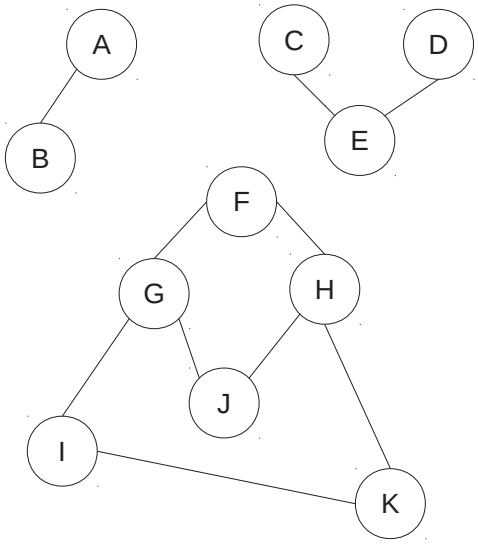

Figure 2.2 – Illustration d'un graphe avec 3 composantes connexes.

Enfin, introduisons la notion de *clique* (dans les graphes non orientés). Il s'agit d'un sous-ensemble de noeuds où toute paire est reliée par une arête, comme illustré dans la figure 2.3.

2.1.3.2 Graphes et matrices

Un graphe avec n noeuds se code en pratique le plus souvent sous la forme d'une matrice d'adjacence[32] carrée $An \times n$ telle que $A_{ij} = 1$ s'il existe un lien entre i et

notamment l'algorithme de Dijkstra et celui de Bellman-Ford.

30. Si le graphe est orienté, on dit qu'il est *fortement connexe*.

31. On parle aussi parfois de *point d'articulation* ou de *pont*.

32. Les matrices sont souvent creuses. Il peut être parfois plus judicieux de coder le graphe sous la forme d'une liste d'adjacence qui renseigne la liste des voisins pour chaque noeud. Le choix de

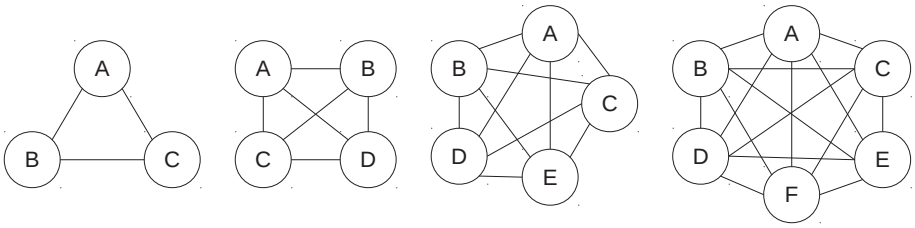

FIGURE 2.3 – Illustration de la notion de clique sur des graphes simples de 3, 4, 5, et 6 noeuds.

j, et 0 autrement. Dans un graphe pondéré, la valeur A_{ij} sera comprise entre 0 et 1, reflétant ainsi le poids de l'arête. Dans un graphe non orienté, $A_{ij} = A_{ji}$ (voir figure 2.4).

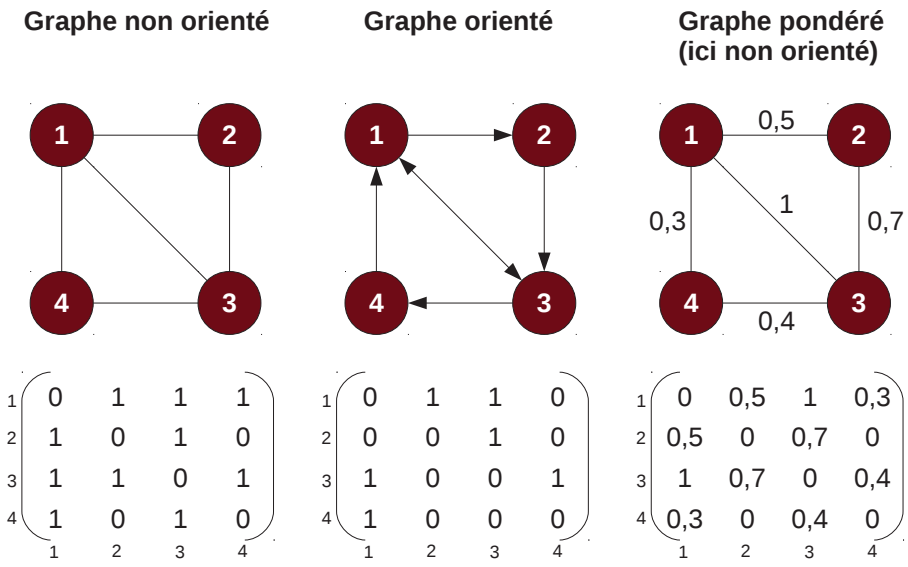

FIGURE 2.4 – Illustration du codage d'un graphe sous la forme d'une matrice d'adjacence.

2.1.3.3 Le graphe du Web

Le Web peut se modéliser sous la forme d'un graphe orienté dont les noeuds sont les pages Web, et les arêtes sont les hyperliens entre ces pages (voir figure 2.5). Il est important de séparer deux types d'hyperliens : les liens transactionnels (par exemple "Ajouter au panier") et les réels liens de navigation. Dans l'idéal, les premiers doivent être ignorés.

la structure dépend du type de graphe à traiter ainsi que du type d'opération à effectuer. Il s'agit

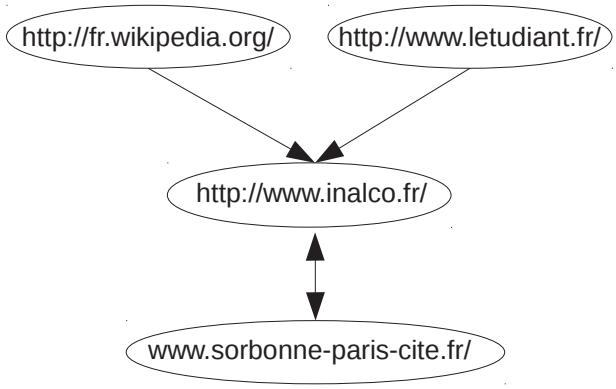

FIGURE 2.5 – Illustration simplifiée d'un sous-graphe du Web. Le Wikipédia et `letudiant.fr` possèdent tous les deux un hyperlien pointant vers `www.inalco.fr`. Ce dernier pointe vers `sorbonne-paris-cite.fr`. Cette dernière relation est réciproque. Pour des raisons de lisibilité, nous nous limitons ici aux noms de domaines.

[Broder 2000] annonce que le graphe du Web n'est pas connexe mais qu'une grosse partie l'est ; le Web contient un agrégat géant vers lequel pointent des milliards de liens et duquel en partent autant. Les pages non reliées à cet ensemble sont relativement rares. Il est fort probable que ces caractéristiques topologiques du Web soient toujours vraies de nos jours mais que leur taille ait fortement augmenté.

Notons cependant que parler du "graphe du Web" est en réalité un raccourci de langage. Il est entendu que personne n'est en mesure de connaître le nombre réel de pages existant sur le Web, et encore moins de les indexer toutes. Google, possédant l'index Web le plus fourni du monde [33], indiquait [34] avoir 26 millions de pages à leurs débuts en 1998. En 2000, ce nombre passait la barre de 1 milliard. En 2008, les index dépassaient le trillion d'URLs uniques, étant tout de même précisé que plusieurs URLs peuvent pointer vers la même page. Le nombre de pages réelles est donc inférieur mais les proportions restent colossales. Au mieux est-il possible de faire des estimations. Par ailleurs, ce nombre variera selon la définition que l'on donne à une page Web : par exemple, doit-on compter les pages au contenu autogénéré ? On parle de Web profond ou de Web invisible pour désigner l'ensemble des ressources encore inexplorées [35].

toujours d'un compromis entre vitesse d'exécution et espace mémoire occupé.

33. La façon de constituer et indexer le graphe du Web est en théorie triviale mais implique des architectures complexes si l'on veut avoir une chance de tenir la distance face au nombre sans cesse croissant de pages qui se crée chaque jour, tout en considérant les problèmes de surcharge de bande passante. Un crawler (*robot d'indexation*) part d'un ensemble de pages bien connectées (appelées *pages pivots*) et suit les hyperliens de façon récursive.

34. `http://googleblog.blogspot.fr/2008/07/we-knew-web-was-big.html`.

35. Certains pages sont même inaccessibles ; elles peuvent en effet être protégées par un mot de passe, ou bien n'être reliées à aucune autre page.

2.1.3.4 Six degrés de séparation et petits mondes

Dans sa nouvelle *Chaînes* [Karinthy 1929], l'écrivain hongrois Frigyes Karinthy introduit l'idée des *six degrés de séparation* : toute personne sur Terre serait reliée à n'importe quelle autre par moins de 6 intermédiaires[36]. [Milgram 1967] et [Travers 1969] tentent de démontrer expérimentalement cette hypothèse : l'étude du "petit monde".

Pour ce faire, des personnes vivant dans deux villes du Nebraska et du Kansas sont choisies comme points de départ et des personnes de Boston comme destinataires. Ces villes sont choisies du fait de leur éloignement géographique et social. On demande alors aux premiers de faire parvenir une lettre aux derniers. S'ils ne connaissent pas l'adresse de la personne en question, il ne peuvent alors que l'envoyer à des connaissances personnelles qu'ils jugent susceptibles de la connaître. On observa alors un nombre moyen de 6 intermédiaires, parmi lesquels se trouvaient certaines personnes récurrentes, les *sociometric superstars* [Milgram 1967].

L'expérience fut vivement critiquée, notamment car le nombre de lettres arrivant à destination était faible. Par ailleurs, l'expérience part du principe que toute personne est capable de déterminer le chemin le plus court entre eux et le destinataire. De nombreux travaux récents cherchent encore à valider l'hypothèse du petit monde. Parmi les plus notables, [Leskovec 2008b] analyse 30 milliards de conversations générées via MSN par 240 millions de personnes. A partir du graphe des conversations, les auteurs constatent un chemin moyen entre tout utilisateur de 6,6. En 2011, Facebook analyse son graphe[37], alors composé de 721 millions de personnes, et évalue une distance moyenne entre toute paire de noeuds de 4,74. Les auteurs signalent que le graphe est de plus en plus connecté et que la distance moyenne diminue au fil des années : elle était de 5,28 en 2008.

Cela nous permet d'introduire la notion de graphe petit monde. [Watts 1998] constate que le graphe du Web possède certaines caractéristiques intéressantes (*small-world phenomenon*). Ces caractéristiques sont les suivantes. Tout d'abord, le diamètre est faible : il existe au moins un court chemin entre toute paire de sommets du graphe. Par ailleurs, il y a une forte tendance pour que 2 voisins d'un sommet soient directement connectés[38]. On constate aussi la présence de *long range nodes* –il s'agit de "raccourcis" reliant des noeuds qui seraient normalement éloignés. Ce sont ces noeuds qui réduisent le diamètre du graphe. Enfin, ces graphes sont aussi

36. *Planet Earth has never been as tiny as it is now. It shrunk - relatively speaking of course - due to the quickening pulse of both physical and verbal communication. [...] One of us suggested performing the following experiment to prove that the population of the Earth is closer together now than they have ever been before. We should select any person from the 1.5 billion inhabitants of the Earth - anyone, anywhere at all. [...] Our friend was absolutely correct : nobody from the group needed more than five links in the chain to reach, just by using the method of acquaintance, any inhabitant of our Planet.*

37. https://www.facebook.com/notes/facebook-data-team/anatomy-of-facebook/10150388519243859.

38. Le coefficient d'agglomération (*clustering coefficient*) est élevé.

hiérarchiques : la probabilité $P_{(k)}$ qu'un sommet ait k voisins décroît comme une loi de puissance.

2.1.3.5 L'algorithme PageRank

Comprendre le graphe du Web permet de mieux appréhender les schémas de diffusion de l'information. Si on applique les notions vues précédemment au graphe du Web, cela signifie que toute information n'a pas à aller très loin pour se diffuser de façon virale au sein des réseaux et que certains noeuds centraux servent de passages obligés. Ces passages obligés sont les sources d'autorité à détecter.

Concrètement, en terme de graphes, calculer l'autorité d'une page Web ou d'un utilisateur au sein d'un réseau social consiste à appliquer des opérations de calcul de centralité des noeuds. Parmi les métriques [39] de centralité les plus connues, citons la *centralité d'intermédiarité* [40] (*betweennes centrality*) qui traduit la faculté d'un noeud à se connecter à d'autres groupes dans le réseau et qui correspond en pratique au nombre de plus courts chemins passant par ce dernier. Les noeuds avec une forte centralité d'intermédiarité jouent le statut de noeud central (passerelle entre groupes denses de noeuds) : si on les retire, la probabilité que le graphe se sépare en plusieurs agrégats distincts est élevée [41].

Citons aussi la *centralité de degré* (*degree centrality*) qui correspond simplement au degré d'un noeud, en donnant la priorité au degré entrant dans le cas d'un graphe orienté. Appliqué au graphe du Web, cela traduit donc trivialement le nombre d'hyperliens vers une page donnée. Suite aux travaux de [Katz 1953] en analyse des réseaux sociaux ou de [Pinski 1976] en bibliométrie, il est admis que le simple calcul des connections traduit uniquement une relation de popularité. Est alors introduite l'idée d'un calcul cumulatif récursif généralisant la centralité de degré ; on ne considère dès lors plus uniquement le nombre de voisins directs mais aussi les voisins indirects. Le calcul permet de rendre compte de relations hiérarchiques : plus les voisins sont puissants, plus l'autorité qu'ils confèrent sera élevée.

L'algorithme PageRank [42] [Page 1999], à l'origine du succès de Google est l'exemple le plus connu de ce type de calculs hiérarchiques. Nous le décrivons en détail ci-dessous.

Description de la méthode Appliqué au graphe du Web, le calcul peut s'interpréter comme la probabilité qu'un internaute arrive sur une page Web donnée en naviguant aléatoirement à travers les liens sortants rencontrés [Brin 1998b]. Au départ, les probabilités sont équitablement distribuées sur l'ensemble des pages Web.

39. Ces métriques sont issues des travaux en analyse des réseaux sociaux.
40. On attribue cette mesure au sociologue Linton Freeman.
41. [Granovetter 1973] observe que ces noeuds permettent de relier plusieurs groupes par des liens faibles. L'utilisation de ces liens faibles est très importante dans un cadre de propagation virale car ils permettent d'atteindre des groupes beaucoup plus larges. En se limitant aux liens forts, on reste enfermés dans quelques groupes denses.
42. Notons qu'il est possible d'avoir une approximation du PageRank d'une page donnée à l'aide de la *Google Toolbar* (`http://www.google.com/toolbar/ie/index.html`).

Si un noeud a un degré sortant de k, il va diffuser $\frac{1}{k}$ à chacun de ses liens sortants. Le graphe est donc ici représenté sous la forme d'une matrice de transition A, pour laquelle $A_{ji} = \frac{1}{d_i}$ s'il existe une arête allant de i à j, et sachant que d_i correspond au degré sortant du noeud i (voir figure 2.6).

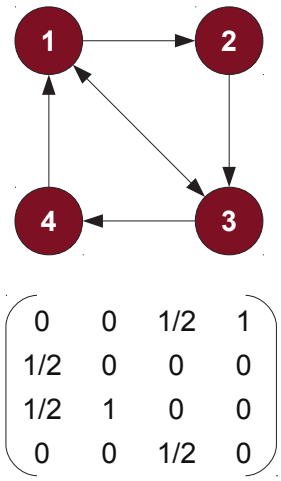

Figure 2.6 – Illustration du codage d'un graphe sous la forme d'une matrice de transition.

A chaque itération, les pages mettent alors à jour leur PageRank qui correspond à la somme des parts qu'elles reçoivent. Cela peut se traduire concrètement par la multiplication de la matrice A avec le vecteur de poids v, initialisé par la distribution équitable des probabilités parmi l'ensemble des noeuds. L'opération est répétée n fois jusqu'à convergence. A la première itération, le vecteur de poids correspond donc à Av. A la seconde itération, on a $A(Av) = A^2v$. A l'itération n, le vecteur de poids équivaut à A^nv. Pour illustrer notre propos, considérons le graphe simple de la figure 2.7.

Le graphe étant composé de 3 noeuds, le vecteur de poids est initialisé à $\frac{1}{3}$. A la première itération, les scores de PageRank sont les suivants :

$$\begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 1 & 0 \end{pmatrix} \bullet \begin{pmatrix} \frac{1}{3} \\ \frac{1}{3} \\ \frac{1}{3} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \frac{2}{3} \end{pmatrix} \tag{2.1}$$

A la seconde et dernière itération, on obtient :

$$\begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 1 & 0 \end{pmatrix} \bullet \begin{pmatrix} 0 \\ 0 \\ \frac{2}{3} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \tag{2.2}$$

Ici, le résultat est complètement contre-intuitif. On obtient des scores nuls pour les trois pages alors que le noeud 3 est cité 2 fois. La définition actuelle du PageRank est donc problématique dans ce genre de situation, mais aussi dans le

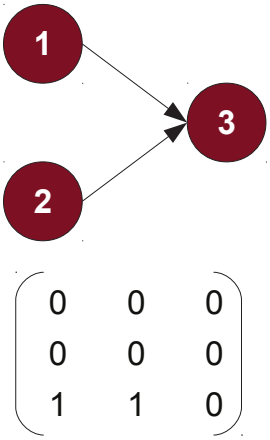

Figure 2.7 – Illustration d'un graphe simple, accompagné de sa matrice de transition.

cas où le graphe n'est pas connexe : l'internaute n'a alors aucun moyen de rejoindre une autre composante connexe que celle dans laquelle il se trouve au départ. Afin d'y remédier, les scores sont équilibrés par un facteur (*damping factor*) qui traduit la probabilité que l'internaute quitte la page et se téléporte vers une autre. La probabilité de téléportation vers une page donnée est équitablement répartie et correspond donc à $\frac{1}{n}$. On parle alors de *marcheur aléatoire* (*random surfer*)[Brin 1998b].

Une autorité universelle Il est très important de préciser que le calcul d'autorité est ici universel, indépendant de la requête. L'algorithme est en effet appliqué à l'ensemble du graphe du Web ; les scores sont calculés (ou réactualisés) lors de l'indexation des pages. C'est ce qui explique en partie la présence quasi-systématique de l'encyclopédie libre en ligne Wikipédia parmi les 10 premiers résultats. Le problème est du même ordre dans un exemple cité par Guilhem Fouetillou[43] : Loic Le Meur, blogueur faisant autorité[44] dans les nouvelles technologies, commence à parler de politique et à publier des articles en faveur de Nicolas Sarkozy, au moment des élections présidentielles de 2007. Dès lors, ce dernier devenait une autorité (toujours au sens Google) en politique.

Un PageRank moins universel De nombreux travaux tentent de rendre le *PageRank* plus dépendant de la requête de l'utilisateur tout en conservant des temps de calculs proches de la version de base, notamment le *Topic-Sensitive PageRank* [Haveliwala 2002], l'*Intelligent Surfer* de [Richardson 2001] ou le *Topical Random Surfer* de [Nie 2006].

43. http://www.csi.ensmp.fr/voxinternet/www.voxinternet.org/spip9276.html?article118%20.
44. Comprendre ici : bien placé dans les classements Google.

Dans [Haveliwala 2002], les auteurs proposent le *Topic-Sensitive PageRank*. A l'instar du PageRank classique, les scores sont calculés en amont au moment de l'indexation des pages. Néanmoins, afin de réduire l'universalité du calcul, chaque page se voit attribuer plusieurs scores distincts en fonction de la thématique abordée. Ces thématiques sont déduites à partir des URLs présentes dans les catégories anglaises les plus générales de l'Open Directory Project [45]. Au moment de la requête, un calcul de la similarité avec chacune de ces thématiques est effectué et la priorité est donnée au PageRank le plus proche thématiquement.

The Intelligent Surfer est introduit dans [Richardson 2001]. Rappelons que l'algorithme initial du PageRank assigne à chaque page un score proportionnel au pourcentage de chance qu'un surfeur visite cette page, si ce dernier surfait aléatoirement de page en page en suivant indifféremment les liens sortants. Afin d'éviter les impasses ou les boucles sans fin, le surfeur va parfois aller vers une page aléatoire. Pour l'*intelligent surfer*, ce passage ne se fait pas à l'aveugle, en donnant une préférence aux liens pointant vers des documents qui contiennent les mots de la requête. L'idée est alors de simuler plus fidèlement le comportement d'un internaute qui a tendance à rester sur des pages thématiquement cohérentes. Le principe est globalement le même pour le *topical random surfer* [Nie 2006].

2.1.3.6 L'algorithme HITS

L'algorithme HITS [Kleinberg 1999a] (*Hyperlinked Induced Topic Search*), dont s'inspire en partie le PageRank, a pour objectif de modéliser l'autorité d'une page tout en prenant en considération la thématique [46] souhaitée par l'utilisateur. A la différence des méthodes présentées précédemment cherchant à rendre le PageRank moins universel, il ne repose pas sur un ensemble de thématiques définies a priori. C'est la raison pour laquelle nous nous baserons sur cette approche pour la suite de nos travaux.

Pour ce faire, les calculs sont effectués non pas en amont sur l'ensemble du graphe du Web, mais sur un sous-graphe bien délimité et construit dynamiquement en fonction de la requête de l'utilisateur.

L'ensemble de la théorie repose sur les notions de *hubs* (annuaires) et d'*authorities* (autorités) ; les annuaires [47] sont des pages pointant vers de nombreuses ressources d'autorité et les autorités sont des pages pointées par de bons annuaires. Du fait de la circularité de la définition, les annuaires et les autorités se renforcent mutuellement. Une page peut être à la fois un hub et une autorité. Pour reprendre une métaphore de Jon Kleinberg [Easley 2010], les autorités sont des bons restaurants, et les annuaires des personnes de bon conseil [48].

45. Ces catégories sont au nombre de 16 : arts, business, computers, games, health, home, kids and teens, news, recreation, reference, regional, science, shopping, society, sports, et world.

46. Le lecteur conviendra que le mot "thématique" est utilisé ici de façon assez approximative.

47. Dans la suite de notre exposé, nous utiliserons indistinctement "hub" et "annuaire".

48. "Suppose you move to a new town and hear restaurant recommendations from a lot of people. After discovering that certain restaurants get mentioned by a lot of people, you realize that certain people in fact had mentioned most of these highly-recommended restaurants when you asked them.

Cette notion de hub est intéressante car elle permet de s'éloigner de la simple popularité : un site "peu connu" pourra se voir mettre en avant s'il est cité par un ou plusieurs hubs de qualité. De plus, dans un cadre de veille, il peut être intéressant de se tenir informé de ces diffuseurs d'information. Ces derniers ont très souvent un PageRank très faible et ont donc peu de chance d'être indiqués par Google.

Description de la méthode L'algorithme suit deux grandes étapes : la constitution d'un sous-graphe du Web pertinent pour la requête de l'utilisateur, et le calcul des scores de hub et d'autorité.

Étape 1 : construire un sous-graphe du Web Il s'agit dans un premier temps de constituer un ensemble de départ de pages Web. Cet ensemble, nommé root set, est construit à l'aide des n premiers résultats d'un moteur de recherche (Altavista dans l'article d'origine), en réponse à la requête. Le root set est ensuite étendu en prenant, pour chacune de ses pages, l'ensemble des liens sortants et m liens entrants (voir figure 2.8). Tous les liens entrants ne sont pas pris en compte car certaines pages peuvent en avoir un nombre considérable, jusqu'à plusieurs dizaines de milliers, ce qui risquerait de faire grossir l'ensemble dans des proportions non souhaitables. On obtient alors ce que l'auteur nomme le base set. Ce dernier est, dans l'idéal, d'une taille "raisonnable" (quelques milliers de pages) et fortement connecté. Bien que de taille restreinte, le postulat de départ est que cet ensemble comportera, par convergence dans les citations, un bon nombre de pages d'autorité.

Cet ensemble forme le sous-graphe sur lequel va être calculée l'autorité. Les noeuds de ce graphe sont les pages Web, et les arêtes représentent les hyperliens entre les pages. Sont ignorés lors de la constitution de cet ensemble les liens purement navigationnels, à savoir ceux pointant vers le même domaine que la page d'origine.

La construction d'un graphe thématiquement cohérent par la seule utilisation des liens sortants et entrants repose sur l'hypothèse qu'il existe une corrélation entre proximité topologique et sémantique : autrement dit, les documents abordant les mêmes sujets seraient localisés dans les mêmes régions [Ghitalla 2004]. L'ensemble des résultats peut être vu comme un cluster de pages se renforçant mutuellement. Certains îlots très compacts mais peu liés aux autres pourraient permettre de repérer des sous-thématiques fortes ou tout simplement mettre en évidence différentes communautés.

Étape 2 : déduire les annuaires et les autorités Le sous-graphe construit à l'étape précédente est supposé riche en autorités et en hubs. Il s'agit désormais d'évaluer chacune des pages de cet ensemble pour en identifier les meilleurs.

Rappelons que les hubs et les autorités se renforcent mutuellement : un bon hub est une page qui pointe vers plusieurs bonnes autorités, et une bonne autorité est une

These people play the role of the high-value lists on the Web, and it's only natural to go back and take more seriously the more obscure restaurants that they recommended, since you now particularly trust their judgment" (p.419).

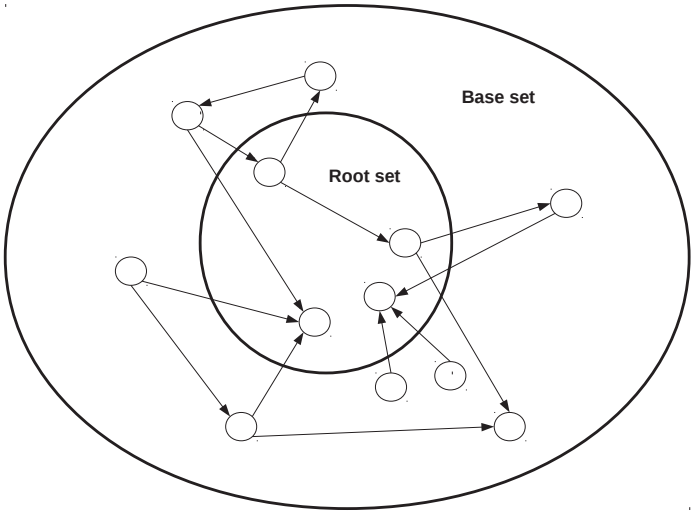

Figure 2.8 – Illustration de la phase d'expansion du root set vers le base set.

page qui est pointée par plusieurs bons hubs. On attribue à chaque page p un score d'autorité X_p et un score d'annuaire Y_p, tous deux initialisés à 1. Respectivement, plus ces valeurs sont élevées, et plus les pages sont susceptibles d'être de bonnes autorités ou de bons hubs. Ces valeurs sont exprimées sous forme de vecteurs de dimension égale au nombre de pages dans le base set. La relation de renforcement mutuelle entre annuaires Y et autorités X peut être formulée comme suit :

$$Xp = Somme\ des\ Yq,\ pour\ chaque\ page\ q\ citant\ la\ page\ p \tag{2.3}$$

$$Yp = Somme\ des\ Xq,\ pour\ chaque\ page\ q\ citee\ par\ la\ page\ p \tag{2.4}$$

Le calcul se fait de manière itérative en appliquant alternativement les formules 2.3 et 2.4. Les scores de Y et X sont normalisés à chaque itération. L'opération est répétée plusieurs fois jusqu'à convergence [49].

En pratique, le base set est un graphe et se code donc sous la forme d'une matrice d'adjacence A de taille $n \times n$ telle que $A_{ij} = 1$ si la page i cite la page j, et $A_{ij} = 0$ autrement. N correspond au nombre de pages du base set. Les deux équations vues précédemment peuvent donc être reformulées en terme de calcul matriciel. Il s'agit du produit de la matrice d'adjacence avec les vecteurs d'autorité et d'annuaire, tous deux initialisés à 1. Sachant que $X(k)$ et $Y(k)$ notent respectivement les vecteurs d'autorité et d'annuaire à l'itération k et que A^T est la matrice transposée de A, on obtient :

$$X(k) = A^t.Y(k) \tag{2.5}$$

49. Nous avons constaté empiriquement qu'il ne faut en moyenne pas plus de 20 itérations à l'algorithme pour converger.

$$Y(k) = A.X(k) \tag{2.6}$$

Pour illustrer le principe, reprenons le graphe présenté dans la figure 2.7. La matrice d'adjacence A est $\left(\begin{smallmatrix}0&0&1\\0&0&1\\0&0&0\end{smallmatrix}\right)$ et la transposée A^T est $\left(\begin{smallmatrix}0&0&0\\0&0&0\\1&1&0\end{smallmatrix}\right)$. Les vecteurs d'autorité et de hub sont initialisés à 1 : $\left(\begin{smallmatrix}1\\1\\1\end{smallmatrix}\right)$.

Les scores d'autorité sont donc calculés de la façon suivante :

$$X = A^T.Y = \begin{pmatrix}0&0&0\\0&0&0\\1&1&0\end{pmatrix} \bullet \begin{pmatrix}1\\1\\1\end{pmatrix} = \begin{pmatrix}0\\0\\2\end{pmatrix} \tag{2.7}$$

Les scores de hub sont :

$$Y = A.X = \begin{pmatrix}0&0&1\\0&0&1\\0&0&0\end{pmatrix} \bullet \begin{pmatrix}0\\0\\2\end{pmatrix} = \begin{pmatrix}2\\2\\0\end{pmatrix} \tag{2.8}$$

Notons que les scores doivent être normalisés à chaque itération[50]. Les scores normalisés d'autorité et de hub sont donc respectivement $\left(\begin{smallmatrix}0\\0\\1\end{smallmatrix}\right)$ et $\left(\begin{smallmatrix}0{,}7\\0{,}7\\0\end{smallmatrix}\right)$. Dans cet exemple, une seule itération suffit pour converger.

Le noeud 3 est la plus grande autorité de ce graphe et les noeuds 1 et 2 sont les hubs.

Principaux problèmes de la méthode Dans sa forme originelle, la méthode présente quelques problèmes plus ou moins gênants.

Tout d'abord, le temps de traitement peut être important, ce qui le rend difficilement utilisable pour une recherche en temps réel ; contrairement à un calcul d'autorité universel, le graphe est construit au moment de la requête et non lors de l'indexation. Néanmoins, dans le contexte d'une application de veille et dans l'optique de la découverte de nouvelles sources à surveiller, ce problème devient mineur.

De plus, le graphe est constitué en l'absence de toute analyse du contenu textuel. Or, on constate en pratique qu'il est primordial de filtrer les liens entrants et sortants selon leur contenu, de façon à éviter l'ajout dans le graphe de pages sortant de la thématique d'origine. Il n'y a en pratique pas nécessairement intersection entre proximité topologique et sémantique. Nous présentons dans la partie suivante de nombreux travaux cherchant à pallier cela. Il s'agit le plus souvent de filtrer (ou pondérer) les hyperliens en fonction de la présence ou non des termes de la requête dans le contexte immédiat du lien (*anchor text*).

Un problème de robustesse de HITS est signalé dans [Ng 2001] ; les résultats renvoyés ne sont pas stables face à des perturbations, parfois mineures, de la structure du graphe. En d'autres termes, l'ajout ou la suppression d'un petit nombre de liens peut parfois changer radicalement le classement des autorités. Cela s'explique en

50. Division de chaque valeur par la racine carrée de la somme des carrés de l'ensemble des valeurs ($\sqrt{\sum_{i=0}^{i=n} x_i^2}$).

partie par la taille des graphes utilisés. HITS est appliqué sur un graphe de quelques milliers de noeuds. Ce dernier est fortement interconnecté dans l'idéal mais assez rarement en pratique ; si les pages du root set contiennent peu de liens, il devient difficile d'en tirer quoi que ce soit. Il convient donc de construire le base set ou le root set différemment.

Faisant écho aux interrogations précédentes quant à la façon de constituer le base set, il convient aussi de s'interroger sur la confiance à accorder aux moteurs de recherche lors de la constitution du graphe. Le moteur de recherche utilisé dans [Kleinberg 1999a], ainsi que dans la plupart des travaux qui ont suivi, était Altavista. Nous l'avons vu, les moteurs de recherche procèdent désormais à un classement selon la structure globale du graphe du Web, auquel ils appliquent de nombreuses heuristiques : classement par l'usage, historique personnel, langue et géolocalisation de l'utilisateur, etc. Dès lors, le root set est fortement biaisé et les pages renvoyées risquent d'être de très mauvais hubs. La priorité est aussi donnée aux informations récentes, pénalisant ainsi potentiellement les pages anciennes. Dans l'optique d'une découverte de sources, il est primordial de constituer le root set de la façon la plus objective possible. Quelques pistes sont envisagées et présentées dans la partie 2.1.6.

Les descendants de HITS L'algorithme HITS a influencé de nombreux travaux et a été maintes fois repris à des fins commerciales. Nous décrivons ci-dessous une sélection de travaux que nous jugeons marquants.

Le projet CLEVER (*CLientside Eigen Vector Enhanced Retrieval* –extraction améliorée de vecteurs propres côté client), réalisé au sein du centre de recherche d'IBM Almaden, avait pour objectif d'utiliser HITS en vue de créer un moteur de recherche complet [51]. Le projet est actuellement abandonné. Des détails d'implémentation sont notamment apportés dans [Kumar 2006a]. De nombreuses améliorations sont effectuées. Tout d'abord, la phase d'expansion du root set se fait sur une profondeur de 2 : sont alors considérés les liens entrants des liens entrants ainsi que les liens sortants des liens sortants. Par ailleurs, le système admet un langage de requêtes plus complexe. Une stop list ainsi qu'une liste de pages exemplaires sont aussi ajoutées, permettant à l'utilisateur respectivement d'ignorer totalement certaines pages ou, au contraire, de partir de ces dernières pour constituer le root set. Enfin, on passe à un graphe de domaines, et non plus de pages Web. Ainsi, deux pages provenant d'un seul et même domaine ne peuvent plus se conférer d'autorité.

Afin d'assurer une meilleure cohérence thématique dans le graphe, une analyse textuelle est aussi effectuée dans CLEVER. Il s'agit alors d'analyser simplement le texte qui se trouve dans une fenêtre autour de l'hyperlien (mots avant et après). Cette fenêtre est appelée l'*anchor text*, introduit par [McBryan 1994]. Concrètement, lorsqu'une page p pointe vers une page q, l'*anchor text* de p autour de l'hyperlien est comparé avec les mots de la requête et on applique à l'arête un poids plus ou moins élevé selon le nombre de mots en commun. Plus un mot est éloigné de l'hyperlien,

51. Le moteur de recherche Teoma [Davison 1999], racheté en septembre 2011 par Ask Jeeves, reposait principalement sur les algorithmes de CLEVER. S'agissant d'une application commerciale, les améliorations apportées n'ont pas été rendues publiques.

plus sa contribution sera faible [Chakrabarti 1998c]. L'analyse des *anchor texts* est préférée à une analyse des pages cibles conformément à l'hypothèse énoncée précédemment selon laquelle les pages ne se décrivent pas de la même façon que des pages tierces. Les *anchor texts* sont plus susceptibles d'utiliser les mêmes mots que ceux de la requête et sont ainsi considérés plus révélateurs du contenu. Cette pondération des arêtes du graphe selon leur proximité avec les mots de la requête est plus ou moins reprise par l'ensemble des travaux ultérieurs.

[Bharat 1998a] étend la requête initiale en concaténant les 1000 premiers mots stemmatisés [Porter 1980] de chacun des documents du root set, en ignorant les mots grammaticaux. L'article signale aussi le problème des liens générés automatiquement ; ces derniers sortent très souvent de la thématique d'origine car il s'agit de liens publicitaires ou de licences. Les auteurs préconisent donc d'ignorer les hyperliens dont le nom de domaine apparaît trop souvent. Par ailleurs, il arrive parfois que plusieurs pages provenant du même domaine pointent vers une page d'un autre domaine. Inversement, une page peut pointer vers plusieurs pages partageant le même domaine. Cela a pour effet d'augmenter virtuellement les scores de hub ou d'autorité. Pour lutter contre ce biais, le poids des hyperliens prend en considération le nombre de pages pour chaque domaine.

SALSA [Lempel 2000] (acronyme de *Stochastic Approach for Link-Structure Analysis* –analyse de la structure des liens par une approche stochastique) remplace le renforcement mutuel par le *random walk*, par ailleurs utilisé dans le PageRank de Google. L'objectif est de lutter contre le *Tightly-Knit Community (TKC) Effect*. Ce problème apparait lorsqu'une communauté est petite mais très fortement interconnectée (les pages sont presque toutes reliées entre elles et chaque hub pointe ainsi vers toutes les autorités). Le renforcement mutuel risque de favoriser à tort ces sites qui vont recevoir toute l'autorité au détriment des autres. Le problème du TKC est aussi pris en compte dans la thèse de [Chikhi 2010].

L'algorithme *Hilltop* [Bharat 2002a] utilise un root set plus précis, exclusivement constitué de ressources "expertes", c'est-à-dire des pages possédant de nombreux liens vers des pages pertinentes. Il s'agit concrètement de pages avec un degré sortant supérieur à un seuil donné ($k = 5$ par exemple) et pointant vers k domaines différents. Les critères peuvent être durcis en appliquant les mêmes conditions sur tout ou parties des pages ainsi découvertes. Selon les auteurs, cela permet de distinguer les réelles ressources expertes des simples listes de liens. Enfin, afin de s'assurer que la ressource experte concerne la bonne thématique, les mots présents dans certaines balises HTML sont analysés ; par exemple, les mots contenus dans un entête vont permettre de qualifier thématiquement l'ensemble des hyperliens qui suivent ; le titre de la page va permettre de qualifier l'ensemble des URLs.

Lancé en 2003, Webfountain (IBM) est une application commerciale de l'algorithme HITS mêlant analyse de contenu et analyse de la topologie des hyperliens. Réalisée à des fins commerciales, relativement peu d'informations précises ont été divulguées.

[Nie 2006] reprend l'idée d'une autorité dépendante du contenu. Les auteurs signalent qu'une page recouvre assez souvent plusieurs thématiques différentes et qu'il

convient de prendre en compte cette hétérogénéité dans l'analyse des liens [52]. Pour ce faire, ils associent à chaque page, en plus des vecteurs d'autorité et de hub, un troisième vecteur défini par le contenu textuel. Il devient dès lors possible de différencier les scores de hubs et d'autorité par thématique. La classification thématique repose sur une classification naïve bayésienne des douze catégories supérieures de l'Open Directory Project. Le principe est repris et amélioré dans [Nie 2008].

Dans [Qi 2006] et [Qi 2008], les auteurs distinguent plusieurs types de relation de voisinage. Quatre types de voisins sont dissociés : parent, enfant, frère et conjoint. Ces relations sont schématisées dans la figure 2.9. Ces rôles ne sont pas exclusifs. Les auteurs énoncent que, dans le cadre d'une classification thématique des documents selon leur voisinage, la relation frère est la plus pertinente.

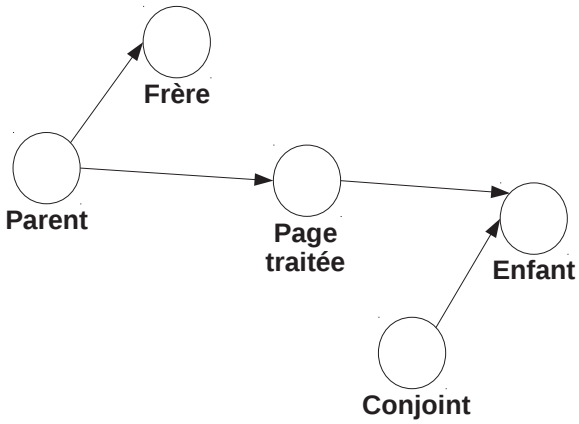

Figure 2.9 – Illustration des quatre types de relations distinguées dans [Qi 2006] et [Qi 2008] : parent, enfant, frère et conjoint.

Bien que légèrement plus éloigné de HITS, citons enfin [Noel 2012] dont l'objectif est de découvrir un ensemble de sources pertinentes et adaptées aux besoins de l'utilisateur. Pour construire le "profil utilisateur", un ensemble de concepts est extrait (à l'aide des bases de données de DBpedia) à partir d'un jeu de sources fourni par ce dernier. Les informations topologiques (liens entrants et sortants, co-citation, etc.) sont exploitées afin de pondérer l'importance des sources détectées.

2.1.4 Autorité et réseaux sociaux

2.1.4.1 Les réseaux sociaux comme caisse de résonance

Nous avons vu dans la partie 1.1.2.3 que les réseaux sociaux occupent de plus en plus de place dans le paysage actuel du Web. Depuis peu, nous constatons que ces derniers influent désormais sur les résultats renvoyés par les moteurs de recherche. Ils présentent en effet quelques caractéristiques intéressantes qui viennent faciliter

52. "A person may also be known in multiple contexts : a successful scientist might also be an amateur musician"[Nie 2006].

le processus de classement. Prenons l'exemple d'un blogueur postant l'URL d'un de ses articles sur des réseaux sociaux. Sur Facebook, l'URL pourra recevoir des *likes*, être commentée ou encore partagée. Sur Twitter, elle pourra notamment être retweetée. Ces diverses interactions sont autant d'empreintes laissées volontairement par des personnes et il est dès lors possible d'estimer, certes grossièrement, la taille du lectorat de la page derrière l'URL en question.

De fait, il est possible de considérer les réseaux sociaux, non pas comme une source d'information (contenu textuel pauvre), mais surtout comme une caisse de résonance, une passerelle entre les internautes et le reste du Web. Autrement dit, ils viennent combler le trou qui existait entre le Web statique et l'expression en temps réel [Easley 2010] : tout article de blog ou de presse pourra par exemple se voir commenté et partagé instantanément.

Ainsi, l'utilisateur se retrouve désormais au centre de toutes les attentions (Web *user-centric*). La plupart des moteurs de recherche prennent déjà en compte ces données "sociales" dans leurs algorithmes de classement. Bing a notamment un partenariat avec Facebook leur permettant de réorienter les documents selon les préférences de l'utilisateur par exemple, mais aussi celles de ses amis [53]. Il n'est pas étonnant de constater que Google a récemment lancé son propre réseau social pour concurrencer Facebook : il lui est dès lors possible de collecter directement l'ensemble des interactions sociales sans passer par un tiers.

Selon une étude réalisée dans [Cha 2010], 92 % des retweets (avec *RT* ou *via*) contiennent une URL, et 97 % d'entre eux contiennent aussi un nom d'utilisateur, ce qui semble confirmer l'intuition du réseau social comme caisse de résonance ; les retweets sont principalement une façon de diffuser du contenu Web et les utilisateurs citent leurs sources. Cette tendance est confirmée dans [Duan 2010]. Twitter est passé du statut de réseau social à celui de réseau d'information ; la question posée à l'utilisateur sur l'écran d'accueil, à l'origine *what are you doing ?*, a plus tard été remplacée par la phrase suivante : *what's happening ?* [54].

Dès lors, de nombreux travaux cherchent à calculer l'autorité des utilisateurs au sein d'un réseau social. Pour reprendre l'exemple de Bing, sont pris en compte le nombre de fois où une page a été (re)tweetée mais aussi l'autorité des utilisateurs qui l'ont (re)tweetée. Les liens inclus dans les tweets ont d'autant plus d'importance s'ils sont émis par un utilisateur faisant autorité. La plupart du temps, c'est Twitter qui est utilisé car il est très populaire mais, surtout, la plupart de ce qui s'y trouve est publiquement accessible [55] (seulement 8 % de comptes privés [Cha 2010]), au contraire de Facebook [56].

53. http://www.webrankinfo.com/dossiers/reseaux-sociaux/impact-referencement.

54. https://blog.twitter.com/2009/whats-happening.

55. Notons néanmoins que Twitter a récemment revu son modèle économique et a une volonté de plus en plus prononcée de restreindre les accès automatiques. Le nombre de requêtes possibles par heure via l'API ne cesse de diminuer et les accès massifs se monnaient désormais.

56. Il est néanmoins possible de savoir ce qu'un utilisateur fait dès lors que l'action a lieu sur une page publique. Par exemple, il est possible de savoir que l'utilisateur A commente l'intervention de B si ce commentaire a lieu sur une page publique (célébrité, marque, etc.), et ce même si le compte de A est privé. Plus d'informations disponibles ici : https://developers.facebook.com/

Il est important de souligner que cette utilisation des réseaux sociaux pour impacter le classement des pages Web est en substance assez similaire au principe du classement par l'usage, notamment utilisé par Google. Rappelons cependant les chiffres suivants [Cheng 2009] : sur Twitter, 65 % des utilisateurs ont moins de 25 ans et 5 % des utilisateurs génèrent 75 % de l'activité. Ainsi, en plus des inconvénients que présente le classement par l'usage, le panel des utilisateurs de réseaux sociaux n'est pas nécessairement représentatif de l'ensemble des utilisateurs du Web. La question reste ouverte.

2.1.4.2 Décompte du nombre d'amis

Une des approches les plus utilisées pour calculer l'autorité d'un utilisateur est l'étude du graphe des connections entre les utilisateurs. Il est intéressant de noter que les graphes des réseaux sociaux respectent globalement la même structure que le reste du Web (voir partie 2.1.3.3). [Cha 2010] remarque sur Twitter une grande composante connexe (94,8 % des utilisateurs –99 % des tweets générés), des singletons (5 % des utilisateurs), et des éléments plus petits (0,2 % des utilisateurs). Similairement, une étude sur Facebook [57] en 2011 montre que 99.91 % des utilisateurs sont interconnectés.

Une approche naïve consiste à tout simplement compter le nombre d'amis d'un utilisateur donné au sein du réseau [Leavitt 2009]. Sur Twitter, cela se traduit par le nombre de followers. En terme de graphe, cela correspond au degré entrant du noeud, étant entendu que les noeuds correspondent aux utilisateurs et que les arêtes [58] représentent les relations d'amitié. C'est la méthode actuellement employée par Twitter pour classer ses utilisateurs. C'est aussi le cas de services tiers [59]. Une autre métrique populaire est le ratio entre le nombre de followers d'un twittos et le nombre de personnes que ce dernier suit. Relativement aisée à mettre en place, la méthode se heurte au mêmes problèmes que le simple décompte du nombre d'hyperliens d'une page [Kleinberg 1999a], [Brin 1998b]. Une analyse plus fine du graphe des connections est envisagée dans les travaux ci-dessous.

[Java 2007] est un des premiers gros travaux en recherche d'information sur les microblogs. Les auteurs appliquent HITS sur le "graphe d'amitié". Les hubs sont alors les utilisateurs qui suivent beaucoup ; les autorités ont, à l'inverse, de nombreux followers. Cette rapide catégorisation permet de dégager trois types d'utilisateurs sur Twitter : ceux qui partagent l'information, ceux qui sont à la recherche de l'information, ceux qui entretiennent des relations d'amitié.

Dans [Weng 2010], c'est un dérivé du PageRank qui est appliqué sur le graphe des connections. Afin de mieux comprendre la façon dont l'information circule au

docs/reference/api/.

57. https://www.facebook.com/notes/facebook-data-team/anatomy-of-facebook/10150388519243859.

58. Le graphe est orienté dans le cadre de Twitter étant donné que la relation d'amitié n'est pas réciproque.

59. Citons notamment Twitaholic (http://twitaholic.com/) et Wefollow (http://wefollow.com/).

sein du réseau, les auteurs prennent aussi en compte les centres d'intérêt des twittos. Ils proposent alors le *TwitterRank*, une extension du PageRank, qui utilise aussi bien la structure des liens que la similarité de contenu entre les tweets des utilisateurs. Le principe est proche du *Topic-Sensitive PageRank* [Haveliwala 2002]. Afin de compenser le problème de la taille des tweets, les auteurs agrègent par thématique l'ensemble des tweets publiés par chaque twittos et y appliquent LDA (*Latent Dirichlet Allocation* [Blei 2003]), une méthode d'apprentissage non supervisée [60]. En comparant ainsi les thématiques abordées par les utilisateurs, les auteurs annoncent une plus forte proximité thématique entre les personnes qui se suivent.

Toujours dans [Weng 2010], les auteurs constatent que 72,4 % des utilisateurs suivent plus de 80 % de leurs followers. Ce phénomène de réciprocité peut s'expliquer de deux façons : les utilisateurs décident de suivre uniquement les personnes partageant des centres d'intérêt communs, ou bien il s'agit d'une convention d'usage qui fait que l'on suit, par politesse, les gens qui nous suivent. La dernière hypothèse remettrait en cause la pertinence du graphe d'amitié.

Par ailleurs, notons qu'un grand nombre d'amis traduit l'audience potentielle d'un utilisateur, sa popularité. Il n'est pas étonnant de constater que les comptes ayant un nombre conséquent de followers correspondent à des personnes connues publiquement (acteurs, hommes politiques, etc.) [61]. Évidemment, cette mesure n'a en réalité de sens que si chaque tweet est effectivement lu par l'ensemble de ses followers [62], ce qui n'est pas nécessairement le cas. [Leavitt 2009] montre qu'il n'existe pas de corrélation entre le nombre de followers et le nombre de fois que l'on est retweeté.

Afin d'estimer l'audience réelle, certains travaux s'intéressent non pas au graphe des connections mais au graphe des interactions entre utilisateurs.

2.1.4.3 Réseaux sociaux et citation

Chaque fois qu'un utilisateur A répond à un utilisateur B, mentionne son nom ou partage son contenu, il y a une interaction entre ces deux utilisateurs. Il est dès lors possible de construire un graphe d'interactions où les arêtes, orientées, traduisent la fréquence d'une interaction donnée entre les deux noeuds. Prenons l'exemple de Facebook. Si A *like* un contenu posté par B, il existera un lien de A vers B, dont la force correspond, par exemple, à la fréquence de cette interaction entre A et B. Si le degré entrant d'un noeud dans le graphe d'amitié traduit son audience potentielle, il se rapproche plus de l'audience réelle dans le graphe d'interaction. Il s'agit d'une preuve que A a interagit avec B. À noter cependant que, à l'instar de l'hyperlien, une interaction sur un réseau social n'est pas nécessairement une preuve d'approbation ou un vote positif.

60. Voir partie 3.1.2.3.

61. Au moment de la rédaction de ce mémoire, plus de 13 millions d'abonnés sur Twitter pour l'acteur Ashton Kutcher (*@aplusk*), 26 millions pour Barack Obama (*@BarackObama*), 33 millions pour Lady Gaga (*@ladygaga*).

62. Certains followers peuvent être inactifs, des bots, des comptes factices (spam).

L'autorité est donc ici définie comme la faculté de susciter des actions par d'autres utilisateurs. Travaillant sur un échantillon de plus d'1 milliard 750 millions de tweets, [Cha 2010] montre que les comptes ayant le plus grand nombre de followers sont très mentionnés, mais pas les plus retweetés. Les plus retweetés sont plutôt les services d'agrégation (Mashable, TwitterTips, etc.) et les sites d'information comme celui du journal *The New York Times*. La différence entre le nombre de followers et la faculté à être retweeté est par ailleurs confirmée dans [Kwak 2010].

Cette façon de classer les twittos selon leurs interactions est notamment utilisée par KLOUT [63], société spécialisée dans l'analyse des utilisateurs sur Twitter. Les réponses, retweets, commentaires et mentions sont pris en compte.

2.1.5 Biais et manipulations

Comme le signale [Le Deuff 2006], le classement des pages Web et des utilisateurs par l'utilisation de la citation parait démocratique et neutre. Néanmoins, gardons à l'esprit que les mécanismes de classement sont, dans les grandes lignes, connus de tous [64] et que les premières places dans le classement des moteurs de recherche valent de l'or [65].

La *search engine optimization* [66] (SEO) illustre parfaitement ce phénomène. Il s'agit d'un ensemble de procédés plus ou moins honnêtes visant à améliorer le positionnement d'une page dans les résultats d'un moteur de recherche. Les méthodes employées par les moteurs sont moins sensibles aux détournements qu'un simple décompte des liens entrants mais peuvent quand même être manipulées, par exemple avec des fermes de liens (*link farms*) ; il s'agit alors d'insérer des liens de façon stratégique dans le graphe du Web. Ces sites factices se pointent mutuellement. Les moteurs de recherche apportent régulièrement des modifications de leurs algorithmes [67] de façon à lutter contre ces détournements mais les mécanismes finissent toujours par être déjoués après un certain temps [68].

Similairement, certaines personnes cherchent à obtenir le plus rapidement possible un nombre conséquent d'amis sur les réseaux sociaux, particulièrement sur

63. http://klout.com/home.

64. Précisons néanmoins que les méthodes de classements de Google ne sont plus si transparentes de nos jours, le PageRank représentant actuellement une part minime dans toutes les heuristiques utilisées par le moteur.

65. Citons notamment les AdWords de Google. Google fait clairement la distinction entre les résultats commerciaux (à droite) et les autres. Tous les moteurs ne le font pas. On parle parfois de "triangle d'or" pour désigner les trois premiers résultats renvoyés.

66. Voir notamment http://searchengineland.com/.

67. Google a notamment lancé début 2011 une importante mise à jour appelée Panda dans le but de pénaliser tous les sites qui monopolisaient abusivement les premières places. Le plus souvent, ces sites présentent très peu d'intérêt et sont remplis presque exclusivement de publicités.

68. L'algorithme TrustRank [Gyongyi 2004] est un exemple de lutte contre ce type de détournement. Il s'agit d'attribuer manuellement des scores de confiance à un ensemble de pages de départ. Cet indice de confiance se propage ensuite de page en page de façon similaire au PageRank. Plus une page est éloignée dans le graphe d'une page de confiance, plus le TrustRank est faible. L'article annonce qu'il est possible de réduire le spam de façon considérable même en n'analysant qu'une petite partie du Web (200 pages).

Twitter, conscientes que le décompte du nombre de followers est une des méthodes les plus utilisées pour classer les utilisateurs. C'est d'autant plus important de nos jours étant donné que les moteurs de recherche prennent de plus en plus en compte dans le classement des pages Web de ce qui se dit au sein des réseaux sociaux. [Gosh 2012] (repris plus tard par [Dugué 2012]) introduit la notion de capitalistes sociaux ; il s'agit d'utilisateurs voulant capitaliser des données sociales afin de gagner en visibilité au sein de Twitter. Le cas est assez similaire pour les faux commentaires sur les forums ou les blogs (générés automatiquement ou par des personnes payées).

2.1.6 Résultats et discussion

L'algorithme HITS servira de point de départ à nos contributions. Nous le jugeons particulièrement adapté car il est centré sur un sous-ensemble pertinent de pages, ce qui est crucial dans un cadre de veille afin de qualifier précisément les émetteurs par domaine, et non simplement par leur visibilité globale sur le Web. Par ailleurs, la méthode distingue deux types de pages : les autorités (générateurs d'information) d'une part, et les hubs (diffuseurs d'information) d'autre part.

2.1.6.1 Notre implémentation

La version de HITS que nous avons implémentée reprend la description qui en est faite dans [Kleinberg 1999b]. Pour la recherche des liens entrants lors de la phase d'expansion du root set, nous avons initialement eu recours à la fonction *inlinks* de Yahoo [69]. Néanmoins, nous avons constaté au moment de la rédaction de ce mémoire que la fonction que nous utilisions n'est plus proposée par Yahoo [70]. Nous utilisons désormais un service proposé par Alexa [71].

Nous prenons aussi en compte de nombreuses évolutions proposées dans les papiers cités plus haut, notamment la lutte contre le renforcement mutuel entre hôtes [Bharat 1998b]. Nous avons aussi recours à une stop list pour ignorer les sites commerciaux très visibles sur le Web [72].

Nous analysons aussi l'*anchor text* pour pondérer les arêtes au sein du graphe. En complément des mots de la requête, nous proposons d'étendre la requête d'origine à l'aide d'un jeu de termes extraits automatiquement à partir des pages du root set [73]. Enfin, nous avons recours à la fonction de filtrage topologique des liens [74] : les

69. Yahoo a été choisi car l'API de Google ne renvoyait qu'un nombre limité de liens entrants.
70. L'arrêt de certaines fonctionnalités s'explique par une migration d'une partie des infrastructures de Yahoo vers Microsoft.
71. `http://www.alexa.com/`.
72. Par exemple :
    — `http://www.amazon.com/` ;
    — `http://jigsaw.w3.org/css-validator/check/referer` ;
    — `http://creativecommons.org/`.

73. Outil de *text mining* développé par AMI Software.
74. Rappelons qu'il s'agit d'une fonctionnalité développée par AMI Software et utilisée au sein du produit. Elle a été décrite brièvement dans la partie 1.1.4.1.

hyperliens hors du bloc principal sont ignorés. Par ailleurs, s'il y a plus de n liens entrants ou sortants vers un même domaine, on les ignore étant donné qu'il s'agit très probablement d'hyperliens générés automatiquement : il s'agit alors très probablement de bannières, sponsors, ou liens publicitaires.

Enfin, nous revoyons légèrement la phase d'expansion du root set car nous avons constaté que le base set obtenu était très rarement assez dense pour être exploitable. La figure 2.10 illustre les structures "en étoiles" typiques des graphes ainsi obtenus. Appliquer HITS sur de tels graphes revient à classer les noeuds en fonction de leur degré entrant.

Figure 2.10 – Illustration de la structure "en étoile" observée sur de nombreux sous-graphes extraits du Web.

Pour obtenir un base set plus dense, plusieurs pistes non exclusives ont été envisagées. La première consiste à passer d'un graphe de pages Web à un graphe de domaines : deux pages d'un même blog par exemple ne constituent plus deux noeuds différents mais sont fusionnées dans un seul. Il peut cependant être intéressant de garder une trace des différentes pages agrégées pour des éventuels traitements ultérieurs. La deuxième piste consiste à augmenter la profondeur lors de la phase d'expansion (passage du root set au base set). Par exemple, avec une profondeur de 2, on ne se contente non plus seulement d'extraire les liens entrants et sortants de chacune des pages du root set, mais on prend aussi les liens entrants et sortants de chacune d'entre elles. Cette solution a déjà été proposée lors du projet *CLEVER*.

2.1.6.2 Validation de l'implémentation

L'évaluation de ce genre d'approche est loin d'être une tâche triviale. Cela implique de connaître a priori l'ensemble des pages pertinentes pour une requête donnée, ce qui est impossible sur le Web. La façon de procéder de l'ensemble des auteurs pour évaluer les résultats est sensiblement la même [Bharat 1998a] [Bharat 2002a] [Chakrabarti 1998c].

La première approche est la comparaison avec des ressources externes. Certains auteurs comparent leurs résultats à ceux d'articles plus anciens (notamment ceux

utilisés par l'article d'origine). Nous pensons que cette approche n'est plus pertinente de nos jours étant donné que le Web a beaucoup évolué depuis. Il est aussi possible de recourir aux résultats renvoyés par d'autres moteurs de recherche [75]. Enfin, il est courant d'avoir recours à des annuaires, ce qui peut poser des problèmes d'exhaustivité et de subjectivité signalés dans la partie 2.1.2.1. Par ailleurs, que cela soit pour les moteurs ou les annuaires, il peut y avoir de nombreux biais éventuels (boîte noire). Enfin, si un résultat est renvoyé par un moteur (voir plusieurs en même temps [76]) et/ou se trouve dans un annuaire, il est possible d'inférer que le résultat est bon, mais l'inverse n'est pas vrai. C'est pourquoi certains auteurs ont recours à une évaluation humaine, impliquant encore une fois subjectivité et problème d'exhaustivité.

Étant donnés les biais exposés précédemment, nous préférons parler de *validation* des résultats, plutôt que de réelle évaluation. Conscients de ces limites, nous avons tout de même comparé nos résultats avec ceux de Google et Yahoo afin de valider l'implémentation. Nous utilisons aussi les résultats de la version française de l'Open Directory Project, qui est un annuaire généraliste. Malgré les réserves émises précédemment, l'ajout des liens dans cet annuaire semble peu influencé par un modèle économique particulier. Enfin, il est très complet et possède de nombreux sites français. C'est le seul annuaire qui, à notre connaissance, répond à l'ensemble de ces critères. Néanmoins, notons que la version française de l'annuaire compte "uniquement" 142 802 URLs [77] et que les catégories sont inégalement représentées.

Sans surprise, les résultats sont bien meilleurs [78] avec le filtrage par l'*anchor text*. Nous constatons par ailleurs une légère amélioration par l'extension de la requête avec les termes extraits automatiquement. Néanmoins, nous ne nous attarderons pas plus sur ce point car nous jugeons qu'il ne s'agit pas tant ici de comparer les performances de notre implémentation de HITS que de réfléchir à la validité et à l'utilité des résultats obtenus. Nous avons testé notre algorithme sur de nombreuses requêtes et nous nous concentrons ci-dessous sur quelques jeux de résultats particulièrement révélateurs des limites.

2.1.6.3 Discussion

L'ensemble des résultats que nous présenterons ci-dessous a été calculé avec les paramètres indiqués dans le tableau 2.1.

Requête : "avortement" La figure 2.11 présente les 10 premières autorités renvoyées par notre version de HITS pour la requête "avortement". Nous constatons que 4 pages sont plutôt en faveur du droit à l'avortement :

— Sosbebe.org : pour les femmes en difficultés avec la grossesse ;

75. Les moteurs les plus utilisés étaient alors Altavista et Yahoo.

76. Notons qu'il faut s'interroger sur l'intérêt de l'approche si les résultats renvoyés sont les mêmes que ceux des moteurs classiques reposant sur un calcul d'autorité universel...

77. Au moment de la rédaction.

78. En terme de bruit. La plupart des résultats renvoyés, en l'absence de toute analyse de contenu, est composée de sites complètement hors sujet.

Taille du root set	200
Moteur utilisé	Yahoo
Profondeur des liens	2

Tableau 2.1 – Paramètres utilisés pour nos expérimentations. On considère un root set d'une taille de 200, constitué à l'aide du moteur de recherche Yahoo. La phase d'expansion vers le base set se fait sur une profondeur de 2.

— Association Nationale des Centres d'Interruption de Grossesse et de Contraception ;
— Avortement ? Interruption de grossesse : pour le droit au libre choix ;
— Mouvement français pour le planning familial.

4 sont informatives :

— Aufeminin.com : la faiseuse d'ange et l'avortement du moyen-âge à nos jours ;
— Wikipedia : Abortion ;
— Le Monde Diplomatique : Le droit à l'avortement dans le monde
— Liberalism.ro : De l'avortement volontaire.

2 sont contre l'avortement :

— Laissez-les-vivre : SOS futures mères ;
— Droit de naître.

Sont donc mis au même plan des éléments difficilement comparables. Comment interpréter ces résultats ? Doit-on en déduire que le Mouvement français pour le planning familial fait plus autorité que les associations anti-IVG comme Droit de naître ?

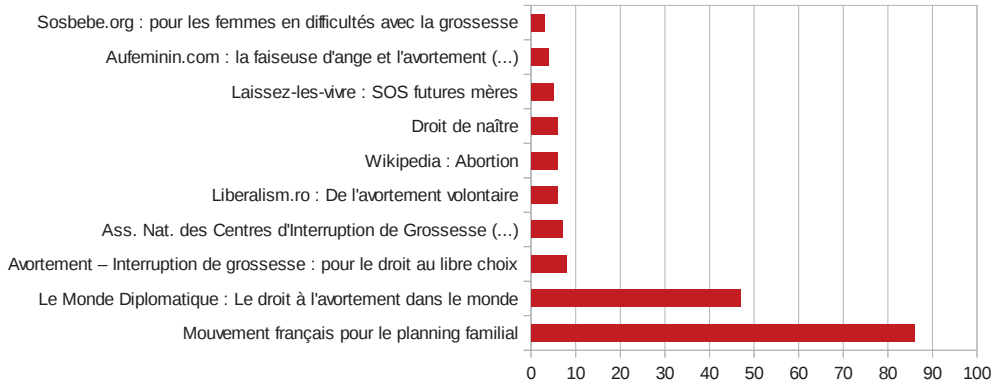

Figure 2.11 – 10 premières autorités renvoyées par notre implémentation de HITS pour la requête "avortement".

Requête : "réforme retraites" Le problème est assez comparable dans les résultats de la figure 2.12. Sont présentés les 10 premiers résultats de notre algorithme en réponse à la requête "réforme retraites". Il s'agit majoritairement de sites officiels du gouvernement. En troisième position, avec un score très bas, se trouvent les propositions du Parti Socialiste concernant les retraites. En vingtième position (hors du tableau) se trouvent les propositions du MoDem. Encore une fois, que penser de ce classement ?

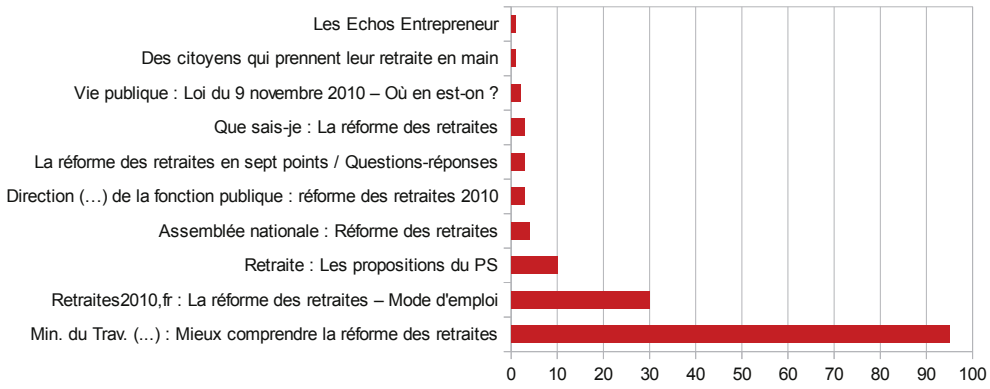

Figure 2.12 – 10 premières autorités renvoyées par notre implémentation de HITS pour la requête "réforme retraites".

Bien entendu, ces résultats sont tout à fait compréhensibles et sont révélateurs de la visibilité des pages en question sur le Web. Le problème est ici principalement interprétatif. La décontextualisation des résultats, qui découle de leur présentation traditionnelle sous forme de listes, fait qu'il est difficile a priori de remettre en cause la pertinence des résultats renvoyés. Comme le signale à juste titre [Broudoux 2007], *du fait de l'absence d'explications claires divulguées au grand public, sur les procédés employés pour générer des résultats, on peut avancer que les moteurs de recherche font autorité par opacité ou omission d'informations.* Le calcul d'un score d'autorité, aussi respectueux soit-il de la requête de l'utilisateur, ne sera jamais qu'un indicateur artificiel.

Nous pensons qu'il est primordial de proposer des outils de recontextualisation de l'information. L'idée n'est pas de remettre en cause le principe même d'indicateur, mais il est nécessaire pour le veilleur d'avoir une vision globale afin d'effectuer les meilleurs choix possible. En d'autres termes, il s'agit de passer du paradigme "rechercher et extraire" à une métaphore "organiser et naviguer" [Savoy 2000].

2.2 Approche proposée : calcul d'autorité et communautés

Selon nous, une interaction itérative entre l'utilisateur et l'outil combinant calcul d'autorité, détection de communautés et outils de visualisation adéquats est une condition sine qua non à la constitution d'un bouquet de sources précises autour d'une thématique donnée.

Les outils de qualification des sources doivent permettre au veilleur d'identifier clairement les émetteurs qui sont des vecteurs potentiels de diffusion ou d'amplification de l'information autour de la thématique souhaitée. Pour ce faire, il est primordial de contextualiser au maximum l'information, de façon à lui proposer toutes les pistes interprétatives nécessaires.

2.2.1 Principe général

Nous décrivons ici l'approche proposée, qui permet une meilleure contextualisation des sources, et aide ainsi le veilleur dans son parcours interprétatif. La figure 2.13 illustre notre approche.

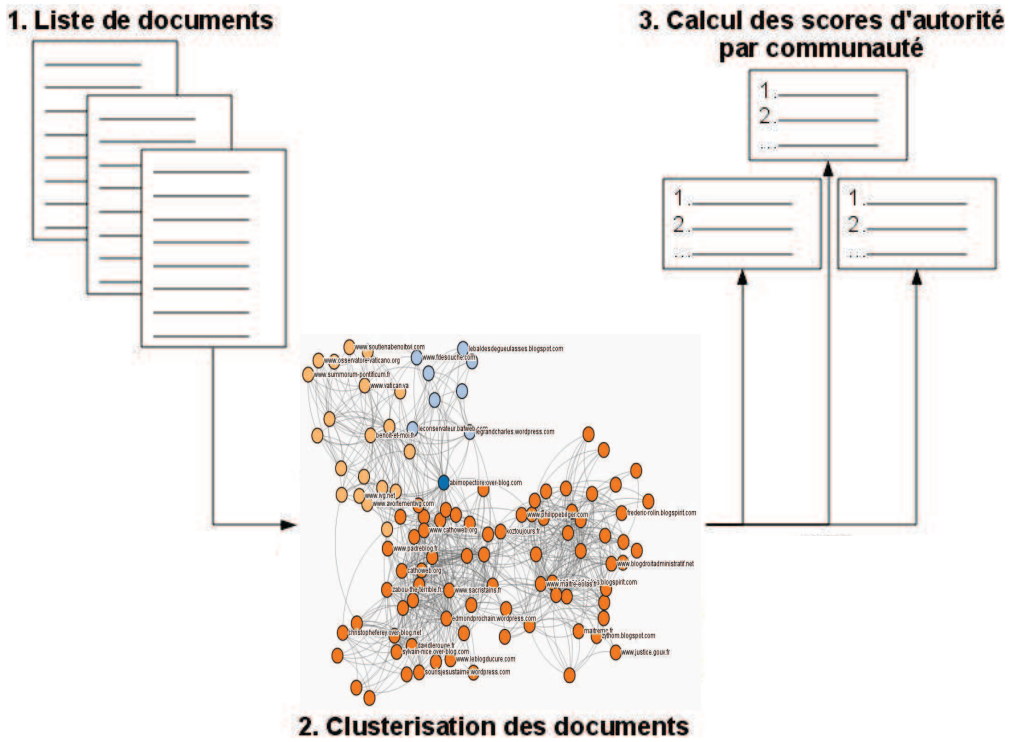

Figure 2.13 – Illustration de l'approche proposée. Un découpage du graphe est effectué en amont de tout calcul d'autorité des sources.

Nous considérons qu'une page n'est d'autorité que pour une communauté donnée. En d'autres termes, il devrait y avoir autant d'ensembles de sources de qualité

qu'il existe de "facettes" pour une requête donnée. Nous proposons donc de calculer l'autorité des sources après détection des communautés. On calcule ainsi un score d'autorité pour une requête donnée, mais aussi pour une communauté donnée.

2.2.2 Détection de communautés

Détecter des communautés en ligne implique d'analyser la topologie des graphes et de regrouper les noeuds proches. Il s'agit donc de faire du partitionnement (*clustering*) de graphe. La façon de regrouper les noeuds dépend de la définition que l'on donne à la communauté mais il s'agit la plupart du temps de détecter les ensembles denses de noeuds, ou de briser les arêtes les plus faibles.

Notons qu'il s'agit ici de dégager des regroupements à l'aide de la topologie des sous-graphes du Web –il ne s'agit pas nécessairement de regroupements thématiques ; la corrélation entre proximité topologique et sémantique reste à démontrer. L'analyse des regroupements thématiques sera abordée dans le chapitre 3.

2.2.2.1 Principales approches

S'agissant d'un domaine particulièrement dynamique, nous ne dresserons pas un état de l'art exhaustif des algorithmes de partitionnement de graphes [Porter 2009] [Fortunato 2010]. En règle générale, on définit une partition comme un groupe de noeuds densément connectés entre eux mais peu connectés avec les autres groupes. [Estivill-Castro 2002] explique le grand nombre d'algorithmes de ce type par le fait que la définition de ce qu'est une partition, et donc sa formalisation mathématique, dépend fortement de l'application et de l'avis de l'auteur : chaque méthode tente de définir plus précisément ce "densément connecté". Nous dressons ci-dessous un résumé concis des méthodes les plus populaires, avant de nous attarder sur quelques méthodes permettant le chevauchement de partitions que nous avons jugées intéressantes pour notre tâche.

Méthodes par partitionnement Les deux approches les plus connues sont les méthodes par partitionnement et les méthodes hiérarchiques. L'algorithme des k-means[79] est la méthode par partitionnement la plus utilisée [Lloyd 1957] [Steinhaus 1957] [MacQueen 1967]. Certains noeuds du graphe sont choisis pour service de barycentre aux partitions. Le choix peut se faire aléatoirement ou selon certains critères, par exemple en prenant les noeuds les plus éloignés les uns les autres. À chaque itération, les noeuds sont affectés au barycentre le plus proche. La distance entre deux noeuds peut se calculer de différentes façons : distance euclidienne, cosinus, etc. Les barycentres de chacune des partitions ainsi obtenues sont recalculés. Il y a convergence lorsque les barycentres restent stables pour chaque partition.

Le partitionnement final dépend du choix initial des barycentres et n'est donc pas déterministe. Néanmoins, il est possible de lancer l'opération sur plusieurs points

79. On parle aussi parfois en français de k-moyennes.

d'origine afin d'obtenir la meilleure solution, c'est-à-dire celle qui minimise au maximum le nombre d'arêtes entre les partitions. Cela permet par ailleurs de détecter la présence de "noyaux durs", c'est-à-dire ceux qui changent peu ou pas dans tous les cas de figure.

De nombreux travaux reprennent la méthode, notamment [Zhong 2005] permettant de meilleures performances pour des résultats comparables. [Cleuziou 2007] permet le chevauchement de partitions (*fuzzy k-Means*). Dans tous les cas, le principal inconvénient de l'approche est le besoin de connaître a priori le nombre de partitions, ce qui est impossible dans notre cas.

Méthodes hiérarchiques Contrairement aux algorithmes par partitionnement, les méthodes hiérarchiques ne nécessitent pas en entrée le nombre et la taille des partitions à extraire. Les travaux de [Newman 2001] et [Girvan 2002] entraînent un regain d'intérêt pour ce genre d'approches.

Il existe deux types de méthodes hiérarchiques : agglomératives et séparatives. Dans les premières, chaque noeud correspond au départ à une partition. Les deux partitions les plus proches sont récursivement fusionnées. Les secondes procèdent dans l'autre sens : chaque noeud appartient à l'origine à une seule et même partition divisée récursivement en brisant les arêtes.

Le fonctionnement même de ces algorithmes implique une structure hiérarchique des communautés sous forme d'arbre que l'on nomme dendogramme ; les noeuds représentent les communautés et les feuilles sont les noeuds du graphe. La figure 2.14 illustre un dendogramme obtenu par agglomération. Des fonctions de qualité attribuant un score à chaque partition sont utilisées pour tenter de déterminer le meilleur partitionnement possible à partir de cette structure hiérarchique. La modularité, introduite dans [Girvan 2002], est de loin la plus utilisée. Néanmoins, l'exercice est loin d'être trivial.

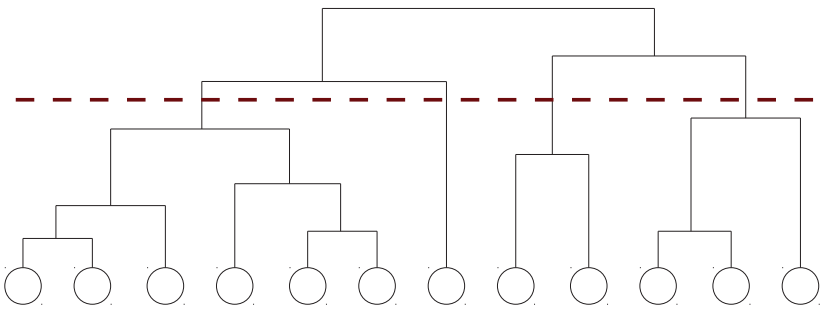

Figure 2.14 – Exemple de dendogramme construit par agglomération. Chaque noeud correspond à une communauté. Déterminer le meilleur partitionnement dans cet ensemble, ici marqué en rouge, n'est pas trivial.

Outre la difficulté à déterminer le meilleur partitionnement du graphe, notons que les résultats dépendent fortement de la mesure de similarité adoptée. De plus, le partitionnement est nécessairement hiérarchique, quelle que soit la structure du

graphe en entrée ; la relation hiérarchique peut ainsi dans certains cas être purement artefactuelle. Enfin, ce genre d'approches est coûteux, entraînant des problèmes de passage à l'échelle.

Méthodes locales Les méthodes présentées précédemment sont globales, c'est-à-dire qu'elles considèrent le graphe dans son ensemble. Toute modification de ce dernier a ainsi une incidence sur les partitions obtenues, même si celles-ci sont composées de noeuds très éloignés. Par ailleurs, la taille des plus petites partitions dépend très fortement de la taille totale du graphe [80]. Enfin, pour des raisons évidentes de complexité algorithmique, leur implémentation a le plus souvent recours à des approximations gloutonnes de la solution souhaitée.

Nous allons désormais nous intéresser aux méthodes permettant le chevauchement entre partitions, c'est-à-dire qu'un noeud peut appartenir à plusieurs partitions à la fois. La plupart de ces méthodes sont locales ; elles traitent chaque noeud en ne considérant que son voisinage immédiat, ce qui semble intuitivement pertinent notamment dans le cas du Web : les communautés se forment et se défont indépendamment de ce qui peut avoir lieu à l'autre bout du graphe. Par ailleurs, étant donné que les noeuds ne s'intéressent qu'à leurs voisins immédiats, ces algorithmes sont particulièrement adaptés à la parallélisation.

[Palla 2005] décrit la *Clique Percolation Method* (CPM), qui est l'un des premiers algorithmes à permettre le chevauchement de partitions. L'approche est celle que nous avons retenue et sera donc décrit plus précisément dans la partie 2.2.3.1.

L'algorithme *RAK* de [Raghavan 2007] repose sur une diffusion virale d'étiquettes entre noeuds. Au départ, une étiquette unique est assignée à chacun des noeuds. À chaque itération, les noeuds prennent pour nouvelle étiquette la plus diffusée parmi l'ensemble de ses voisins. En cas de concurrence, le choix est aléatoire, rendant l'algorithme non déterministe. Il y a convergence si aucune étiquette ne change. Pour contourner l'aspect non déterministe de la méthode, les auteurs préconisent de lancer plusieurs fois l'opération et proposent des heuristiques pour agréger les différentes solutions. Cela permet de détecter les noeuds à cheval entre plusieurs partitions. [Leung 2009] optimise l'algorithme et corrige notamment le problème des *pestes galopantes*, qui se traduisait par la formation d'énormes communautés à cause de certains noeuds infectant un nombre trop élevé de voisins. La version de [Gregory 2010] (algorithme *COPRA*) permet de prendre en compte plus explicitement le chevauchement de partitions en attribuant un poids à chacune des étiquettes. L'inconvénient principal de cette approche, l'originale comme ses successeurs, est son non-déterminisme, bien que les solutions soient très souvent assez proches les unes des autres. Néanmoins, l'algorithme présente l'avantage d'être rapide ; le nombre d'itérations requises est assez faible et peu dépendant de la taille du graphe à partitionner.

Enfin, une approche intéressante est proposée dans [Gregory 2009]. L'auteur

80. On parle alors de *resolution limit problem*.

décrit une méthode universelle permettant à n'importe quel algorithme préexistant de prendre en compte le chevauchement de communauté.

Propagation de l'information et évolution dans le temps Étant donné la difficulté à partitionner un graphe à un instant t, il n'est pas étonnant de constater le peu de travaux abordant l'évolution des graphes dans le temps. Analyser l'évolution de la structure communautaire des graphes permet de mieux comprendre comment les noeuds interagissent entre eux, comment les groupements se forment, ouvrant ainsi la voie vers des modèles de prédiction des schémas d'évolution de ces derniers [Palla 2007a].

[Leskovec 2008a], entre autres, applique ce type de modèles prédictifs sur des sous-graphes issus du Web afin de mieux comprendre comment l'information se propage au sein des réseaux. La diffusion d'une information est ici vue comme une épidémie qui se diffuse de noeud en noeud. Les données étant datées dans le temps, il est possible d'inférer le graphe de propagation.

2.2.2.2 La question de l'évaluation

La constitution de jeux de test (*benchmark graphs*) génériques et fiables pour comparer les méthodes entre elles est un réel enjeu [Lancichinetti 2008]. Ces graphes doivent contenir des partitions naturelles que n'importe quel algorithme devrait être en mesure de détecter. Néanmoins, en l'absence de consensus sur la définition d'une partition, la tâche est d'autant plus complexe [Estivill-Castro 2002]. Au mieux est-il possible de s'appuyer sur des fonctions de qualité pour évaluer la "pertinence" des jeux de test, mais il s'agit ici d'un niveau d'approximation supplémentaire.

L'utilisation de plusieurs méthodes sur les mêmes graphes permet de déterminer si les résultats dépendent ou non de l'algorithme choisi. Si les partitions sont par ailleurs connues en amont, il est possible d'évaluer le taux de réussite dans le partitionnement. En tout cas, les tests peuvent être effectués sur deux types de graphes : générés automatiquement ou issus de situations réelles.

[Girvan 2002] introduit une méthode permettant de générer automatiquement un graphe de test composé de 128 noeuds divisés équitablement en 4 groupes. Malgré la grande popularité de la méthode, les graphes ainsi générés ressemblent très peu à des graphes réels et présentent les inconvénients suivants : petite taille, noeuds ayant quasiment tous le même degré, et communautés de taille égale. Quelques méthodes plus récentes, notamment [Lancichinetti 2008], permettent la génération de graphes plus proches des graphes réels tant dans la taille des groupes que dans la distribution des noeuds.

Les graphes issus de situations réelles[81] présentent l'avantage de tester l'algorithme de façon moins abstraite. Bien entendu, le choix du graphe réel utilisé pour évaluer une méthode dépend idéalement des objectifs de chacun et du type de

81. Quelques exemples sont disponibles à l'adresse suivante : `http://www-personal.umich.edu/~mejn/netdata/`.

données à analyser.

Enfin, [Estivill-Castro 2002] rappelle à juste titre la part de subjectivité inhérente à l'évaluation de la qualité du partitionnement sur graphes réels. Dans une étude réalisée par [Macskassy 1998], dix personnes devaient partitionner manuellement un petit jeu de documents renvoyés par un moteur de recherche. Au final, il y avait très peu de similarité entre les regroupements effectués. Le partitionnement manuel d'un graphe dépend de l'interprétation de chacun. Par ailleurs, il convient de garder à l'esprit que la simple structure du graphe peut dans certains cas n'être pas suffisante, ou être artefactuelle. Par exemple, le fait que deux utilisateurs ne soient pas amis sur un réseau social ne signifie pas nécessairement que les deux personnes ne se connaissent pas ou qu'ils ne partagent pas les mêmes centres d'intérêt, et inversement.

2.2.3 Approche retenue

Le choix des algorithmes à utiliser devait selon nous reposer sur les critères suivants :

- Le partitionnement ne doit pas nécessairement être exhaustif. Malgré les filtres textuels mis en place, le bruit parmi les pages reste assez important. En ne retenant que les regroupements forts, on exclut idéalement ces pages.
- Le chevauchement doit être permis : une page peut appartenir à plusieurs communautés.
- Les résultats ne doivent pas être trop rigides et doivent permettre un certain niveau de granularité.
- L'analyse doit requérir le moins possible de données en entrée ; il est en effet difficile de connaître a priori le nombre de communautés à rechercher ou leur taille.
- Enfin, les temps de calcul doivent être acceptables.

Suivant ces critères, deux algorithmes en particulier ont retenu notre attention et ont été implémentés : l'algorithme *RAK* proposé par [Raghavan 2007] puis étendu par [Gregory 2010] afin de permettre le chevauchement de communautés, et la *Clique Percolation Method* de [Palla 2005]. Nous utiliserons finalement ce dernier algorithme car c'est celui qui répond au mieux aux critères choisis et qui nous a apporté empiriquement les résultats les plus satisfaisants.

2.2.3.1 Description de l'approche

La *Clique Percolation Method* (CPM) est décrite la première fois dans [Palla 2005]. L'algorithme a depuis été amendé de nombreuses fois, notamment par [Palla 2007b] et [Farkas 2007] pour prendre en compte, respectivement, les graphes dirigés et la pondération des arêtes. [Kumpula 2008] développe une implémentation plus rapide de cet algorithme, alors appelée *Sequential Clique Percolation Algorithm* (*SCP*).

Une partition est ici définie comme l'union de toutes les *k-cliques* qui peuvent être atteintes par toutes les autres via des *k-cliques* adjacentes. Rappelons qu'une clique est, dans un graphe non orienté, un sous-ensemble de noeuds pour lesquels il existe une arête entre chaque paire ; ce sous-ensemble est un graphe complet. Une *k-clique* est une clique à k sommets. Un triangle est un exemple de *3-clique*. Deux *k-cliques* sont adjacentes si elles partagent *k-1* noeuds. Par exemple, une *partition 3-clique* est l'union des triangles reliés par une série de liens en commun. Plus k est élevé, moins il y a de partitions mais plus leur cohésion est forte.

Rechercher naïvement l'ensemble des *k-cliques* est une tâche coûteuse. L'algorithme, schématisé dans la figure 2.15, extrait donc tout d'abord l'ensemble des cliques maximales[82], c'est-à-dire les sous-graphes complets du graphe qui ne font pas partie d'un sous-graphe complet plus large. Une fois ces cliques détectées, la matrice A de chevauchement des cliques est construite. A_{ij} traduit le nombre de noeuds que la clique i a en commun avec la clique j. Les diagonales marquent la taille des cliques. Les *k-cliques* sont extraites en mettant à 0 les éléments non diagonaux plus petit que $k-1$, à 0 les éléments diagonaux plus petits que k, et à 1 les éléments restants. Les partitions correspondent à chaque sous-graphe séparé.

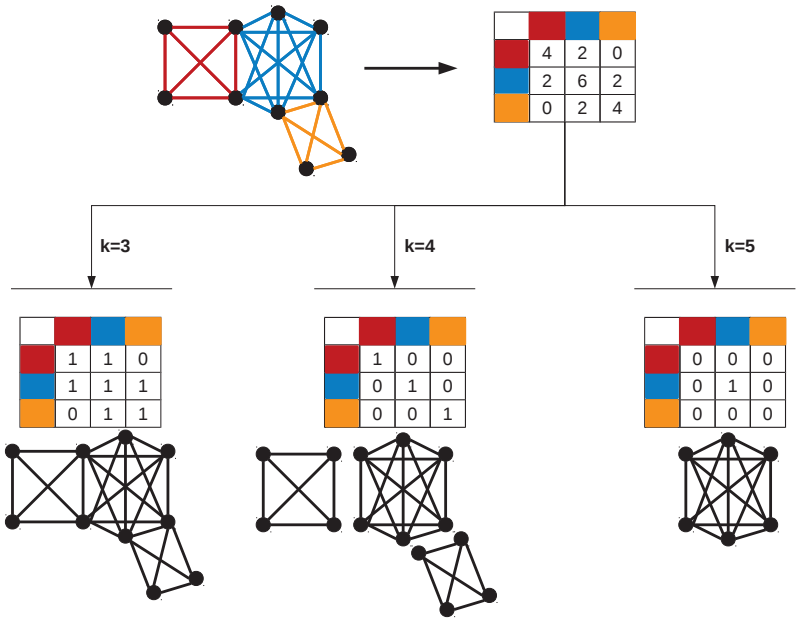

Figure 2.15 – Illustration de l'algorithme employé dans [Palla 2005] pour extraire les *k-cliques* d'un graphe, pour $k = 3$, $k = 4$, et $k = 5$.

Le principal inconvénient de la méthode est, selon nous, la rigidité de la notion de clique qui peut parfois empêcher de détecter des modules denses mais pas assez bien connectés pour entrer dans la définition. Les résultats et le temps de traitement sont par ailleurs trop fortement dépendants de la présence ou non de

82. L'énumération de l'ensemble des cliques d'un graphe est, en soi, un axe de recherche très riche. Citons notamment [Bron 1973], [Tsukiyama 1977], ou encore [Cazals 2008].

cliques dans le graphe. D'autres problèmes, souvent relevés dans la littérature, sont le choix a priori arbitraire de la valeur de k ainsi que la non-inclusion de certains noeuds dans le partitionnement. Nous pensons cependant que le premier point n'est pas gênant dans notre cas : tout noeud n'a pas à appartenir nécessairement à une communauté.

2.2.3.2 Résultats

Nous avons, lors de nos phases de test, utilisé le logiciel CFinder [83], une implémentation de l'algorithme *CPM* directement issue des travaux de [Palla 2005] car elle offre en outre des outils de visualisation simples d'utilisation.

Requête : "avortement" La figure 2.16 illustre quelques communautés extraites du graphe construit avec la requête "avortement" (pour k=5).

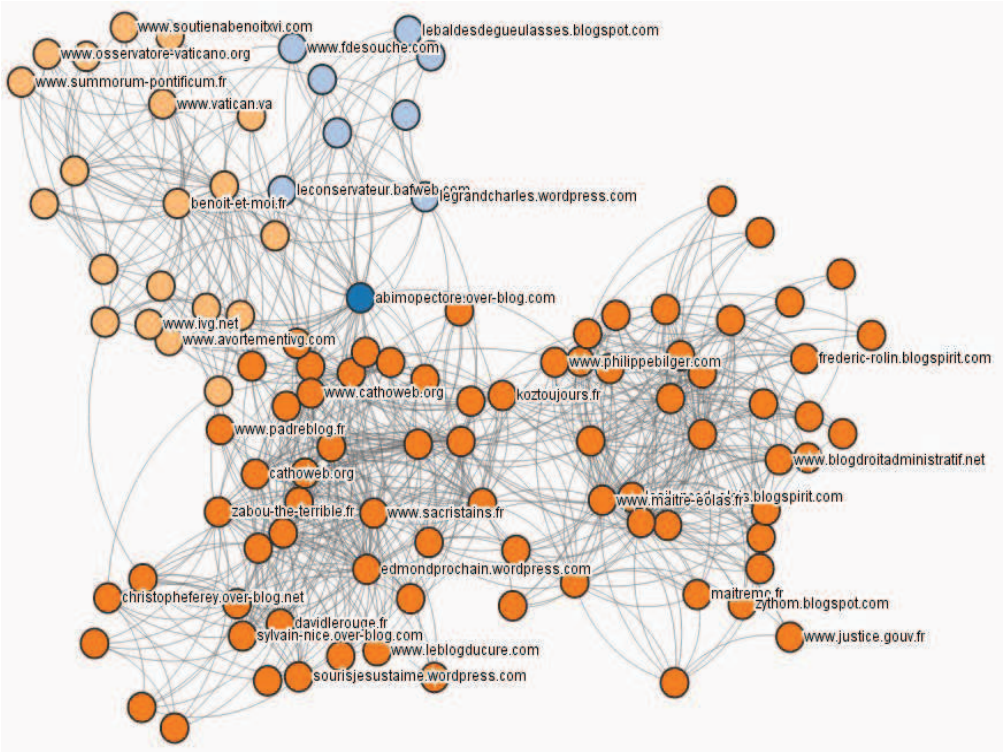

FIGURE 2.16 – Illustration de communautés extraites à partir du base set pour la requête "avortement".

On observe 3 communautés se chevauchant en partie. Cette segmentation est intéressante, car elle met clairement en avant certaines tendances dans les propos

83. http://angel.elte.hu/cfinder/.

tenus par les auteurs : la communauté en bleu est fortement réactionnaire, la communauté en beige parle du Vatican et de la vie de l'Église en général, tandis que la orange aborde des sujets du quotidien vus à travers des yeux de jeunes catholiques modérés (une dizaine de sites est d'ailleurs tenue par des prêtres). Ce découpage est confirmé par la présence fréquente de bannières affichant clairement l'appartenance des auteurs à tel ou tel réseau : label "certifié réacosphère" en bleu, le réseau "Riposte-Catholique"[84] ainsi que le soutien au Pape Benoît XVI[85] en beige, et enfin le réseau "Sacristains"[86] en orange.

Les figures 2.17, 2.18, et 2.19 présentent les résultats du classement par autorité pour chacune des communautés citées précédemment.

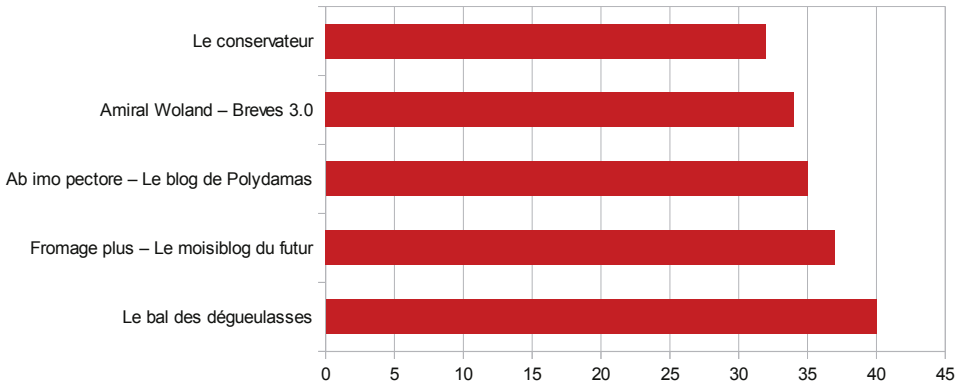

Figure 2.17 – 5 premières autorités parmi la communauté indiquée en bleu dans la figure 2.16. Cette communauté se caractérise par des sites majoritairement réactionnaires. Les 4 premiers affichent le label "certifié réacosphère".

84. http://www.riposte-catholique.fr/.
85. http://www.soutienabenoitxvi.com/.
86. http://www.sacristains.fr/.

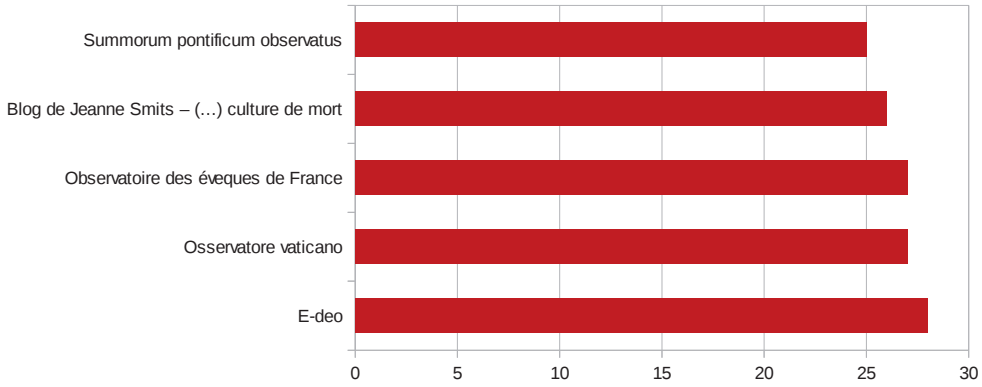

Figure 2.18 – 5 premières autorités parmi la communauté indiquée en beige dans la figure 2.16. Cette communauté est composée de sites autour du Vatican et de la vie de l'Église en général. Les 5 sites revendiquent leur lien au groupe Riposte-Catholique ; `www.riposte-catholique.fr` est d'ailleurs le deuxième annuaire renvoyé par le système.

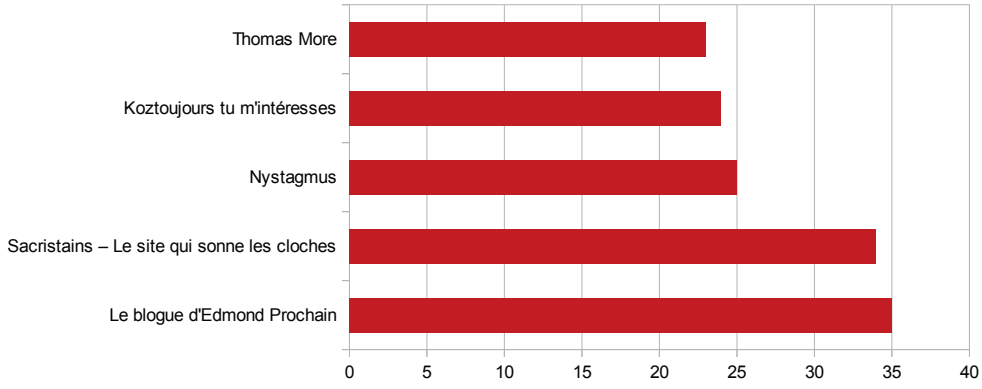

Figure 2.19 – 5 premières autorités parmi la communauté indiquée en orange dans la figure 2.16. Cette communauté aborde principalement des sujets du quotidien à travers les yeux de jeunes catholiques modérés.

Requête : "réforme retraites" Les figures 2.20 et 2.21 illustrent les résultats sur 2 communautés extraites pour la requête "réforme retraites". Il est intéressant de constater que les sites officiels de l'UDF et du MoDem ne sont pas des autorités mais des annuaires influents (respectivement le 1^{er} et le 3^{e}), comme illustré dans la figure 2.22.

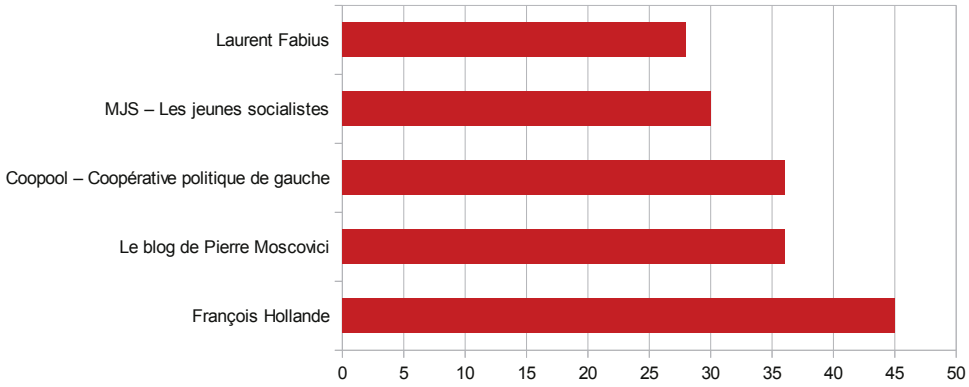

Figure 2.20 – 5 premières autorités parmi la communauté construite autour du site www.parti-socialiste.fr/.

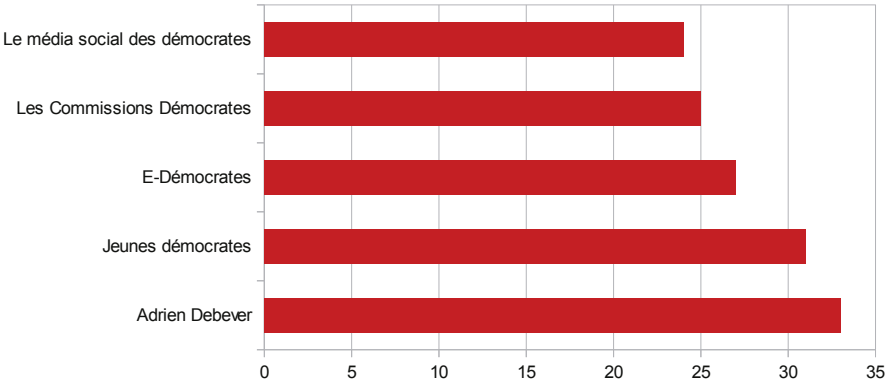

Figure 2.21 – 5 premières autorités parmi la communauté construite autour du site www.mouvementdemocrate.fr/.

2.2.3.3 Discussion

Visualiser la structure réelle du graphe permet d'interpréter le contenu de ces sites selon leur contexte d'origine. Par exemple, les sites www.avortementivg.com et www.ivg.net sont en apparence des sites objectifs concernant la question de l'avortement, mais ils sont en réalité très fortement liés aux sites de la communauté beige (Vatican et vie de l'Église), comme illustré dans la figure 2.23.

La visualisation du graphe permet aussi de relativiser les résultats obtenus. En effet, la communauté orange de la figure 2.16 est en réalité mal découpée : la partie droite est composée de pages ayant attrait au domaine juridique (par exemple des

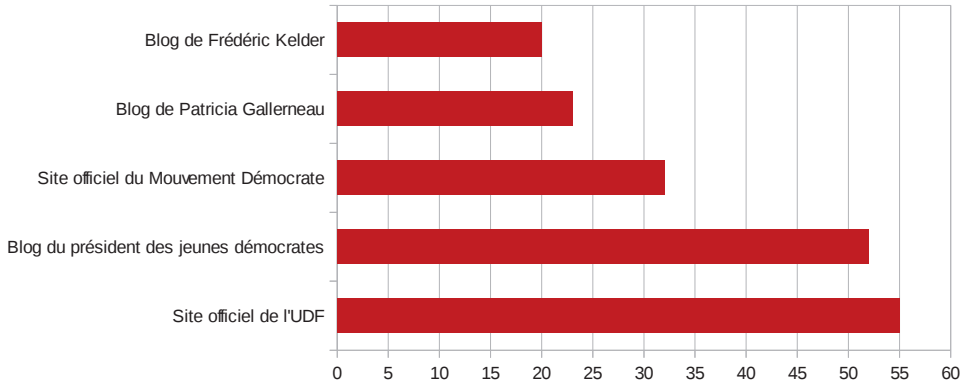

FIGURE 2.22 – 5 premiers annuaires parmi la communauté construite autour du site www.mouvementdemocrate.fr/.

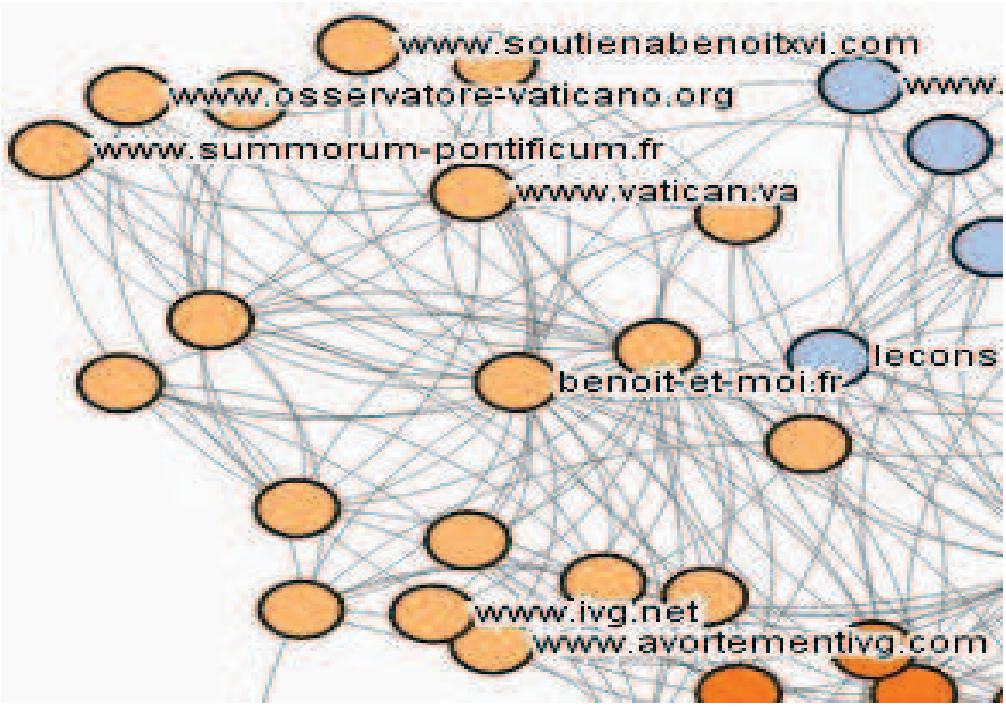

FIGURE 2.23 – Agrandissement de la communauté beige. On distingue deux sites (www.ivg.net et www.avortementivg.com) au contenu prétendument neutre sur la question de l'avortement mais en réalité fortement liés à des sites catholiques anti-avortement.

pages d'avocats). Cela s'explique par la présence de koztoujours.fr, blog tenu par un avocat chrétien. De fait, ce dernier joue le rôle de noeud central reliant les deux communautés.

L'approche que nous proposons se veut itérative. En effet, tout ou partie de ces communautés peuvent être ignorées ou au contraire détaillées selon les besoins de l'utilisateur. Dans ce cas, les résultats peuvent alors permettre de constituer un nouveau root set à partir duquel relancer l'algorithme, et ainsi permettre potentiellement de découvrir de nouvelles communautés [87] ou d'enrichir celles déjà existantes afin de dégager de nouvelles autorités, et ainsi de suite. Nous schématisons plus précisément l'approche proposée dans la figure 2.24 :

1. Interrogation d'un moteur de recherche, pour une requête donnée de façon à constituer un premier root set.
2. Expansion de cet ensemble vers le base set.
3. Extraction de communautés à partir du graphe que forme le base set. Idéalement, le veilleur doit pouvoir modifier manuellement, si besoin, les communautés calculées automatiquement.
4. Calcul d'autorités et d'annuaires pour chacune des communautés.
5. Interaction du veilleur avec les résultats (scores et graphes).
6. Constitution manuelle d'un nouveau root set à l'aide de certaines parties du graphe jugées pertinentes.

Ce processus itératif de sélection de contextes permet selon nous l'élaboration d'un catalogue de sources de qualité, recentrées non seulement sur une thématique donnée mais aussi et surtout sur un groupe d'auteurs entretenant des liens forts. Cette opération est répétable jusqu'à convergence : pour chaque communauté, il ne devrait plus être possible après un certain nombre d'itérations de détecter de nouveaux éléments. Il s'agit d'une phase d'exploration des sous-graphes du Web. Toutefois, il est important de souligner que la phase d'expansion du root set vers le base set est coûteuse en temps. Les itérations ne peuvent donc pas être réalisées en temps réel.

Pour conclure, nous présentons les pistes qui restent à explorer. Tout d'abord, l'algorithme *CPM* que nous utilisons fournit des résultats satisfaisants dans ces exemples, mais le principe de découpage par cliques (certes allégées) est assez restrictif. Nous ne sommes pour le moment pas en mesure d'évaluer le nombre de noeuds pertinents qui auraient été évincés lors de la constitution des communautés. Il serait intéressant de réfléchir à des moyens d'assouplir la notion de communauté utilisée dans l'algorithme.

Par ailleurs, nous nous étions interrogés dans la partie 2.1.3.6 sur le bien-fondé de la façon de constituer le base set. Dans les tests effectués, nous nous sommes

87. Via les noeuds périphériques. Rappelons la puissance des liens faibles signalée par [Granovetter 1973].

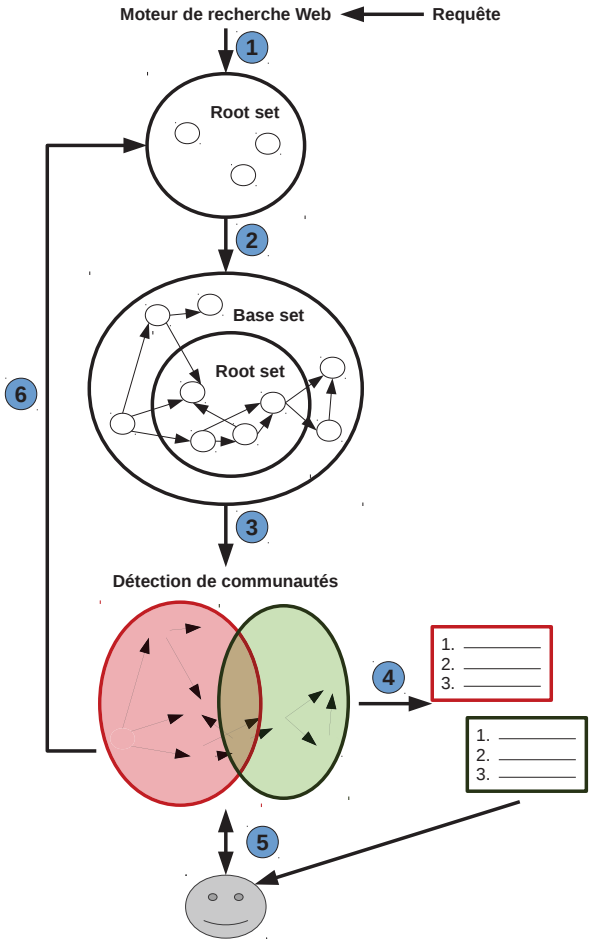

Figure 2.24 – Schéma détaillé de l'approche proposée.

servis du moteur de recherche Yahoo. Les pages du root set sont donc influencées par les nombreuses heuristiques utilisées par ce dernier (historique personnel, date des documents, etc.). Il conviendrait de trouver un moyen de former un ensemble de pages de départ de la façon la plus objective possible. Une première piste consiste à recourir aux URLs publiées sur les réseaux sociaux, par exemple les tweets. Néanmoins, rappelons les problèmes éventuels de représentativité des utilisateurs sur les réseaux sociaux. La question reste ouverte.

Chapitre 3

Analyse textuelle du buzz

Sommaire

Le chapitre précédent a décrit les différentes méthodes d'analyse de la topologie des sous-graphes du Web. Nous avons présenté des premières pistes permettant une meilleure qualification des émetteurs ainsi qu'une meilleure compréhension de la dynamique dans laquelle ils s'inscrivent.

Il est désormais temps de traiter l'analyse du contenu textuel afin d'analyser le contenu des buzz. Rappelons que les buzz ne sont pas connus a priori. Le domaine

de la veille introduit en effet une difficulté inhabituelle par rapport aux domaines d'application classiques des techniques de fouille de textes : repérer des choses potentiellement inattendues, ce qui n'est jamais apparu ou bien ce qui disparaît brutalement [Jacquenet 2004]. Des centaines, voir des milliers d'articles de presse et de blog parlent quotidiennement de François Hollande par exemple, mais cela ne signifie pas nécessairement qu'il y a un évènement marquant à signaler au veilleur. Il convient ainsi de proposer des méthodes, plus complexes que la simple analyse par mots-clés, permettant dans un premier temps d'identifier les buzz, puis de les filtrer et les regrouper selon leur contenu, qualifier leur intensité relative, et voir comment ils évoluent dans le temps.

Il s'agit de la dernière brique de notre étude : une analyse conjointe du contenu textuel et du graphe des émetteurs permet de mieux appréhender la façon dont le contenu se propage et ainsi de remonter potentiellement aux personnes à l'origine d'un buzz donné. Il convient donc de détecter des éléments assez discriminants pouvant être aisément tracés.

Nous retenons deux approches. La première consiste à ramener la tâche de détection de buzz à celle d'extraction de thématiques ; chaque buzz correspond ainsi à une thématique englobant un nombre important de documents. Ancrée dans les traditions de la lexicométrie et de la linguistique de corpus, la méthode linguistique que nous proposons est une alternative aux méthodes statistiques classiques de classification non supervisée de documents. Elle repose sur l'analyse d'un graphe de cooccurrents de deuxième ordre. Le formalisme de la théorie des graphes nous permet d'exprimer des relations sémantiques assez fines entre les mots de chaque thématique. Ces graphes peuvent par ailleurs être visualisés afin, encore une fois, d'aider le veilleur dans son interprétation.

On peut considérer qu'il s'agit là d'une analyse macroscopique du buzz. Les thématiques détectées permettent une bonne vision d'ensemble de ce dont les internautes parlent mais nous verrons qu'elles restent relativement statiques, malgré quelques pistes intéressantes, pour permettre un niveau de granularité plus fin. Par ailleurs, ce genre d'approches est assez difficilement applicable sur des textes courts comme c'est le cas des données issues des réseaux sociaux.

Nous proposons ainsi une seconde approche en complément reposant sur l'extraction de citations : il s'agit de chaînes de caractères très courantes, évoluant de jour en jour et aisément attribuables à un auteur et à une date. L'approche permet une vision microscopique, en faisant apparaître plusieurs petites histoires quotidiennes au sein de chacune des thématiques. Une analyse diachronique de ces quasi-duplications sera par ailleurs proposée afin de mieux comprendre la façon dont le sens peut se déformer au fil du temps.

Nous pensons que ces deux méthodes [1] sont complémentaires car elles travaillent

1. Précisons que nous nous sommes concentrés dans cette thèse sur des documents rédigés en français et en anglais. Néanmoins, nous pensons que nos approches sont aisément généralisables à d'autres langues, modulo quelques légères modifications ; par exemple, dans le cas des langues à système d'écriture sans séparateur comme le japonais, il conviendrait de trouver en amont une

à des niveaux de granularité différents. Nous présenterons pour chacune d'entre elles les travaux existants. Nous aborderons aussi quelques pistes afin d'appliquer les analyses sur un axe temporel.

3.1 Analyse du buzz par extraction de thématiques

La première approche que nous proposons consiste à ramener la tâche de détection de buzz à celle d'extraction de thématiques. On peut ainsi définir le buzz comme une thématique englobant un nombre conséquent de documents.

Avant de présenter les méthodes existantes, nous aborderons les différents prétraitements applicables au texte. Nous détaillerons ensuite notre approche. Après avoir présenté nos premiers résultats, nous discuterons des moyens d'adapter le traitement en vue d'une analyse temporelle des thématiques extraites.

3.1.1 Prétraitements

Dans la plupart des méthodes, le corpus est représenté selon le modèle vectoriel [Salton 1975] (on parle parfois aussi de sémantique vectorielle). il s'agit d'une matrice $m \times n$ mots-documents où chaque document est un vecteur à m dimensions correspondant à l'ensemble de ses vocables (mots graphiques uniques) [2]. Ce type de représentation est souvent appelé "sac de mots" (*bag of words*) étant donné que l'on réduit un document à un ensemble de mots, ignorant l'ordre entre ces derniers ainsi que toute relation syntaxique [3].

Dans la version la plus simple, les vecteurs peuvent n'être remplis que de 0 et de 1 selon la présence ou l'absence du mot dans le document. Néanmoins, on a le plus souvent une liste de poids correspondant assez classiquement au nombre d'occurrences du mot dans le document (TF –*Term Frequency*), ou encore la fréquence du mot dans le document multipliée par l'inverse de sa fréquence dans l'ensemble du corpus (TF.IDF –*Term Frequency-Inverse Document Frequency*). Le TF.IDF du mot i dans le document j se calcule ainsi :

$$TF \cdot IDF(m_i, d_j) = TF(m_i, d_j) \times IDF(m_i) \tag{3.1}$$

Sachant que :

$$IDF(m_i) = log\left(\frac{N}{n_i}\right) \tag{3.2}$$

Où N correspond au nombre total de documents du corpus et n_i renvoie au nombre de documents où le mot i apparaît.

solution de segmentation adaptée.

2. A noter que les mots sont parfois remplacés par des n-grammes de mots ou de caractères.

3. [Blei 2009a] défend ce point de vue en énonçant qu'un tel degré de précision n'est pas nécessaire : même si l'on mélange tous les mots d'un texte, il reste souvent facile de déterminer le contenu global du document. En d'autres termes, le processus n'a pas nécessairement de sens dans l'absolu mais a du sens pour la tâche à accomplir (propos tenus lors d'une présentation de LDA –voir partie 3.1.2.3).

La quantification de l'objet d'étude se fait au prix d'une série de pré-traitements appliqués sur les documents. Dans l'idéal, les réductions effectuées impactent peu les faits lexicaux les plus importants tout en permettant une réduction salutaire du nombre d'entrées dans la matrice.

3.1.1.1 Nettoyage

La phase de nettoyage consiste à retirer les caractères spéciaux et la ponctuation, et normaliser la casse. Les documents sont ensuite découpés en unités minimales (*tokens*). Il s'agit le plus souvent de mots graphiques mais certains travaux considèrent des n-grammes –on parle parfois de *termes*, *syntagmes* ou *chunks*. L'intérêt est une meilleure prise en compte des unités polylexicales ("pomme de terre" par exemple) au prix parfois d'associations erronées pouvant fausser les analyses futures. Il est ensuite possible d'appliquer un étiquetage morphosyntaxique[4]. Les stop words qui sont des mots graphiques dont on juge qu'ils n'apportent que peu à l'analyse, peuvent être retirés. Ils peuvent être tout simplement recensés au sein d'une liste (stop list) ou filtrés après étiquetage morphosyntaxique –les catégories morphosyntaxiques conservées varient selon les travaux mais on ne conserve en général que les mots "pleins" (substantifs, adjectifs, adverbes, verbes). Le primat est très souvent donné aux substantifs (et au syntagme nominal).

3.1.1.2 Racinisation ou lemmatisation

Les mots peuvent ensuite être racinisés ou lemmatisés. La racinisation (*stemming* ou désuffixation) consiste à retirer l'ensemble des flexions d'un mot pour ne garder que le *stemme*, ou la racine. Il est important de noter que le retrait des flexions se fait selon des critères uniquement statistiques et le stemme ne correspond pas nécessairement à la racine grammaticale du mot mais la racine la plus probable pour une langue donnée. La méthode la plus utilisée est l'algorithme de [Porter 1980] qui repose sur une liste d'affixes pour chaque langue, accompagnée de règles de racinisation. Citons aussi [Adamson 1974] reposant sur des n-grammes, [Krovetz 1993] et [Paice 1996] combinant n-grammes et analyse linguistique. Enfin, [Savoy 1993] se base sur des lexiques afin de valider ou non une tentative de transformation d'un mot en radical.

La lemmatisation, quant à elle, vise à obtenir la forme canonique du mot –son lemme. Cette dernière repose sur un lexique et utilise le contexte afin de déterminer la catégorie morphosyntaxique des mots. Tandis que la stemmatisation gommait à la fois les variations flexionnelles et dérivationnelles, la lemmatisation se limite à la morphologie flexionnelle. Prenons la forme *animaux* pour illustrer la différence entre les deux. Son stemme sera *anim*, créant ainsi une potentielle ambigüité avec le mot *animer*. Son lemme sera *animal*. Il est communément admis que la lemmatisation est plus fiable mais la racinisation est beaucoup plus simple et rapide à implémenter.

4. TreeTagger est probablement l'étiqueteur le plus utilisé : `http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/`.

Par ailleurs, cette dernière est moins sensible aux fautes d'orthographe (la forme fléchie doit exister dans le lexique pour une lemmatisation réussie). Certains travaux [Geffroy 1973] [Brunet 2001] mettent en garde contre une lemmatisation systématique, étant entendu que l'identité de sens entre les différentes formes n'est pas toujours certaine et que le contexte est parfois nécessaire afin de préciser le sens de deux formes morphologiquement proches. Néanmoins, il est important de souligner que le choix de la lemmatisation (ou racinisation), à l'instar des traitements cités précédemment, est contextuel : il convient de prendre en compte la singularité des données à traiter ainsi que la tâche à accomplir.

3.1.1.3 Élagage

On a enfin une dernière phase d'élagage qui consiste à retirer les mots avec un poids trop faible ou, au contraire, trop élevé. La loi de Zipf nous enseigne que le nombre de formes différentes est inversement proportionnel au nombre d'occurrences de la forme. Autrement dit, les mots très fréquents sont peu nombreux et il existe un nombre conséquent de mots occurrant une ou deux fois seulement. La relation est de même ordre si l'on considère le TF.IDF au lieu de la simple fréquence. Le nombre de vocables statistiquement significatifs est donc relativement faible. [Mauceri 2007] annonce qu'un texte reste compréhensible même si l'on remplace par un même signe tous les mots occurrant moins de 5 fois.

Avant de continuer, précisons que, comme le rappelle [Valette 2010], le mot est un concept linguistique fragile de par ses frontières théoriques et matérielles imprécises. Il s'agit pourtant souvent du point de départ de l'analyse linguistique. [Saussure 1916] reconnaît de nouveaux palliers de découpage, dont le morphème qui est la plus petite unité (indécomposable) de sens. Ce que l'on nomme "mot" correspond ainsi plutôt à la lexie, qui est l'élément unitaire du lexique pouvant être constitué de morphèmes (potentiellement uniques) ou de plusieurs lexies. Dans le cadre de ce mémoire, nous parlerons de "mot", qui désignera le plus souvent le mot graphique[5], c'est-à-dire une chaîne de caractères située entre deux séparateurs[6]. Une dénomination plus précise sera ponctuellement utilisée au besoin.

3.1.2 Méthodes d'extraction de thématiques

Dans la littérature, on nomme *thématique* ou *topic* des groupes de mots "sémantiquement proches" qui, ensemble, permettent de paraphraser grossièrement le contenu thématique des documents traités. Concrètement, il s'agit de regrouper en classes un ensemble d'éléments, en l'occurrence les mots des documents. On parle alors de classification non supervisée de documents (*text clustering* ou *document clustering*). La classification est dite non supervisée car les thématiques à extraire ne sont pas connues a priori. C'est un point particulièrement intéressant dans notre

5. Il s'agit du signifiant privilégié en TAL.
6. Les séparateurs utilisés dépendent des segmenteurs utilisés.

cadre d'identification de buzz. La différence entre une classification non supervisée (*clustering*) et une classification supervisée (*classification*) est schématisée dans la figure 3.1. Dans la suite de ce mémoire, nous utiliserons indifféremment les mots *clustering* et *classification* pour désigner la classification non supervisée.

A noter que les classes peuvent être disjointes ou chevauchantes (*fuzzy partitions* ou *soft partitions*), c'est-à-dire qu'un même élément peut appartenir à plusieurs classes différentes. Le chevauchement s'avère important dans notre étude car nous verrons que cela permet de prendre en considération les mots polysémiques.

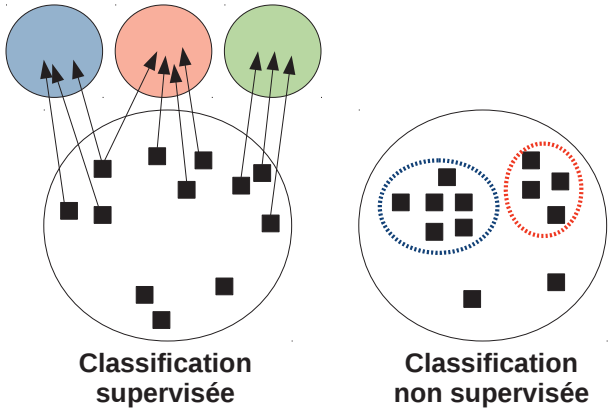

Figure 3.1 – Différence entre classification supervisée (à gauche) et non supervisée (à droite). Dans le premier cas, on dispose d'éléments appartenant à des classes connues a priori ; l'opération consiste donc à classer tout nouvel élément dans une (ou plusieurs) de ces classes. Dans le second cas, il s'agit de regrouper des éléments proches les uns des autres, et comparativement très différents des autres éléments. Nous nous situons dans le cas de droite : il s'agit d'extraire des documents un ensemble de thématiques non connues a priori.

Nous distinguons 4 grands types de méthodes d'extraction de thématiques : le rapprochement par distance, la factorisation de matrices, les modèles de thématiques, et enfin le partitionnement de graphe. Nous décrivons ci-dessous chacune d'entre elles.

3.1.2.1 Méthodes à base de calcul de distances

Nous rangeons dans cette catégorie l'ensemble des méthodes reposant sur un calcul de distances entre éléments, principalement les k-means [MacQueen 1967] et les méthodes hiérarchiques [Girvan 2002]. Ces travaux ont déjà été décrits dans la partie 2.2.2.1 mais quelques adaptations sont nécessaires pour appliquer ces méthodes à des documents : on travaille sur la matrice $m \times n$ mots-documents et une distance entre chaque paire de documents est calculée. Ces méthodes servant à l'origine à calculer une distance entre des documents, l'extraction de thématiques n'est de fait pas leur vocation première.

Citons notamment [Pincemin 1999a] qui s'inspire dans ses travaux des nuées dynamiques [Diday 1971], une généralisation de l'algorithme des k-means. Plus récemment, [Cleuziou 2007] propose l'*Overlapping K-Means* (OKM) qui permet d'attribuer un document à plusieurs classes. La méthode sera reprise dans [Rizoiu 2010].

3.1.2.2 Méthodes à base de factorisation de matrices

Ce type d'approches est dérivé de la sémantique vectorielle étant donné qu'elles représentent aussi le corpus sous la forme d'une matrice mots-documents, pondérée le plus souvent par un TF.IDF.

En algèbre linéaire, la factorisation (décomposition) de matrices est une opération visant à réécrire [7] une matrice sous la forme du produit de deux matrices qui va permettre le rapprochement entre des documents, des mots et des thématiques [8]. Concrètement, l'opération permet l'obtention, à partir de la matrice mots-documents, de deux matrices W et H :

$$A \approx W \times H \tag{3.3}$$

La matrice mots-documents A est décomposée en deux matrices. H constitue un nouvel espace et traduit la probabilité d'appartenance des mots aux thématiques : on parle alors d'*espace sémantique latent*. W est la projection des documents dans ce nouvel espace sémantique où chaque élément traduit le degré d'appartenance d'un document à une thématique.

L'opération de factorisation de matrices peut se faire de plusieurs façons. La LSA [9] (*Latent Semantic Analysis* –analyse sémantique latente) [Deerwester 1990], à l'origine de ce genre d'approches [10] procède par décomposition en valeurs singulières (SVD –*Singular Value Decomposition*). Plus tard, [Lee 1999] et [Lee 2001] recourent à la factorisation non-négative de matrice (NMF –*Non-negative Matrix Factorization*) pour contourner le problème des valeurs négatives dans la matrice qui posaient des problèmes dans l'interprétabilité des résultats de la LSA.

Ces calculs complexes traduisent en réalité tout simplement des relations de cooccurrence à l'échelle du document. Ce genre d'approches postule donc que les mots sémantiquement proches apparaîssent souvent ensemble dans les textes. Des documents partageant un nombre réduit de mots pourront ainsi être regroupés s'ils ont en commun de nombreux cooccurrents. Nous reviendrons plus en détail sur ce point dans la partie traitant de la cooccurrence (3.1.3.1).

Parmi les inconvénients de ces approches, notons que le besoin de spécifier le nombre de dimensions sémantiques à extraire est réellement problématique. Par ailleurs, la polysémie est assez mal gérée étant donné que chaque mot est représenté

7. Il s'agit en réalité plutôt d'une approximation de cette matrice.
8. [Deerwester 1990] parle originellement de *concept* pour désigner ce que l'on nomme désormais *thématique* ou *topic*. On parle aussi parfois de *dimension sémantique*.
9. On parle aussi parfois de LSI (*Latent Semantic Indexing* –indexation sémantique latente). Le choix de la dénomination dépend de l'utilisation que l'on fait de la méthode.
10. La LSA a été brevetée en 1988.

comme un point unique dans l'espace ; la valeur du mot dans la matrice correspond à la moyenne de tous ses usages. Enfin, l'opération est très coûteuse [11].

La LSA a notamment été utilisée dans la thèse de [Mauceri 2007] qui, à l'instar de [Pincemin 1999a] et de [Rossignol 2005] cherche à informatiser au maximum les enseignements théoriques de la sémantique interprétative [Rastier 1996].

3.1.2.3 Modèles de thématiques

Les modèles de thématiques (*topic models*) sont des méthodes probabilistes ayant pour objectif la découverte de thématiques au sein d'un corpus de documents [12]. Ils sont rapidement devenus très populaires.

Ces modèles sont génératifs, c'est-à-dire qu'ils permettent de générer des documents qui respectent la distribution lexicale réelle d'un corpus. En utilisant le processus inverse de génération, il est possible de déterminer la probabilité (vraisemblance) qu'aurait le système de générer un document similaire à ceux que l'on a dans le corpus.

Le premier modèle de thématiques, la PLSA (*Probabilistic Latent Semantic Analysis* –analyse sémantique latente probabiliste) [13], a été proposé par [Hofmann 2002]. Il s'agit d'une évolution de la LSA. À la différence de cette dernière, un document peut appartenir à plusieurs thématiques, dans des proportions quantifiables. Les thématiques sont définies par une distribution sur les mots du vocabulaire : les thématiques sont des ensembles de mots pondérés par des probabilités d'apparition. La polysémie est ainsi mieux prise en compte : il est donc possible pour un mot donné d'avoir de fortes probabilités dans plusieurs thématiques différentes. LDA (*Latent Dirichlet Allocation* –allocation de dirichlet latente) proposée dans [Blei 2003] est une généralisation de la PLSA. Il s'agit du modèle de thématiques le plus utilisé à ce jour. Ce dernier est décrit plus en détail dans [Blei 2009b] et [Blei 2011].

Comme c'est le cas de la plupart des méthodes d'extraction de thématiques, il est nécessaire de spécifier en entrée le nombre de *topics* à détecter. On ignore souvent ce problème en renseignant volontairement un nombre élevé, ce qui est rendu possible par l'existence des *topics* "poubelle" qui regroupent l'ensemble des mots "en trop". En effet, le clustering est exhaustif et tous les mots doivent nécessairement être assignés à au moins une thématique.

---

11. La complexité algorithmique sur la SVD est $O(n^2k^3)$, où n correspond au nombre de vocables et k au nombre de dimensions dans l'espace sémantique.

12. Bien que les modèles de thématiques aient été décrits et implémentés originellement dans un cadre d'analyse textuelle, il est précisé dans [Blei 2009a] qu'ils peuvent s'appliquer à n'importe quel type de données : images, réseaux sociaux, musique, historique d'achats, lignes de code, données génétiques, etc. Plus généralement, ces modèles permettent d'organiser, comprendre, chercher et résumer automatiquement de larges archives numériques.

13. On parle aussi parfois de PLSI (*Probabilistic Latent Semantic Indexing* –indexation sémantique latente probabiliste). Ici aussi, la dénomination dépend de ce pourquoi on utilise le modèle.

3.1.2.4 Méthodes à base de partitionnement de graphe

Nous avons vu que les méthodes décrites précédemment se basent sur le modèle vectoriel [Salton 1975] ; le corpus est représenté sous forme d'une matrice dont les lignes et les colonnes sont respectivement les documents et les mots. Nous allons désormais présenter quelques travaux récents qui représentent les documents sous forme de graphes lexicaux. Il devient dès lors possible de représenter les relations entre les thématiques extraites, ce qui peut s'avérer utile pour une meilleure interprétation des résultats.

Les graphes lexicaux Les graphes lexicaux sont des graphes dont les noeuds sont les mots du document et les arêtes représentent des relations entre eux. Le graphe est orienté ou non selon le type de relation que traduisent les arêtes. Ces dernières sont le plus souvent pondérées de façon à représenter la proximité entre les mots. Un exemple de graphe lexical est illustré dans la figure 3.2.

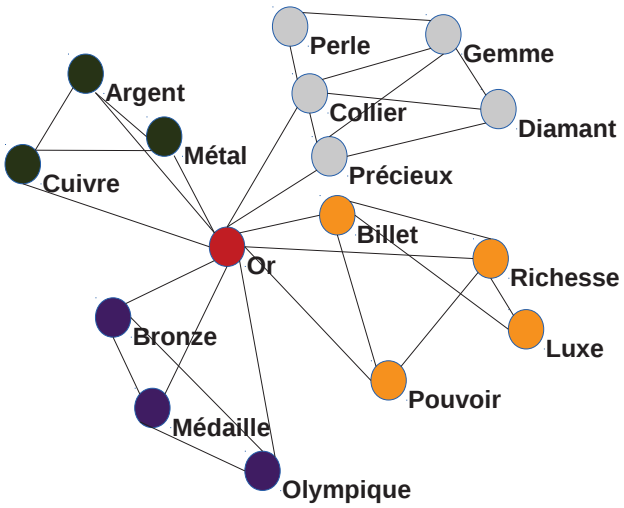

Figure 3.2 – Exemple de graphe lexical, inspiré d'une illustration tirée de [Palla 2005].

Il convient de réfléchir à la sémantique des arêtes. Il est tout d'abord envisageable de recourir à une approche exogène, c'est-à-dire en faisant appel à des ressources externes aux documents. Il peut s'agir de dictionnaires [Kozima 1993], de thésaurus [Morris 1991], ou encore de WordNet [14]. WordNet est un lexique sémantique généraliste développé à l'origine par le laboratoire des sciences cognitives de l'université de Princeton [15]. Il s'agit d'une structure arborescente hiérarchisée (ontologie) qui organise le lexique selon des relations d'hyperonymie/hyponymie. Le parcours de cet arbre permet de rendre compte de relations de (quasi-)synonymie, antonymie ou méronymie selon la position des noeuds par rapport aux autres noeuds. Comme

14. http://wordnet.princeton.edu/.
15. Il n'est donc pas étonnant de constater que la version anglaise est la plus fournie.

le rappelle [Mauceri 2007], cette vision ontologique du lexique repose sur le présupposé aristotélicien énonçant que le mot se définit au travers du triangle sémiotique, dont les trois sommets sont le mot, la chose et le concept [16]. On a alors des liens conceptuels figés reposant sur une représentation close du monde, selon des relations hiérarchiques définies en amont des mécanismes réels de la langue [Slodzian 2000] [Valette 2009] [Valette 2010] [17].

Il est aussi possible d'exploiter un autre type de relations externes : les relations de cooccurrences lexicales calculées à partir de gros corpus de référence [Ferret 1998]. En d'autres termes, il s'agit de détecter les mots ayant tendance à apparaître ensemble au sein d'une fenêtre contextuelle donnée [18]. On s'affranchit ainsi de la référence. Les liens lexicaux extraits vont au-delà des relations traditionnelles que sont la synonymie, l'antonymie ou la méronymie. Pour l'interprétation d'un texte, ces dernières sont par ailleurs moins pertinentes que des relations du type *oiseau/voler* ou *enfant/énergie* qui sont difficilement définissables selon les typologies classiques [Morris 2004] [Adam 2009]. Dans le même ordre d'idées, il semble naturel de rapprocher *caviar* et *champagne* qui sont tous deux des mets festifs [Valette 2006] ; ces derniers sont pourtant très éloignés dans Wordnet, l'un étant solide, l'autre liquide. *Caviar* est d'ailleurs en relation d'hyponymie avec les aliments issus de la mer, au même titre que le poisson pané qui est aux antipodes des mets festifs.

La taille que doit alors avoir le corpus est une question intéressante. [Lin 2001] énonce qu'un corpus de 50 à 100 millions de mots commence à pouvoir être considéré comme statistiquement représentatif de la langue étudiée. Cette taille pouvant diminuer jusqu'à 1 million, voir 100 000 mots dans le cas de langues de spécialité [19]. Le problème est qu'il s'agit alors de relations générales, pas nécessairement pertinentes pour un autre corpus donné. La constitution de ce genre de ressources est par ailleurs chronophage.

Une approche purement endogène est plus intéressante dans notre cas ; il s'agit d'exploiter uniquement les connaissances issues du corpus. On construit alors dynamiquement le graphe selon les relations de cooccurrences observées dans les documents à traiter, permettant de définir le sens de chaque mot par opposition aux autres mots du lexique. D'un point de vue purement pratique, il est préférable de ne pas avoir à constituer et entretenir de lourdes ressources sémantiques (approche *knowledge poor*).

Quelques travaux Les travaux que nous présentons ici codent tous les documents sous la forme d'un graphe lexical auquel ils appliquent des algorithmes de segmentation (*clustering*) de graphe. Ces méthodes n'ont pas toutes nécessairement

16. Signalons par ailleurs que les concepts sont jugés généraux et ne varient donc pas selon la langue.

17. Les lexiques sémantiques généralistes comme WordNet reposent sur l'hypothèse selon laquelle il existerait une "langue générale".

18. Il peut s'agir de la phrase, du paragraphe, ou encore du document. Nous reviendrons en détail sur ce point dans la partie 3.1.3.1.

19. Langue utilisée entre professionnels d'un même domaine de connaissances. On parle aussi de langue spécialisée.

pour objectif premier l'extraction de thématiques mais il s'agit de bonnes introductions aux notions que nous utiliserons plus tard. Par ailleurs, elles peuvent être aisément adaptées pour y parvenir.

[Widdows 2002] propose une méthode non supervisée à base de graphes lexicaux, dans le but de créer des ressources lexicales. Ces ressources peuvent alors être utilisées pour des tâches de désambigüisation sémantique[20] par exemple. A partir d'un corpus étiqueté avec un étiqueteur morphosyntaxique, les auteurs extraient l'ensemble des substantifs apparaîssant dans le contexte suivant : *N et/ou N*[21]. Les noeuds du graphe sont constitués des noms ainsi extraits et il existe une arête entre deux noeuds s'ils cooccurrent dans ce contexte bien particulier. Le poids des arêtes correspond au nombre de fois que la relation a lieu. Les clusters sont formés de façon incrémentale ; les mots sont ajoutés aux clusters déjà existants uniquement s'ils sont reliés à un certain nombre des voisins, de façon à éviter les "infections". Les mots se trouvant à la frontière entre plusieurs clusters sont polysémiques. [Dorow 2005] étend le principe et se concentre tout particulièrement sur la question de la désambigüisation des mots polysémiques.

Les travaux de [Bordag 2003a] et [Bordag 2006] se situent dans une problématique assez similaire. Un graphe de cooccurrence est construit en retenant les (200) cooccurrents les plus pertinents pour chaque mot. Les mots sont rapprochés s'il existe une intersection importante entre les cooccurrents directs de chacun d'entre eux. Les clusters correspondent aux différents usages des mots. A noter qu'un mot ne peut appartenir qu'à un seul cluster.

L'auteur fait remarquer que les graphes de cooccurrence ainsi extraits sont des graphes petits mondes [Watts 1998]. Rappelons que ces derniers ont des noeuds "raccourcis" (*long-range nodes*) qui ont de fortes connexions avec de nombreux clusters. Dans le cadre d'un graphe lexical, il s'agit par exemple de verbes très courants. De fait, les clusters sont très proches les uns des autres : il est possible d'atteindre n'importe quel cluster en traversant un nombre restreint de noeuds. [Ferrer i Cancho 2001] est le premier à appliquer la notion de graphe petit monde sur des graphes lexicaux. Le fait que les graphes lexicaux soient des graphes petit monde présente un intérêt certain en psycholinguistique concernant les mécanismes cognitifs d'acquisition et d'accès au lexique[22] ; ces régularités structurelles pourraient être la preuve d'une organisation linguistique et cognitive sous-jacente. Ce phénomène pourrait expliquer la raison pour laquelle il est possible de tout exprimer avec quelques milliers de mots seulement. C'est notamment le cas des pidgins par exemple, dont le lexique dépasse pourtant rarement le millier de mots

20. Il s'agit d'extraire les différents sens d'un mot donné –ou plutôt ses usages. On parle alors de *Word Sense Discrimination* (WSD), parfois de *Word Sense Induction* (WSI).

21. D'autres relations grammaticales sont envisagées par les auteurs mais ces derniers ont préféré les ignorer dans un premier temps étant donné que les relations asymétriques ajoutent un degré de complexité supplémentaire. Ces relations étaient les suivantes : *N+V* (relation sujet), *V+N* (relation objet), *Adj+N*, *N+N* (en anglais, le premier substantif modifie le suivant).

22. [Palermo 1964] et [Meyer 1975] constatent notamment que les sujets associent plus rapidement certains mots entre eux si ces derniers sont précédés de quelques mots proches.

[Romaine 1992]. Cela se fait néanmoins au prix d'une redondance accrue. De façon purement pragmatique, les caractéristiques du petit monde peuvent être utilisées pour améliorer les performances des algorithmes de segmentation [Gaume 2006] [Chee 2007].

[Véronis 2003] se donne aussi pour objectif l'énumération des différents usages des mots, cette fois-ci dans un contexte de recherche d'information sur Internet. Les noeuds du graphe sont les substantifs et les adjectifs lemmatisés. A l'aide d'une stop list, les mots-outils, les mots trop génériques ainsi que les mots courants du Web sont filtrés. Seuls les mots occurrant au moins 10 fois sont retenus afin de réduire la combinatoire. Les arêtes du graphe traduisent une relation de cooccurrence à l'échelle du paragraphe. Le découpage du graphe se fait en détectant des composantes denses formées autour de noeuds racines. Pour ce faire, les noeuds sont classés par ordre décroissant de fréquence. Pour chacun d'entre eux, on vérifie s'il a au moins 6 voisins propres (il ne doit pas s'agir d'un voisin d'un autre noeud racine détecté précédemment) et que le poids moyen entre ce dernier et les 6 premiers voisins est inférieur ou égal à un seuil donné. Si c'est le cas, il s'agit d'un noeud racine. Il est alors éliminé de la liste (avec l'ensemble de ses voisins) et on passe au noeud suivant. Dans cette méthode aussi, un mot ne peut appartenir qu'à un seul cluster.

Les travaux de [Rossignol 2005] sont proches de [Pichon 1999] et [Pichon 2000] ; il s'agit d'une analyse cooccurrentielle (à l'échelle du paragraphe) en vue de l'extraction de thématiques, à la différence que l'auteur a recours à des graphes. L'extraction des thématiques consiste alors à rechercher les parties fortement connexes du graphe de cooccurrence, et à les retirer successivement du graphe. Ici non plus, le multi-classage n'est pas permis : un mot ne peut appartenir qu'à une seule thématique.

Plus récemment, [Tauveron 2012] analyse la relation de cooccurrence généralisée au sein de graphes afin de tenter d'expliquer et déceler les différences pouvant exister entre différentes traductions.

Enfin, citons [Ferret 2004] et [Ferret 2006]. Cette approche ayant très fortement influencé nos travaux, nous la décrirons dans la partie 3.1.3.2.

3.1.3 Approche proposée : analyse du graphe des plus proches voisins cooccurrentiels

La méthode que nous proposons repose sur l'analyse d'un graphe de cooccurrences car nous pensons que la visualisation des thématiques sous forme de graphe est une aide précieuse à l'interprétation des résultats. Nous reviendrons plus en détail sur les raisons de ce choix dans la partie 3.1.3.2.

Avant de présenter notre approche, nous allons définir quelques notions impor-

tantes pour la suite : notion de cooccurrence, taille de l'empan contextuel, calcul de la force d'attraction, etc.

3.1.3.1 La cooccurrence : définitions

Depuis [Firth 1957] [23] puis [Harris 1957] et [Halliday 1976], le caractère statistique du vocabulaire commence à être reconnu ; de nombreux travaux, en particulier issus d'équipes françaises, ont abordé la problématique de la détection et de la pondération de ces relations de cooccurrences, que cela soit autour d'un mot-pôle donné ou à l'échelle de l'ensemble des mots présents dans le corpus [Viprey 1997] [Martinez 2003] [Mayaffre 2008a] [Luong 2010].

On peut définir la cooccurrence comme la coprésence de deux unités linguistiques (types/formes) au sein d'un même co(n)texte déterminé. Cette coprésence est statistiquement pertinente, c'est-à-dire qu'elle ne relève pas du hasard ; on parle de stabilisation du sens lorsqu'un usage devient dominant.

La cooccurrence est une unité calculable. [Mayaffre 2008b] va plus loin dans la définition de la cooccurrence en considérant qu'il s'agit ainsi de la forme minimale du contexte, et donc du sens, car le sens naît toujours du contexte (chaque mot se contextualise à l'aide de l'autre).

Fenêtre/empan contextuel(le) Nous avons vu que la cooccurrence peut se définir comme la coprésence significative de deux unités linguistiques au sein d'un même contexte. On parle aussi parfois de fenêtre contextuelle ou d'empan. Plusieurs types de contexte peuvent être considérés.

Comme le signale [Manning 1999], [Choueka 1985] a montré que les humains discriminent beaucoup mieux les mots ambigus si quelques mots (même peu) du contexte adjacent sont montrés avec le mot à désambigüiser : un contexte plus grand améliore peu les performances humaines de désambigüisation. Cependant, cela ne veut pas pour autant signifier qu'un contexte plus large est inutile pour une machine ; [Gale 1992] a montré que de nombreuses informations précieuses dans une fenêtre de 50 mots autour du mot ambigu peuvent aider leur algorithme à désambigüiser. Par ailleurs, certaines informations se trouvant à des milliers de mots peuvent encore aider.

Distinguons deux principaux types de relation, selon la taille de la fenêtre contextuelle. Les mots qui ont tendance à apparaître ensemble dans une fenêtre contextuelle étroite mettent en avant des relations de type syntagmatique [24]. Ces dernières

23. *You shall know a word by the company it keeps.*

24. Par abus de langage, on parle souvent de collocation dans ce cas. Dans la littérature anglophone, cooccurrence et collocation sont d'ailleurs quasiment synonymes. La collocation est une forme de cooccurrence dont les constituants sont régis par des restrictions strictes, impliquant une certaine idée de figement (relation syntaxique, distributionnelle). Citons par exemple les lexies composées (*chemin de fer*, *malade mental*) ou encore les syntagmes semi-figés (*salaire de misère*, *gravement malade*). La notion est donc, en français, plus proche de la traductologie et de la lexicographie (on peut par exemple définir un terme comme étant une collocation extraite dans un domaine technique).

font référence à l'axe syntagmatique [Saussure 1916], qui concerne l'enchaînement linéaire des signifiants et rend ainsi compte de relations positionnelles et/ou fonctionnelles (syntaxiques). Le syntagme et la phrase sont des exemples de contextes permettant d'extraire ce type de relations. La phrase est une unité contextuelle facile à identifier grâce à la ponctuation [25] mais le syntagme implique une analyse syntaxique de qualité. Il est possible de recourir à une fenêtre de taille variable (n mots) mais il y a alors le risque de briser artificiellement les liens entre éléments. À noter que ce problème peut être compensé par l'utilisation de fenêtres chevauchantes.

Les relations à plus longue distance sont plutôt paradigmatiques, mettant ainsi en avant une certaine commutabilité ; une unité ne prend sa valeur que relativement aux autres. Citons la fenêtre du paragraphe, pour laquelle il semble y avoir un consensus assez marqué. Notons aussi la section logique ou la partition, le document, ou encore le corpus dans son ensemble.

[Grefenstette 1994], entre autres, distingue ainsi deux types de cooccurrents. Les cooccurrents de premier ordre (*first-order affinities*) renvoient aux cooccurrents se trouvant dans le contexte immédiat ; il s'agit de ce fait de relations plutôt syntagmatiques. Les cooccurrents de deuxième ordre (*second-order affinities*) renvoient à des mots reliés, bien que n'apparaîssant pas nécessairement ensemble. Ces derniers partagent en effet un environnement cooccurrentiel similaire et concernent plutôt l'axe paradigmatique.

Pour simplifier, les petites fenêtres ont donc tendance, en pratique, à identifier des expressions figées ou autres relations fonctionnelles. Les fenêtres plus larges vont mettre en évidence des "classes sémantiques". Il convient de garder à l'esprit que le choix dépend surtout du type de textes à traiter, mais aussi de l'application visée.

Force d'attraction des éléments Comme signalé plus haut, la coprésence des unités linguistiques doit être statistiquement pertinente [Brunet 2003] [Labbe 2003].

Dans la littérature anglo-saxonne, l'information mutuelle [Church 1990] est très utilisée. Cette formule, consiste à évaluer la probabilité jointe d'observer x et y ensemble ($P_{(x,y)}$) par rapport à la probabilité d'observer x et y séparément ($P_{(x)}P_{(y)}$) :

$$I_{(x,y)} \equiv log_2 \frac{P_{(x,y)}}{P_{(x)}P_{(y)}} \tag{3.4}$$

Le score sera positif s'il y a association statistiquement pertinente entre x et y. Sinon, $I_{(x,y)}$ est proche de 0. Dans le cas plus rare où les mots auraient une distribution complémentaire (c'est-à-dire qu'ils ne peuvent jamais apparaître ensemble dans le même contexte), le score sera négatif.

En textométrie française, c'est surtout la formule de [Lafon 1984] qui est reprise. Il s'agit alors de calculer, selon le modèle hypergéométrique, la probabilité de voir apparaître k fois un mot :

25. Il s'agit ici d'une définition pragmatique. Les frontières sémantiques de la phrase sont en réalité assez floues et arbitraires.

$$Prob_{(x=k)} = \frac{f!(T-f)!t!(T-t)!}{k!(f-k)!(t-k)!(T-f-t+k)!T!} \tag{3.5}$$

Sachant que :
— T = nombre de mots du corpus ;
— t = nombre de mots du texte ;
— f = fréquence du mot dans le corpus ;
— k = fréquence du mot dans le texte.

Toutes ces formules sont très dépendantes du volume textuel. Elles seront mises en échec sur de petits corpus non représentatifs.

3.1.3.2 Description de notre approche

Positionnement Rappelons que l'objectif est d'assister le veilleur dans deux tâches :

1. dégager des thématiques du corpus (aide à la détection, la lecture et l'interprétation des buzz) ;
2. ranger chaque texte dans une ou plusieurs de ces thématiques afin de faciliter le tri et le retour au texte.

Nous avons vu dans la partie 3.1.2 les approches statistiques les plus fréquemment utilisées pour cette tâche d'extraction de thématiques : la LSA [Deerwester 1990], la NMF [Lee 1999] et surtout les *topic models* [Hofmann 2002] [Blei 2003].

Le cadre concret de veille et d'intelligence économique à partir de données issues du Web dans lequel nous nous plaçons implique des contraintes bien particulières :

1. la nécessité de considérer des corpus de taille moyenne (entre 100 000 et 500 000 mots) ; ces derniers sont réputés trop petits pour les méthodes statistiques, mais trop grands pour être analysés manuellement ;
2. les thématiques que l'on cherche à extraire peuvent être très proches les unes des autres et risquent de partager un nombre conséquent de mots : il ne s'agit en effet pas ici d'opposer des documents médicaux et juridiques par exemple, mais de dégager des thématiques (buzz) toutes relatives à un même sujet général.

Nous souhaitons par ailleurs formaliser la notion de thématique avec un point de vue linguistique, plus fortement ancré dans les traditions de la lexicométrie et de la linguistique de corpus. L'algorithme proposé doit ainsi respecter a minima certaines propriétés qui, selon nous, font sens d'un point de vue linguistique [Pincemin 1999b].

— Le sens n'est pas dans les mots, mais entre les mots. Nous faisons donc l'hypothèse que des thématiques peuvent être modélisées par des regroupements de mots apparaîssant dans des contextes similaires. La classification repose ainsi principalement sur la répartition des mots dans le corpus, et moins sur leur fréquence.

— L'algorithme doit être indépendant de la configuration des regroupements de mots. Ce point est important car nous avons constaté que les clusters peuvent grandement varier en taille et/ou en densité selon leur représentativité dans le corpus.
— Un mot peut n'appartenir à aucune thématique. L'attribution systématique d'un cluster à chaque mot risque de faire perdre en cohérence certains regroupements : on ne s'intéresse qu'aux liens les plus forts.
— Un mot peut appartenir à plusieurs thématiques à la fois, ce qui peut par exemple traduire une relation de polysémie, d'homographie, ou toute nuance de sens plus fine.

De plus, à notre connaissance, la notion de thématique n'a jamais été réellement formalisée. Les *topic models*, par exemple, considèrent qu'une thématique est un ensemble de mots pondérés par des probabilités d'apparition. En prenant le parti de recourir à une approche de clustering sur des graphes de cooccurrences, nous espérons formaliser la notion de cohérence sémantique de façon plus poussée. En effet, le formalisme de la théorie des graphes nous permet d'exprimer concrètement des relations pondérées entre les mots de chaque thématique ; en d'autres termes, un mot m_1 peut entretenir des liens forts avec les mots m_2 et m_3 mais quasiment nuls avec le reste de la thématique. Nous jugeons qu'il s'agit là d'une aide précieuse à l'interprétation des résultats.

Enfin, précisons que nous préférons éviter tout recours à des ressources sémantiques extérieures afin de qualifier les relations entre les mots (dictionnaires, thésaurus, ontologies –Wordnet par exemple) car elles sont difficilement applicables sur le Web. Nous nous basons sur les cooccurrents des mots du corpus car nous estimons qu'ils représentent la forme minimale du contexte [Mayaffre 2008b] , et donc du sens [Rastier 1987]. Les regroupements se font donc de façon dynamique et dépendent uniquement du corpus analysé, non de relations universelles définies en amont ; notre approche reste indépendante du domaine, ce qui est crucial lorsque l'on travaille sur le Web.

Description détaillée Nous avons vu dans la partie 3.1.2.4 quelques travaux ayant recours à des méthodes de découpage de graphe à des fins d'analyse textuelle. Néanmoins, aucun d'entre eux ne cumule une analyse non exhaustive et la possibilité d'attribuer plusieurs classes à un même élément. Rappelons que la Clique Percolation Method [Palla 2005], décrite dans la partie 2.2.2.1, répond à ces deux exigences, mais l'utilisation de la notion de cliques (certes légèrement allégée) rend la méthode trop rigide : certains regroupements denses (mais pas assez pour former des cliques) peuvent ainsi être ignorés.

L'algorithme SNN (*Shared Nearest Neighbours* –plus proches voisins partagés) de [Jarvis 1973] a retenu notre attention. Les auteurs évaluent la similarité entre deux noeuds à l'aide du nombre de plus proches voisins que ces derniers partagent ; dès lors, on ne calcule pas uniquement la similarité entre tous les points pris deux à deux (ce type de relations binaires retranscrit mal les interactions linguistiques

complexes), mais on met l'accent sur des regroupements de points considérés simultanément. Par ailleurs, le découpage du graphe reste ainsi théoriquement indépendant de l'échelle et de la configuration du graphe : on considère le plus proche voisinage, indépendamment de sa densité et de sa taille. Enfin, SNN procède par nettoyages successifs d'un graphe de similarité (filtrage des relations les moins importantes), afin de ne conserver que les regroupements les plus pertinents. En ce sens, le clustering n'est pas exhaustif.

Les trois premiers critères parmi les quatre cités plus haut sont jusqu'ici respectés, mais le chevauchement de clusters n'est pas permis. SNN sera repris par [Ertoz 2003] et [Ferret 2006], qui vont l'appliquer sur des données textuelles, mais ces travaux partent toujours du postulat qu'un noeud ne peut appartenir qu'à un seul cluster.

Notre méthode reprend l'algorithme de [Jarvis 1973] auquel nous apportons quelques évolutions permettant la prise en compte de la possible appartenance d'un mot à plusieurs thématiques.

Nous créons une matrice C $(v \times v)$ de cooccurrences, tel que $C_{ij} = Freq_{i,j}$, sachant que v correspond au nombre de vocables du corpus et que $Freq_{i,j}$ est le nombre de fois que les mots aux indices i et j apparaîssent ensemble dans un même contexte. Nous choisissons le paragraphe comme fenêtre d'analyse afin de mettre l'accent sur des relations à longue distance et ainsi éviter les éventuelles "interférences" syntaxiques [26]. Les relations de cooccurrence ne sont pas orientées, car cela n'a réellement de sens que dans les contextes plus petits (la phrase par exemple). Les mots apparaîssant moins de 10 fois dans l'ensemble du corpus sont ignorés afin de réduire la combinatoire. Les valeurs de la matrice sont ensuite remplacées par la mesure de dissimilarité présentée dans [Véronis 2003] :

$$W_{A,B} = 1 - max[p(A|B), p(B|A)] \tag{3.6}$$

$p(A|B)$ correspond à la probabilité conditionnelle d'observer A dans le même contexte que B, et inversement pour $p(B|A)$. Concrètement, $p(A|B) = Freq_{A,B}/Freq_B$ et $p(B|A) = Freq_{A,B}/Freq_A$. Le score est compris entre 0 (association systématique) et 1 (jamais d'association entre les mots).

Cette matrice forme un graphe de cooccurrences où les noeuds sont les mots, et les arêtes représentent les relations (non nulles) de cooccurrences entre ces mots, pondérées selon la formule vue précédemment. On y applique l'algorithme SNN de [Jarvis 1973], qui ramène le problème du découpage de graphe à celui de la détection de composantes comparativement denses dans un graphe. Pour ce faire, SNN procède par nettoyages successifs des arêtes les moins fortes, jusqu'à briser le graphe en plusieurs composantes connexes [27] ; chacune de ces composantes connexes

26. Nous préférons le paragraphe à une fenêtre glissante de n mots car il s'agit d'un découpage humain qui, a priori, fait sens. L'échelle du document nous semble trop large. Néanmoins, la question de la délimitation du contexte reste encore ouverte.

27. Rappelons que, dans un graphe non orienté G, une composante connexe est un sous-graphe dans lequel il existe un chemin entre toute paire de noeuds.

correspond à un cluster. Plus précisément, l'algorithme d'origine se divise en trois grandes étapes :

1. filtrage de tous les liens sauf des plus forts : on obtient le graphe des plus proches voisins, qui représente les cooccurrents de premier ordre (liste des voisins directs d'un mot [Grefenstette 1994]) ;
2. remplacement du poids des arêtes par le nombre de voisins que les noeuds ont en commun afin d'obtenir le graphe des plus proches voisins partagés. A cette étape, il est possible de créer de nouveaux liens (par exemple : les mots A et B qui ne sont à l'origine pas reliés vont le devenir par transitivité s'ils partagent le voisin C). Cette étape est intéressante car elle permet une transition vers des cooccurrents de second ordre (mots partageant un même environnement). On passe donc de relations plutôt syntagmatiques à des relations paradigmatiques propices à des regroupements sémantiques ;
3. on procède à un filtrage des liens inférieurs à un seuil donné, fixé empiriquement à 5 ; les clusters correspondent aux composantes connexes du graphe simplifié ainsi obtenu. En d'autres termes, SNN considère que deux noeuds appartiennent au même cluster s'ils partagent un nombre suffisant de voisins.

Nous avons constaté empiriquement sur nos corpus que cette définition est problématique. Nous avons un nombre conséquent de mots susceptibles d'appartenir à plusieurs thématiques ; le risque de fusionner plusieurs clusters à cause de ces noeuds est élevé : notre graphe est donc à ce stade encore connexe [28] et on obtient donc, le plus souvent, un seul gros cluster.

Nous proposons ci-dessous de nouvelles étapes simples afin de prendre en compte le chevauchement de clusters.

— On remplace à nouveau le poids des liens par le nombre de voisins que les noeuds partagent, ce qui rapproche encore une fois par transitivité certains mots.

— Il s'agit désormais d'extraire les composantes comparativement denses du graphe, et d'isoler les mots susceptibles d'appartenir à plusieurs thématiques. Pour ce faire, nous considérons une définition de cluster légèrement différente de celle de SNN : nous considérons désormais que deux noeuds appartiennent au même cluster thématique s'ils partagent la majorité de leurs voisins respectifs. En d'autres termes, on ne compare plus uniquement le nombre absolu de voisins communs mais le nombre de voisins communs relativement au nombre total de voisins des noeuds en question, comme indiqué dans l'équation ci-dessous. Les arêtes vers les noeuds multiclasses sont ainsi pénalisées car ces derniers ont un nombre total de voisins largement supérieur au nombre de voisins qu'ils partagent avec chaque thématique prise séparément. Cela présente aussi l'avantage de normaliser le poids des arêtes.

$$\frac{{C_{ij}}^2}{(N_i - 1) \cdot (N_j - 1)} \tag{3.7}$$

28. Un graphe non orienté est connexe s'il existe un chemin entre toutes ses paires de noeuds.

C_{ij} correspond au nombre de voisins que partagent i et j ; N_i et N_j renvoient respectivement au nombre de voisins qu'ont les noeuds i et j. On retire 1 à chacune de ces valeurs car i ne peut évidemment pas partager j avec ce dernier. Ce score est compris entre 0 et 1.

— On filtre les arêtes ayant un poids inférieur à un seuil donné. Ce dernier est choisi volontairement bas afin que la définition d'un cluster ne soit pas trop restrictive (risque de scinder à tort certaines thématiques). Nous avons constaté que 0.5 et 0.6 donnent en général les meilleurs résultats. Nos thématiques sont les composantes connexes de ce nouveau graphe ainsi obtenu.

— On réintègre les arêtes n'ayant pas « survécu » à l'étape précédente. Un noeud appartient à un cluster s'il a des liens avec une majorité de mots de ce cluster (seuil fixé empiriquement à 80 %). Cela permet aussi de fusionner certains petits clusters isolés à tort (quand tous les mots du cluster sont liés à un autre).

Ces 4 nouvelles étapes sont schématisées dans la figure 3.3.

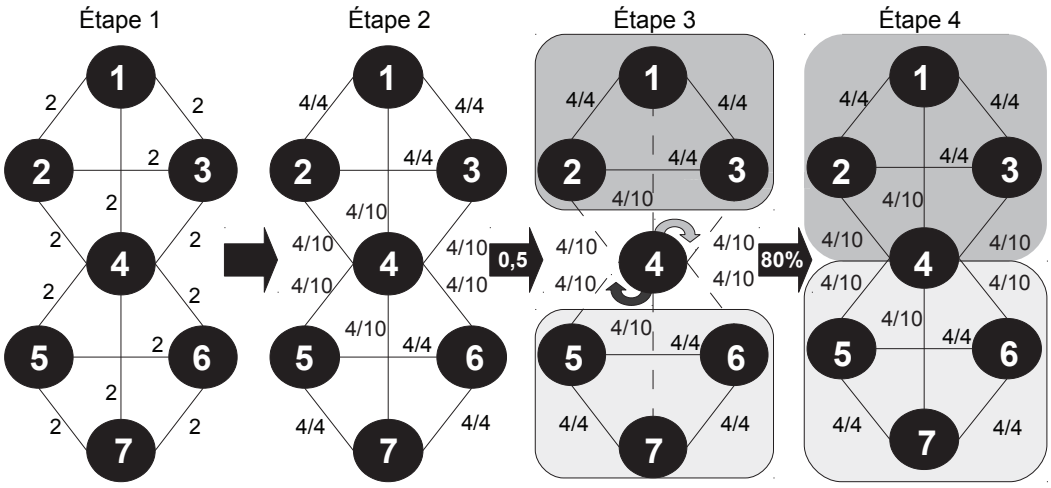

Figure 3.3 – Les 4 nouvelles étapes ajoutées à SNN. A l'étape 1, on entrevoit deux regroupements : $[1, 2, 3, 4]$ et $[4, 5, 6, 7]$. On constate que le noeud 4 est à la croisée de ces deux clusters. Chaque noeud partage 2 voisins avec n'importe quel autre noeud du graphe. La pondération des arêtes est modifiée à l'étape 2 afin d'isoler les noeuds multiclasses comme 4. A l'étape 3, on procède à la détection des composantes connexes du graphe, en ignorant les arêtes ayant un poids inférieur à 0.5 : on obtient donc deux clusters : $[1, 2, 3]$ et $[5, 6, 7]$. A l'étape 4, les arêtes ignorées à l'étape précédente reprennent leur place : 4 est relié à plus de 80 % des noeuds de chacun des clusters détectés à l'étape 3. Il les rejoint donc.

3.1.4 Résultats et discussion

3.1.4.1 Corpus de test

Notre corpus de test a été collecté à l'aide de AMI-EI en réponse à la requête "nucléaire". Nous n'avons considéré que les articles de presse rédigés en français entre le 17/04/2011 et le 16/05/2011 inclus. Cette période a été choisie pour son intérêt dans un cadre de veille : quelle image a le nucléaire en France un mois après l'incident survenu à Fukushima le 11 mars 2011 ? Après filtrage manuel [29], le corpus comporte 471 articles uniques, 170 437 mots et 12 070 vocables. Le corpus a été étiqueté avec Cordial [30]. Les mots trop fréquents ou fortement liés au Web ont été filtrés. Seuls les substantifs ont été conservés ; ces derniers sont lemmatisés afin de favoriser les rapprochements.

Il existe de nombreux corpus de test couramment utilisés pour l'évaluation de ce genre de méthodes, par exemple composés d'articles du Wikipédia, d'articles scientifiques ou de presse. Néanmoins, nous avons jugé préférable d'utiliser un "scénario réel" de collecte d'entreprise. Il est courant d'avoir des corpus de veille dans cet ordre de grandeur (moins d'un million de mots). Il est ainsi primordial de ne pas négliger ce cas de figure.

3.1.4.2 Présentation des thématiques

Rappelons que le but de ces clusters est de permettre une bonne vision d'ensemble du corpus, et servir de premières pistes d'exploration et d'interprétation des buzz de la part du veilleur. Ils ont surtout une valeur exploratoire, et non probatoire. 11 thématiques sont renvoyées par notre système ; 10 d'entre elles sont facilement interprétables. Nous leur donnons ci-dessous un nom issu de l'interprétation des mots [31] :

1. la hausse des prix de l'électricité en France ;
2. Tchernobyl ;
3. la centrale de Mühleberg et le nucléaire suisse en général ;
4. écologie, société et politique ;
5. reportage La Zone à propos des familles vivant aux alentours de Tchernobyl ;
6. incident dans un brise-glace russe ;
7. bourse et entreprises (rachats, fusions, etc.) ;
8. mouvements anti-EPR ;
9. candidature de Nicolas Hulot ;

29. Textes hors sujet ou devenus inaccessibles (liens morts).

30. `http://www.synapse-fr.com/Cordial_Analyseur/Presention_Cordial_Analyseur.htm`.

31. De nombreux travaux cherchent à attribuer automatiquement un nom aux classes (*topic labeling* –nommage de thématiques), par exemple [Mei 2007]. Il est courant de recourir à un triplet composé des mots les plus significatifs. L'intérêt du nommage de classe par un triplet est de recontextualiser au maximum les mots isolés de leur contexte d'origine. C'est notamment la méthode utilisée par [Rossignol 2005].

10. nucléaire iranien.

Pour des raisons de place, les résultats sont présentés dans trois tableaux (3.1, 3.2, et 3.3). Une bonne connaissance du domaine est bien entendu requise pour interpréter au mieux ces rapprochements. Le cas échéant, ces ensembles servent de points d'amorce à des recherches plus poussées sur le sujet.

1. Prix de l'électricité	*EDF, PRIX, EUROPÉEN*, électricité, hausse, augmentation, tarif, facture, euro, centime, inflation, marché, consommateur, ménage, client, particulier, impact, gouvernement, besson, eric, ue, union, loi, vigueur, arenh, nome, commission, champsaur, cre, fixation, taxe, concurrence, concurrent, opérateur, fournisseur, ouverture, contrat, direct, énergie, poweo, suez, gdf, proglio, mégawattheure, mégawatt/heure, mwh, kwh, type, rang, canal, vertu, rythme, garde, voisin, différence, période, lien, départ, moment, moyenne, détail, organisation
2. Tchernobyl	*CENTRALE, TCHERNOBYL*, russie, dmitri, drame, victime, pompier, employé, liquidateur, mort, monument, mémoire, bougie, scientifique, science, connaissance, médecine, bandajevski, auteur, étude, rapport, institut, radioprotection, risque, dépôt, particule, échelle, niveau, césium, environnement, évolution, nature, organisme, organe, cerveau humanité, population, enfant, femme, famille, communauté, alsace, homologue, ministre, angela merkel, françois, fillon, manuel, josé, président, présidence, conférence, travail, représentant, construction, béton, bouygues, vinci, pression, résistance, perte, exigence, principe, ressource, action, fin, voyage, sorte, exclusion, poids, haut, cinquantaine

Tableau 3.1 – Présentation des thématiques concernant le prix de l'électricité et Tchernobyl. Les mots appartenant à plusieurs thématiques à la fois sont indiqués en majuscule.

La thématique 5 ne concerne qu'un seul document dans l'ensemble du corpus. Les très petites thématiques ne sont pas un problème en soi, mais cette dernière est beaucoup trop proche de celle sur Tchernobyl. Dans l'idéal, elles devraient être fusionnées. La thématique 6 concerne un sujet d'actualité qui n'est présent dans la presse qu'un seul jour (4 articles en tout) ; ce niveau de granularité est encourageant, préfigurant la possibilité de considérer la dimension temporelle dans le processus, par exemple : quelles sont les thématiques du jour, comment évoluent-elles, etc. Enfin, certaines thématiques ont une certaine propension à

3. Nucléaire suisse (centrale de Mühleberg)	mühleberg, peter, hufschmied, mandat, wildi, fmb, ifsn, sécurité, inspection, indépendance, administration, fonction, membre, conseil, ats, awp
4. Écologie, société et politique (débats)	*CENTRALE, ÉCOLOGIE, FOSSILE, BIODIVERSITÉ, AUTONOMIE, MONDE, EUROPÉEN, SYSTÈME, PRIX, PRODUCTION, LOBBY, RENDEMENT, INVESTISSEMENT, PROJET, ÉCONOMIE, POMPE, VOITURE, PHASE, CONTRAIRE, SCORPION*, planète, espace, terre, humain, peuple, pays, co2, serre, révolution, conscience, utopie, choix, dimension, combat, priorité, vision, changement, défi, progrès, espèce, survie, agriculture, réalité, inquiétude, pollution, avion, transport, commerce, profit, coût, compétitivité, secteur, compétence, capital, capitalisme, politique, moyen, mot, bout, instant, dernier, méthode, clé, mode, course, extraction, produit

Tableau 3.2 – Présentation de deux autres thématiques : le nucléaire suisse et l'écologie. Les mots appartenant à plusieurs thématiques à la fois sont indiqués en majuscule.

apparaître ensemble. C'est par exemple le cas des thématiques 1 et 7. Il en est de même pour la 4 et la 10 : la plupart des documents parlant des candidatures aux présidentielles de Nicolas Hulot et Eva Joly traitent d'écologie (mais l'inverse n'est pas nécessairement vrai).

L'appartenance d'un mot à plusieurs thématiques est très intéressante dans un cadre de veille. Par exemple, *EDF* est ici fortement lié aux problématiques de la hausse des prix de l'électricité et aux mouvements anti-EPR. Le cas du mot *système* est aussi intéressant. Ce dernier est à la fois présent dans la thématique sur l'écologie (4) et dans celle sur l'incident dans un brise-glace (6), mais avec un sens différent. Dans un cas, il s'agit principalement du système politique ou économique, dans l'autre du système mécanique (voir tableau 3.4).

Pour des raisons d'exhaustivité, les résultats des tableaux 3.1, 3.2 et 3.3 ont été présentés sous forme de listes. Rappelons que nos thématiques sont des sous-graphes et que les mots entretiennent donc des relations plus ou moins fortes entre eux. Certains mots entretiennent des relations privilégiées avec d'autres, esquissant ainsi des sous-regroupements intéressants. Il devient dès lors plus aisé de naviguer au sein de chaque thématique et de mieux les interpréter. La figure 3.4 représente la thématique sur Tchernobyl sous forme de graphe. Trois sous-thématiques se démarquent clairement [32] :

1. construction du sarcophage de confinement ;

32. Ces sous-regroupements sont bien séparés lorsque l'on augmente le seuil à 0.6.

5. Reportage "La Zone"	*CENTRALE, TCHERNOBYL, BIODIVERSITÉ, SCORPION*, zone, reportage, bruno, guillaume, herbaut, photographe, lieu, histoire, livre
6. Incident dans un brise-glace russe	*SYSTÈME*, rosatomflot, brise-glace, navire, réacteur, fuite, incident, évènement, cas, agence, arctique, barents, mer
7. Bourse et entreprises	*ÉNERGIE, PRODUCTION, FOSSILE, PROJET*, bourse, titre, ebitda, points, part, offre, affaire, chiffre, dollar, salaire, prime, objectif, baisse, trimestre, cadre, analyste, producteur, entreprise, acteur, areva, constellation, exelon, leader, fusion, partenariat, partenaire, consensus, rassemblement, opération, processus, activité, pétrole, solaire, alternative, recours, chauffage, temps, matière, appareil, plan, passage, appel, outil, hauteur, etats-unis, allemand, twh
8. Mouvements anti-EPR	*EDF*, epr, flamanville, greenpeace, yannick, rousselet, andré, militant, moratoire, asn, manche, oeuvre, approvisionnement, exploitation, accès, chantier, port, site, grue, véhicule, grille, installation, plate-forme, maison, poste, maître, gendarmerie, commune, possibilité, capacité, proximité, réseau, association, correspondant
9. Présidentielles et écologie	*FOSSILE, ÉCOLOGIE*, nicolas hulot, eva joly, joly, écolo, eelv, primaire, campagne, électeur, droite, stéphane, ps, pacte, discours, proposition, europe, formation, interdiction, barrage, essence, carbone, média, télé, ami, conflit, intérêt, éthique
10. Iran	iran, téhéran, ashton, programme, enrichissement, arme, sanction, discussion, dialogue, négociation, lettre, grande-bretagne, onu, syrie, istanbul, puissance

Tableau 3.3 – Présentation des 6 autres thématiques renvoyées par le système sur ce corpus. Les mots appartenant à plusieurs thématiques à la fois sont indiqués en majuscule.

2. radioactivité, santé et pollution ;
3. mort et commémorations.

L'allure générale du graphe présenté dans la figure 3.4 nous renseigne à la fois sur certaines sous-thématiques mais aussi sur leur degré de maturité. On distingue des zones lexicalement pauvres et d'autres beaucoup plus denses. En périphérie du graphe, des îlots de forte cohérence lexicale donnent à penser qu'il s'agit de formes sémantiques stabilisées tant elles sont aisément restituables (par exemple, *bougie*

Écologie, société et politique (4)	**Incident dans un brise-glace russe (6)**
Il faut donc agir maintenant et remplacer le ***système*** *économique actuel (...)*	*Une faible augmentation de la radioactivité dans l'air a été constatée dans le* ***système*** *de ventilation dans la salle du réacteur (...)*
Sortir du nucléaire, c'est d'abord sortir du ***système*** *actuel*	*Le* ***système*** *du réacteur sera arrêté et le processus de refroidissement commencera.*
Notre ***système*** *est périmé, car bâti sur le principe d'une énergie bon marché (...)*	*La cause probable de l'incident est une perte d'étanchéité des* ***systèmes*** *de la première enceinte du réacteur*
Car au delà de ce problème d'énergie, c'est tout le ***système*** *de production intensive, biens de consommation, agriculture, etc. qu'il faut mettre en question.*	*Une légère fuite radioactive a été détectée dans les* ***systèmes*** *de ventilation du propulseur nucléaire du brise-glace russe (...)*

Tableau 3.4 – Différents usages du mot *système* dans deux thématiques différentes.

et *mémoire* ; *chape* et *sarcophage*) et s'apparentent parfois à des figements syntagmatiques (*centrale [de] Tchernobyl*, *Tchernobyl [en] Ukraine*, *niveau [de] risque*). L'épaisseur des liens atteste d'ailleurs de fréquences remarquables. A l'inverse, les mots qui opèrent des jonctions entre les sous-thématiques sont des éléments génériques susceptibles d'être partagés par plusieurs formes sémantiques (*1986*, *catastrophe*, *explosion*, *radiation*, *mort*). Ils constituent des éléments structurants du graphe ; ils en assurent la cohésion générale. Enfin, la densité de la sous-thématique liée à la construction du sarcophage de confinement, en bas du graphe, est l'indice de son importance dans le corpus, mais aussi de la relative instabilité des formes sémantiques qui le composent : le vocabulaire est diversifié en raison de la vitalité du thème tandis que les liens, denses et variés, témoignent quant à eux de combinaisons lexicales riches. A l'inverse, les formes sémantiques liées aux niveaux de risques, aux victimes et à la catastrophe proprement dite sont, comme nous l'avons vu, lexicalement appauvries parce qu'elles sont stabilisées, leur lexicalisation est achevée.

Malheureusement, notons que la plupart des 11 thématiques extraites sont extrêmement denses et que les liens entre les mots deviennent alors triviaux (voir figure 3.5) : dans ces cas, la représentation sous forme de graphe n'apporte rien par rapport aux listes classiques telles qu'on peut voir notamment en sortie des *topic models*. Nous expliquons ce phénomène par notre première étape, qui consiste à remplacer une nouvelle fois le poids des liens par le nombre de voisins que les noeuds partagent, de façon à permettre de nouveaux rapprochements.

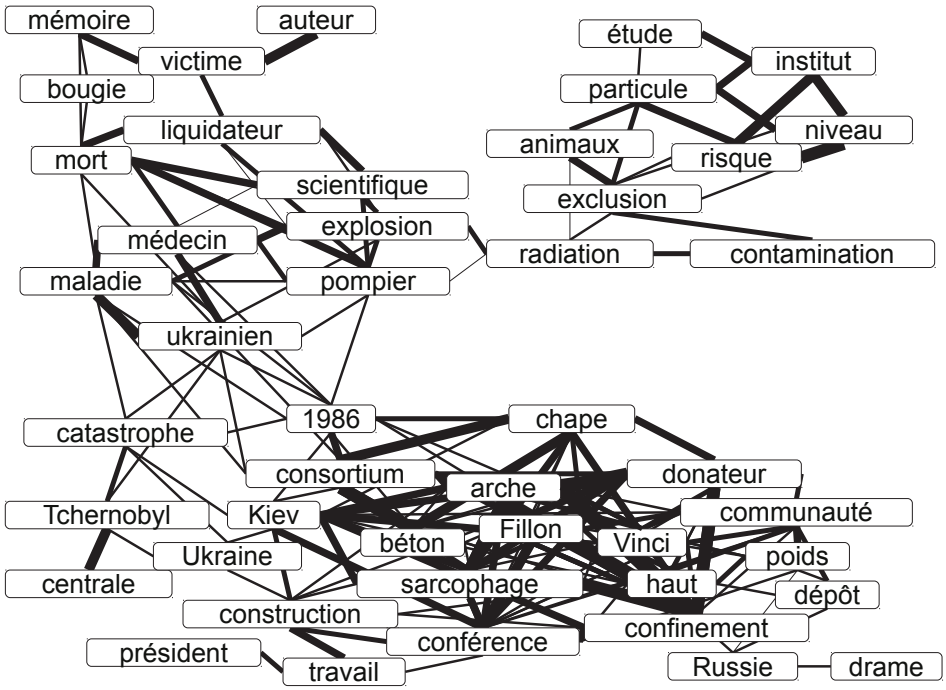

Figure 3.4 – Thématique sur Tchernobyl sous forme de graphe. Pour des raisons de lisibilité, le graphe a été allégé en retirant tous les mots apparaîssant moins de 15 fois. On constate des liens privilégiés entre certains mots, conduisant à des sous-regroupements. Par exemple : *mémoire*, *victime*, *bougie*, et *mort* entretiennent des liens privilégiés entre eux, tandis que *radiation* est naturellement plus proche de *particule*, ou de *contamination*.

Enfin, revenons sur la thématique que nous n'avons pas encore décrite. Cette dernière est illustrée dans le tableau 3.5. La raison pour laquelle nous obtenons cette thématique difficilement interprétable parmi les 11 renvoyées est évidente lorsque l'on visualise la configuration du graphe : les thématiques apparaîssent toutes nettement en périphérie du graphe tandis que les mots fortement multiclasses se retrouvent au centre (au croisement des thématiques concernées). Nous avons constaté que les noeuds de ce cluster problématique se trouvent tous au centre du graphe : il s'agit donc de noeuds ayant une très forte densité et rattachés à l'ensemble des thématiques, mais reliés à moins de 80 % des noeuds de chacune d'entre elles. Nous sommes donc confrontés à une thématique "poubelle" malgré la non-exhaustivité de notre méthode de clustering.

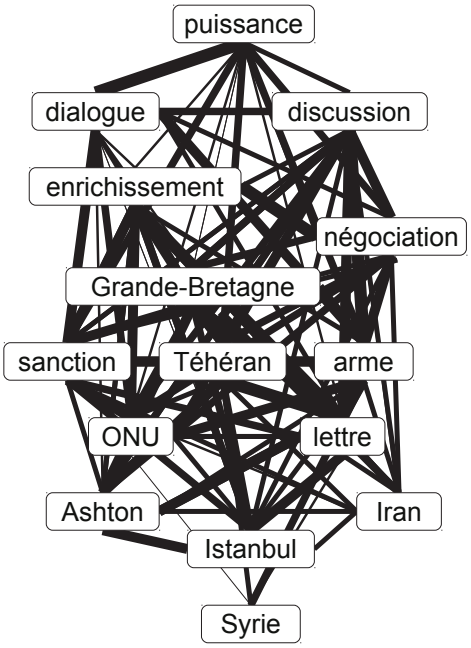

Figure 3.5 – Thématique sur l'Iran. Tous les mots sont fortement liés entre eux. La représentation sous forme de graphe n'apporte ici rien par rapport à la représentation sous forme de liste.

ÉNERGIE, PRIX, PRODUCTION, LOBBY, FOSSILE, AUTONOMIE, POMPE, RENDEMENT, VOITURE, CONTRAIRE, MONDE, BIODIVERSITÉ, INVESTISSEMENT, PROJET, ÉCONOMIE, PHASE, développement, nation, contribution, giec, but, génération, penly, étranger, village, apple, coopérative, fond, série, frontière, émission, fonctionnement, achat, bataille, chaîne, efficacité, composant, compagnie, place, bouchehr, economiesuisse, fissure, séisme, université, forêt, nuage, radioactivité, dampierre, dose, prestataire, combustible, tritium, sécheur [33], indice

Tableau 3.5 – Présentation de la thématique "poubelle" renvoyée par notre système sur ce corpus. Les mots appartenant à plusieurs thématiques à la fois sont indiqués en majuscule.

3.1.4.3 Évaluation des résultats

Méthodes d'évaluation Il existe deux types de mesures permettant l'évaluation d'une méthode de *clustering* : des mesures internes et des mesures externes.

Les premières permettent d'évaluer la qualité des partitions en l'absence de connaissances extérieures. Citons par exemple les mesures de cohésion (*cohesive-*

ness), vraisemblance (*likelihood*) et perplexité (*perplexity*). Les mesures externes, au contraire, évaluent la qualité des thématiques extraites par rapport à une classification de référence (réalisée manuellement). L'entropie, la pureté, et le F-score en sont des exemples. Le F-score, qui est la mesure la plus utilisée, est calculée à l'aide du rappel (r) et de la précision (p) :

$$\text{F-score}_i = \frac{2p_i r_i}{p_i + r_i} \tag{3.8}$$

Rappelons les formules pour calculer le rappel (*recall*) et la précision (*precision*). Le premier correspond au nombre de documents pertinents trouvés, par rapport au nombre total de documents pertinents. Il permet d'évaluer le silence : un faible score de rappel traduit un silence élevé :

$$\text{Rappel}_i = \frac{\text{documents correctement attribués à i}}{\text{nombre de documents appartenant à i}} \tag{3.9}$$

La précision renvoie au nombre de documents pertinents trouvés parmi l'ensemble des documents renvoyés par le système. Cette mesure évalue ainsi le bruit.

$$\text{Précision}_i = \frac{\text{documents correctement attribués à i}}{\text{nombre de documents attribués à i}} \tag{3.10}$$

Une évaluation multi-classe revient à calculer la moyenne des rappels et des précisions des n classes :

$$\text{Rappel} = \frac{\sum_{i=1}^{n} \text{rappel}_i}{n} \tag{3.11}$$

$$\text{Précision} = \frac{\sum_{i=1}^{n} \text{précision}_i}{n} \tag{3.12}$$

Évaluation de nos résultats Les deux types de mesures présentés précédemment ont leurs avantages et inconvénients. Le recours à une segmentation humaine sur un corpus de référence peut entrainer des problèmes de fiabilité dans l'évaluation étant donné une certaine subjectivité dans la tâche. Ce point peut être compensé en se reposant sur les fortes intersections entre les jugements. En tout cas, l'évaluation est difficilement reproductible. Les mesures internes permettent d'évaluer les résultats plus formellement, et aussi complètement automatiquement. Néanmoins, il n'y a pas nécessairement corrélation entre ces mesures et les jugements humains d'interprétabilité.

Nous optons pour la comparaison avec une classification effectuée manuellement. Chaque document du corpus a été rangé parmi les thématiques renvoyées par notre approche, en gardant à l'esprit qu'un document peut traiter de plusieurs thématiques à la fois, et peut n'appartenir à aucune des thématiques détectées[34].

L'affectation automatique d'un document aux thématiques renvoyées, quant à elle, fait appel à une méthode simple ; nous considérons qu'un document appartient

34. 26 % des documents n'a pas pu être rangé parmi les catégories détectées.

à une thématique s'il possède au moins n mots du cluster (aucune pondération selon la taille), fidèlement à l'idée selon laquelle tout ensemble de mots apparaîssant ensemble identifie un sujet de façon univoque.

La figure 3.6 indique les scores de rappel[35] et de précision (ainsi que le F-score associé) pour cette tâche de classification thématique. On constate que la précision atteint un très haut score et commence à stagner à partir de $n = 7$. $n = 4$ semble un bon compromis entre rappel et précision.

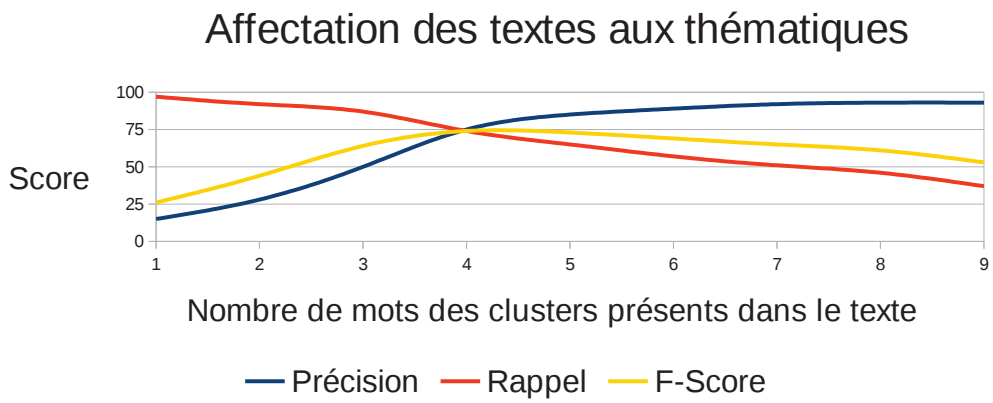

Figure 3.6 – Précision, rappel et F-Score pour la tâche d'affectation des documents aux thématiques, en fonction du nombre de mots des clusters présents dans le texte.

La majeure partie du silence provient ici des très petits documents (une phrase ou deux –c'est le cas des dépêches) étant donné que la valeur de n ne prend pas en compte la taille des textes. Le bruit est souvent dû aux mêmes mots assez peu "discriminants". C'est par exemple le cas de *dialogue* ou *discussion* dans la thématique autour de l'Iran. Afin de contourner ce problème, les mots devraient idéalement être pondérés au sein de chacun des clusters ; on verrait alors apparaître des mots centraux entourés de mots périphériques (sous-classe de mots pour lesquels le sème est afférent). La meilleure façon de pondérer les noeuds au sein de chaque cluster reste encore à déterminer, mais quelques pistes reposant sur la topologie des graphes peuvent être envisagées : nombre de liens forts intra-cluster, nombre de voisins inter-cluster, etc.

3.1.4.4 Comparaison avec les résultats de LDA

Étant donnée la difficulté à évaluer les résultats de ce type d'approches, nous avons aussi choisi de comparer nos résultats avec ceux produits par LDA sur le

35. Dans un cadre de veille, il peut être préférable de légèrement privilégier le rappel à la précision car il est important de perdre le moins d'information possible.

même corpus ; cette dernière étant l'approche la plus populaire d'extraction de thématiques. L'objectif n'est pas tant ici de comparer la pertinence des thématiques extraites que de valider, à titre indicatif, les résultats de notre approche.

Pour ce faire, nous avons utilisé l'implémentation en Java JGibbLDA [36], avec les paramètres par défaut. Nous demandons à extraire 10 thématiques, illustrées partiellement dans les tableaux 3.6 et 3.7.

Les thématiques extraites sont assez proches des nôtres. LDA présente l'avantage de pondérer l'importance de chacun des mots au sein des thématiques mais les relations entre ces derniers ne sont pas explicitées, rendant l'interprétation des résultats potentiellement plus complexes.

3.1.4.5 Perspectives

La méthode que nous avons proposée, reposant sur les plus proches voisins partagés dans un graphe de cooccurrences cumule des propriétés qui nous semblent primordiales d'un point de vue linguistique : classification non exhaustive et multi-classe. De plus, la visualisation sous forme de graphe facilite le parcours interprétatif des thématiques détectées par le veilleur. Les premiers résultats sont encourageants mais certaines pistes restent à approfondir.

Tout d'abord, l'impact de la mesure de dissimilarité employée pour pondérer les relations de cooccurrences dans le graphe reste à analyser, en la comparant avec d'autres mesures comme l'information mutuelle de [Church 1990].

Ensuite, nous avons pris le parti de nous limiter, dans un premier temps, aux substantifs car il s'agit de la catégorie la plus "marquée sémantiquement". Cependant, nous souhaitons dans l'avenir montrer l'influence de chacune des catégories sur les thématiques obtenues : nous projetons donc de prendre en compte, à court terme, les adjectifs ainsi que les verbes. Nous comptons aussi évaluer notre méthode en ignorant les catégories morphosyntaxiques des mots. Des premiers tests incluant les adjectifs ont été réalisés. On constate que les résultats sont très similaires et que les mots représentant les thématiques sont majoritairement des substantifs. Néanmoins, ce constat s'explique probablement par le filtrage en amont par fréquence appliqué au corpus. Par ailleurs, il pourrait être intéressant de travailler sur des termes et pas uniquement des mots graphiques : Cordial fait des regroupements pour certaines entités nommées comme "Nicolas Hulot" mais cela reste assez marginal.

Enfin, les noeuds au sein de chaque thématique devraient idéalement être pondérés afin de mieux filtrer certains mots encore trop génériques, mais aussi pour faciliter davantage l'interprétation des résultats. En d'autres termes, il s'agirait ici de distinguer les mots périphériques des mots plus discriminants [37]. De nombreux critères peuvent être envisagés pour la pondération : le nombre d'occurrences au sein du corpus, le nombre de voisins au sein de la thématique ou des thématiques voisines, calculs de centralité, etc.

36. http://jgibblda.sourceforge.net/.
37. Opposition entre sèmes afférents et sèmes inhérents ?[Valette 2010]

hulot (0.0224)	nucléaire (0.0384)	électricité (0.0523)
nicolas (0.0165)	chantier (0.0180)	prix (0.0491)
candidat (0.0153)	france (0.0178)	edf (0.0344)
parti (0.0132)	réacteur (0.0175)	euro (0.0319)
question (0.0130)	epr (0.0161)	gouvernement (0.0259)
écologie (0.0129)	association (0.0160)	tarif (0.0216)
europe (0.0115)	lundi (0.0139)	1er (0.0157)
ps (0.0097)	militant (0.0131)	hausse (0.0137)
sorti (0.0090)	greenpeace (0.0119)	marché (0.0135)
vert (0.0084)	test (0.0116)	concurrent (0.0135)
écologiste (0.0081)	personne (0.0112)	juillet (0.0121)
gauche (0.0078)	afp (0.0107)	énergie (0.0115)
débat (0.0074)	fessenheim (0.0105)	loi (0.0115)
nucléaire (0.0068)	jour (0.0104)	mwh (0.0100)
eelv (0.0063)	site (0.0085)	commission (0.0098)
énergie (0.0542)	centrale (0.0581)	rapport (0.0200)
nucléaire (0.0399)	projet (0.0181)	année (0.0173)
monde (0.0140)	réacteur (0.0133)	million (0.0151)
france (0.0123)	site (0.0118)	ministre (0.0134)
an (0.0113)	niveau (0.0115)	rente (0.0130)
développement (0.0112)	déchet (0.0113)	enfant (0.0104)
pays (0.0105)	installation (0.0108)	vie (0.0096)
consommation (0.0105)	combustible (0.0105)	belgique (0.0094)
système (0.0071)	eau (0.0103)	étude (0.0094)
production (0.0068)	mesure (0.0084)	commission (0.0092)
suisse (0.0068)	cas (0.0082)	bnb (0.0090)
année (0.0065)	région (0.0076)	cancer (0.0090)
fait (0.0063)	comité (0.0071)	mercredi (0.0083)
technologie (0.0063)	arrêt (0.0064)	danger (0.0081)
vie (0.0060)	vendredi (0.0062)	risque (0.0077)

Tableau 3.6 – Extrait de 6 des 10 thématiques détectées par LDA sur notre corpus. Pour des raisons de lisibilité, seuls les 15 premiers mots sont affichés, accompagnés de leur poids arrondi au sein de la thématique.

Pour conclure avec cette approche, il nous reste à aborder la question de l'analyse diachronique de ces thématiques. A ce sujet, citons les travaux autour du *Topic Detection and Tracking* (TDT), notamment [Allan 1998a], [Allan 2002], [Binsztok 2002], et [Makkonen 2009]. Ce programme de recherche, initié en 1997 par le DARPA [38], avait pour ambition la réalisation d'un système permettant l'extraction automatique d'évènements thématiques au sein d'un flux continu d'informa-

38. La *Defense Advanced Research Projects Agency* est une unité de recherche et développement du département de la Défense des Etats-Unis.

euro (0.0347)	iran (0.0163)
groupe (0.0261)	conseil (0.0138)
milliard (0.0258)	sécurité (0.0138)
chiffre (0.0131)	programme (0.0135)
million (0.0125)	président (0.0122)
areva (0.0123)	arme (0.0084)
edf (0.0110)	chine (0.0077)
affaire (0.0111)	ukraine (0.0075)
dollar (0.0103)	chef (0.0074)
jeudi (0.0102)	uranium (0.0068)
production (0.0086)	puissance (0.0063)
croissance (0.0084)	négociation (0.0063)
objectif (0.0083)	inde (0.0059)
total (0.0078)	kiev (0.0059)
trimestre (0.0078)	accord (0.0056)
gaz (0.0381)	tchernobyl (0.0535)
schiste (0.0177)	centrale (0.0290)
information (0.0140)	an (0.02668)
exploitation (0.0107)	catastrophe (0.0261)
environnement (0.0105)	réacteur (0.0167)
amp (0.0102)	avril (0.0158)
gouvernement (0.0097)	zone (0.01540)
sécurité (0.0088)	ukraine (0.0139)
loi (0.0086)	million (0.0126)
droit (0.0077)	1986 (0.0116)
jour (0.0075)	travail (0.0114)
matière (0.0072)	pays (0.0010)
service (0.0064)	mardi (0.0093)
source (0.0063)	président (0.0088)
demande (0.0059)	kiev (0.0088)

Tableau 3.7 – Extrait des 4 dernières thématiques détectées par LDA sur notre corpus. Pour des raisons de lisibilité, seuls les 15 premiers mots sont affichés, accompagnés de leur poids arrondi au sein de la thématique.

tion. Trois notions sont alors distinguées : l'évènement, l'histoire, et la thématique [39]. L'évènement est un fait qui s'est produit à un instant et un endroit précis. L'histoire est un ensemble d'évènements autour d'un sujet donné. Enfin, la thématique regroupe toutes les histoires composées d'évènements similaires. Il ne s'agit donc pas uniquement de détecter des éléments saillants, mais aussi d'inférer le lien entre ces derniers et de suivre leur évolution dans le temps.

Un document peut appartenir à plusieurs thématiques à la fois, de façon plus

39. Traductions littérales de l'anglais : *event*, *story*, et *topic*.

ou moins forte. Les difficultés inhérentes à la TDT sont le nombre potentiellement restreint de documents qui parlent d'un évènement donné, l'absence de changement majeur dans le vocabulaire au fil du temps pour une même histoire, et enfin la difficulté à faire la distinction entre deux évènements similaires : des articles parlant par exemple de deux accidents d'avion différents partagent en effet un nombre conséquent de mots ; il faut alors s'appuyer sur les entités nommées, les lieux, ou encore surtout les dates [40]. Les articles parlant d'un évènement donné sont en effet le plus souvent publiés dans un intervalle de temps restreint [41].

Il peut aussi être intéressant, à l'instar de [Fukumoto 2000], de séparer le vocabulaire propre aux évènements, histoires et thématiques. Par exemple, pour la thématique traitant du séisme de 1995 à Kobe, *Kobe*, *Japon* et *séisme* seraient plutôt associés à la thématique tandis que les mots *secours* et *travailleurs* caractériseraient un évènement précis. Partant du postulat que les nouvelles histoires sont marquées plus fortement par les mots inhabituels, d'autres travaux comme [Swan 2000] calculent plutôt la probabilité d'apparition d'un mot à une date donnée.

Faisant suite à cette idée, nous pouvons tenter d'adapter notre méthode et analyser l'évolution dans le temps des thématiques extraites. La principale difficulté est que de nouveaux évènements apparaîssent et disparaîssent en permanence, sans changement majeur du vocabulaire employé. Afin de pallier ce problème, nous envisageons d'analyser conjointement le profil lexical et le profil temporel des mots. On pense ici à un calcul des spécificités lexicales et des accroissements spécifiques [Lafon 1980] [Lebart 1994] pour chaque fenêtre temporelle (par exemple le jour) afin de dégager ce qui caractérise réellement chacune d'entre elles et ainsi détecter au sein de chaque thématique les évènements inattendus ; certains mots sont banals, c'est-à-dire qu'ils ne sont jamais discriminants tout au long de la thématique tandis que certains ont une utilisation beaucoup plus ponctuelle : on peut estimer qu'un groupe de cooccurrents spécifiques à une période donnée marque de façon assez certaine l'arrivée d'un nouvel évènement au sein de la thématique. En d'autres termes, on considère un nouvel évènement comme un ajout soudain de vocabulaire spécifique à un instant t, relié dans le graphe de cooccurrences à un certain nombre de mots appartenant au vocabulaire de base.

Le calcul des spécificités permet d'estimer de façon probabiliste le degré de surreprésentation ou de sous-représentation d'une forme donnée dans un sous-corpus. Pour nos premiers tests, nous nous sommes appuyés sur l'implémentation qui en est faite dans le logiciel Lexico3 [42] qui rassemble des outils de statistiques textuelles et de lexicométrie. Faute de temps, la méthode envisagée n'a malheureusement pas été complètement implémentée et évaluée au cours de cette thèse.

40. Date de publication des documents ou bien dates apparaîssant au sein de ces derniers.
41. [Sayyadi 2009] confirme cette hypothèse à partir de 18 000 articles de presse : 90 % des histoires ont une durée de vie inférieure à 11 jours.
42. http://www.tal.univ-paris3.fr/lexico/.

3.2 Analyse du buzz par extraction de reprises

Les thématiques extraites avec la méthode précédente permettent une bonne vision d'ensemble de ce dont les internautes parlent. Néanmoins, malgré quelques premières pistes prometteuses, la méthode semble relativement peu adaptée à l'analyse en temps réel d'histoires quotidiennes. Par ailleurs, elle risque d'être mise en échec sur des textes courts pourtant de plus en plus fréquents sur le Web avec les réseaux sociaux [43]

Nous proposons donc en complément une seconde approche, plus fine. Cette dernière s'appuie sur quelques travaux récents de détection de groupes de citations proches. Nous allons dans un premier temps décrire les travaux en question et mettrons en avant l'intérêt de l'analyse des citations dans un cadre d'identification et de suivi de buzz. Après une réflexion sur la notion même de citation, nous présenterons nos adaptations par rapport à la méthode originale, qui consistent principalement à étendre le principe à la détection de tout contenu quasi-dupliqué, ce qui est particulièrement utile pour l'analyse textuelle des réseaux sociaux. Nous présenterons et discuterons ensuite nos résultats. Enfin, une analyse diachronique de ces quasi-duplications sera proposée afin d'essayer de comprendre la façon dont le sens peut se déformer au fil du temps.

3.2.1 Travaux antérieurs

Depuis [Leskovec 2009] [44], un certain intérêt est né dans la communauté scientifique autour de la détection et l'analyse automatique de familles de citations. Ces familles sont des groupes de citations proches les unes des autres (faibles variantes textuelles), sachant qu'une citation est ici définie tout simplement comme une chaîne de caractères entre guillemets. Les auteurs présentent une méthode capable d'identifier et regrouper ces familles de citations.

Rappelons le contexte actuel de l'information en continu [45] (24h news cycle). Les histoires circulent et sont mises à jour instantanément aussi bien à travers les médias classiques que les nouveaux acteurs de l'information que sont les blogs et les réseaux sociaux.

Pour analyser le cycle de vie d'une information, les auteurs discutent les pistes suivantes. Tout d'abord, ils estiment (et cela va dans le sens de notre ressenti) que l'extraction de thématiques permet surtout la découverte de thématiques générales. Il y a donc un problème de granularité : ces dernières parviennent difficilement à rendre compte des modifications quotidiennes. Les auteurs préconisent alors une analyse simplifiée : il convient d'extraire des fragments textuels, se répétant d'article en article (pouvoir discriminant), qui soient faciles à détecter et traquer. Ces

43. [Ben Jabeur 2012] signale que les tweets les plus fréquents ont une longueur de 4 mots, et que la longueur moyenne est de 11.

44. Les données utilisées ainsi qu'une démonstration de leur outil sont disponibles aux adresses suivantes : `www.memetracker.org` et `www.blogcascades.org`.

45. Une description plus détaillée du contexte est donnée dans l'intervention de [Kleinberg 2009].

fragments doivent par ailleurs permettre de rendre compte des évolutions quotidiennes ; c'est la raison pour laquelle ils considèrent que les entités nommées et les n-grammes [46] ne sont pas adaptés car ces derniers apparaîssent quotidiennement sans que l'on soit réellement capable de déterminer ce qu'on en dit de nouveau. Ces histoires quotidiennes (*story lines*) sont difficiles à analyser car elles n'impliquent pas nécessairement de changement majeur du vocabulaire.

Les auteurs s'intéressent alors à la citation. Cette dernière est très courante, surtout dans la presse. Les citations changent de jour en jour, sont facilement identifiables et peuvent être rattachées à une date et à un auteur. Ils considèrent la citation comme une *empreinte génétique* permettant une identification univoque d'une histoire, une idée, ou encore un mème [Dawkins 1976] [47]. Il est dès lors possible d'effectuer des calculs volumétriques de ces citations, faire des estimations temporelles, analyser les déformations, ou encore inférer statistiquement la façon dont l'information s'est propagée de façon virale.

Les citations sont susceptibles d'être altérées, quelles qu'en soient les raisons : elles peuvent être citées de mémoire, être tronquées pour des raisons stylistiques ou pour mettre l'accent sur un élément que l'on juge plus marquant, etc. Par conséquent, une même citation peut avoir une multitude de variantes textuelles plus ou moins proches, formant ainsi une famille. Dans la suite de l'exposé, nous nommerons ces familles *citations*, et les variantes de chaque citation seront désignées en tant qu'*instances*. Une citation peut donc avoir plusieurs instances plus ou moins longues (nombre de mots graphiques) et plus ou moins éloignées (distance d'édition).

Afin de prendre en compte ce phénomène, les auteurs préconisent la construction d'un graphe orienté dont les noeuds sont les instances, et dont les arêtes traduisent une relation d'inclusion approximative : il existe un lien de l'instance p vers l'instance q si p est approximativement incluse dans q. Étant donnée cette relation d'inclusion, p a donc une taille strictement inférieure à q. Le graphe ainsi constitué est acyclique [48]. Les arêtes sont pondérées selon la distance d'édition entre les instances, mais aussi par leur fréquence. Seules les instances supérieures à une taille et à une fréquence données sont incluses dans le graphe [49]. La figure 3.7 [50] illustre ce type de graphes.

Il s'agit désormais de découper le graphe de façon à isoler les différentes citations. Afin de regrouper convenablement entre elles les instances, les auteurs introduisent la notion de *noeud racine* (*root node*) ; ce sont des noeuds dont le degré sortant est nul et qui correspondent donc aux instances les plus longues de chaque citation.

46. En vue par exemple d'une détection d'unités polylexicales.

47. Idée simple propagée très vite massivement sur le Web par imitation et/ou émulation. Il peut s'agir par exemple d'une vidéo, une anecdote, ou encore une phrase.

48. Un graphe acyclique, comme son nom l'indique, ne comporte pas de *cycle*. Un cycle est un chemin dont l'origine et la destination sont identiques.

49. A noter par ailleurs le recours à une stop list constituée par exemple des répliques cultes de films, ou des titres de chansons.

50. Le corpus utilisé par les auteurs était constitué d'articles de presse et de blogs écrits pendant les 3 mois ayant précédé les élections présidentielles américaines de 2008 (données extraites de Spinn3r –`spinn3r.com`).

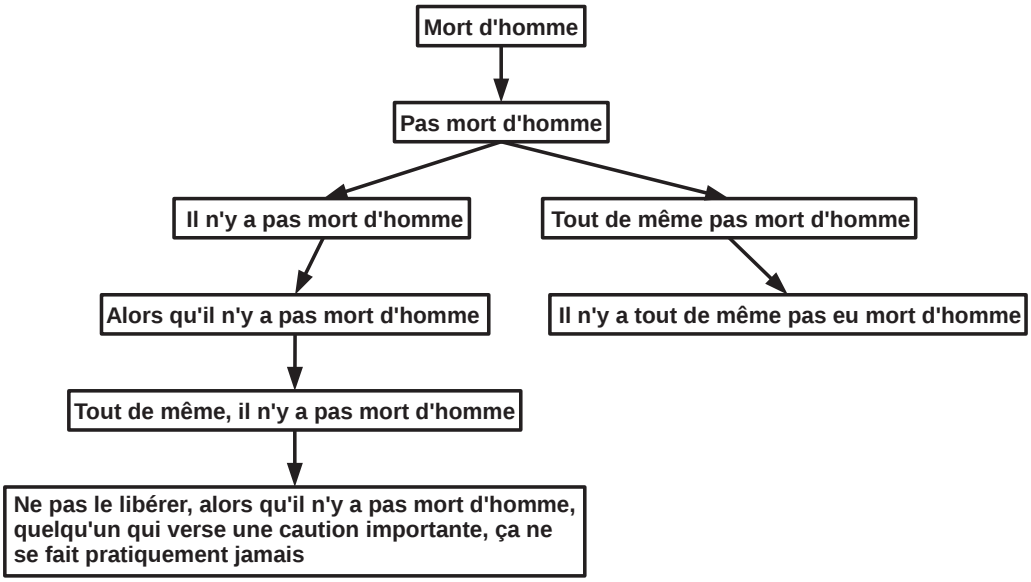

Figure 3.7 – Illustration du graphe orienté de citations, tel que décrit dans [Leskovec 2009]. Les noeuds sont les instances. Les arêtes traduisent une relation d'inclusion approximative entre ces dernières, pondérées selon leur distance d'édition ainsi que leur fréquence.

Cela permet de définir une citation comme un sous-graphe dont tous les chemins aboutissent à un noeud racine unique (ensemble des instances de la citation incluses dans l'instance la plus longue). Il s'agit donc de supprimer les arêtes les plus faibles de façon à découper le graphe en un ensemble de composantes connexes ayant toutes un noeud racine unique. Le problème étant NP difficile, les auteurs proposent des heuristiques leur permettant de partitionner le graphe. Le principe est schématisé dans la figure 3.8.

La méthode permet aux auteurs de constater certains phénomènes intéressants permettant de mieux comprendre la façon dont l'information se propage sur le Web. Par exemple, la plupart des blogs semblent tirer leurs informations des médias classiques (information postée en moyenne 2h30 après la presse), à l'exception de certains blogs professionnels diffusant l'information bien avant la presse.

[Simmons 2011] reprend la méthode et analyse en parallèle le graphe des hyperliens de façon à estimer le degré de variations entre les différentes instances des citations selon le type et l'autorité de la source. Les auteurs constatent que les citations sont beaucoup moins altérées dans la presse, qui a néanmoins tendance à les rétrécir pour des raisons de style.

Dans [Omodei 2012], un nouvel algorithme de regroupement de citations est introduit. La méthode s'inspire fortement de [Leskovec 2009] mais repose sur des analyses plus linguistiques. Les mots sont lemmatisés à l'aide de TreeTagger et les mots vides sont filtrés. A la place de l'inclusion approximative, les auteurs ont recours

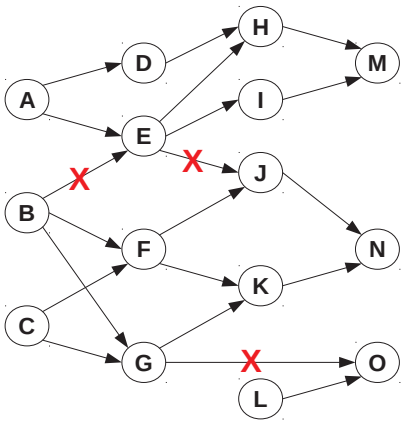

Figure 3.8 – Illustration du partitionnement du graphe orienté acyclique des citations. L'objectif est de supprimer le moins possible d'arêtes de façon à obtenir un ensemble de composantes connexes ayant chacune une racine unique (noeud avec un degré sortant nul).

à la distance de Levenshtein [Levenshtein 1966] pour relier les instances entre elles. Cette mesure est légèrement adaptée de façon à prendre en compte la fréquence d'occurrence des mots dans le calcul –l'objectif est de donner plus de poids aux mots rares [51], considérés plus informatifs.

3.2.2 La citation

Rappelons que la citation est pragmatiquement définie dans les travaux ci-dessus comme une chaîne de caractères située entre guillemets.

La citation est une forme d'intertextualité très utilisée dans le domaine journalistique qui permet une recontextualisation d'un discours énoncé par un tiers. L'usage de guillemets permet de bien identifier le discours repris et permet ainsi une certaine forme d'objectivité et de neutralité dans les propos rapportés. [Poulard 2008a] rappelle néanmoins que la délimitation de cet objet linguistique est loin d'être triviale étant donné que ses bornes ne sont pas toujours évidentes.

Le cas le plus simple est la reprise du discours au style direct. Le passage repris, placé entre guillemets, est en théorie retranscrit fidèlement, outre quelques légères modifications morphosyntaxiques à des fins stylistiques : meilleure concordance des temps avec le texte avant la citation, anaphores, etc. A noter que, bien que cela soit la marque privilégiée afin de marquer la citation, le guillemet peut être concurrencé par une modification typographique (principalement l'italique). Dans un contexte d'extraction automatique de citations, le guillemet reste tout de même un marqueur beaucoup moins ambigu.

Plus complexe à détecter automatiquement : le style indirect simple. Le discours rapporté est modifié de façon à s'intégrer directement dans la nouvelle situation

51. Les mots sont classiquement pondérés par un TF.IDF.

d'énonciation. La citation n'est alors plus délimitée par un marqueur particulier –au mieux quelques indices tels que *verbe + que*.

Citons enfin le style indirect libre, à mi-chemin entre les deux cas signalés précédemment. C'est le style le plus utilisé par les journalistes. Il s'agit d'un style majoritairement indirect, agrémenté de quelques courts passages au discours direct [52]. [Jackiewicz 2006] parle dans ce cas de *citation interprétée* : l'énonciateur ne cite alors que ce qu'il juge nécessaire et suffisant pour retranscrire l'intention communicative originelle.

La typologie des citations effectuée par [Jackiewicz 2006] est intéressante en ce qu'elle permet de nuancer l'idée, énoncée dans [Leskovec 2009], de la citation comme empreinte génétique permettant une identification univoque d'une histoire : *certains discours appellent, en raison notamment de l'absence d'un bagage culturel ou d'un savoir commun entre le locuteur d'origine et le co-énonciateur, un commentaire, une explication ou une reformulation mettant en évidence leur teneur effective. Nous dirons donc que, selon la nature de ses composants, une citation peut être qualifiée de minimale, authentifiée, interprétée ou dialogique.* Nous pensons néanmoins qu'il s'agit d'une piste pertinente dans le cadre d'un suivi de buzz.

3.2.3 Notre implémentation

Notre implémentation s'inspire fortement de la méthode décrite dans [Leskovec 2009]. Nous préférons néanmoins la relation de ressemblance à celle d'inclusion car elle nous semble plus souple et tolérante. La mesure de ressemblance utilisée est issue de travaux en détection de plagiat et permet la détection de quasi-duplications. Il s'agit d'une technique connue sous le nom de *shingling* [Broder 1997].

Le texte est dans un premier temps divisé en tokens –en l'occurrence des mots graphiques. On nomme *shingle* toute séquence contigüe de tokens, de taille w, contenue dans un document D. $S_w(D)$ (w-*shingles*) traduit l'ensemble de tous les *shingles* uniques de taille w que contient le document D. Pour reprendre l'exemple classique pour décrire ce genre de méthodes, le *4-shingles* de *[a rose is a rose is a rose]* correspond à :

```
{(a,rose,is,a), (rose,is,a,rose), (is,a,rose,is)}
```

Les 2 premiers *shingles* apparaîssent 2 fois dans l'exemple. En fonction de la proximité entre les w-*shingles*, il est possible d'évaluer le degré de quasi-duplication entre deux documents. Concrètement, on évalue la ressemblance r entre deux documents A et B en utilisant le taux d'intersection et d'union entre les w-*shingles* de chacun d'entre eux comme suit :

$$r_w(A,B) = \frac{|S_w(A) \cap S_w(B)|}{|S_w(A) \cup S_w(B)|} \tag{3.13}$$

52. Citons cet exemple, tiré de [Poulard 2008a] : *Elle était bien l'"organisatrice" du concert. Ce concert était une activité de "service public". Les agents qui ont commis des fautes disposaient d'un "pouvoir de représentation" de la ville.*

La ressemblance [53] est une valeur comprise entre 0 et 1 et $r(A, A) = 1$ est toujours vrai.

Nous adoptons fidèlement la méthode décrite ci-dessus, à la seule différence que nous calculons des taux de quasi-duplication entre instances de citations, et non entre documents. Le choix de la valeur de w est important ; une valeur trop grande ignorerait les petites citations et une valeur trop petite rapprocherait à tort des instances à cause d'expressions (semi-)figées du type "c'est-à-dire-" ou "mais en réalité". Nous avons constaté empiriquement que $w = 4$ est un bon compromis, à condition de disposer d'une stop list des expressions (semi-)figées les plus courantes de la langue. Notons enfin que nous appliquons une opération de racinisation [54] sur les instances avant tout traitement.

La phase d'extraction des citations est assez similaire à la méthode originelle : le découpage du graphe se fait de proche en proche en partant des noeuds racines et en suivant les liens dans le sens inverse. On associe à chaque noeud la citation majoritaire de ses descendants. A égalité, on prend la citation du descendant qui a l'arête avec le poids le plus élevé. Chacune des citations ainsi extraites est caractérisée par plusieurs attributs :

— nombre d'instances différentes ;
— fréquence totale de la citation (toutes instances confondues) ;
— dates d'occurrence et identifiants des documents dont elles sont issues (à des fin volumétriques) ;
— un label qui sert de "représentant" à l'ensemble de la citation. Le choix du label est un compromis entre la taille de l'instance et le nombre d'occurrences ; nous jugeons que ce dernier doit être assez fréquent, et assez petit pour pouvoir être affiché confortablement [55].

3.2.4 Discussion

Nous pensons que l'utilisation de la citation comme *empreinte génétique* d'un buzz présente un intérêt certain. Néanmoins, rappelons que la méthode se limite ici à la citation au discours direct notée entre guillemets. Si cette dernière est très présente dans les articles journalistiques, qu'en est-il des autres types de documents issus du Web ? Il est tentant de chercher à étendre cette définition de la citation en prenant en compte les autres cas de figure présentés dans la partie 3.2.2 mais la détection automatique de ces éléments est loin d'être triviale. C'est la raison pour laquelle nous préférons plutôt étendre le principe à la détection de tout contenu quasi-dupliqué issu des réseaux sociaux. La duplication de contenu est en effet un des principaux mécanismes de propagation virale au sein de ces derniers : tout message posté par un tiers peut être partagé à l'identique à sa liste d'amis. C'est d'autant plus vrai

53. Trivialement, la distance entre deux documents se définit ainsi : $d(A, B) = 1 - r(A, B)$.
54. La racinisation est ici préférée à la lemmatisation car la méthode est complètement intégrée dans le produit d'AMI Software. Il n'existe pas, à notre connaissance, d'outil de lemmatisation qui puisse être utilisé librement à des fins commerciales.
55. Il s'agit donc tout simplement du nombre d'occurrences de l'instance multiplié par sa taille.

sur Twitter avec la notion centrale de retweet. En plus du simple ajout du mot-clé marquant le retweet, tout message repris peut être altéré par l'utilisateur. C'est en pratique très utilisé pour ajouter un bref commentaire en début ou en fin de tweet. Prenons pour exemple ce tweet de @lemondefr :

```
Décentralisation : comment le gouvernement renforce les régions
http://t.co/pqA3MwYMr1
```

Ce dernier a été retweeté de la façon suivante :

```
RT @lemondefr : ‘‘Décentralisation : comment le gouvernement renforce
les \#régions’’ cc @jphuchon @iledefrancefr
http://www.lemonde.fr/tiny/4405834/
```

On voit ici par l'utilisation des guillemets que l'utilisateur a pris le parti de bien séparer le contenu repris de son commentaire personnel[56]. Néanmoins, cela n'est pas systématique.

Appliquer la méthode sur un ensemble de tweets est trivial. Il suffit de construire le graphe d'origine à l'aide de tweets et non de citations. Étant donné le nombre très élevé de tweets, il est cependant préférable d'appliquer en plus un filtre par fréquences, de façon à éviter la construction d'un graphe trop volumineux.

Il est délicat d'évaluer la qualité des résultats de ce genre d'approches. D'un côté, il est possible de mesurer la qualité des regroupements en analysant le degré de cohésion des partitions obtenues. De l'autre, on peut demander à un ensemble d'évaluateurs humains d'estimer la qualité subjective des regroupements. Néanmoins, comment évaluer le silence ? Nous présenterons les résultats de notre implémentation appliquée sur des cas concrets dans les parties 4.2.2 et surtout 4.3.

3.2.5 Analyse qualitative des déformations

Les variations pouvant exister entre les différentes instances ont été analysées automatiquement dans les papiers présentés précédemment : relation entre le taux de variation et la taille de la citation [Simmons 2011], degré de stabilité d'un mot selon la catégorie morphosyntaxique [Omodei 2012], etc. Néanmoins, à notre connaissance, une analyse linguistique détaillée des phénomènes de déformations n'a jamais été effectuée.

Nous pensons qu'étudier et catégoriser plus précisément la façon dont les instances d'une même citation évoluent et se déforment dans le temps présente un intérêt certain en TAL. En effet, cela permet de fournir des pistes visant à étendre et améliorer ce type d'algorithmes de détection de citations proches et, dans une moindre mesure, les méthodes d'identification de paraphrases. Cela présente par ailleurs un intérêt d'un point de vue (socio-)linguistique, en apportant une meilleure compréhension des schémas de diffusion et de déformation du discours rapporté.

56. Signalons tout de même l'ajout de hashtags par rapport au message d'origine...

3.2.5.1 Description de l'approche

De par la taille des données à traiter, une observation complètement manuelle des variations n'est pas envisageable. Afin de nous aider dans cette tâche, nous avons recours à des techniques d'alignement de séquences couramment employées en biologie computationnelle pour identifier des régularités parmi des séquences d'ADN ou de protéines [Mount 2004] [57].

Les alignements sont le plus souvent représentés à l'aide d'une matrice, dont les lignes correspondent aux séquences. Les éléments identiques sont alignés sur les mêmes colonnes. Des *trous* (*gaps*) peuvent être introduits au besoin quand un élément n'a aucune correspondance (*match*) dans les autres séquences (une colonne ne peut être composée uniquement de trous). Plutôt qu'une correspondance entre résidus chimiques, nous nous intéressons à une correspondance entre les mots des différentes instances d'une citation, mais la façon de procéder est néanmoins assez similaire [Irving 2004] : lorsqu'un mot n'a pas de correspondance, un trou est inséré. Les mots identiques sont placés sur la même colonne. Un exemple simple d'un alignement d'instances est présenté dans le tableau 3.8.

-	-	-	-	-	-	offered	a	compelling	and	(...)
-	-	-	-	-	-	-	-	compelling	and	(...)
-	-	-	-	-	-	-	-	compelling	and	(...)
the	complainant	in	this	case	has	offered	a	compelling	and	(...)
the	victim	-	-	-	has	given	a	compelling	and	(...)
-	-	-	-	-	-	offered	-	compelling	and	(...)
-	-	-	-	-	-	-	a	compelling	and	(...)
-	-	-	-	-	-	-	-	compelling	and	(...)
she	-	-	-	-	-	offered	a	compelling	and	(...)

Tableau 3.8 – Illustration de l'alignement global de 9 instances d'une citation. Les derniers mots sont tronqués pour que les instances restent affichées sur une seule ligne. Les tirets traduisent les trous (le mot n'est pas présent dans l'instance). On observe ici deux remplacements : *complainant/victim* et *offered/given*.

En utilisant la méthode décrite dans [Needleman 1970], il est trivial d'obtenir l'alignement global optimal pour deux séquences données. L'algorithme suit 3 étapes. Pour aligner une séquence $S1$ de taille m avec une séquence $S2$ de taille n, il faut tout d'abord initialiser une matrice $m+1 \times n+1$ de similarité telle que :

$$M_{(i,0)} = M_{(0,j)} = 0 \quad \forall i \forall j \tag{3.14}$$

57. Il est amusant de constater que la métaphore biologique est filée dans l'ensemble des papiers présentés précédemment. [Leskovec 2009] parle d'*empreinte génétique* permettant de spécifier un *mème*, défini par [Dawkins 1976] comme une sorte de *gène* culturel capable d'autoréplication. [Omodei 2012] utilise à son tour une terminologie issue de la biologie, en réponse aux travaux en mémétique. De façon tout à fait fortuite, nous en sommes arrivés à nous servir de méthodes issues de travaux en bio-informatique !

L'objectif de cette matrice est de renseigner les meilleurs scores pour tout alignement. Nous utilisons empiriquement la pondération suivante : +1 pour les correspondances (*match* –m), −1 de pénalité pour les remplacements (r) et les trous (*gap* –g). La matrice est remplie en suivant la formule suivante :

$$M_{(i,j)} = \begin{cases} M_{(i-1,j-1)} + m & \text{si } S1_{(i)} = S2_{(j)} \\ max(0, M_{(i-1,j)} - g, M_{(i,j-1)} - g, M_{(i-1,j-1)} - r) & \text{sinon} \end{cases} \tag{3.15}$$

Comme toute méthode de programmation dynamique, la matrice est ensuite parcourue de bas en haut en suivant les meilleurs scores. Un trou dans $S2$ est inséré chaque fois que l'on doit monter dans la matrice et dans $S1$ si on doit aller vers la gauche. Un parcours en diagonal (quand plusieurs directions sont possibles) traduit une correspondance ou un remplacement.

Il est, en théorie, possible de généraliser l'approche pour permettre l'alignement de plus de deux séquences, en considérant un hypercube à la place d'une matrice. L'extrême complexité algorithmique rend néanmoins cette généralisation inutilisable. Des heuristiques sont alors utilisées. On distingue deux types d'approches : les méthodes progressives et les méthodes itératives [Edgar 2004]. Les premières consistent en l'alignement de paires de séquences de proche en proche. Néanmoins, les choix localement optimaux ne garantissent plus l'obtention de l'alignement global optimal (algorithme glouton) [Sze 2006] [Notredame 2000]. Les secondes procèdent de façon similaire, mais réalignent à chaque étape les séquences initiales de façon à éviter au maximum la propagation d'erreurs. Elles fournissent en général de meilleurs résultats mais sont plus coûteuses en temps. C'est la raison pour laquelle les méthodes progressives sont les plus populaires, en particulier l'implémentation ClustalW [Thompson 1994] dont nous nous sommes inspirés. Tout d'abord, un alignement de toutes les paires d'instances est effectué et la distance entre chacune d'entre elles est stockée dans une matrice de distances. Nous construisons ensuite un arbre suivant une méthode de *neighbour joining* [Saitou 1987]. Cet arbre indique l'ordre dans lequel les séquences devront être comparées.

Sans surprise, l'alignement que nous obtenons est bruité, surtout dans les citations dont les instances sont assez disparates. Il nous a été alors nécessaire de corriger manuellement les erreurs. L'opération a néanmoins été beaucoup moins chronophage qu'un alignement totalement manuel.

Nous décrivons nos résultats, ainsi que le jeu de données utilisé, dans la partie suivante.

3.2.5.2 Résultats et discussions

Présentation du jeu de données Notre corpus a été constitué au moment où l'ancien directeur du Fonds monétaire international était accusé d'agression sexuelle, en 2011 à New York. Les documents, qui sont des articles de presse rédigés en anglais, ont été collectés à l'aide d'AMI-EI pour la requête "dsk OR strauss". Au total, la taille du corpus s'élève à 27 439 articles. Les citations ont ensuite été extraites, nous limitant volontairement aux chaînes de caractères entre guillemets afin de limiter

le bruit (voir partie 3.2.2). Pour chacune d'entre elles, nous stockons son nombre d'occurrences, les jours (ainsi que le nombre d'occurrences pour chaque jour) où elles sont citées, et l'identifiant unique des documents dont elles sont issues : il peut parfois être nécessaire de visualiser la citation dans son contexte d'origine.

Les instances de chaque citation sont ensuite regroupées selon la méthode décrite dans la partie 3.2.3 [58]. A partir des 27 439 documents, 22 099 citations sont détectées, la plupart d'entre elles n'ayant qu'une seule instance. Pour des raisons de charge de travail, nous avons pris le parti dans cette étude de nous concentrer sur les 100 meilleures citations, sachant que nous définissons le poids d'une citation comme la somme des occurrences de l'ensemble de ses instances. Au total, ces 100 citations représentent 1039 instances différentes, et 13 958 occurrences.

Notons enfin que 16 % de ces citations sont en réalité des traductions, principalement du français. C'est un cas très intéressant. Bien que les propos aient été tenus dans une autre langue, ils sont reproduits au discours direct, probablement pour des raisons de distanciation journalistique (expression d'une certaine neutralité).

Nous proposons ici une première piste de réflexion. Notons que d'autres jeux de données devront être utilisés par la suite afin de valider ou non les observations que nous décrivons dans les parties qui suivent. Par ailleurs, nous sommes conscients de la possible influence du genre du corpus sur les résultats : il serait intéressant de procéder, en complément, à l'analyse d'articles de blogs, ainsi que de tweets.

Analyse diachronique Avant de décrire les résultats de notre analyse manuelle des variations et suppressions, nous procédons à quelques analyses quantitatives automatiques, en utilisant les métadonnées que nous stockons pour chaque instance (nombre d'occurrences, etc.). L'objectif est de vérifier s'il existe une corrélation entre le nombre de mots et le degré de reprise. La figure 3.9 montre la relation entre le nombre de mots d'une instance et son nombre d'occurrences. Elle présente aussi sa durée de vie (nombre de jours différents où elle apparaît). Les versions les plus courtes ont une meilleure espérance de vie. Par ailleurs, il est important de signaler que la plupart des instances apparaîssent le même jour. Plutôt qu'une version originale suivie de n reprises, on a n instances qui naissent en parallèle à divers endroits du Web [59]. Ces résultats viennent confirmer les constats préalablement effectués dans [Leskovec 2009].

Variations selon la taille Les instances étant alignées, il est possible de décompter le nombre de suppressions (*gaps*) et variations (remplacements). La figure 3.10 montre la relation entre la taille d'une citation (nombre de mots de son instance

58. Signalons que l'opération de racinisation a lieu lors de la constitution du graphe. Néanmoins, afin d'éviter toute neutralisation artefactuelle des différences possibles entre instances, nous conservons et utilisons les formes originales lors de l'analyse manuelle.

59. La transmission ne se fait donc pas nécessairement de façon linéaire et descendante. Voir partie 1.2.2.

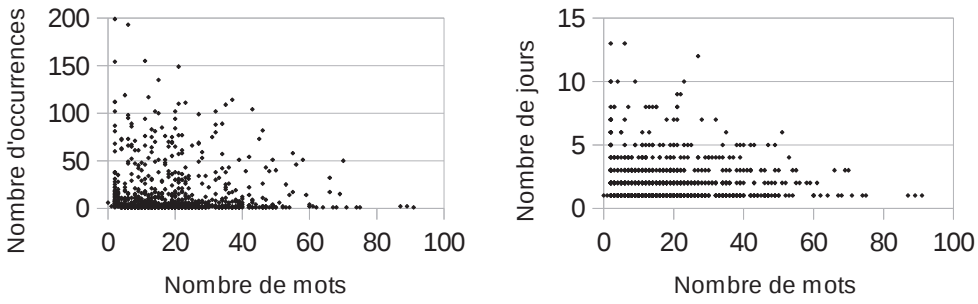

Figure 3.9 – Corrélation entre le nombre de mots d'une instance et (à gauche) son nombre d'occurrences, (à droite) le nombre de jours différents pendant lesquels elle est mentionnée. Les versions les plus courtes sont utilisées plus souvent et plus longtemps.

la plus longue) et son degré d'altération. Les citations les plus longues ont tendance à être plus altérées. C'est assez étonnant étant donnée que l'on s'attendrait à ce que les longues citations soient plutôt copiées-collées et non citées de mémoire [Simmons 2011].

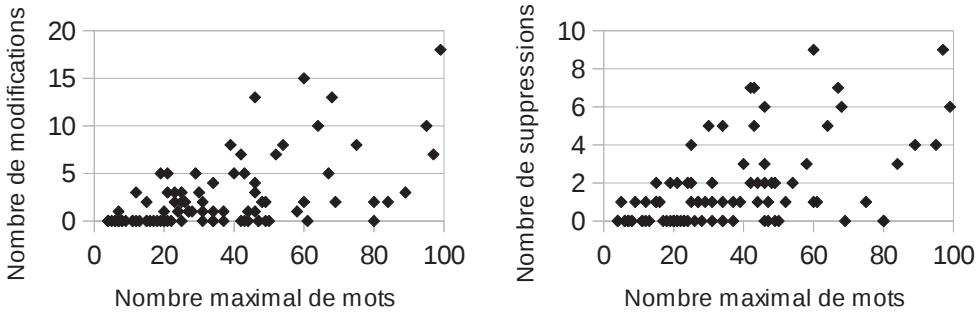

Figure 3.10 – Corrélation entre le nombre de mots de la plus longue instance d'une citation, et son degré d'altération : nombre de variations à gauche, et nombre de suppressions à droite.

Convergence Nous avons observé une certaine tendance, pour les citations les plus populaires, à "converger" vers des syntagmes de 2-3 mots (voir tableau 3.9).

Ces syntagmes sont tout ce qui reste de la citation (3-4 mois après la première occurrence) et sont, semble-t-il, assez puissants pour recontextualiser l'ensemble : la situation d'énonciation d'origine est alors évidente et ces courtes instances se

Durée d'utilisation	Taux de convergence	Distribution totale
Inférieure à un mois	4,11 %	64,39 %
Supérieure à un mois	69,23 %	35,61 %

Tableau 3.9 – Taux de "convergence" vers des syntagmes de 2-3 mots observée sur les citations les plus populaires. Près de 70 % des citations ayant une durée de vie de plus d'un mois sont concernées –seulement 4 % pour les autres, qui représentent pourtant les deux tiers des citations détectées.

suffisent. En d'autres termes, on peut considérer que la citation a un certain bagage sémantique susceptible de se figer en un "concept". Des exemples sont montrés ci-dessous :

— *He said he was leaving his IMF post with* "infinite sadness" *so that he could devote full time to proving his innocence.*
— *That night he is in the custody of the New York Police Department facing the humiliating* "perp walk".
— *How then, did she go from* "compelling and unwavering" *to having her case dismissed due to lack of credibility ?*

Analyse des variations Dans notre jeu de données, 215 variations ont été observées. Le tableau 3.10 en présente les différents types. Nous allons nous concentrer sur les trois plus importants types de variations : la synonymie, la co-référence, et la reformulation.

La *variation synonymique* est le type de variations le plus fréquent. Le phénomène touche principalement les verbes et les substantifs [60]. Il est intéressant de noter que près de la moitié des variations synonymiques observées proviennent de citations traduites [61]. Par ailleurs, le nombre de variantes concurrentes est beaucoup plus important pour les citations traduites, allant jusqu'à 6 :

— *It was a moral failing/failure/weakness/error/mistake/fault.*
— *He tried to open/undo/remove my jeans.*
— *It's important for a politician/man in politics/political man to be able to seduce.*

La *co-référence* est un phénomène linguistique très étudié : plusieurs signifiants peuvent alors correspondre au même signifié (référent). La plupart du temps, les mentions d'un élément précédemment introduit sont plus simples (pronoms). Par exemple :

60. Cela va à l'encontre de [Omodei 2012], qui constatait un degré assez similaire de stabilité indépendamment de la catégorie morphosyntaxique des mots (hormis les interjections et les noms propres).

61. Autrement, ces dernières ont des comportements assez similaires (en terme de types de variations et de suppressions) aux autres.

Type de variation	Sous-type	Exemple
Synonymes (28.37 %)	Verbes (40.98 %)	*happen/occur*
	Substantifs (37.70 %)	*relationship/liaison*
	Adverbes (18.03 %)	*gravely/seriously*
	Adjectifs (3.28 %)	*unjust/unfair*
Co-référence (16.74 %)	Pers. : pronom (41.67 %)	*dsk/he*
	Pers. : reformulation (25 %)	*woman/victim*
	Pers. : abbr. (16.67 %)	*district attorney/DA*
	Action (8.33 %)	*what happened/it*
	Chose (8.33 %)	*this incident/it*
Reformulation (12.09 %)	Paraphrase (57.69 %)	*has no idea/doesn't know*
	Var. syntaxique (42.31 %)	*a man with/this man has*
Orthographe (11.16 %)	Faute (37.5 %)	*whatsoever/what so ever*
	UK/US (37.5 %)	*honour/honor*
	Faute de frappe (25 %)	*candidate/cadidate*
Determinants (8.84 %)	Art. def./dem. (57.89 %)	*the/this*
	Art def./poss. (15.79 %)	*the/my*
	Art indef./dem. (10.53 %)	*a/this*
	Art indef./quant. (5.26 %)	*a/one*
	Art def./art. indef. (5.26 %)	*the/a*
	Dem./dem. (5.26 %)	*this/that*
Conjugaison (8.37 %)	Temps (83.33 %)	*have been/were*
	Personne (11.11%)	*are/is*
	Mode (5.55 %)	*put/puts*
Liaisons (6.51 %)	Prep./prep. (71.42 %)	*on/in*
	Subord./subord. (21.43 %)	*although/even if*
	Coord./coord. (7.14 %)	*or/and*
Contractions (4.19 %)	/	*was not/wasn't*
Nombres (1.86 %)	/	*skill/skills*
Inversions (1.86 %)	/	*still is/is still*

Tableau 3.10 – Catégorisation manuelle des variations observées, triées par nombre d'occurrences. La plupart des catégories peut etre subdivisée. Un exemple est donné à chaque fois.

— *I'm rather proud of my husband reputation as a seducer.*

Dans d'autres instances, la réputation de séducteur est introduite au discours indirect :

— *Anne Sinclair seemed forgiving of his reputed behavior. "No, I'm rather proud of it !" she told.*

Parfois, la co-référence a lieu pour des raisons stylistiques, de façon à éviter les répétitions d'un mot, par exemple "man" à la place de "defendant". Volontairement

ou nom, certaines nuances peuvent alors être introduites : “victim” n’a pas les mêmes implications que “complainant”.

Nous distinguons deux types de *reformulations* : les simples variations syntaxiques, et la paraphrase. La première inclut de nombreux phénomènes différents, par exemple la transition du discours direct au discours indirect ou une modification dans l’ordre des constituants :

— *I just want to know if I need a lawyer* vs. *do I need a lawyer ?*
— *We expect him to be released tomorrow* vs. *we expect he will be released tomorrow.*
— *Of which I am not proud* vs. *and I’m not proud of it.*

La paraphrase va plus loin que la simple modification syntaxique :

— *I have no doubt* vs. *I am certain.*
— *She has no idea what* vs. *she doesn’t know what (...)*
— *There were many reasons to believe* vs. *we continue to believe (...)*
— *It was not just* vs. *it was more than (...)*

Analyse des suppressions 129 suppressions ont été observées, présentées dans le tableau 3.11. Nous avons pris le parti d’ignorer les suppressions au début et à la fin des instances car l’opération de recadrage est très courante dans le monde journalistique –discours indirect libre mentionné dans la partie 3.2.2. Nous nous concentrons ici sur le phénomène le plus récurrent, à savoir la suppression des mots exprimant la modalité.

La *modalité* est l’expression de la subjectivité par un locuteur. Il n’y a pas de réel consensus autour de la catégorisation de la modalité, mais la plupart des travaux semblent s’accorder sur les deux types suivants : l’expression de la nécessité, et de la possibilité. Nous nous inspirons, dans notre catégorisation de la modalité, de travaux comme [Frawley 2006] ou [Portner 2009].

La modalité *déontique* désigne la permission et l’obligation morale. La modalité *aléthique* concerne la possibilité (ou impossibilité) ainsi que la nécessité logique. Enfin, la modalité *épistémique* indique le jugement du locuteur. Les modalités aléthiques et épistémiques sont souvent regroupées car il peut être étrange d’opposer ce qui est logiquement vrai, et ce que le locuteur juge vrai [Palmer 1986].

Nous constatons que les mots servant à exprimer la modalité ont une forte tendance à disparaître :

— ”Forensic evidence (we believe) will not be consistent with a forcible account.“
— ”He is (obviously) not in a position to run the IMF.“
— ”(I think) it was a moral failing.“

Type de suppression	Sous-type	Exemple
Modalité (20.15 %)	Épistemique (80.77 %)	*(I think) it is*
	Aléthique (11.54 %)	*(may) have*
	Déontique (3.85 %)	*(have to) face*
	Affective (3.85 %)	*I'm (sorry I'm) not*
Modifieurs (18.60 %)	Adjectifs (50 %)	*(physical) evidence*
	Synt. adj (41.67 %)	*influence (througout the world)*
	Compl. nom. (8.33 %)	*(selection) process*
Liaisons (15.50 %)	Coord. (90 %)	*my children (and) my friends*
	Adv. conj. (10 %)	*(indeed), we were intent on*
Déterminants (10.85 %)	Art. def (50%)	*to (the) prosecutors*
	Possessifs (28.57 %)	*our guest and (our) staff*
	Art. indef. (21.43 %)	*with (a) complete conviction*
Complétives (10.07 %)	Verbes (53.85 %)	*I felt (that) I*
	Adjectifs (30.77 %)	*important (that) the*
	Noms (15.38 %)	*the idea (that) she*
Enumérations (7.75 %)	/	*my strength (and all my energy)*
Temps (6.98 %)	/	*I feel compelled (today) to*
Répétitions (5.43 %)	Même référent (71.43 %)	*this man (Mr. Strauss-Kahn)*
	Même mot (28.57 %)	*a very (very) defensible case*
Intensité (4.65 %)	/	*changed (a single) thing*

Tableau 3.11 – Liste des suppressions observées, triées par nombre d'occurrences.

Chapitre 4

Industrialisation : intégration de nos travaux

Sommaire

Cette thèse a été réalisée en entreprise. Le caractère applicatif de nos travaux est donc une évidence. Après une nouvelle présentation du produit principal de AMI Software, nous allons montrer dans ce chapitre la place de nos travaux au sein de ce dernier. Ce sera aussi l'occasion de revenir sur des détails d'implémentation (algorithmes, performances, etc.) de nos méthodes qui auraient été omis, pour des raisons de lisibilité, dans le reste du mémoire.

Enfin, une analyse rétrospective de quelques buzz ayant eu lieu pendant l'élaboration de cette thèse sera effectuée afin de démontrer à nouveau l'intérêt d'une analyse conjointe de la topologie du Web et du contenu textuel pour identifier un buzz, qualifier les émetteurs, et comprendre sa propagation au sein des réseaux.

4.1 Présentation de la plateforme AMI-EI

AMI-EI (*AMI Enterprise Intelligence*) est le produit principal de AMI Software[1]. Il s'agit d'une solution logicielle destinée à satisfaire l'ensemble du cycle de veille des entreprises dans des contextes divers tels que l'intelligence économique, la veille technologique, ou l'analyse comportementale et l'e-réputation.

Une plateforme AMI-EI consiste en une suite de modules indépendants, complémentaires et communicants. Ils permettent de constituer une chaîne fonctionnelle cohérente et de couvrir l'ensemble des besoins d'un processus de veille : l'acquisition de l'information, la capitalisation, l'analyse, et enfin le partage et la diffusion. Nous avons déjà décrit dans la partie 1.1.4 ces différentes étapes, que nous rappelons dans la figure 4.1. Nous allons ici aborder brièvement la façon dont ces dernières s'imbriquent concrètement au sein du produit.

Figure 4.1 – Présentation des différentes étapes d'un processus de veille. Une fois les documents recherchés et collectés, ces derniers peuvent être triés, filtrés, modifiés, puis analysés. Les résultats de la veille peuvent alors être diffusés.

Le partage et la diffusion des informations acquises se fait principalement à travers deux points d'accès principaux : les modules *Partager* et *Mon espace* (voir la figure 4.2 pour une capture d'écran de ce dernier). Le premier est un portail de consultation, permettant la recherche et le partage des informations organisées par thématique, avec une gestion des droits d'accès à partir de profils prédéfinis. Le portail, qui dispose d'un moteur de recherche intégré, permet une navigation par dossiers ou par métadonnées. Il permet aussi aux utilisateurs de s'abonner à des flux RSS afin d'accéder aux contenus de la plateforme à partir d'applications tierces.

Le module *Mon espace* peut être utilisé comme point d'entrée à toute l'application. Il permet de personnaliser, pour chaque utilisateur, son accès à la plateforme AMI-EI. Ainsi, un veilleur peut suivre son activité de veille, les documents collectés,

1. Une description plus exhaustive de AMI-EI est disponible à l'adresse suivante : `http://www.amisw.com/fr/uploads/SPD-AMI-EI7.0-1.0.pdf`.

les documents en attente de validation chez des collègues, les documents dernièrement publiés, l'état des indexations sur les sources, les volumes d'information produits par chaque source, les statistiques d'utilisation, etc. En pratique, il permet à l'utilisateur de construire un tableau de bord personnalisé à l'aide de *widgets interactifs*, qui sont des composants graphiques légers entièrement paramétrables.

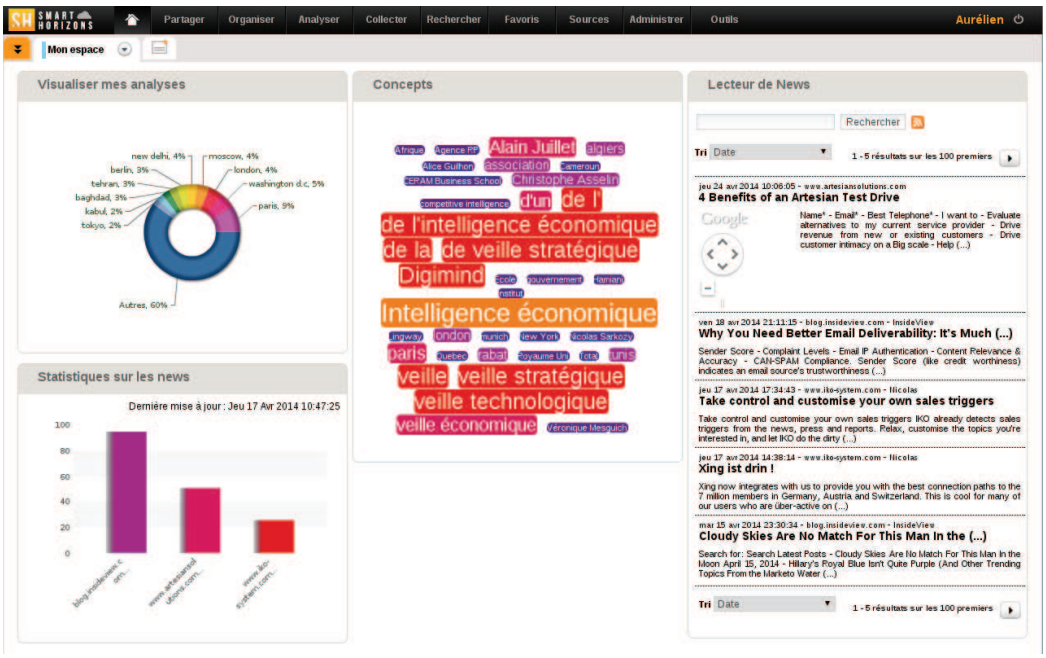

Figure 4.2 – Module *Mon espace* de AMI-EI. Quatre widgets sont visibles : *Visualiser mes analyses*, *Statistiques sur les news*, *Concepts*, et *Lecteur de news*.

4.2 Intégration de nos travaux dans le produit

L'ensemble des outils que nous avons développé dans le cadre de cette thèse prend la forme de widgets spécifiques : *Topologie des sources*, *Mon Twitter influence*, et enfin *Suivi de buzz*, que nous décrivons respectivement dans les parties 4.2.3, 4.2.1, et 4.2.2.

Précisons que les jeux de données des parties suivantes sont utilisés uniquement à titre indicatif. Ils ne seront donc pas particulièrement commentés. Une analyse détaillée sera effectuée dans la partie 4.3.

4.2.1 Widget : *Mon Twitter influence*

Ce widget concrétise nos recherches abordées dans la partie 2, en particulier le point 2.1.4. Il s'agit ici de procéder à des calculs d'autorité à partir d'un ensemble de tweets possédant certains mots-clés et rédigés pendant une période de temps choisie.

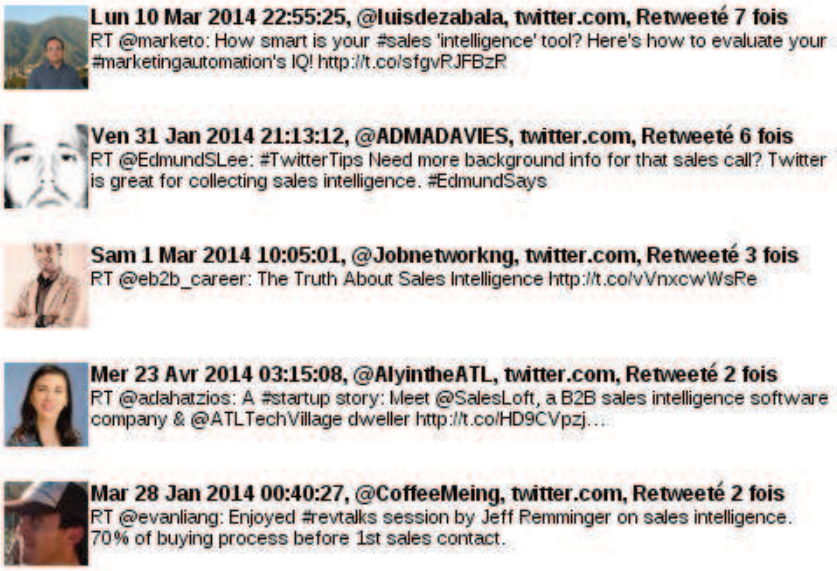

Figure 4.3 – Illustration de la liste des tweets les plus diffusés : classement par nombre de retweets uniques.

Les résultats prennent plusieurs formes. Tout d'abord, une liste des n tweets les plus retweetés est proposée, comme illustrée dans la figure 4.3. La figure 4.4 illustre quant à elle un classement des utilisateurs en fonction du nombre de tweets postés et du nombre de fois qu'ils ont été retweetés, un classement des hashtags les plus utilisés, ainsi que la volumétrie des tweets dans le temps.

Afin de pouvoir analyser ces données chiffrées dans leur contexte d'origine, le graphe des relations entre utilisateurs est aussi proposé (voir figures 4.5 et 4.6 [2]).

Les noeuds représentent les utilisateurs tandis que les arêtes traduisent les relations de retweet entre ces derniers. L'orientation de la relation est symbolisée par un code de couleurs (non montré dans l'image) lorsqu'un noeud est survolé par le curseur : les liens entrants en verts, les liens sortants en orange. Le survol de la souris permet aussi l'affichage du profil de l'utilisateur et la liste des tweets postés par ce dernier (voir figure 4.7).

La taille des noeuds indique, au choix, le nombre de tweets postés par l'utilisateur ou le nombre de fois où il a été retweeté. Leur couleur est fonction des hashtags employés par l'utilisateur, créant ainsi un premier lien entre structure topologique et contenu thématique.

L'analyse de la propagation de l'information dans le temps est possible à l'aide d'un curseur temporel situé en bas du graphe. Ce dernier permet d'afficher ou cacher les noeuds en fonction de leur date d'apparition ; seul le premier utilisateur à avoir posté un tweet dans le contexte choisi est visible lorsque le curseur est à

2. L'affichage de ces graphes est effectué à l'aide de la librairie JavaScript D3 : `http://d3js.org/`.

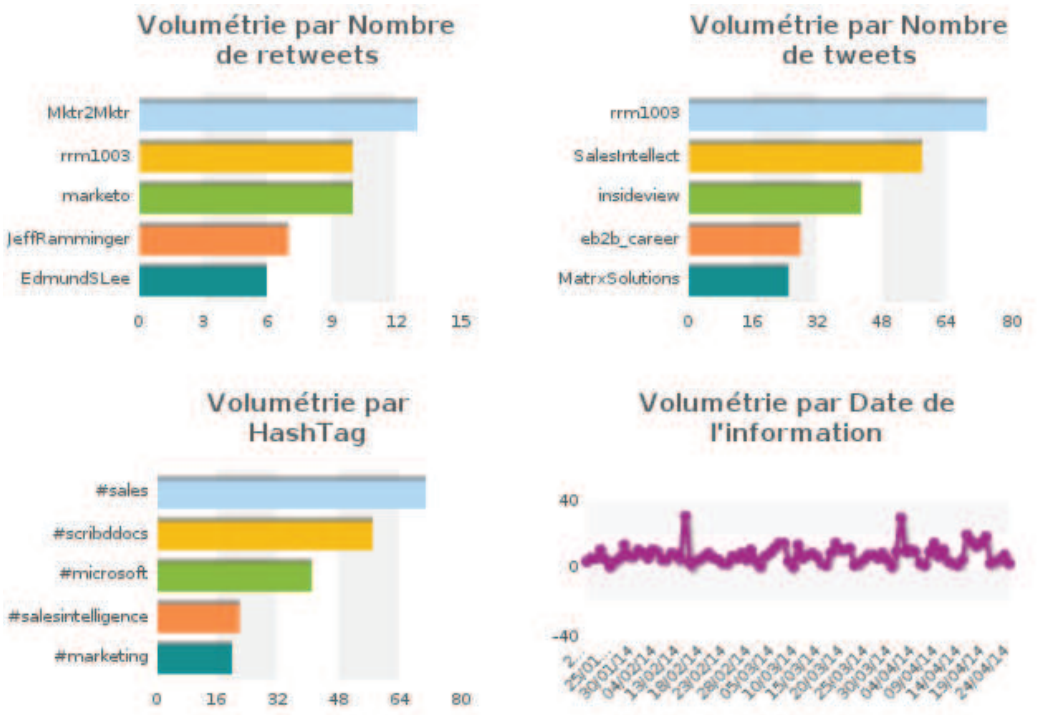

Figure 4.4 – Illustration de calculs volumétriques effectués à partir des tweets analysés : classement des hashtags les plus employés, des utilisateurs les plus prolifiques et les plus influents, et volumétrie des tweets dans le temps.

l'extrémité gauche. Les autres noeuds apparaîssent au fur et à mesure qu'il est déplacé vers la droite. Un exemple est proposé dans la figure 4.8.

L'analyse conjointe des données volumétriques et du graphe des utilisateurs permet, à l'instar de ce que nous disions dans la partie 2.2.3.3, de relativiser les calculs automatiques. Prenons pour exemple la figure 4.9 qui montre un certain Flogazan comme étant l'utilisateur faisant le plus autorité. Un rapide coup d'oeil sur le graphe des utilisateurs permet aisément de nuancer ce constat. Son profil montre qu'il s'agit d'un journaliste radio. S'agissant d'une personne célèbre, son nombre d'amis sur Twitter est conséquent, augmentant ainsi sa probabilité d'être retweeté. En pratique, lui et ses amis ne sont pas reliés au reste du graphe. L'interprétation suggère donc ici qu'il ne s'agit pas d'un réel influenceur, simplement d'une personne plus largement écoutée que les autres.

Gardons néanmoins à l'esprit qu'une visualisation de données n'est jamais complètement neutre. La visualisation peut être contrainte par certaines difficultés techniques et ergonomiques, influant ainsi potentiellement sur l'interprétation que l'on fait des résultats. Dans notre cas, seuls les 600 noeuds les plus importants (nombre de retweets) sont affichés pour des raisons de performance. Ce filtre peut cacher artificiellement certaines relations. Prenons pour exemple le cas où plusieurs noeuds

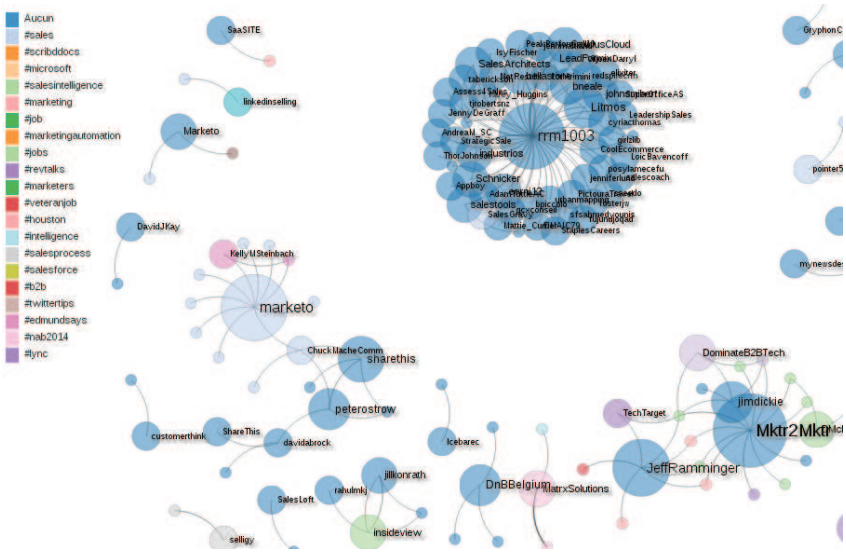

FIGURE 4.5 – Illustration du graphe des relations entre utilisateurs. Notons que la spatialisation des noeuds dépend uniquement des algorithmes de force utilisés et n'est donc pas réellement pertinente : les noeuds non reliés se repoussent mutuellement.

FIGURE 4.6 – La structure communautaire des graphes d'utilisateurs est parfois très fortement marquée.

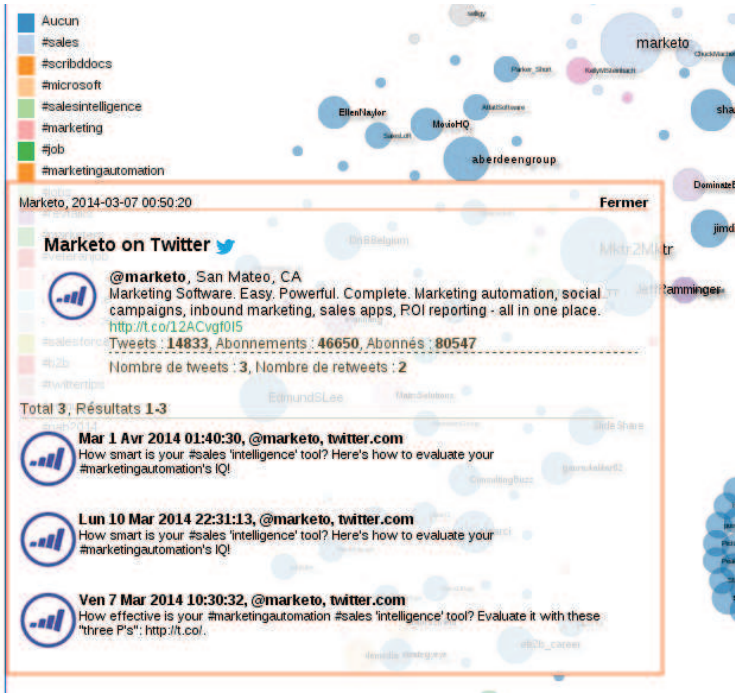

Figure 4.7 – Informations affichées lorsqu'un noeud est survolé par le curseur de la souris : profil utilisateur et liste des tweets postés.

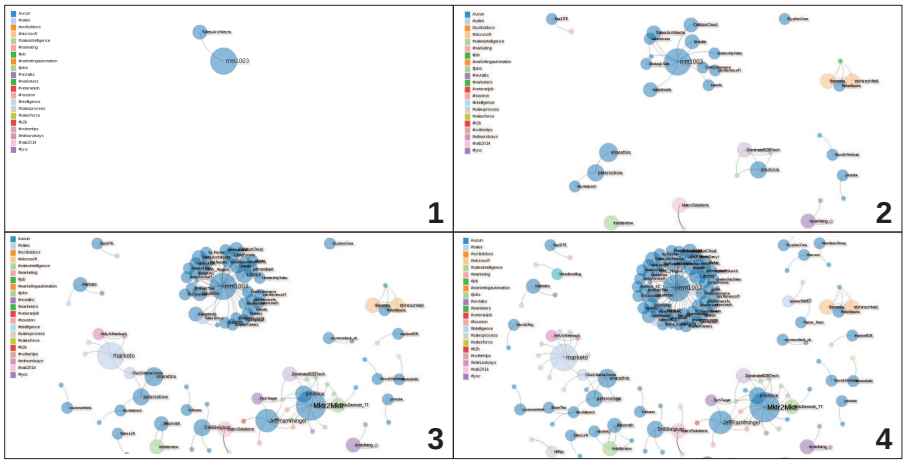

Figure 4.8 – Illustration de l'apparition au fil du temps des utilisateurs. La date associée à chaque noeud correspond à la date du premier tweet posté par ce dernier. L'état du graphe suggère ici une diffusion non virale (de noeud en noeud) de l'information.

importants sont reliés entre eux par un ensemble de noeuds faibles.

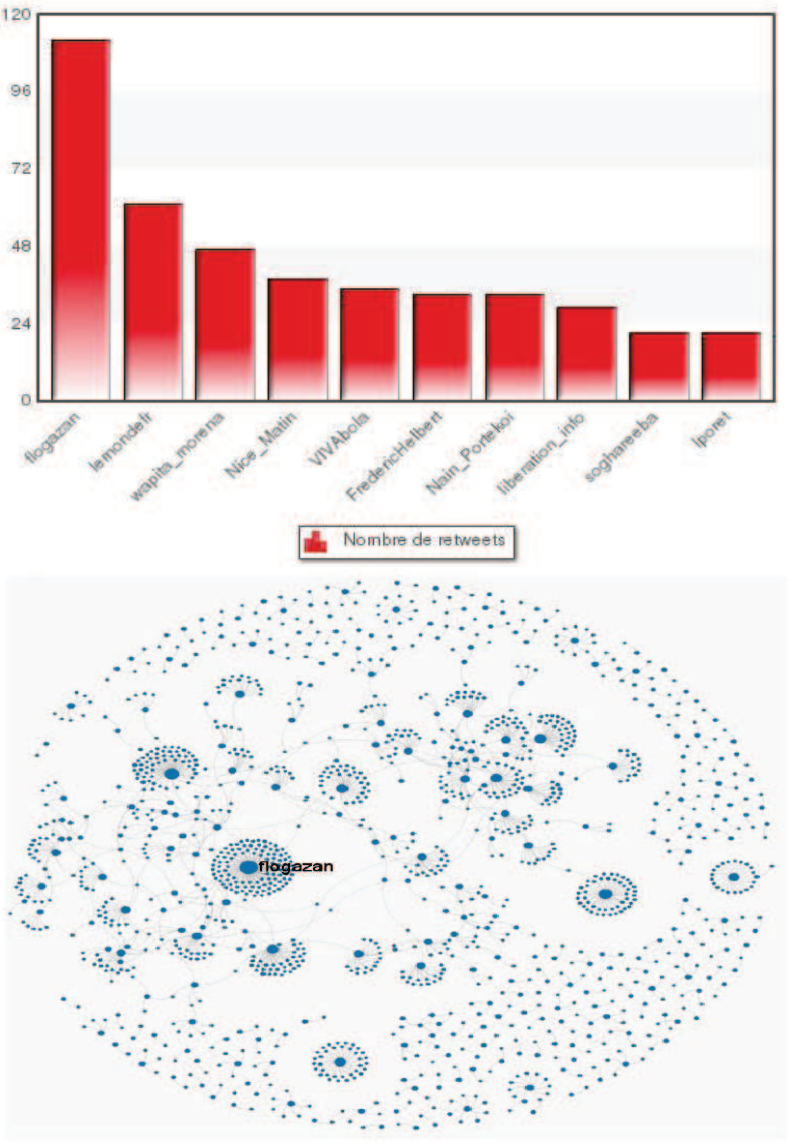

Figure 4.9 – Visualisation conjointe des données volumétriques (en haut) et des relations topologiques (en bas), permettant une meilleure interprétation.

4.2.2 Widget : *Suivi de buzz*

Ce widget illustre nos travaux de la partie 3.2. Il a pour objectif l'identification de buzz et l'analyse de leur propagation par une analyse textuelle. Il permet l'extraction de groupes de citations proches, chacun d'eux traduisant un buzz. L'extraction peut se faire sur des documents de presse ou issus de blogs. Il est aussi possible de sortir du cadre de la simple citation ; sur des données issues de Twitter, il est en effet

possible de choisir pour unité de base l'ensemble du tweet à la place de la citation. L'implémentation est fidèle à la description faite en 3.2.3.

Ici aussi, les résultats sont de formats multiples. Tout d'abord, la liste chronologique des buzz est proposée à l'utilisateur sous la forme d'un nuage de citations (voir figure 4.10). Chaque bloc correspond à un buzz. Ces derniers sont disposés en ordre chronologique d'apparition. La taille traduit leur ampleur (taux de reprise de la citation, toutes instances confondues [3]). La couleur utilisée est ici simplement cosmétique.

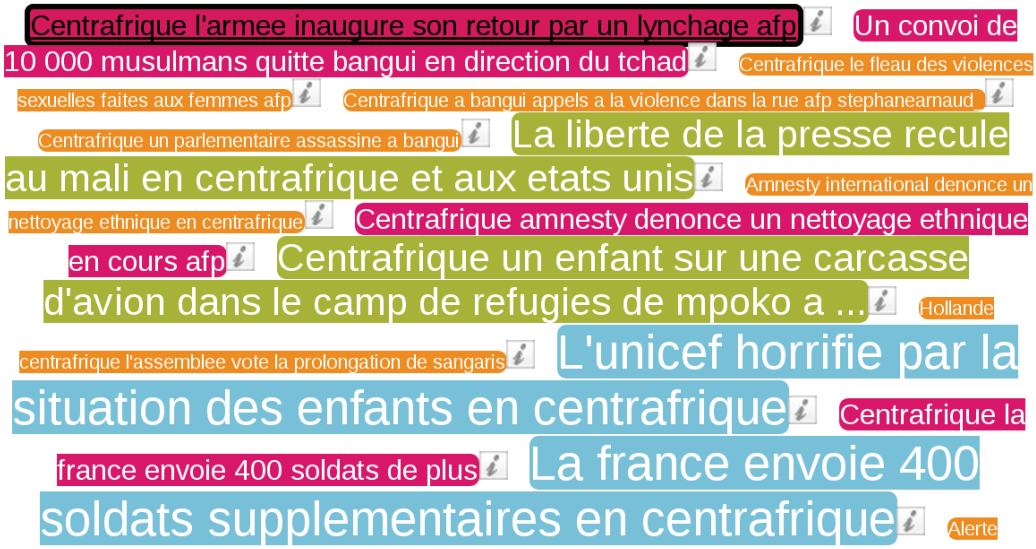

Figure 4.10 – Extrait de la liste chronologique des buzz détectés par le widget pour un sujet donné.

Chaque buzz est sélectionnable, et sa volumétrie peut être affichée (voir figure 4.11). A l'aide des algorithmes d'alignement de séquences décrits en 3.2.5.1, les éventuelles déformations que subissent les citations dans le temps peuvent être visualisées pour chacun des buzz. Rappelons néanmoins que ce type d'algorithme ne garantit pas l'obtention d'un alignement optimal. En pratique, l'alignement proposé est encore assez fortement bruité. La figure 4.12 illustre un exemple d'alignement obtenu pour un buzz donné.

Afin de découvrir les principaux acteurs de ces buzz et faire la passerelle entre analyse du contenu et qualification des émetteurs, un autre graphique est proposé à l'utilisateur (voir figure 4.13). Pour le moment, ce dernier est uniquement disponible pour les sources de type Twitter mais il est en théorie applicable aux autres sources Web (sites d'actualité, blogs, etc.). Chaque point dans l'espace correspond à un émetteur : la taille correspond à l'autorité et la couleur au buzz auquel il est rattaché. L'axe des abscisses est l'axe temporel [4] tandis que les ordonnées traduisent le nombre

3. Rappelons que nous appelons *instances* les variantes textuelles de chaque citation. Une *citation* désigne l'ensemble des instances.

4. L'unité de temps en abscisses est calculée automatiquement selon le corpus analysé. L'uti-

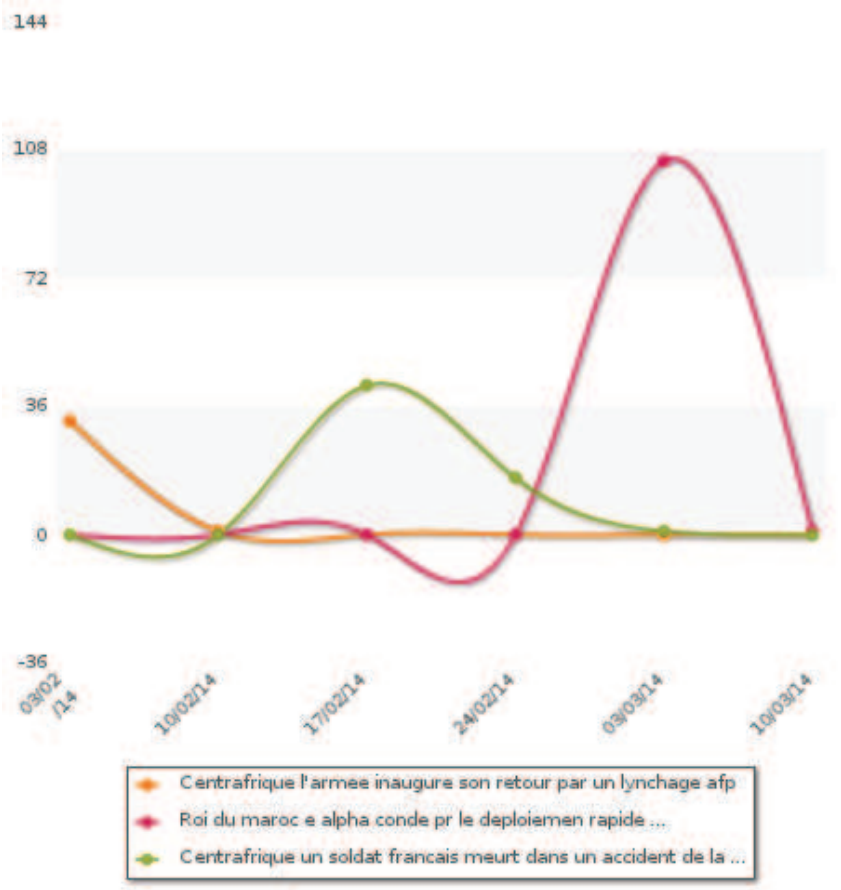

FIGURE 4.11 – **Volumétrie comparative de trois buzz sélectionnés par l'utilisateur.**

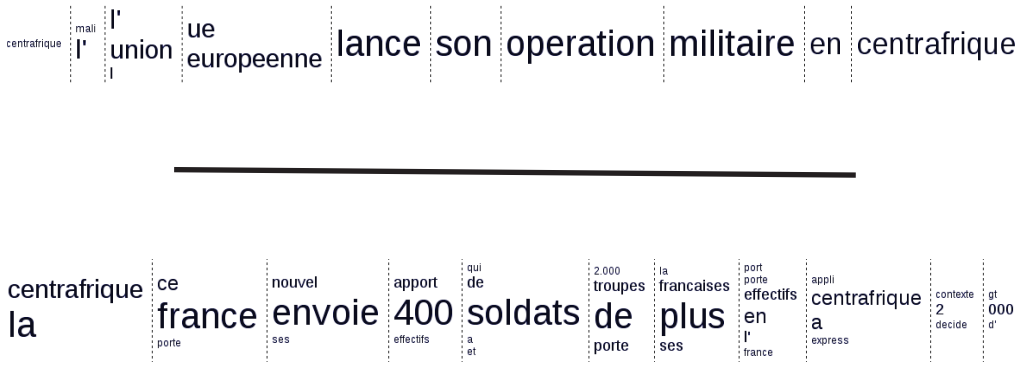

FIGURE 4.12 – **Alignement de deux buzz, correct en haut et bruité en bas. Chaque colonne correspond à un mot graphique, ou plusieurs lorsque différentes alternatives sont possibles. La taille de chacun des mots correspond au nombre d'instances dans lesquelles ce dernier apparaît.**

lisateur peut néanmoins choisir ce qui lui convient le mieux : secondes, minutes, heures, jours, semaines, mois, années.

de messages postés par l'émetteur. Les points les plus en haut sont les plus prolixes. L'intérêt de ce graphique est de permettre à l'utilisateur d'inférer quelles sont les émetteurs à l'origine d'un buzz donné, ceux qui ont aidé à sa diffusion et à son amplification.

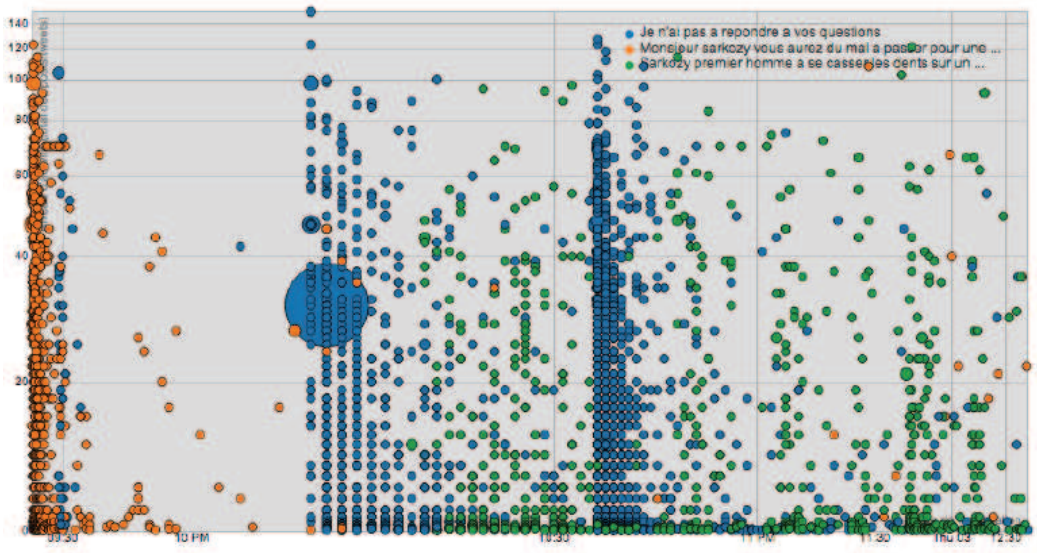

Figure 4.13 – Graphique de propagation des buzz. Dans cet exemple, trois buzz sont sélectionnés par l'utilisateur. En orange, on constate un buzz très ponctuel. En vert, un buzz beaucoup plus diffus. Le buzz bleu connaît plusieurs pics d'intensité ; après un début modeste, ce dernier gagne en importance. La présence à cet instant précis d'un émetteur d'autorité n'est probablement pas anodine : il est possible d'inférer que ce dernier joue ici le rôle de caisse de résonance.

4.2.3 Widget : *Topologie des sources*

Ce dernier widget concerne aussi la partie 2, mais s'appuie en particulier sur les points 2.1.3.6, 2.1.6, et 2.2. Il s'agit de notre implémentation de l'algorithme HITS ; il permet donc l'extraction d'autorités et de hubs à partir d'un certain nombre de sources de départ (root set) et d'une requête. Au moment de la rédaction de cette thèse, le widget est encore à l'état de prototype et le découpage en communautés n'est pas encore implémenté. Le calcul d'autorités et de hubs se fait donc sur la totalité du graphe.

Le classement des 10 meilleurs autorités et hubs est présenté à l'utilisateur, comme illustré respectivement dans les figures 4.14 et 4.15. Les exemples présentés ici ont été calculés à partir de la requête "market intelligence", et avec pour root set quelques pages renvoyées par le connecteur AMI-EI au moteur de recherche Bing. Rappelons que l'implémentation de ce widget n'est pas terminée et que les résultats ne sont donc ici présentés qu'à titre indicatif.

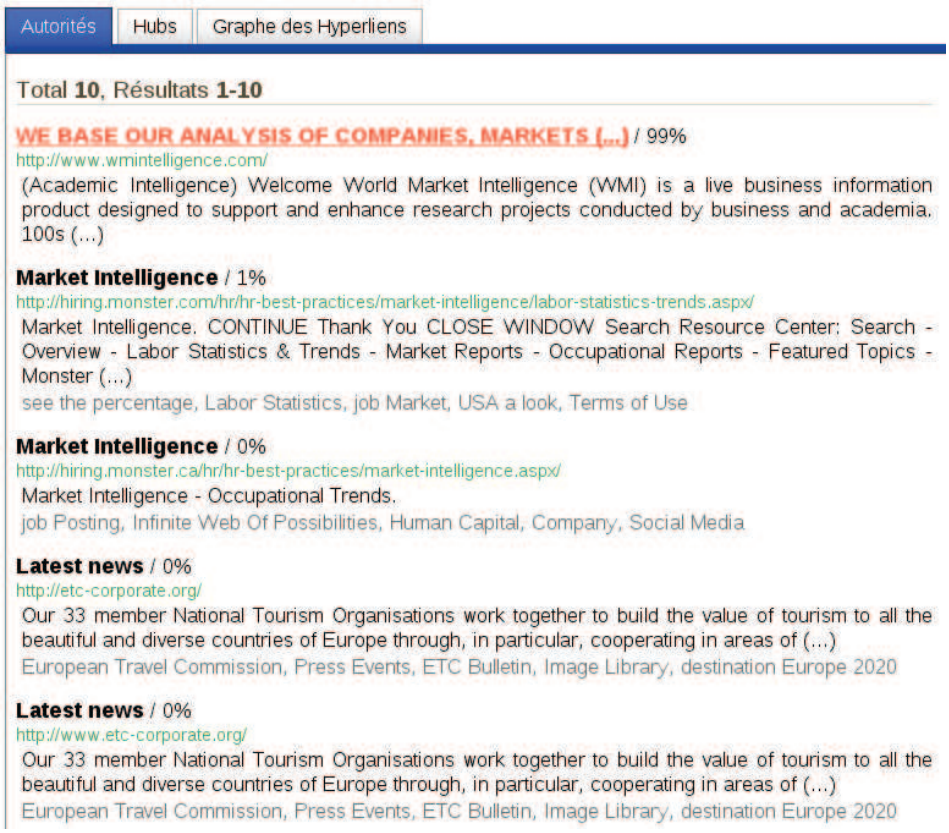

Figure 4.14 – Extrait des 10 meilleures autorités calculées pour la requête "market intelligence".

Comme illustré dans la figure 4.16, le graphe des hyperliens peut être visualisé de façon à mieux interpréter les classements. Nous nous retrouvons ici aussi confrontés aux problématiques de visualisation signalées à la fin de la partie 4.2.1. Pour des raisons de performance, les noeuds dont la valeur d'autorité/hub est inférieure à un seuil donné sont filtrés. Un seuil trop élevé cache ici la majorité des noeuds.

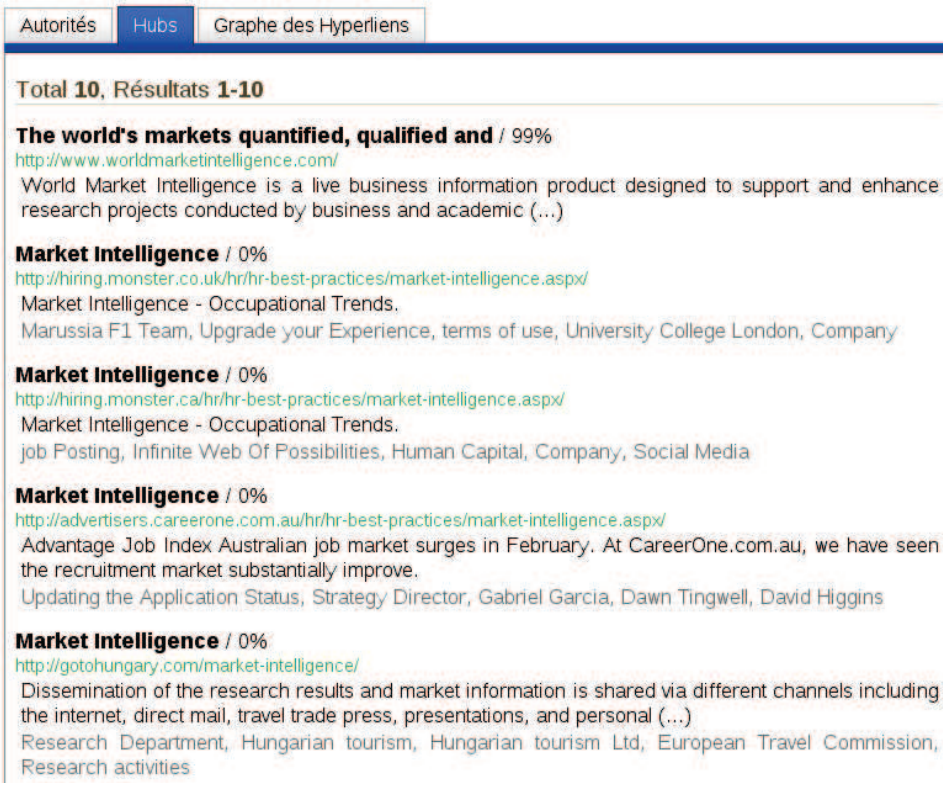

FIGURE 4.15 – Extrait des 10 meilleurs hubs calculés pour la requête “market intelligence”.

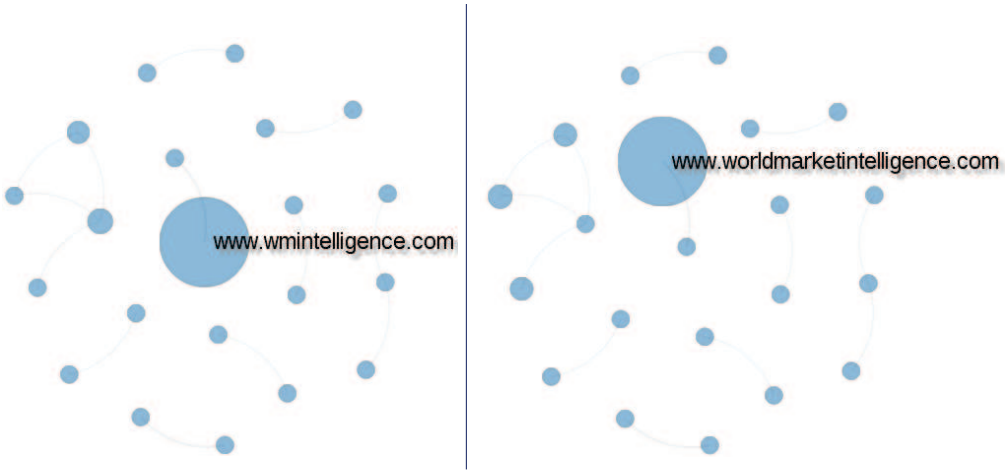

FIGURE 4.16 – Extrait du graphe des hyperliens calculé pour la requête “market intelligence”. La taille des noeuds peut traduire, au choix, l’autorité (à gauche) ou le score de hub (à droite).

4.3 Étude de cas

Pour conclure ce mémoire, nous proposons une analyse rétrospective de quelques buzz ayant eu lieu pendant l'élaboration de cette thèse. Ces cas concrets [5] nous serviront à montrer de nouveau l'intérêt d'une analyse conjointe de la topologie du Web et du contenu textuel afin d'identifier un buzz, qualifier les émetteurs, et comprendre sa propagation au sein des réseaux.

Pour chacun d'entre eux, nous procèderons à une remise en contexte des problématiques abordées et expliciterons les méthodes employées pour collecter les documents. Nous présenterons ensuite les résultats de nos analyses effectuées à l'aide des widgets décrits précédemment.

4.3.1 Arrestation et procès de Dominique Strauss-Kahn

4.3.1.1 Présentation des données

La première étude de cas traite du buzz dont nous avions parlé dans la partie 3.2.5.2 : il concerne les accusations d'aggression sexuelle présumées envers l'ancien directeur du Fonds Monétaire International, en 2011 à New York. L'ensemble du corpus a été constitué en réponse à la requête "dsk OR strauss" sur les principaux moteurs de recherche, sans restriction de langue. Nous procédons néanmoins à un filtre ultérieur car nous nous concentrons ici uniquement sur les sources rédigées en langue française. Pour les articles issus de la presse et des blogs, l'intervalle de date s'étend du 18/05/2011 au 23/09/2011. Autrement dit, les collectes ont été lancées 4 jours après le début de l'affaire et stoppées un peu moins d'une semaine après la prise de parole télévisée de l'accusé suite à son retour en France. Cela correspond à 10 573 documents. Concernant les réseaux sociaux, étant donné le nombre très important de tweets à ce sujet, nous nous sommes limités aux messages rédigés au cours des premiers jours : du 15/05/2011 au 24/05/2011, soient 296 791 tweets. Les données relatives à ces corpus sont résumées dans le tableau 4.1.

Type	Presse et blogs	Tweets
Langue	FR	FR
Période	du 18/05/2011 au 23/09/2011	du 15/05/2011 au 24/05/2011
Taille	10 573 documents	296 791 tweets

Tableau 4.1 – Caractéristiques des corpus sur l'affaire DSK : les articles de presse et de blog à gauche, les tweets à droite.

Rappelons brièvement quelques évènements marquants [6]. Le 14 mai 2011, une femme de chambre du Sofitel de New York dépose une plainte pour agression

5. Il s'agit d'exemples de veilles effectuées au sein de l'entreprise, suite à des demandes précises de clients.

6. Pour une chronologie plus détaillée, lire cet article : http://www.lemonde.fr/dsk/article/2011/07/01/le-film-de-l-affaire-strauss-kahn_1543285_1522571.html.

sexuelle ; l'accusé est le directeur général du Fonds monétaire international (FMI). Ce dernier est interpellé par la police dans la journée. Très tôt, un tweet [7] est rédigé par un militant UMP après avoir appris la nouvelle d'un ami à New York, lançant ainsi les rumeurs d'un complot politique visant un potentiel candidat à la présidentielle française. Le 18 mai, l'accusé démissionne de son poste de directeur du FMI. Libéré sous caution, il est assigné à résidence le lendemain. Le 1^{er} juillet, cette assignation à résidence est annulée et la caution restituée suite à l'arrivée d'éléments entachant la crédibilité de la plaignante. Le 4 juillet, la journaliste française Tristane Banon décide de porter plainte contre Dominique Strauss-Kahn pour tentative de viol en 2003. Le 25 juillet, la plaignante new-yorkaise sort de son silence. Le 21 aout, le procureur chargé de l'affaire demande l'abandon des poursuites contre l'accusé. Début septembre, ce dernier rentre en France et décide de prendre la parole publiquement au journal télévisé de 20h.

4.3.1.2 Analyses

Nous allons désormais appliquer certains des outils que nous avons développés au cours de cette thèse afin de voir ce qu'ils nous permettent de détecter comme buzz.

Graphe de cooccurrences Bien que la méthode n'ait pas été intégrée pour le moment dans AMI-EI, nous avons construit un graphe de cooccurrences suivant la description faite en 3.1.3.2 afin d'avoir une première vision d'ensemble des thématiques abordées. Ce dernier est montré dans la figure 4.17. 4 grandes thématiques se distinguent. A droite, la succession au poste de directeur général au FMI. En haut à gauche, il est question de politique française : les élections, les répercussions de l'affaire au sein du Parti Socialiste. En bas : le procès et l'incarcération de l'accusé. Entre ces deux thématiques, l'affaire Tristane Banon. Nous avons volontairement laissé le graphe non découpé car il permet de mettre en avant les relations de proximité entre les différentes thématiques.

Widget *Suivi de buzz* A l'aide du widget *suivi de buzz*, nous espérons désormais obtenir un aperçu plus fin des différents buzz ayant eu lieu au sein de chacune de ces thématiques (voir figure 4.18). S'agissant d'un évènement très médiatisé et touchant au monde de la politique, l'utilisation de la citation pour détecter les buzz semble ici particulièrement pertinente.

Passées les premières indignations et la réputation de séducteur de l'accusé, tous les regards se tournent vite vers la crédibilité de la plaignante et, dans une moindre mesure, les présidentielles en France. Il est aussi question de la plainte de Tristane Banon. L'accusé est libéré, retourne en France et s'explique publiquement. Si la liste des buzz détectés par le système semble assez complète, il est difficile de s'assurer

7. https://twitter.com/j_pinet/status/69507272040136704.

Figure 4.17 – Graphe de cooccurrences construit à partir des articles de presse et de blogs.

de l'exhaustivité. Par ailleurs, une bonne connaissance de la situation d'énonciation d'origine est importante pour interpréter certains d'entre eux.

Analyse de la topologie : widget *Twitter influence* Parmi les tweets les plus populaires obtenus par cette analyse (voir figure 4.19), certains sont humoristiques comme c'est souvent le cas sur Twitter. D'autres s'étonnent de l'effervescence médiatique autour de ce sujet malgré la crise nucléaire de Fukushima. Comme le montre la figure 4.20, les acteurs les plus importants semblent ici être les professionnels de la presse : comptes officiels de journaux ou envoyés spéciaux suivant l'affaire de près. Étant donné la nature de l'affaire, cela n'est pas étonnant. Seules certains professionnels de l'information peuvent assister au procès et divulguer des informations à ce sujet.

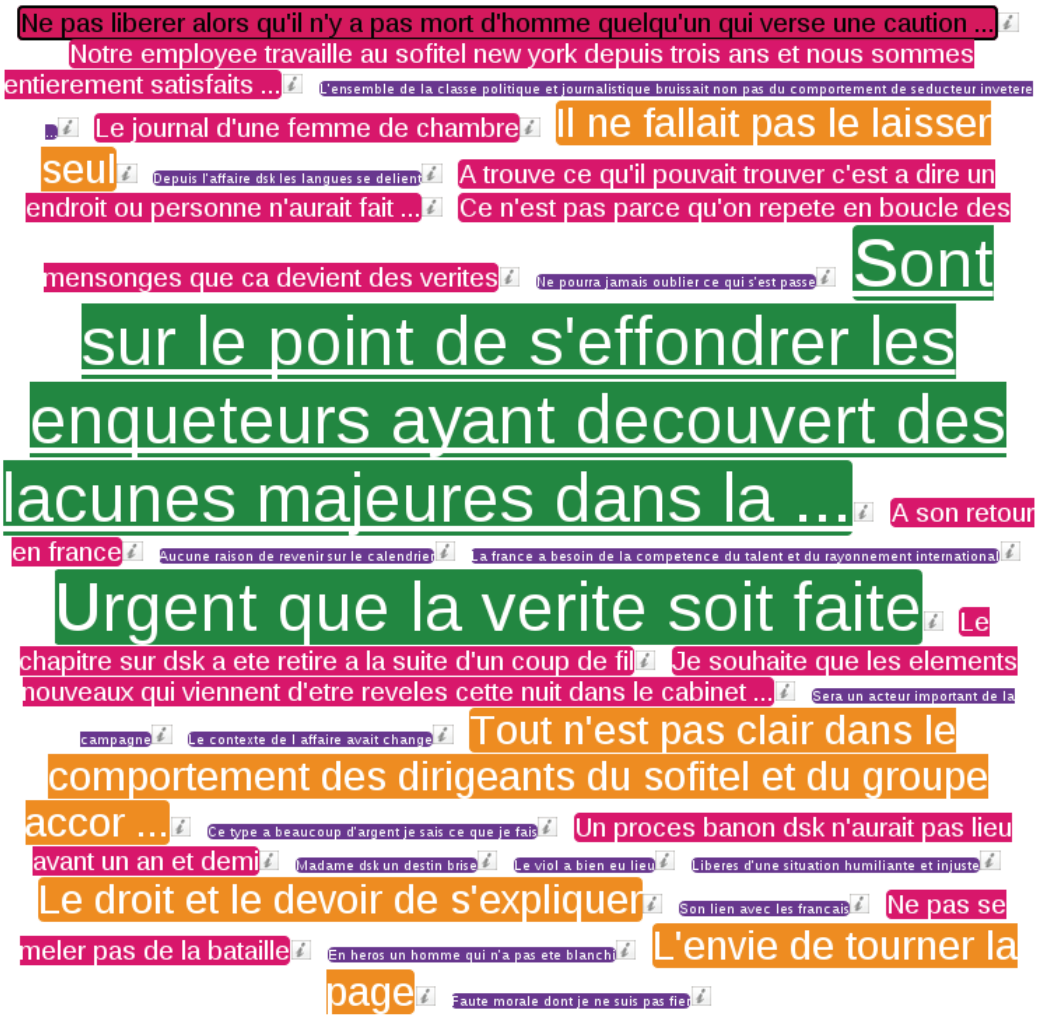

FIGURE 4.18 – Liste des buzz extraits à l'aide du widget *suivi de buzz* sur le corpus DSK.

Jeu 19 Mai 2011 21:17:33, @CSMBeggen, twitter.com, Retweeté 920 fois
RT @arretsurimages: Pendant #DSK, 3 réacteurs de Fukushima sont en fusion. Et les médias regardent ailleurs http://goo.gl/.

Dim 15 Mai 2011 23:21:00, @marcvama, twitter.com, Retweeté 728 fois
RT @dufourdufour: #DSK. Selon un rapport de police reçu par la reporter du LA Times à côté de moi, la victime présumée est noire.

Mar 17 Mai 2011 10:01:55, @mcastagnet, twitter.com, Retweeté 432 fois
RT @ldeboissieu: "2011 sera pour DSK une année géniale: à 62 ans, c'est l'année de sa vie" Élizabeth Teissier (astrologue/arnaquologue), Paris Match, 12/.

Mer 18 Mai 2011 14:39:57, @yeca, twitter.com, Retweeté 313 fois
RT @le_hiboo: pendant que tout le monde s'occupe des exces hormonaux de DSK, 3 réacteurs ont fondu à Fukushima.

Dim 15 Mai 2011 12:31:49, @9suricate3, twitter.com, Retweeté 280 fois
RT @maxime: Aucun témoignage prévu de la femme de ménage : son corps a été immédiatement immergé selon les rites mexicains.

Dim 15 Mai 2011 19:47:48, @adrienth, twitter.com, Retweeté 224 fois
RT @lemondefr: Comment un tweet sème le doute sur l'arrestation de DSK http://lemde.fr/.

Mar 17 Mai 2011 18:19:49, @Dali_Dalilaa, twitter.com, Retweeté 195 fois
RT @notstephan: Après Fukushima ma TL se composait d'experts en nucléaire, avec DSK ils sont devenus experts en droit américain. Trop polyvalent le twitto.

Mer 18 Mai 2011 15:42:50, @ronnie974, twitter.com, Retweeté 195 fois
RT @Maitre_Eolas: Vous voulez écrire à DSK ? Dominique Strauss-Kahn NYSID 09132366L Rikers Island West Facility.

Jeu 19 Mai 2011 07:31:44, @MartinBaumer, twitter.com, Retweeté 180 fois
RT @LANDEYves: Dominique Strauss-Kahn, la position.

Jeu 19 Mai 2011 21:09:02, @guybirenbaum, twitter.com, Retweeté 174 fois
RT @balasseNY: Le juge ne veut pas se décider aujourd'hui sur mise en liberté #.

FIGURE 4.19 – Liste des 10 tweets les plus retweetés pendant la première semaine de l'affaire DSK.

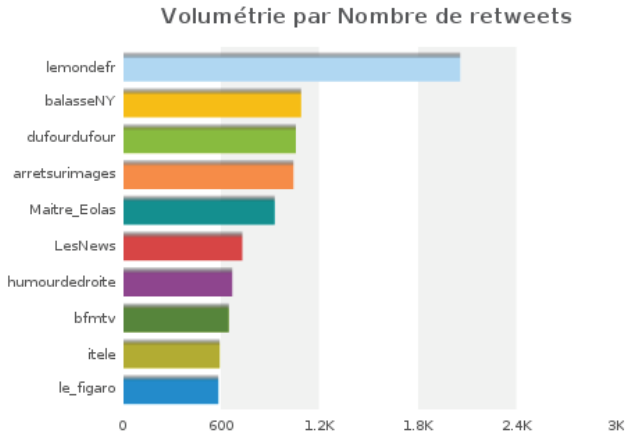

FIGURE 4.20 – Analyse volumétrique des acteurs de l'affaire DSK sur Twitter. Il s'agit ici majoritairement de professionnels de la presse.

4.3.2 Fukushima et le nucléaire en France

4.3.2.1 Présentation des données

La deuxième étude de cas concerne le sujet que nous avions utilisé dans la partie 3.1.4.1 : quel est l'impact de l'incident de Fukushima sur l'image du nucléaire en France ? Le tableau 4.2 résume les données relatives au corpus utilisé.

Type	Presse	Tweets
Langue	FR	FR
Période	du 17/04/2011 au 16/05/2011	du 14/03/2011 au 29/04/2011
Taille	471 documents	83 889 tweets

Tableau 4.2 – Caractéristiques des corpus sur le nucléaire en France.

4.3.2.2 Analyses

Graphe de cooccurrences Nous avons vu en 3.1.4.2 les thématiques abordées dans la presse. Tel que l'illustre la figure 4.21, les regards étaient principalement portés sur l'anniversaire de l'incident de Tchernobyl et la construction du sarcophage de protection, le prix de l'électricité, les débats de société autour de l'énergie nucléaire, ainsi que les candidatures des écologistes aux présidentielles.

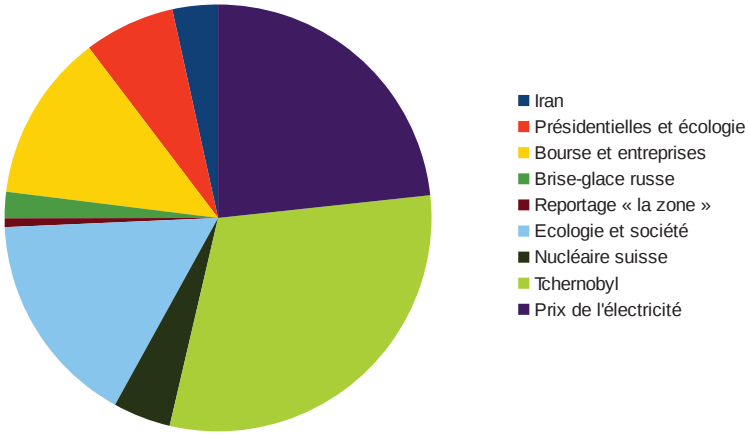

Figure 4.21 – Répartition des différentes thématiques extraites autour du nucléaire par analyse des plus proches voisins cooccurrentiels.

Widget *Suivi de buzz* Nous allons dorénavant nous concentrer sur des données Twitter. La figure 4.22 recense les buzz extraits par le widget *suivi de buzz*. Rappelons que l'unité d'analyse est ici le tweet, et non la citation.

On constate la reprise de certaines des thématiques abordées dans la presse comme le prix de l'électricité ou encore Tchernobyl. Un questionnement autour de

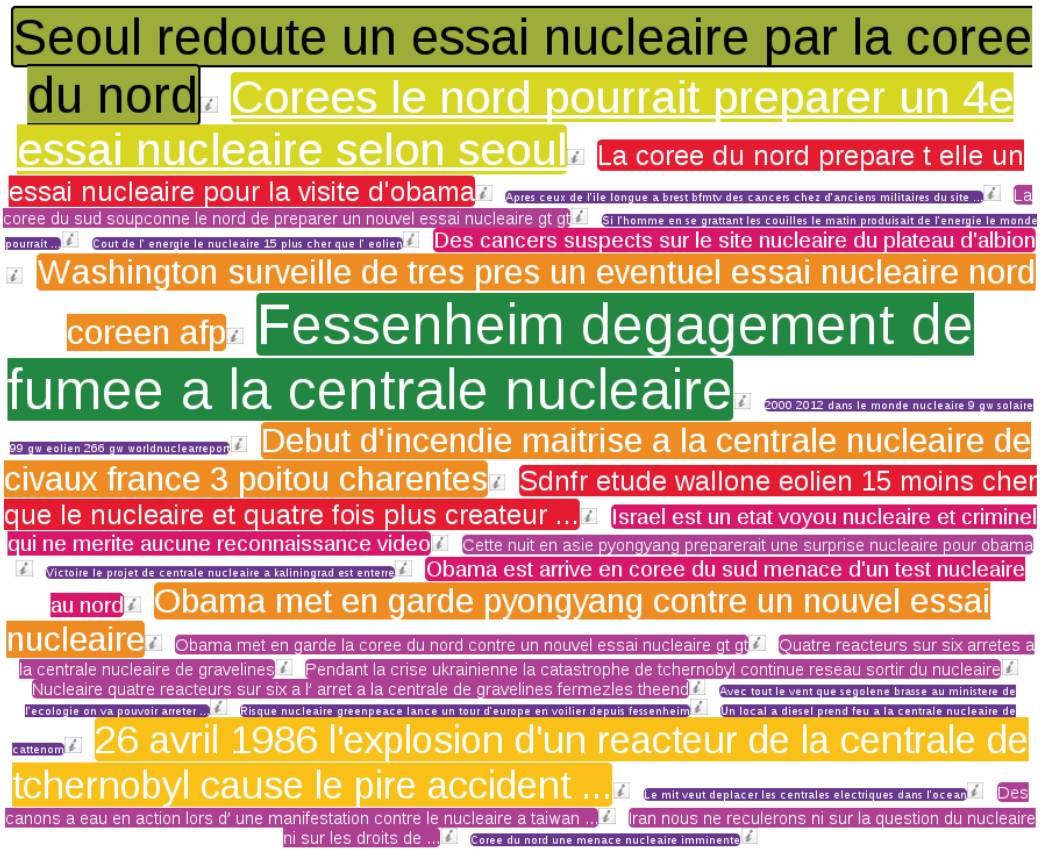

Figure 4.22 – Liste des buzz extraits à l'aide du widget *Suivi de buzz* sur les tweets concernant le nucléaire.

la sécurité du nucléaire en France est aussi palpable, que cela soit la peur d'un accident dans une centrale ou les risques indirects sur la santé. Néanmoins, ce qui semble passionner le plus les twittos est la crainte d'essais nucléaires par la Corée du nord.

Analyse de la topologie : widget *Twitter influence* Comme illustré dans la figure 4.23, on constate que la problématique du nucléaire semble principalement intéresser les acteurs traditionnels de la presse, relayant leurs articles sur Twitter via leur comptes officiels. Afin d'avoir une meilleure vision d'ensemble et de qualifier plus finement ces acteurs, il peut être intéressant de visualiser le graphe des relations entre twittos. Néanmoins, on constate dans la figure 4.24 que ce dernier illustre les problématiques de visualisation dont nous parlions à la fin du point 4.2.1 ; sa structure extrêmement dense rend l'ensemble peu lisible.

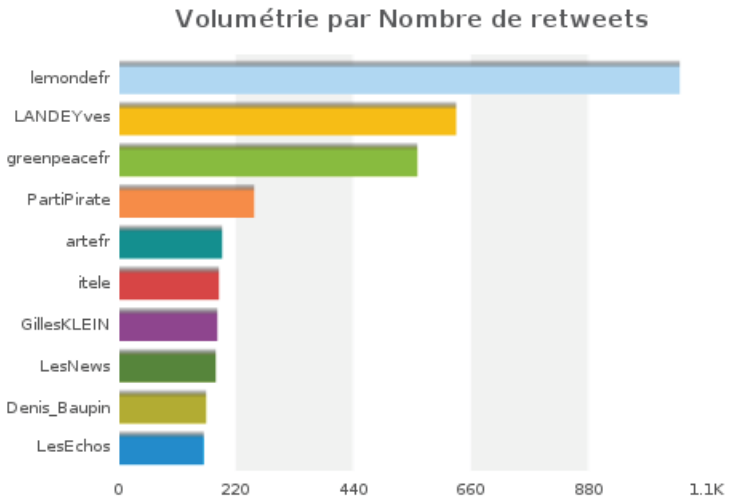

Figure 4.23 – Liste des 10 twittos les plus influents autour du nucléaire. Il s'agit principalement de comptes officiels de presse traditionnelle.

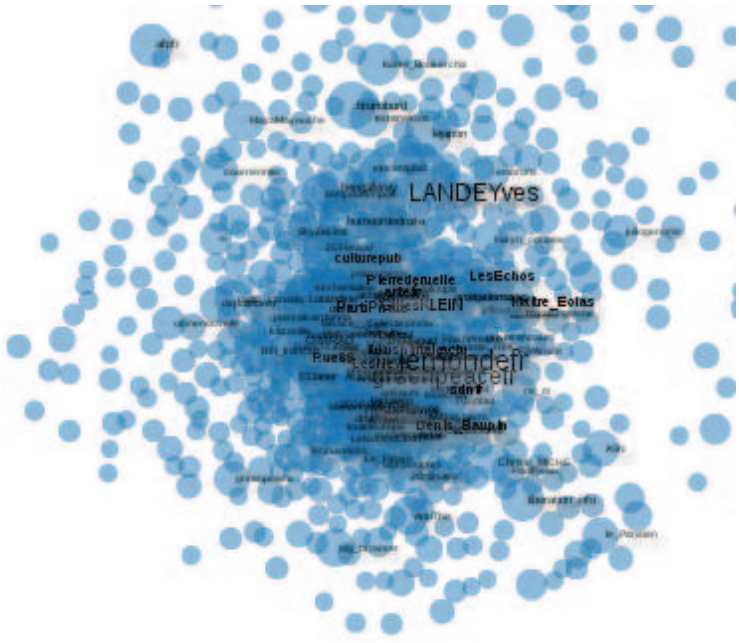

Figure 4.24 – Graphe des relations entre les utilisateurs les plus influents autour du nucléaire.

4.3.3 Le hashtag #unbonjuif

4.3.3.1 Présentation des données

Cette dernière étude de cas est intéressante en ce qu'elle met tout particulièrement en lumière l'intérêt d'une analyse topologique pour comprendre la façon dont un buzz se propage.

Début octobre 2012, une polémique a lieu suite à l'apparition parmi les *trending topics*[8] du hashtag #unbonjuif. Le premier tweet, publié le 10 octobre, est une invitation à la plaisanterie sur le sujet : "#UnBonJuif ?". Très rapidement, les amis de l'auteur répondent à l'invitation et la surenchère démarre.

Les caractéristiques du corpus utilisé sont résumées dans le tableau 4.3.

Type	Tweets
Langue	FR
Période	du 10/10/2012 au 11/11/2012
Taille	35 588 tweets

Tableau 4.3 – Caractéristiques des corpus sur #unbonjuif.

4.3.3.2 Analyses

Toutes les analyses pour ce cas ont été réalisées avec le widget *Twitter influence*. Commençons par une analyse volumétrique des tweets (voir figure 4.25). On observe un pic des volumes de publications à partir du 14 octobre, dépassant de loin ceux de la journée initiale. Ce n'est qu'à partir du 16 octobre que l'on redescend sous le seuil de la journée du 10 octobre. Comment expliquer ce soudain regain d'intérêt ?

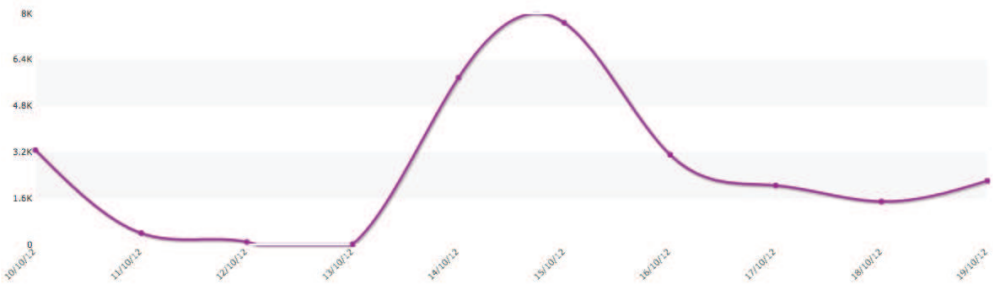

Figure 4.25 – Volumétrie des tweets sur #unbonjuif.

Une analyse des auteurs des tweets publiés avant le 14 met clairement en évidence une communauté de twittos connectés entre eux avec des signes d'appartenance à des "Teams" : #TeamPSG, #TeamMaroc, #TeamMuslim, #Palesteam ou #TeamRTsi[9]. Rapidement, la presse s'intéresse à la question ; l'entrée du hashtag parmi les dix meilleurs *trending topics* suscite des réactions de consternation ou d'indignation aussi bien dans les médias que sur Twitter. En moins de 3 heures, 3350 tweets seront émis par 1700 tweetos différents. Néanmoins, ce buzz relatif s'essouffle rapidement puisque, sur les 3 jours suivants, le hashtag n'est plus

8. Rappelons que les *trending topics* sont des hashtags massivement utilisés sur une période donnée. Cet usage important signifie un réel intérêt pour le sujet en question, mais il peut parfois être le fruit d'un effort réfléchi de plusieurs utilisateurs.

9. Lire l'article suivant à ce sujet : `http://www.lesinrocks.com/2012/10/01/medias/decouvrez-la-team-rtsi-11308611/`.

mentionné que dans 555 tweets dont 400 sur la journée du 11 octobre.

Analysons désormais les données à partir du 14 octobre pour tenter d'expliquer cette soudaine augmentation des publications à ce sujet. Une journaliste et blogueuse influente interpelle Twitter France :

```
@twitter\_fr combien de temps allez-vous laissez faire cela
 en toute impunité?
```

Son réseau d'influence sur le réseau social est suffisamment conséquent (presque 4000 followers) pour que le sujet se diffuse rapidement au sein d'une nouvelle communauté des personnes indignées. Vers 17 heures du même jour, le compte Twitter officiel de l'Union des Étudiants Juifs de France (UEJF) annonce son intention de demander un rendez-vous à Twitter, puis son intention de porter plainte. L'information circule vite sur le réseau et les premiers articles sont publiés sur des sites d'actualités, repris par leur compte Twitter. Dans le camp de personnes indignées, l'activité devient dès lors plus intense et on assiste vite à une vague de contributions sur la liberté d'expression, reprenant l'argument du journal Charlie Hebdo pour justifier la publication des caricatures de Mahomet :

```
Hashtag #UnBonJuif = antisémitisme !! Caricatures CharlieHebdo
 Mahomet = liberté d'expression !! (Ps: pour information
 j'suis Chrétien)

Si on me poursuit pour le # #UnBonJuif, Je dirais que Je travaille
 pour Charlie Hebdo

2 poids 2 mesures: Charlie Hebdo peut insulter les catholiques
 et les musulmans, mais @Besbar n'a pas le droit de plaisanter
 avec #UnBonJuif

J'me ferai une joie de rejoindre les gens qui dénoncent le hashtag
 #UnBonJuif, lorsque ceux-ci s'indigneront aussi contre l'islamophobie

S'indigner du hashtag #UnBonJuif c'est bien, mais il faudrait
 l'être également quand il y a des dérapages sur d'autres communautés

#ACauseDunBonJuif ou GRÂCE à #UnBonJuif, on peut voir que cette
 fameuse "LIBERTÉ D'EXPRESSION" ne s'applique pas pour tout et
 tout le (...)
```

Les figures 4.26, 4.27 et 4.28 montrent l'évolution dans le temps des deux communautés citées précédemment.

Pour conclure sur ce sujet, il est intéressant de remarquer que le phénomène de buzz a été considérablement amplifié par les réactions indignées, plus que par la communauté d'origine : ce sont donc les personnes indignées qui ont redonné vie au sujet.

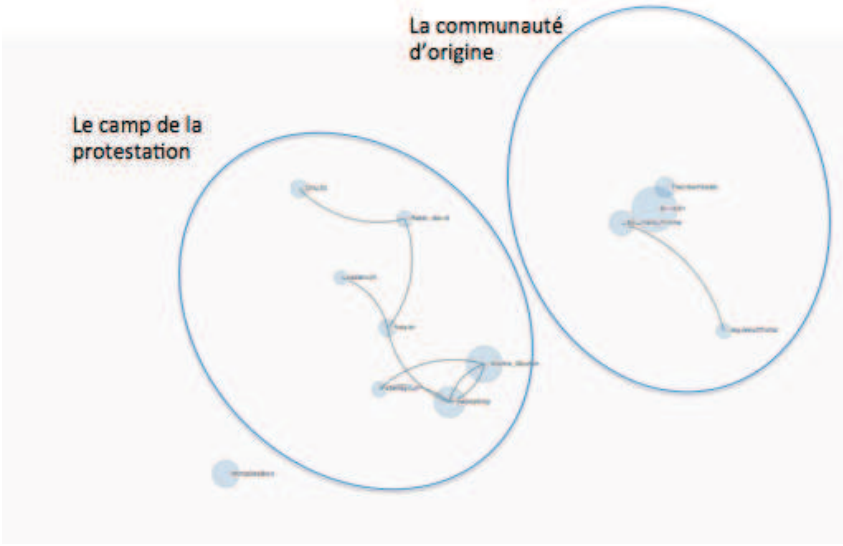

FIGURE 4.26 – Graphe topologique des interactions entre twittos le 14 octobre en fin de matinée. Un découpage communautaire apparaît ici clairement : d'un côté les twittos d'origine, de l'autre les personnes indignées à ce sujet.

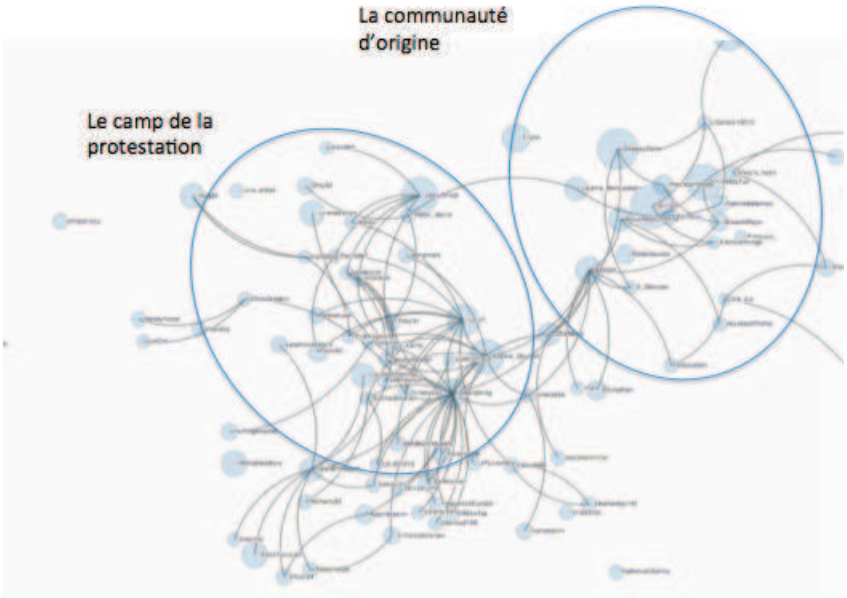

FIGURE 4.27 – Évolution du graphe présenté dans la figure 4.26, ici à 17h. Des liens entre les deux communautés commencent à apparaître : les initiateurs du hashtag reprennent les tweets de protestation comme un signe de succès de l'opération initiale.

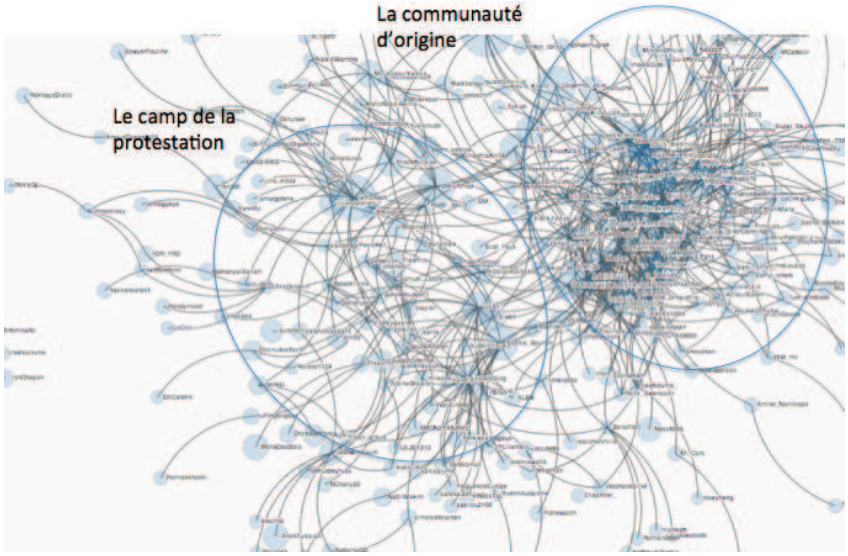

FIGURE 4.28 – Illustration de la topologie des twittos sur le thème #unbonjuif sur des tweets rédigés entre le 14 et le 16 octobre. Le réseau est dense, traduisant une réelle diffusion virale du buzz, ainsi qu'un découpage communautaire toujours bien marqué.

Conclusion

Au cours de cette thèse, nous avons abordé quelques axes de réflexion et proposé diverses approches afin de mieux appréhender le cheminement d'un buzz sur le Web. Nous nous sommes appuyés sur l'hypothèse que la compréhension de ce dernier ne peut se faire sans une analyse conjointe de la topologie du Web et du contenu textuel. Il ne s'agit donc pas uniquement de savoir ce qui se dit, et dans quelles proportions, mais aussi de qualifier plus précisément les émetteurs et détecter les influenceurs.

En vue d'une analyse topologique du Web, l'algorithme HITS nous semble pertinent car il permet de modéliser l'autorité d'une page Web tout en prenant en considération des thématiques. Par ailleurs, les notions de hubs (diffuseurs d'information) et d'autorité (générateurs d'information) permettent selon nous de mieux qualifier les émetteurs.

Afin de recontextualiser au maximum l'information et ainsi permettre une meilleure interprétation des résultats, nous avons proposé une approche itérative combinant calcul d'autorité, détection de communautés et outils de visualisation. L'approche est très prometteuse au vue des expérimentations réalisées. Néanmoins, quelques axes restent à améliorer, notamment la robustesse de la méthode de détection de communautés : malgré les filtres proposés, les sous-graphes du Web extraits sont bruités. Signalons cependant que cette phase de détection de communautés pourrait idéalement, avec des outils de visualisation assez puissants, être réalisée manuellement.

L'analyse des sous-graphes du Web est trop dépendante des heuristiques des grands moteurs de recherche : l'accès objectif aux données n'est pas garanti. Notons toutefois que certains grands réseaux sociaux ouverts, Twitter en particulier, offrent un accès aisé à leurs données. Ces dernières sont moins bruitées étant donné qu'elles répondent à un format plus normé et qu'elles suivent certaines conventions d'usage. Les réseaux sociaux faisant office de caisse de résonance du Web classique (citation d'hyperliens), des analyses plus robustes sont alors envisageables.

Pour l'analyse du contenu textuel, nous avons proposé deux approches que nous jugeons complémentaires. La première, qui consiste à ramener la tâche de détection de buzz à celle d'extraction de thématiques, repose sur l'analyse d'un graphe de cooccurrents de deuxième ordre. Une telle analyse permet d'offrir la possibilité de visualiser les relations entre les thématiques, ainsi que celles qu'entretiennent les mots au sein de chacune d'entre elles. Les premiers résultats sont très encourageants mais les graphes de cooccurrences générés sont souvent trop denses, si bien que la représentation présente dans certains cas peu d'avantages par rapport aux listes classiques renvoyées par les méthodes d'extraction de thématiques.

La deuxième approche repose sur l'extraction de citations, l'hypothèse de départ

étant que ces dernières sont des empreintes génétiques permettant une identification univoque d'un buzz. L'approche présente des résultats particulièrement intéressants. Néanmoins, si les résultats sont clairs et robustes sur des corpus journalistiques, l'utilisation de la citation rend l'approche trop dépendante du type de données. Afin de nous affranchir de cette limitation, nous proposons d'étendre l'approche aux messages courts rédigés sur des réseaux sociaux (tweets par exemple). Pour les autres types de données, il serait intéressant d'assouplir la notion de citation en détectant par exemple les citations au style indirect via une analyse linguistique.

De plus, nous avons proposé une première approche vers une analyse linguistique qualitative de l'évolution de ces citations dans le temps. Ainsi, nous avons pu observer l'émergence de certaines régularités statistiques, notamment le phénomène de figement de certaines citations en un concept. Ces phénomènes constituent des pistes intéressantes à suivre pour une meilleure compréhension des schémas de diffusion des messages en ligne.

S'agissant d'une thèse réalisée dans un cadre de recherche applicative en entreprise, nous avons cherché à adapter au mieux ces enseignements théoriques aux contraintes industrielles. Précisons la volonté de l'entreprise AMI Software de ne pas se limiter à un domaine en particulier étant donné que l'objet d'étude est le Web dans son ensemble. Les outils proposés se devaient d'être le plus généralisables et robustes possible. Nous pensons que nos développements répondent à ce critère. Demeurent toutefois des problèmes généraux de visualisation des données topologiques. Ces derniers sont d'autant plus importants si on inclut la dimension temporelle.

Bibliographie

[Adam 2009] Clémentine Adam et François Morlane-Hondère. *Détection de la cohésion lexicale par voisinage distributionnel : application à la segmentation thématique*. In Actes de RECITAL'09, Senlis, 2009. (Cité en page 76.)

[Adamson 1974] George Adamson et Jillian Boreham. *The use of an association measure based on character structure to identify semantically related pairs of words and document titles*. Information storage and retrieval, vol. 10, no. 7-8, pages 253–260, 1974. (Cité en page 70.)

[Ahn 2009] Yong-Yeol Ahn, James Bagrow et Sune Lehmann. *Link communities reveal multiscale complexity in networks*. Nature, vol. 466, no. 7307, pages 761–764, 2009. (Non cité.)

[Allan 1998a] James Allan, Jaime Carbonell, George Doddington, Jonathan Yamron et Yiming Yang. *Topic detection and tracking pilot study : final report*. In Proceedings of the DARPA broadcast news transcription and understanding workshop, pages 194–218, Lansdowne, 1998. (Cité en page 96.)

[Allan 1998b] James Allan, Ron Papka et Victor Lavrenko. *On-line new event detection and tracking*. In SIGIR'98 - Proceedings of the 21st annual international ACM SIGIR conference on research and development in information retrieval, pages 37–45, Melbourne, 1998. (Non cité.)

[Allan 2001] James Allan, Rahul Gupta et Vikas Khandelwal. *Temporal summaries of news topics*. In SIGIR'01 - Proceedings of the 24th annual international ACM SIGIR conference on research and development in information retrieval, pages 10–18, New Orleans, 2001. (Non cité.)

[Allan 2002] James Allan. Topic detection and tracking : event-based information organization. Springer, 2002. (Cité en page 96.)

[Allport 1951] Gordon Allport et Joseph Postman. Psychology of Rumor. Russel and Russel, 1951. (Cité en page 16.)

[Anaya-Sánchez 2008] Henry Anaya-Sánchez, Aurora Pons-Porrata et Rafael Berlanga-Llavori. *A new document clustering algorithm for topic discovering and labeling*. In Proceedings of the 13th Iberoamerican congress on pattern recognition (CIARP'08), pages 161–168, Berlin, 2008. (Non cité.)

[Andrews 2007] Nicholas Andrews et Edward Fox. *Recent developments in document clustering*. Rapport technique, 2007. (Non cité.)

[Arendt 1989] Hannah Arendt. La crise de la culture. 1989. (Cité en page 21.)

[Auber 2003] David Auber, Yves Chiricota, Fabien Jourdan et Guy Melançon. *Multiscale visualization of small world networks*. In INFOVIS'03 - Proceedings of the 9th annual IEEE conference on information visualization, pages 75–81, Seattle, 2003. (Non cité.)

[Balog 2006] Krisztian Balog, Gilad Mishne et Maarten de Rijke. *Why are they excited ? Identifying and explaining spikes in blog mood levels.* In EACL'06 - Proceedings of the 11th conference of the European chapter of the association for computational linguistics, pages 207–210, Trento, Avril 2006. (Non cité.)

[Baumes 2005] Jeffrey Baumes, Mark Goldberg, Mukkai Krishnamoorthy, Malik Magdon-Ismail et Nathan Preston. *Finding communities by clustering a graph into overlapping subgraphs.* In AC 2005 - Proceedings of the IADIS international conference on applied computing, pages 97–104, Algarve, 2005. (Non cité.)

[Bearman 2004] Peter Bearman, James Moody et Katherine Stovel. *Chains of affection : the structure of adolescent romantic and sexual networks.* American journal of sociology, vol. 110, no. 1, pages 44–91, 2004. (Non cité.)

[Becchetti 2008] Luca Becchetti, Carlos Castillo, Debora Donato, Ricardo Baeza Yates et Stefano Leonardi. *Link analysis for web spam detection.* ACM transactions on the web (TWEB), vol. 2, no. 1, 2008. (Non cité.)

[Ben Jabeur 2012] Lamjed Ben Jabeur, Lynda Tamine et Mohand Boughanem. *Intégration des facteurs temps et autorité sociale dans un modèle bayésien de recherche de tweets.* In CORIA 2012 - Actes de la 9ème conférence en recherche d'information et applications, Bordeaux, 2012. (Cité en page 99.)

[Benoist 2014] Jean-Marie Benoist. *La veille ne fait que commencer*, Février 2014. (Cité en page 10.)

[Bertels 2005] Ann Bertels. *Les spécificités en contexte : comment étudier la polysémie dans un corpus technique ?* In Journées scientifiques du réseau de chercheurs Lexicologie, Terminologie et Traduction (LTT), Bruxelles, 2005. (Non cité.)

[Bestgen 2006] Yves Bestgen et Sophie Piérard. *Comment évaluer les algorithmes de segmentation automatique ? Essai de construction d'un matériel de référence.* In TALN 2006 - Actes de la 13ème conférence sur le traitement automatique des langues naturelles, pages 407–414, Leuven, Belgique, 2006. (Non cité.)

[Bezdek 1981] James Bezdek. Pattern recognition with fuzzy objective function algorithms. Springer, 1981. (Non cité.)

[Bharat 1998a] Krishna Bharat et Monika R. Henzinger. *Improved algorithms for topic distillation in a hyperlinked environment.* In Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '98, pages 104–111, New York, New York, USA, Août 1998. ACM Press. (Cité en pages 27, 41 et 48.)

[Bharat 1998b] Krishna Bharat et Monika R. Henzinger. *Improved algorithms for topic distillation in a hyperlinked environment.* In SIGIR'98 - Proceedings of the 21st annual international ACM SIGIR conference on research and development in information retrieval, pages 104–111, Melbourne, Août 1998. ACM Press. (Cité en page 47.)

[Bharat 2002a] Krishna Bharat et George A. Mihaila. *When experts agree : using non-affiliated experts to rank popular topics.* ACM transactions on information systems, vol. 20, no. 1, pages 47–58, Janvier 2002. (Cité en pages 41 et 48.)

[Bharat 2002b] Krishna Bharat et George A. Mihaila. *When experts agree : using non-affiliated experts to rank popular topics.* ACM Transactions on Information Systems, vol. 20, no. 1, pages 47–58, Janvier 2002. (Non cité.)

[Biber 1993a] Douglas Biber. *Co-occurrence patterns among collocations : a tool for corpus-based lexical knowledge acquisition.* Computational Linguistics, vol. 19, no. 3, pages 531–538, Septembre 1993. (Non cité.)

[Biber 1993b] Douglas Biber. *Co-occurrence patterns among collocations : a tool for corpus-based lexical knowledge acquisition.* Computational linguistics, vol. 19, no. 3, pages 531–538, Septembre 1993. (Non cité.)

[Binsztok 2002] Henri Binsztok et Patrick Gallinari. *Un algorithme en ligne pour la détection de nouveauté dans un flux de documents.* In JADT 2002 - Proceedings of the 6th international conference on the statistical analysis of textual data, Saint-Malo, 2002. (Cité en page 96.)

[Biskri 2004] Ismaïl Biskri, Jean-Guy Meunier et Sylvain Joyal. *L'extraction des termes complexes : une approche modulaire semi-automatique.* In JADT 2004 - Proceedings of the 7th international conference on the statistical analysis of textual data, Louvain-la-neuve, 2004. (Non cité.)

[Blei 2003] David Blei, Andrew Y. Ng et Michael I. Jordan. *Latent dirichlet allocation.* Journal of machine learning research, vol. 3, pages 993–1022, 2003. (Cité en pages 45, 74 et 81.)

[Blei 2009a] David Blei. *Topic models (Video : http://videolectures.net/mlss09uk_blei_tm/)*, 2009. (Cité en pages 69 et 74.)

[Blei 2009b] David Blei et John Lafferty. *Topic models.* In Text mining : classification, clustering, and applications, page 328. Chapman and Hall/CRC, 2009. (Cité en page 74.)

[Blei 2011] David Blei. *Introduction to probabilistic topic models.* Communications of the ACM, 2011. (Cité en page 74.)

[Bolshakov 2001] Igor Bolshakov et Alexander Gelbukh. *Text segmentation into paragraphs based on local text cohesion.* In TSD'01 - Proceedings of the 4th international conference on text, speech and dialogue, pages 158–166, Zelezna Ruda, 2001. (Non cité.)

[Bonato 2008] Anthony Bonato. A course on the web graph. American Mathematical Society, 2008. (Non cité.)

[Bordag 2003a] Stefan Bordag. *Sentence co-occurrences as small-world graphs : a solution to automatic lexical disambiguation.* In CICLing'03 - Proceedings of the 4th International conference on computational linguistics and interlligent text processing, pages 329–332, Mexico City, 2003. (Cité en page 77.)

[Bordag 2003b] Stefan Bordag, Gerhard Heyer et Uwe Quasthoff. *Small worlds of concepts and other principles of semantic search.* In IICS 2003 - Proceedings of the 3rd international workshop in innovative internet community systems, Leipzig, 2003. (Non cité.)

[Bordag 2006] Stefan Bordag. *Word sense induction : triplet-based clustering and automatic evaluation.* In EACL'06 - Proceedings of the 11th conference of the European chapter of the association for computational linguistics, Trento, 2006. (Cité en page 77.)

[Bordag 2008] Stefan Bordag. *A comparison of co-occurrence and similarity measures as simulations of context.* In CICLing'08 - Proceedings of the 9th International Conference on Computational Linguistics and interlligent text processing, pages 52–63, Haifa, Février 2008. (Non cité.)

[Borodin 2005] Allan Borodin, Gareth O. Roberts, Jeffrey S. Rosenthal et Panayiotis Tsaparas. *Link analysis ranking : algorithms, theory, and experiments.* ACM transactions on internet technology, vol. 5, no. 1, pages 231–297, Février 2005. (Non cité.)

[Bourigault 1994] Didier Bourigault. *Lexter, un logiciel d'extraction de terminologie. Application à l'acquisition des connaissances à partir de textes.* PhD thesis, École des hautes Etudes en Sciences Sociales, 1994. (Non cité.)

[Bourigault 2002] Didier Bourigault. *UPERY : un outil dâanalyse distributionnelle étendue pour la construction dâontologies à partir de corpus.* In TALN 2002 - Actes de la 9ème conférence sur le traitement automatique des langues naturelles, Nancy, 2002. (Non cité.)

[Bourigault 2007] Didier Bourigault. *Un analyseur syntaxique opérationnel : SYNTEX.* Habilitation, Université Toulouse 2 - Le Mirail, 2007. (Non cité.)

[Bourion 2001] Evelyne Bourion. *L'aide à l'interprétation des textes électroniques.* PhD thesis, Université Paris 10, 2001. (Non cité.)

[BRIN 1998a] S BRIN et L PAGE. *The anatomy of a large-scale hypertextual Web search engine.* Computer Networks and ISDN Systems, vol. 30, no. 1-7, pages 107–117, Avril 1998. (Non cité.)

[Brin 1998b] Sergey Brin et Lawrence Page. *The anatomy of a large-scale hypertextual web search engine.* Computer networks and ISDN systems, vol. 30, no. 1-7, pages 107–117, Avril 1998. (Cité en pages 27, 33, 35 et 44.)

[Broder 1997] Andrei Broder, Steven Glassman, Mark Manasse et Geoffrey Zweig. *Syntactic clustering of the web.* Computer networks and ISDN systems, vol. 29, no. 8-13, pages 1157–1166, 1997. (Cité en pages 14 et 103.)

[Broder 2000] Andrei Broder, Ravi Kumar, Farzin Maghoul, Prabhakar Raghavan, Sridhar Rajagopalan, Raymie Stata, Andrew Tomkins et Janet Wiener. *Graph structure in the Web.* Computer networks : the international journal of computer and telecommunications networking, vol. 33, no. 1-6, pages 309–320, 2000. (Cité en page 31.)

[Bron 1973] Coen Bron et Joep Kerbosch. *Finding all cliques of an undirected graph.* Communications of the ACM, vol. 16, no. 9, pages 575–577, 1973. (Cité en page 58.)

[Broudoux 2007] Evelyne Broudoux. *Construction de l'autorité informationnelle sur le web.* In A document (re)turn, pages 265–278. Peter Lang GmbH, 2007. (Cité en pages 21, 22 et 51.)

[Brunet 2001] Etienne Brunet. *Qui lemmatise dilemme attise.* Lexicometrica, vol. 2, page 19, 2001. (Cité en page 71.)

[Brunet 2003] Etienne Brunet. *Peut-on mesurer la distance entre deux textes ?* Corpus, vol. 2, pages 1–19, 2003. (Cité en page 80.)

[Brunet 2008] Etienne Brunet. *Les séquences (suite).* In JADT 2008 - Proceedings of the 9th international conference on the statistical analysis of textual data, pages 253–266, Lyon, 2008. (Non cité.)

[Caillet 2004] Marc Caillet, Jean-François Pessiot et Patrick Gallinari. *Unsupervised learning with term clustering for thematic segmentation of texts.* In Actes de la 7è conférence en recherche d'information assistée par ordinateur (RAIO '04), pages 1–11, 2004. (Non cité.)

[Carrière 1997a] S.Jeromy Carrière et Rick Kazman. *WebQuery : searching and visualizing the Web through connectivity.* Computer networks and ISDN systems, vol. 29, no. 8-13, pages 1257–1267, Septembre 1997. (Non cité.)

[Carrière 1997b] S.Jeromy Carrière et Rick Kazman. *WebQuery : searching and visualizing the Web through connectivity.* Computer Networks and ISDN Systems, vol. 29, no. 8-13, pages 1257–1267, Septembre 1997. (Non cité.)

[Cazals 2008] Frederic Cazals et Chinmay Karande. *A note on the problem of reporting maximal cliques.* Theoretical computer science, vol. 407, no. 1-3, pages 564–568, 2008. (Cité en page 58.)

[Cha 2009] Meeyoung Cha, Alan Mislove et Krishna Gummadi. *A measurement-driven analysis of information propagation in the Flickr social network.* In WWW'09 - Proceedings of the 18th international conference on world wide web, pages 721–730, Madrid, 2009. (Non cité.)

[Cha 2010] Meeyoung Cha, Hamed Haddadi, Fabricio Benevenuto et Krishna Gummadi. *Measuring user influence in twitter : the million follower fallacy.* In ICWSM 2010 - Proceedings of the 4th international AAAI conference on weblogs and social media, Washington DC, 2010. (Cité en pages 9, 43, 44 et 46.)

[Chakrabarti 1998a] Soumen Chakrabarti, Byron Dom, David Gibson, Ravi Kumar, Prabhakar Raghavan et Sridhar Rajagopalan. *Experiments in topic distillation.* Rapport technique, 1998. (Non cité.)

[Chakrabarti 1998b] Soumen Chakrabarti, Byron Dom et Piotr Indyk. *Enhanced hypertext categorization using hyperlinks.* In Proceedings of the 1998 ACM SIGMOD international conference on management of data, Seattle, 1998. (Non cité.)

[Chakrabarti 1998c] Soumen Chakrabarti, Byron Dom, Prabhakar Raghavan, Sridhar Rajagopalan, David Gibson et Jon Kleinberg. *Automatic resource compilation by analyzing hyperlink structure and associated text.* Computer networks and ISDN systems, vol. 30, no. 1-7, pages 65–74, Avril 1998. (Cité en pages 27, 41 et 48.)

[Chakrabarti 1999a] Soumen Chakrabarti, Byron Dom, David Gibson, Jon Kleinberg, Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan et Andrew Tomkins. *Mining the link structure of the world wide web.* IEEE computers, vol. 32, pages 60–67, 1999. (Cité en pages 26 et 27.)

[Chakrabarti 1999b] Soumen Chakrabarti, Martin Van den Berg et Byron Dom. *Distributed hypertext resource discovery through examples.* In VLDB'99 - Proceedings of the 25th international conference on very large data bases, pages 375–386, Edinburgh, 1999. (Non cité.)

[Chakrabarti 1999c] Soumen Chakrabarti, Martin Van den Berg et Byron Dom. *Focused crawling : a new approach to topic-specific Web resource discovery.* In WWW'99 - Proceedings of the 8th international conference on world wide web, pages 1623–1640, Toronto, 1999. (Non cité.)

[Champclaux 2009] Yaël Champclaux. *Un modèle de recherche d'information basé sur les graphes et les similarités structurelles pour l'amélioration du processus de recherche d'information.* PhD thesis, Université Toulouse 3, 2009. (Non cité.)

[Chauvin 2005] Sophie Chauvin. *Visualisations heuristiques pour la recherche et l'exploration de données dynamiques : l'art informationnel en tant que révélateur de sens.* PhD thesis, Université Paris 8, 2005. (Non cité.)

[Chee 2007] Brant Chee et Bruce Schatz. *Document clustering using small world communities.* In JCDL'07 - Proceedings of the 7th ACM/IEEE-CS joint conference on digital libraries, pages 53–62, Vancouver, 2007. (Cité en page 78.)

[Cheng 2009] Alex Cheng et Mark Evans. *An in-depth look inside the Twitter world.* Rapport technique, 2009. (Cité en page 44.)

[Chikhi 2008] Nacim Fateh Chikhi, Bernard Rothenburger et Nathalie Aussenac-Gilles. *Combining link and content information for scientific topics discovery.* In ICTAI - Proceedings of the 20th IEEE international conference on tools with artificial intelligence, pages 211–214, Dayton, Novembre 2008. IEEE. (Non cité.)

[Chikhi 2010] Nacim Fateh Chikhi. *Calcul de centralité et identification de structures de communautés dans les graphes de documents.* PhD thesis, Université Paul Sabatier, 2010. (Cité en page 41.)

[Choi 2001] Freddy Choi, Peter Wiemer-Hastings et Johanna Moore. *Latent semantic analysis for text segmentation.* In EMNLP 2001 - Proceedings of the conference on empirical methods in natural language processing, Pittsburgh, 2001. (Non cité.)

[Choueka 1985] Yaacov Choueka et Serge Lusignan. *Disambiguation by short contexts*. Computers and the Humanities, vol. 19, no. 3, pages 147–157, 1985. (Cité en page 79.)

[Church 1990] Kenneth Ward Church et Patrick Hanks. *Word association norms, mutual information, and lexicography*. Computational linguistics, vol. 16, no. 1, pages 22–29, Mars 1990. (Cité en pages 80 et 95.)

[Clauset 2004] Aaron Clauset, Mark Newman et Cristopher Moore. *Finding community structure in very large networks*. Physical review E, vol. 70, 2004. (Non cité.)

[Cleuziou 2007] Guillaume Cleuziou. *A generalization of k-means for overlapping clustering*. Rapport technique, 2007. (Cité en pages 54 et 73.)

[Condamines 2003] Anne Condamines. *Vers la définition de genres interprétatifs*. In TIA 2003 - Actes de la 5ème conférence internationale terminologie et acquisition de connaissances à partir de textes, Strasbourg, 2003. (Non cité.)

[Conrad 2008] Jack G. Conrad, Jochen Leidner et Frank Schilder. *Professional credibility : authority on the web*. In WICOW'08 - Proceeding of the 2nd ACM workshop on information credibility on the web, page 85, Napa Valley, Octobre 2008. ACM Press. (Non cité.)

[Coutrot 2008] Laurence Coutrot. *Sur l'usage récent des indicateurs bibliométriques comme outil d'évaluation de la recherche scientifique*. Bulletin de méthodologie sociologique, vol. 100, pages 45–50, 2008. (Cité en page 27.)

[Cutting 1992] Douglass Cutting, David Karger, Jan Pedersen et John Tukey. *Scatter/gather : a cluster-based approach to browsing large document collections*. In SIGIR'92 - Proceedings of the 15th annual international ACM SIGIR conference on research and development in information retrieval, pages 318–329, Copenhagen, 1992. (Non cité.)

[Dai 2010] Na Dai et Brian D. Davison. *Freshness matters : in flowers, food, and web authority*. In SIGIR'10 - Proceeding of the 33rd international ACM SIGIR conference on research and development in information retrieval, page 114, Geneva, Juillet 2010. ACM Press. (Cité en page 23.)

[Damak 2012] Firas Damak, Karen Pinel-Sauvagnat, Guillaume Cabanac et Mohand Boughanem. *Recherche de microblogs : quelles sources d'évidences pour raffiner les résultats des moteurs usuels de RI ?* In CORIA 2012 - Actes de la 9ème conférence en recherche d'information et applications, Bordeaux, 2012. (Non cité.)

[Davison 1999] Brian D. Davison, Apostolos Gerasoulis, Konstantinos Kleisouris, Yingfang Lu, Hyun-ju Seo, Wei Wang et Baohua Wu. *DiscoWeb : applying link analysis to web search*. In Proceedings of the 8th World Wide Web Conference, 1999. (Cité en page 40.)

[Davison 2000a] Brian D. Davison. *Topical locality in the Web*. In Proceedings of the 23rd annual international ACM SIGIR conference on Research and

development in information retrieval - SIGIR '00, pages 272–279, New York, New York, USA, Juillet 2000. ACM Press. (Cité en page 27.)

[Davison 2000b] Brian D. Davison. *Topical locality in the web.* In SIGIR'00 - Proceedings of the 23rd annual international ACM SIGIR conference on research and development in information retrieval, pages 272–279, Athens, Juillet 2000. ACM Press. (Non cité.)

[Davison 2003] Brian D. Davison. *Unifying text and link analysis.* In IJCAI'03 - Workshop on text mining and link analysis (TextLink), Acapulco, 2003. (Non cité.)

[Dawkins 1976] Richard Dawkins. The selfish gene. Oxford University Press, 1976. (Cité en pages 100 et 106.)

[Deerwester 1990] Scott Deerwester, Susan Dumais, George Furnas, Thomas Landauer et Richard Harshman. *Indexing by latent semantic analysis.* Journal of the American society for information science, vol. 41, no. 6, pages 391–407, 1990. (Cité en pages 73 et 81.)

[Degenne 2004] Alain Degenne et Michel Forsé. Les réseaux sociaux. Une approche structurale en sociologie. Armand Colin, 2004. (Non cité.)

[Dias 2000] Gaël Dias, Sylvie Guilloré et José Gabriel Pereira Lopes. *Extraction automatique d'associations textuelles à partir de corpora non traités.* In JADT 2000 - Proceedings of the 5th international conference on the statistical analysis of textual data, Lausanne, 2000. (Non cité.)

[Dias 2005] Gaël Dias et Elsa Alves. *Discovering topic boundaries for text summarization based on word cooccurrence.* In Recent advances in natural language processing (RANLP 2005), pages 187–191, Borovets, Bulgaria, 2005. (Non cité.)

[Diday 1971] Edwin Diday. *Une nouvelle méthode en classification automatique et reconnaissance des formes : la méthode des nuées dynamiques.* Revue de statistique appliquée, vol. 19, no. 2, pages 19–33, 1971. (Cité en page 73.)

[Diligenti 2000] Michelangelo Diligenti, Frans Coetzee, Steve Lawrence, Lee Giles et Marco Gori. *Focused crawling using context graphs.* In VLDB'00 - Proceedings of the 26th international conference on very large data bases, pages 527–534, Cairo, 2000. (Non cité.)

[Donetti 2004] Luca Donetti et Miguel Muñoz. *Detecting network communities : a new systematic and efficient algorithm.* Journal of statistical mechanics : theory and experiment, vol. 2004, no. 10, 2004. (Non cité.)

[Donetti 2005] Luca Donetti et Miguel Muñoz. *Improved spectral algorithm for the detection of network communities.* In Computational and statistical physics - Proceedings of the 8th Granada seminar, pages 1–2, 2005. (Non cité.)

[Dorow 2005] Beate Dorow, Dominic Widdows, Katarina Ling, Jean-Pierre Eckmann, Danilo Sergi et Elisha Moses. *Using curvature and Markov clustering in graphs for lexical acquisition and word sense discrimination.* In Workshop MEANING 2005, 2005. (Cité en page 77.)

[Duan 2010] Yajuan Duan, Long Jiang, Tao Qin, Ming Zhou et Heung-Yeung Shum. *An empirical study on learning to rank of tweets.* In COLING'10 - Proceedings of the 23rd international conference on computational linguistics, pages 295–303, Beijing, 2010. (Cité en page 43.)

[Dugué 2012] Nicolas Dugué et Anthony Perez. *Les capitalistes sociaux sur Twitter : détection via des mesures de similarité.* In Actes de EGC'2013 (Extraction et Gestion des Connaissances), pages 329–334, Toulouse, 2012. (Cité en page 47.)

[Dunning 1993] Ted Dunning. *Accurate methods for the statistics of surprise and coincidence.* Computational linguistics, vol. 19, no. 1, pages 61–74, 1993. (Non cité.)

[Duteil-Mougel 2004] Carine Duteil-Mougel. *Introduction à la sémantique interprétative.* Texto !, vol. 9, no. 4, pages 1–59, 2004. (Non cité.)

[Easley 2010] David Easley et Jon Kleinberg. Networks, crowds, and markets : reasoning about a highly connected world. Cambridge University Press, 2010. (Cité en pages 36 et 43.)

[Edgar 2004] Robert Edgar. *MUSCLE : multiple sequence alignment with high accuracy and high throughput.* Nucleic acids research, vol. 32, no. 5, pages 1792–1797, 2004. (Cité en page 107.)

[Efron 2010] Miles Efron. *Hashtag retrieval in a microblogging environment.* In SIGIR'10 - Proceeding of the 33rd international ACM SIGIR conference on research and development in information retrieval, pages 787–788, Geneva, 2010. (Non cité.)

[Ertoz 2003] Levent Ertoz, Michael Steinbach et Vipin Kumar. *Finding topics in collections of documents : a shared nearest neighbor approach.* In Clustering and information retrieval, pages 83–104. Kluwer Academic Publishers, 2003. (Cité en page 83.)

[Estivill-Castro 2002] Vladimir Estivill-Castro. *Why so many clustering algorithms : a position paper.* ACM SIGKDD explorations newsletter, vol. 4, no. 1, pages 65–75, 2002. (Cité en pages 53, 56 et 57.)

[Farkas 2007] Illés Farkas, Daniel Abel, Gergely Palla et Tamas Vicsek. *Weighted network modules.* New journal of physics, vol. 9, no. 6, 2007. (Cité en page 57.)

[Ferrer i Cancho 2001] Ramon Ferrer i Cancho et Richard Solé. *The small world of human language.* Proceedings of the royal society, vol. 268, no. 1482, pages 2261–2265, 2001. (Cité en page 77.)

[Ferret 1998] Olivier Ferret, Brigitte Grau et Nicolas Masson. *Thematic segmentation of texts : two methods for two kinds of texts.* In Actes de ACL-COLING'98, pages 392–396, Montreal, 1998. (Cité en page 76.)

[Ferret 2001] Olivier Ferret. *Utiliser des corpus pour amorcer une analyse thématique.* TAL, vol. 42, no. 2, pages 517–545, 2001. (Non cité.)

[Ferret 2002] Olivier Ferret. *Segmenter et structurer thématiquement des textes par lâutilisation conjointe de collocations et de la récurrence lexicale.* In TALN 2002 - Actes de la 9ème conférence sur le traitement automatique des langues naturelles, pages 155–165, Nancy, 2002. (Non cité.)

[Ferret 2004] Olivier Ferret. *Découvrir des sens de mots à partir d'un réseau de cooccurrences lexicales.* In TALN 2004 - Actes de la 11ème conférence sur le traitement automatique des langues naturelles, pages 183–194, Fès, 2004. (Cité en page 78.)

[Ferret 2006] Olivier Ferret. *Approches endogène et exogène pour améliorer la segmentation thématique de documents.* TAL, vol. 47, no. 2, 2006. (Cité en pages 78 et 83.)

[Firth 1957] John Rupert Firth. Papers in linguistics, 1934-1951. Oxford University Press, 1957. (Cité en page 79.)

[Fiscus 2002] Jonathan Fiscus et George Doddington. *Topic detection and tracking evaluation overview.* In Topic detection and tracking : event-based Information organization, pages 17–31. Springer, 2002. (Non cité.)

[Flajolet 1985] Philippe Flajolet et Nigel Martin. *Probabilistic counting algorithms for data base applications.* Journal of computer and system sciences, vol. 31, no. 2, pages 182–209, 1985. (Non cité.)

[Forestier 2009] Mathilde Forestier et Julien Velcin. *Un cadre formel pour la veille numérique sur la presse en ligne.* In EGC 2009 - Atelier veille numérique, Strasbourg, 2009. (Non cité.)

[Fortunato 2004] Santo Fortunato, Vito Latora et Massimo Marchiori. *Method to find community structures based on information centrality.* Physical review E, vol. 70, no. 5, 2004. (Non cité.)

[Fortunato 2010] Santo Fortunato. *Community detection in graphs.* Physics reports, vol. 486, no. 3-5, pages 75–174, 2010. (Cité en page 53.)

[Frakes 1992] William Frakes. *Stemming algorithms.* In Information retrieval, pages 131–160. Prentice-Hall, Inc., 1992. (Non cité.)

[Frawley 2006] William Frawley, Erin Eschenroeder, Sarah Mills et Thao Nguyen. The expression of modality. Mouton de Gruyter, 2006. (Cité en page 112.)

[Froissart 2003] Pascal Froissart. *Rumeurs sur Internet : un champ d'investigation pour l'entreprise.* Revue francophone de @management, 2003. (Non cité.)

[Froissart 2004] Pascal Froissart. *Des théories sur la rumeur : pour quoi faire ?* Les cahiers du GRÉDAM, 2004. (Cité en page 16.)

[Froissart 2007] Pascal Froissart. *Buzz, bouffées d'audience et rumeur sur Internet.* Médiamorphoses, 2007. (Cité en pages 16 et 17.)

[Frutiger 1983] Adrian Frutiger. L'Homme et ses signes - signes, symboles, signaux. 1983. (Non cité.)

[Fukumoto 2000] Fumiyo Fukumoto et Yoshimi Suzuki. *Event tracking based on domain dependency.* In Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '00, pages 57–64, New-York City, 2000. ACM Press. (Cité en page 98.)

[Furnas 1988] George Furnas, Scott Deerwester, Susan Dumais, Thomas Landauer, Richard Harshman, Lynn Streeter et Karen Lochbaum. *Information retrieval using a singular value decomposition model of latent semantic structure.* In SIGIR'88 - Proceedings of the 11th annual international ACM SIGIR conference on research and development in information retrieval, pages 465–480, Grenoble, 1988. (Non cité.)

[Gale 1992] William Gale, Kenneth Ward Church et David Yarowsky. *A method for disambiguating word senses in a large corpus.* Computers and the Humanities, vol. 26, pages 415–439, 1992. (Cité en page 79.)

[Garfield 1955] Eugene Garfield. *Citation indexes for science : a new dimension in documentation through association of ideas.* Science, vol. 122, no. 3159, pages 108–111, 1955. (Cité en page 27.)

[Garfield 1972] Eugene Garfield. *Citation analysis as a tool in journal evaluation.* Science, vol. 178, no. 4060, pages 471–479, 1972. (Cité en page 27.)

[Gaume 2006] Bruno Gaume. *Cartographier la forme du sens dans les petits mondes lexicaux.* In JADT 2006 - Proceedings of the 8th international conference on the statistical analysis of textual data, Besançon, 2006. (Cité en page 78.)

[Geffroy 1973] Annie Geffroy, Pierre Lafon et Maurice Tournier. *L'indexation minimale - Plaidoyer pour une non-lemmatisation.* In Communication au colloque sur l'analyse des corpus linguistiques : "Problèmes et méthodes de l'indexation maximale", Strasbourg, 1973. (Cité en page 71.)

[Geraci 2006] Filippo Geraci, Marco Pellegrini, Marco Maggini et Fabrizio Sebastiani. *Cluster generation and cluster labelling for web snippets : a fast and accurate hierarchical solution.* In SPIRE'06 - Proceedings of the 13th international conference on string processing and information retrieval, pages 25–36, Glasgow, 2006. (Non cité.)

[Ghitalla 2004] Franck Ghitalla. *La géographie des agrégats de documents sur le Web.* Rapport technique, 2004. (Cité en pages 2 et 37.)

[Gibson 1998] David Gibson, Jon Kleinberg et Prabhakar Raghavan. *Inferring web communities from link topology.* In HYPERTEXT'98 - Proceedings of the 9th ACM conference on hypertext and hypermedia, pages 225–234, Pittsburgh, 1998. (Non cité.)

[Girvan 2002] Michelle Girvan et Mark Newman. *Community structure in social and biological networks.* Proceedings of the national academy of sciences, vol. 99, no. 12, pages 7821–7826, 2002. (Cité en pages 54, 56 et 72.)

[Gladwell 2002] Malcolm Gladwell. The tipping point : how little things can make a big difference. Back Bay Books, 2002. (Cité en page 26.)

[Gordon 2003] Thomas Gordon. Éduquer sans punir - Apprendre l'autodiscipline aux enfants. Marabout, Éditions d édition, 2003. (Cité en page 21.)

[Gosh 2012] Saptarshi Gosh, Bimal Viswanath, Farshad Kooti, Naveen Kumar Sharma, Gautam Korlam, Fabricio Benevenuto, Niloy Ganguly et Krishna Gummadi. *Understanding and combating link farming in the Twitter social network.* In Proceedings of the 21st international conference on World Wide Web (WWW'12), pages 61–70, 2012. (Cité en page 47.)

[Goyal 2010] Amit Goyal, Francesco Bonchi et Laks Lakshmanan. *Learning influence probabilities in social networks.* In WSDM'10 - Proceedings of the 3rd ACM international conference on web search and data mining, pages 241–250, New-York City, 2010. (Non cité.)

[Granovetter 1973] Mark Granovetter. *The strength of weak ties.* The American journal of sociology, vol. 78, no. 6, pages 1360–1380, 1973. (Cité en pages 33 et 64.)

[Grefenstette 1994] Greg Grefenstette. *Corpus derived first, second and third-order word affinities.* In Proceedings of EURALEX'94, pages 279–290, Amsterdam, 1994. (Cité en pages 80 et 84.)

[Gregory 2009] Steve Gregory. *Finding overlapping communities using disjoint community detection algorithms.* Complex networks, vol. 207, pages 47–61, 2009. (Cité en page 55.)

[Gregory 2010] Steve Gregory. *Finding overlapping communities in networks by label propagation.* New journal of physics, vol. 12, no. 10, 2010. (Cité en pages 55 et 57.)

[Guimerà 2005] Roger Guimerà et Luis Amaral. *Functional cartography of complex metabolic networks.* Nature, vol. 433, pages 895–900, 2005. (Non cité.)

[Guiraud 1959] Pierre Guiraud. Problèmes et méthodes de la statistique linguistique. Presses Universitaires de France, 1959. (Non cité.)

[Gyongyi 2004] Zoltan Gyongyi, Hector Garcia-Molina et Jan Pedersen. *Combating web spam with TrustRank.* In VLDB'04 - Proceedings of the 30th international conference on very large data bases, volume 30, pages 576–587, Toronto, 2004. (Cité en page 46.)

[Habert 1985] Benoit Habert. *L'analyse des formes "spécifiques".* Mots, vol. 11, pages 127–154, 1985. (Non cité.)

[Habert 1997] Benoit Habert, Adeline Nazarenko et André Salem. Les linguistiques de corpus. Armand Colin/Masson, 1997. (Non cité.)

[Habert 2000] Benoit Habert. *Des corpus représentatifs : de quoi, pour quoi, comment ?* Cahiers de l'université de Perpignan, vol. 31, pages 11–58, 2000. (Non cité.)

[Habert 2002] Benoit Habert et Pierre Zweigenbaum. *Contextual acquisition of information categories : what has been done and what can be done automatically ?* The legacy of Zellig Harris : language and information into the 21st century, vol. 2, pages 203–231, 2002. (Non cité.)

[Halliday 1976] Michael Halliday et Ruqaiya Hasan. Cohesion in English. Longman Pub Group, 1976. (Cité en page 79.)

[Hammouda 2004] Khaled Hammouda et Mohamed Kamel. *Efficient phrase-based document indexing for web document clustering.* IEEE transactions on knowledge and data engineering, vol. 16, no. 10, pages 1279–1296, 2004. (Non cité.)

[Harel 2001] David Harel et Yehuda Koren. *On clustering using random walks.* Lecture notes in computer science, vol. 2245, pages 18–41, 2001. (Non cité.)

[Harris 1957] Zellig Harris. *Cooccurrence and transformation in linguistic structure.* Language, vol. 33, pages 283–340, 1957. (Cité en page 79.)

[Haveliwala 2002] Taher H. Haveliwala. *Topic-sensitive PageRank.* In WWW'02 - Proceedings of the 11th international conference on world wide web, page 517, Honolulu, Mai 2002. ACM Press. (Cité en pages 35, 36 et 45.)

[Hearst 1997] Marti Hearst. *TextTiling : segmenting text into multi-paragraph subtopic passages.* Computational linguistics, vol. 23, no. 1, 1997. (Non cité.)

[Henzinger 2005a] Monika Henzinger. *Hyperlink analysis on the world wide web.* In HYPERTEXT'05 - Proceedings of the 16th ACM conference on hypertext and hypermedia, pages 1–3, Salzburg, Septembre 2005. ACM Press. (Cité en page 27.)

[Henzinger 2005b] Monika Henzinger. *Hyperlink analysis on the world wide web.* In Proceedings of the sixteenth ACM conference on Hypertext and hypermedia - HYPERTEXT '05, pages 1–3, New York, New York, USA, Septembre 2005. ACM Press. (Non cité.)

[Higgins 1988] Desmond Higgins et Paul Sharp. *CLUSTAL : a package for performing multiple sequence alignment on a micro computer.* Gene, vol. 73, no. 1, pages 237–244, 1988. (Non cité.)

[Hofmann 2002] Thomas Hofmann. *Probabilistic latent semantic indexing.* In SIGIR'99 - Proceedings of the 22nd annual international ACM SIGIR conference on research and development in information retrieval, Berkeley, 2002. (Cité en pages 74 et 81.)

[Hotho 2003] Andreas Hotho, Steffen Staab et Gerd Stumme. *Wordnet improves text document clustering.* In SIGIR'03 - Semantic web workshop, Toronto, 2003. (Non cité.)

[Huberman 1998] Bernardo Huberman, Peter Pirolli, James Pitkow et Rajan Lukose. *Strong regularities in world wide web surfing.* Science, vol. 280, pages 95–97, 1998. (Non cité.)

[Huberman 2001] Bernardo Huberman. The laws of the web : patterns in the ecology of information. The MIT Press, 2001. (Non cité.)

[Ide 1998] Nancy Ide et Jean Véronis. *Word sense disambiguation : the state of the art.* Computational linguistics, vol. 24, pages 1–40, 1998. (Non cité.)

[Irving 2004] Robert Irving. *Plagiarism and collusion detection using the Smith-Waterman algorithm*. Rapport technique, 2004. (Cité en page 106.)

[Jackiewicz 2006] Agata Jackiewicz. *Relations intersubjectives dans les discours rapportés*. TAL, vol. 47, no. 2, 2006. (Cité en page 103.)

[Jackoway 2011] Alan Jackoway, Hanan Samet et Jagan Sankaranarayanan. *Identification of live news events using Twitter*. In LBSN'11 - Proceedings of the 3rd ACM SIGSPATIAL international workshop on location-based social networks, pages 25–32, Chicago, 2011. (Non cité.)

[Jacquenet 2004] François Jacquenet, Christine Largeron et Stéphanie Chapaux. *Veille technologique assistée par la fouille de textes*. Revue des nouvelles technologies de l'information, vol. 2, pages 429–440, 2004. (Cité en page 68.)

[Jacquenet 2005] François Jacquenet et Christine Largeron. *Extraction automatique d'information inattendue à partir de textes*. Revue des nouvelles technologies de l'information, pages 235–252, 2005. (Non cité.)

[Jakobiak 1990] François Jakobiak. Pratique de la veille technologique. Editions d'Organisation, 1990. (Cité en page 6.)

[Jarvis 1973] R.A. Jarvis et E.A. Patrick. *Clustering Using a Similarity Measure Based on Shared Near Neighbors*. IEEE Transactions on Computers, vol. C-22, no. 11, pages 1025–1034, Novembre 1973. (Cité en pages 82 et 83.)

[Java 2007] Akshay Java, Xiaodan Song, Tim Finin et Belle Tseng. *Why we twitter : understanding microblogging usage and communities*. In WebKDD/SNA-KDD'07 - Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on web mining and social network analysis, pages 56–65, 2007. (Cité en page 44.)

[Jobbins 1998] Amanda Jobbins et Lindsay Evett. *Text segmentation using reiteration and collocation*. In COLING'98 - Proceedings of the 17th international conference on computational linguistics, pages 614–618, 1998. (Non cité.)

[Karinthy 1929] Frigyes Karinthy. Chains (Láncszemek). 1929. (Cité en page 32.)

[Katz 1953] Leo Katz. *A new status index derived from sociometric analysis*. Psychometrika, vol. 18, no. 1, pages 39–43, 1953. (Cité en pages 26 et 33.)

[Katz 1955] Elihu Katz, Paul Lazarsfeld et Elmo Roper. Personal influence : the part played by people in the flow of mass communications. Macmillan, 1955. (Non cité.)

[Kempe 2003] David Kempe, Jon Kleinberg et Eva Tardos. *Maximizing the spread of influence through a social network*. In KDD'03 - Proceedings of the 9th ACM SIGKDD international conference on knowledge discovery and data mining, pages 137–146, Washington DC, 2003. (Non cité.)

[Kessler 1963] Maxwell Mirton Kessler. *Bibliographic coupling between scientific papers*. Journal of the American society for information science, vol. 14, no. 1, pages 10–25, 1963. (Cité en page 27.)

[Kleinberg 1999a] Jon M. Kleinberg. *Authoritative sources in a hyperlinked environment.* Journal of the ACM, vol. 46, no. 5, pages 604–632, Septembre 1999. (Cité en pages 26, 36, 40 et 44.)

[Kleinberg 1999b] Jon M. Kleinberg. *Authoritative sources in a hyperlinked environment.* Journal of the ACM, vol. 46, no. 5, pages 604–632, Septembre 1999. (Cité en pages 26 et 47.)

[Kleinberg 1999c] Jon M. Kleinberg. *Hubs, authorities, and communities.* ACM computing surveys, vol. 31, no. 4es, pages 5–es, Décembre 1999. (Non cité.)

[Kleinberg 1999d] Jon M. Kleinberg, Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan et Andrew S. Tomkins. *The web as a graph : measurements, models, and methods.* In COCOON'99 - Proceedings of the 5th annual international conference on computing and combinatorics, pages 1–17, Tokyo, Juillet 1999. (Cité en page 27.)

[Kleinberg 2000] Jon Kleinberg. *The small-world phenomenon : an algorithm perspective.* In Proceedings of the 32nd annual ACM symposium on theory of computing (STOC '00), pages 163–170, 2000. (Non cité.)

[Kleinberg 2009] Jon Kleinberg. *Meme-tracking, diffusion, and the flow of on-line information (Video : http://videolectures.net/icwsm09_kleinberg_mtdfoi/)*, 2009. (Cité en page 99.)

[Kovach 1999] Bill Kovach et Tom Rosenstiel. Warp Speed : America in the Age of Mixed Media. The Century Foundation, 1999. (Cité en page 17.)

[Kozima 1993] Hideki Kozima. *Text segmentation based on similarity between words.* In ACL'93 - Proceedings of the 31st annual meeting on association for computational linguistics, pages 286–288, Columbus, 1993. (Cité en page 75.)

[Kritikopoulos 2006] Apostolos Kritikopoulos, Martha Sideri et Iraklis Varlamis. *BlogRank : ranking weblogs based on connectivity and similarity features.* In AAA-IDEA'06 - Proceedings of the 2nd international workshop on advanced architectures and algorithms for internet delivery and applications, page 8, Pisa, Octobre 2006. ACM Press. (Cité en page 27.)

[Krovetz 1993] Robert Krovetz. *Viewing morphology as an inference process.* In SIGIR'93 - Proceedings of the 16th annual international ACM SIGIR conference on research and development in information retrieval, pages 191–202, Pittsburgh, 1993. (Cité en page 70.)

[Kumar 1999] Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan et Andrew Tomkins. *Trawling the web for emerging cyber-communities.* In WWW'99 - Proceedings of the 8th international conference on world wide web, pages 1481–1493, Toronto, 1999. (Non cité.)

[Kumar 2000] Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, D Sivakumar, Andrew Tomkins et Eli Upfal. *The Web as a graph.* In PODS'00 - Proceedings of the nineteenth ACM SIGMOD-SIGACT-SIGART symposium on principles of database systems, pages 1–10, Dallas, 2000. ACM Press. (Cité en page 27.)

[Kumar 2005] Ravi Kumar, Jasmine Novak, Prabhakar Raghavan et Andrew Tomkins. *On the bursty evolution of blogspace.* World wide web, vol. 8, no. 2, pages 159–178, 2005. (Non cité.)

[Kumar 2006a] Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan et Andrew Tomkins. *Core algorithms in the CLEVER system.* ACM Transactions on Internet Technology, vol. 6, no. 2, pages 131–152, Mai 2006. (Cité en page 40.)

[Kumar 2006b] Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan et Andrew Tomkins. *Core algorithms in the CLEVER system.* ACM transactions on internet technology, vol. 6, no. 2, pages 131–152, Mai 2006. (Non cité.)

[Kumpula 2008] Jussi Kumpula, Mikko Kivelä, Kimmo Kaski et Jari Saramäki. *Sequential algorithm for fast clique percolation.* Physical review E, vol. 78, no. 2, 2008. (Cité en page 57.)

[Kurland 2005] Oren Kurland et Lillian Lee. *PageRank without hyperlinks : structural re-ranking using links induced by language models.* In SIGIR'05 - Proceedings of the 28th annual international ACM SIGIR conference on research and development in information retrieval, page 306, Salvador, Août 2005. ACM Press. (Non cité.)

[Kurland 2006] Oren Kurland et Lillian Lee. *Respect my authority ! HITS without hyperlinks, utilizing cluster-based language models.* In SIGIR'06 - Proceedings of the 29th annual international ACM SIGIR conference on research and development in information retrieval, page 83, Seattle, Août 2006. ACM Press. (Non cité.)

[Kwak 2010] Haewoon Kwak, Changhyun Lee, Hosung Park et Sue Moon. *What is Twitter, a social network or a news media ?* In WWW'10 - Proceedings of the 19th international conference on world wide web, pages 591–600, Raleigh, 2010. (Cité en page 46.)

[Labbe 2001] Cyril Labbe et Dominique Labbe. *Que mesure la spécificité du vocabulaire ?* Lexicometrica, vol. 3, 2001. (Non cité.)

[Labbe 2003] Cyril Labbe et Dominique Labbe. *La distance intertextuelle.* Corpus, vol. 2, 2003. (Cité en page 80.)

[Lafon 1980] Pierre Lafon. *Sur la variabilité de la fréquence des formes dans un corpus.* Mots, vol. 1, pages 127–165, 1980. (Cité en page 98.)

[Lafon 1981a] Pierre Lafon. *Analyse lexicométrique et recherche des cooccurrences.* Mots, vol. 3, pages 95–148, 1981. (Non cité.)

[Lafon 1981b] Pierre Lafon. *Statistiques des localisations des formes d'un texte.* Mots, vol. 2, pages 157–188, 1981. (Non cité.)

[Lafon 1983] Pierre Lafon et André Salem. *L'inventaire des segments répétés d'un texte.* Mots, vol. 6, pages 161–177, 1983. (Non cité.)

[Lafon 1984] Pierre Lafon. Dépouillements et statistiques en lexicométrie. 1984. (Cité en page 80.)

[Lancichinetti 2008] Andrea Lancichinetti, Santo Fortunato et Filippo Radicchi. *Benchmark graphs for testing community detection algorithms.* Physical review E, vol. 78, no. 4, 2008. (Cité en page 56.)

[Lancichinetti 2009] Andrea Lancichinetti, Santo Fortunato et Janos Kertész. *Detecting the overlapping and hierarchical community structure in complex networks.* New journal of physics, vol. 11, no. 3, 2009. (Non cité.)

[Lauf 2011] Aurélien Lauf, Leila Khouas et Mathieu Valette. *Calcul de l'autorité des pages Web au sein de leurs communautés respectives – Propositions pour une contextualisation de l'information.* In IC 2011 - Atelier ExCoCo, Chambéry, 2011. (Non cité.)

[Lauf 2012] Aurélien Lauf, Mathieu Valette et Leila Khouas. *Analyse du graphe des cooccurrents de deuxième ordre pour la classification non-supervisée de documents.* In JADT 2012 - Proceedings of the 12th international conference on the statistical analysis of textual data, pages 577–589, Liège, 2012. (Non cité.)

[Lauf 2013] Aurélien Lauf, Mathieu Valette et Leila Khouas. *Analyzing variation patterns in quotes over time.* In CICLing'13 - Proceedings of the 14th International Conference on Computational Linguistics and interlligent text processing, Samos, 2013. (Non cité.)

[Le Deuff 2006] Olivier Le Deuff. *Autorité et pertinence vs popularité et influence : réseaux sociaux sur Internet et mutations institutionnelles.* In Séminaire des doctorants du Cersic, Rennes, 2006. (Cité en pages 21, 26 et 46.)

[Leavitt 2009] Alex Leavitt, Evan Burchard, David Fisher et Sam Gilbert. *The influentials : new approaches for analyzing influence on Twitter.* Rapport technique, 2009. (Cité en pages 44 et 45.)

[Lebart 1994] Ludovic Lebart et André Salem. Statistique textuelle. Dunod, 1994. (Cité en page 98.)

[Leblanc 2006] Jean-Marc Leblanc et William Martinez. *L'analyse contrastive des réseaux de cooccurrence.* In JADT 2006 - Proceedings of the 8th international conference on the statistical analysis of textual data, Besançon, 2006. (Non cité.)

[Lee 1999] Daniel Lee et Sebastian Seung. *Learning the parts of objects by non-negative matrix factorization.* Nature, vol. 401, no. 6755, pages 788–791, 1999. (Cité en pages 73 et 81.)

[Lee 2001] Daniel Lee et Sebastian Seung. *Algorithms for non-negative matrix factorization.* Advances in neural information processing systems, vol. 13, no. 1, pages 556–562, 2001. (Cité en page 73.)

[Lee 2008] Ryong Lee, Daisuke Kitayama et Kazutoshi Sumiya. *Web-based evidence excavation to explore the authenticity of local events.* In WICOW'08 - Proceeding of the 2nd ACM workshop on information credibility on the web, page 63, Napa Valley, Octobre 2008. ACM Press. (Cité en page 27.)

[Lempel 2000] R. Lempel et S. Moran. *The stochastic approach for link-structure analysis (SALSA) and the TKC effect.* Computer networks, vol. 33, no. 1-6, pages 387–401, Juin 2000. (Cité en page 41.)

[Leskovec 2008a] Jure Leskovec, Lars Backstrom, Ravi Kumar et Andrew Tomkins. *Microscopic Evolution of Social Networks.* In ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), Las Vegas, 2008. (Cité en page 56.)

[Leskovec 2008b] Jure Leskovec et Eric Horvitz. *Planetary-scale views on a large instant-messaging network.* In Proceedings of the 17th International World Wide Web Conference (WWW2008), Beijing, 2008. (Cité en page 32.)

[Leskovec 2009] Jure Leskovec, Lars Backstrom et Jon Kleinberg. *Meme-tracking and the dynamics of the news cycle.* In KDD'09 - Proceedings of the 15th ACM SIGKDD international conference on knowledge discovery and data mining, pages 497–506, Paris, 2009. (Cité en pages 99, 101, 103, 106 et 108.)

[Leskovec 2010] Jure Leskovec. *Meme-tracking and the dynamics of the news cycle (Video : http://www.youtube.com/watch?v=Ex6muzQzqI8)*, 2010. (Non cité.)

[Leung 2009] Ian Leung, Pan Hui, Pietro Lio et Jon Crowcroft. *Towards real-time community detection in large networks.* Physical review E, vol. 79, no. 6, 2009. (Cité en page 55.)

[Levenshtein 1966] Vladimir Levenshtein. *Binary codes capable of correcting deletions, insertions, and reversals.* Soviet physics doklady, vol. 10, pages 707–710, 1966. (Cité en page 102.)

[Lin 1998] Dekang Lin. *An information-theoretic definition of similarity.* In ICML'98 - Proceedings of the 15th international conference on machine learning, pages 296–304, Madison, 1998. (Non cité.)

[Lin 2001] Dekang Lin et Patrick Pantel. *Induction of semantic classes from natural language text.* In KDD'01 - Proceedings of the 7th ACM SIGKDD international conference on knowledge discovery and data mining, pages 317–322, San Francisco, 2001. (Cité en page 76.)

[Lin 2006] Yu-ru Lin, Hari Sundaram, Yun Chi, Jun Tatemura et Belle Tseng. *Discovery of blog communities based on mutual awareness.* In WWW'06 - Proceedings of the 3rd annual workshop on the weblogging ecosystems : aggregation, analysis and dynamics, Edinburgh, 2006. (Non cité.)

[Lloyd 1957] Stuart Lloyd. *Least squares quantization in PCM.* IEEE transactions on information theory, vol. 28, no. 2, pages 129–137, 1957. (Cité en page 53.)

[Luong 2010] Xuan Luong, Etienne Brunet, Dominique Longrée, Damon Mayaffre, Sylvie Mellet et Céline Poudat. *La cooccurrence, une relation asymétrique ?* In JADT 2010 - Proceedings of the 10th international conference on the statistical analysis of textual data, Rome, 2010. (Cité en page 79.)

[MacQueen 1967] James MacQueen. *Some methods for classification and analysis of multivariate observations.* In Proceedings of 5th Berkeley symposium on mathematical statistics and probability, pages 281–297, Berkeley, 1967. (Cité en pages 53 et 72.)

[Macskassy 1998] Sofus Macskassy, Arunava Banerjee, Brian D. Davison et Haym Hirsh. *Human performance on clustering web pages : a preliminary study.* In KDD'98 - Proceedings of the 4th ACM SIGKDD international conference on knowledge discovery and data mining, New-York City, 1998. (Cité en page 57.)

[Makino 2004] Kazuhisa Makino et Takeaki Uno. *New algorithms for enumerating all maximal cliques.* In SWAT'04 - Proceedings of the 9th Scandinavian workshop of algorithm theory, pages 260–272, Humlebæk, 2004. (Non cité.)

[Makkonen 2009] Juha Makkonen. *Semantic classes in topic detection and tracking.* Rapport technique, 2009. (Cité en page 96.)

[Manning 1999] Christopher Manning et Hinrich Schutze. Foundations of statistical natural language processing. MIT Press, 1999. (Cité en page 79.)

[Manning 2008] Christopher Manning, Prabhakar Raghavan et Hinrich Schutze. Introduction to information retrieval. Cambridge University Press, 2008. (Cité en page 13.)

[Martinez 2000] William Martinez. *Mise en évidence de rapports synonymiques par la méthode des cooccurrences.* In JADT 2000 - Proceedings of the 5th international conference on the statistical analysis of textual data, Lausanne, 2000. (Non cité.)

[Martinez 2003] William Martinez. *Contribution à une méthodologie de l'analyse des cooccurrences lexicales multiples dans les corpus textuels.* PhD thesis, Université de la Sorbonne nouvelle, 2003. (Cité en page 79.)

[Martinez 2008] William Martinez. *Répulsions lexicales : expériences autour de la cooccurrence négative.* In JADT 2008 - Proceedings of the 9th international conference on the statistical analysis of textual data, Lyon, 2008. (Non cité.)

[Mauceri 2007] Christian Mauceri. *Indexation et isotopie : vers une analyse interprétative des données textuelles.* PhD thesis, Ecole Nationale Supérieure des Télécommunications de Bretagne, 2007. (Cité en pages 71, 74 et 76.)

[Mayaffre 2008a] Damon Mayaffre. *De l'occurrence à l'isotopie - Les co-occurrences en lexicométrie.* Syntaxe et sémantique - Textes, documents numériques, corpus. Pour une science des textes instrumentée, vol. 9, pages 53–72, 2008. (Cité en page 79.)

[Mayaffre 2008b] Damon Mayaffre. *Quand "travail", "famille", "patrie" co-occurrent dans le discours de Nicolas Sarkozy. Etude de cas et réflexion théorique sur la co-occurrence.* In JADT 2008 - Proceedings of the 9th international conference on the statistical analysis of textual data, Lyon, 2008. (Cité en pages 79 et 82.)

[McBryan 1994] Oliver McBryan. *GENVL and WWWW : tools for taming the Web.* In Proceedings of the 1rst World Wide Web Conference, 1994. (Cité en page 40.)

[McGlohon 2008] Mary McGlohon. *Graph mining techniques for social media analysis (Video : http://videolectures.net/icwsm08_mcglohon_gmtsma/)*, 2008. (Non cité.)

[Mei 2005] Qiaozhu Mei et ChengXiang Zhai. *Discovering evolutionary theme patterns from text : an exploration of temporal text mining.* In KDD'05 - Proceedings of the 11th ACM SIGKDD international conference on knowledge discovery and data mining, pages 198–207, Chicago, 2005. (Non cité.)

[Mei 2007] Qiaozhu Mei, Xuehua Shen et ChengXiang Zhai. *Automatic labeling of multinomial topic models.* In Proceedings of the 13th ACM SIGKDD international conference on knowledge discovery and data mining, pages 490–499, San Jose, 2007. (Cité en page 86.)

[Mejova 2009] Yelena Mejova. *Event intensity tracking in weblog collections (Video : http://videolectures.net/icwsm09_mejova_eitwc/)*, 2009. (Non cité.)

[Meyer zu Eissen 2005] Sven Meyer zu Eissen. *The suffix tree document model revisited.* In I-KNOW'05 - Proceedings of the 5th international conference on knowledge management, Graz, 2005. (Non cité.)

[Meyer 1975] David Meyer, Roger Schaneveldt et Margaret Ruddy. *Loci of contextual effects on visual word-recognition.* In Attention and performance, chapitre 8, pages 98–118. Academic Press, Londres, 1975. (Cité en page 77.)

[Milgram 1967] Stanley Milgram. *The small world problem.* Psychology Today, vol. 1, no. 61, 1967. (Cité en page 32.)

[Miller 1995] George Miller. *WordNet : a lexical database for english.* Communications of the ACM, vol. 38, no. 11, pages 39–41, 1995. (Non cité.)

[Mori 2006] Masaki Mori, Takao Miura et Isamu Shioya. *Topic detection and tracking for news web pages.* In WI'06 - Proceedings of the 2006 IEEE/WIC/ACM international conference on web intelligence, pages 338–342, Hong Kong, 2006. (Non cité.)

[Morris 1991] Jane Morris et Graeme Hirst. *Lexical cohesion computed by thesaural relations as an indicator of the structure of text.* Computational linguistics, vol. 17, no. 1, pages 21–48, 1991. (Cité en page 75.)

[Morris 2004] Jane Morris et Graeme Hirst. *Non-classical lexical semantic relations.* In HLT-NAACL 2004 - Proceedings of the computational lexical semantics workshop, pages 46–51, Boston, 2004. (Cité en page 76.)

[Mount 2004] David Mount. Bioinformatics : sequence and genome analysis. Cold Spring Harbor Laboratory Press, 2004. (Cité en page 106.)

[Muller 1973] Charles Muller. Initiation aux méthodes de la statistique linguistique. Champion, 1973. (Non cité.)

[Nagmoti 2010] Rinkesh Nagmoti, Ankur Teredesai et Martine De Cock. *Ranking approaches for microblog search.* In WI-IAT'10 - Proceedings of the 2010 IEEE/WIC/ACM international conference on web intelligence and intelligent agent technology, pages 153–157, Toronto, 2010. (Non cité.)

[Najork 2007] Marc A. Najork, Hugo Zaragoza et Michael J. Taylor. *Hits on the web : how does it compare ?* In SIGIR'07 - Proceedings of the 30th annual international ACM SIGIR conference on research and development in information retrieval, page 471, Amsterdam, Juillet 2007. ACM Press. (Non cité.)

[Najork 2009] Marc Najork, Sreenivas Gollapudi et Rina Panigrahy. *Less is more : sampling the neighborhood graph makes SALSA better and faster.* In WSDM'09 - Proceedings of the 2nd ACM international conference on web search and data mining, page 242, Barcelona, Février 2009. ACM Press. (Non cité.)

[Needleman 1970] Saul Needleman et Christian Wunsch. *A general method applicable to the search for similarities in the amino acid sequence of two proteins.* Journal of molecular biology, vol. 48, no. 3, pages 443–453, 1970. (Cité en page 106.)

[Newman 2001] Mark Newman. *Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality.* Physical review E, vol. 64, no. 1, 2001. (Cité en page 54.)

[Newman 2004] Mark Newman. *Fast algorithm for detecting community structure in networks.* Physical review E, vol. 69, no. 6, 2004. (Non cité.)

[Newman 2008] Mark Newman. *The physics of networks.* Physics today, pages 33–38, 2008. (Non cité.)

[Ng 2001] Andrew Y. Ng, Alice X. Zheng et Michael I. Jordan. *Stable algorithms for link analysis.* In SIGIR'01 - Proceedings of the 24th annual international ACM SIGIR conference on research and development in information retrieval, pages 258–266, New Orleans, Septembre 2001. ACM Press. (Cité en page 39.)

[Nie 2006] Lan Nie, Brian D. Davison et Xiaoguang Qi. *Topical link analysis for web search.* In SIGIR'06 - Proceedings of the 29th annual international ACM SIGIR conference on research and development in information retrieval, page 91, Seattle, Août 2006. ACM Press. (Cité en pages 27, 35, 36, 41 et 42.)

[Nie 2007a] Lan Nie, Brian D. Davison et Baoning Wu. *From whence does your authority come ? : utilizing community relevance in ranking.* pages 1421–1426, Juillet 2007. (Non cité.)

[Nie 2007b] Lan Nie, Brian D. Davison et Baoning Wu. *From whence does your authority come ? Utilizing community relevance in ranking.* In AAAI'07 - Proceedings of the 22nd national conference on artificial intelligence, pages 1421–1426, Vancouver, Juillet 2007. (Non cité.)

[Nie 2007c] Lan Nie, Baoning Wu et Brian D. Davison. *Incorporating trust into web search*. Rapport technique, 2007. (Non cité.)

[Nie 2008] Lan Nie et Brian D. Davison. *Separate and inequal : preserving heterogeneity in topic authority flows*. In SIGIR'08 - Proceedings of the 31st annual international ACM SIGIR conference on research and development in information retrieval, page 443, Singapore, Juillet 2008. ACM Press. (Cité en page 42.)

[Noel 2012] Romain Noel. *Automatic relevant source discovery over the internet based on user profile*. In CORIA 2012 - Actes de la journée jeunes chercheurs, pages 401–406, Bordeaux, 2012. (Cité en page 42.)

[Notredame 2000] Cédric Notredame, Desmond Higgins et Jaap Heringa. *T-coffee : a novel method for fast and accurate multiple sequence alignment*. Journal of molecular biology, vol. 302, no. 1, pages 205–217, 2000. (Cité en page 107.)

[O'Connor 2010] Brendan O'Connor, Ramnath Balasubramanyan, Bryan Routledge et Noah Smith. *From tweets to polls : linking text sentiment to public opinion time series*. In ICWSM 2010 - Proceedings of the 4th international AAAI conference on weblogs and social media, Washington DC, 2010. (Non cité.)

[Ohsawa 1998] Yukio Ohsawa, Nels Benson et Masahiko Yachida. *KeyGraph : automatic indexing by co-occurrence graph based on building construction metaphor*. In ADL'98 - Proceedings of the advances in digital libraries conference, Santa Barbara, 1998. (Non cité.)

[Omodei 2012] Elisa Omodei, Thierry Poibeau et Jean-Philippe Cointet. *Multilevel modeling of quotation families morphogenesis*. In Proceedings of the 2012 ASE/IEEE international conference on social computing (SocialCom 2012), Amsterdam, 2012. (Cité en pages 101, 105, 106 et 110.)

[Osinski 2003] Stanislaw Osinski. *An algorithm for clustering of web search results*. PhD thesis, PoznaÅ University of Technology, 2003. (Non cité.)

[Page 1999] Lawrence Page, Sergey Brin, Motwani Rajeev et Terry Winograd. *The PageRank citation ranking : bringing order to the web*. Rapport technique, 1999. (Cité en pages 26 et 33.)

[Paice 1996] Chris Paice. *Method for evaluation of stemming algorithms based on error counting*. Journal of the American society for information science, vol. 47, no. 8, pages 632–649, 1996. (Cité en page 70.)

[Pal 2009] Anshika Pal, Deepak Singh Tomar et S.C Shrivastava. *Effective focused crawling based on content and link structure analysis*. International journal of computer science and information security (IJCSIS), vol. 2, no. 1, 2009. (Non cité.)

[Pal 2011] Aditya Pal et Scott Counts. *Identifying topical authorities in microblogs*. In WSDM'11 - Proceedings of the 4th ACM international conference on web search and data mining, pages 45–54, Hong Kong, 2011. (Non cité.)

[Palermo 1964] David Palermo et James Jenkins. Word associaiton norms. University of Minnesota Press, Minneapolis, 1964. (Cité en page 77.)

[Palla 2005] Gergely Palla, Imre Derényi, Illés Farkas et Tamas Vicsek. *Uncovering the overlapping community structure of complex networks in nature and society.* Nature, vol. 435, pages 814–818, 2005. (Cité en pages 55, 57, 58, 59, 75 et 82.)

[Palla 2007a] Gergely Palla, Albert-Laszlo Barabasi et Tamas Vicsek. *Quantifying social group evolution.* Nature, vol. 446, pages 664–667, 2007. (Cité en page 56.)

[Palla 2007b] Gergely Palla, Illés Farkas, Péter Pollner, Imre Derényi et Tamas Vicsek. *Directed network modules.* New journal of physics, vol. 9, no. 6, 2007. (Cité en page 57.)

[Palmer 1986] Frank Palmer. Mood and modality. Cambridge University Press, 1986. (Cité en page 112.)

[Palmer 2002] Christopher Palmer, Phillip Gibbons et Christos Faloutsos. *ANF : a fast and scalable tool for data mining in massive graphs.* In Proceedings of the 8th ACM SIGKDD international conference on knowledge discovery and data mining, pages 81–90, Edmonton, 2002. (Non cité.)

[Pariser 2012] Eli Pariser. The filter bubble : what the internet is hiding from you. Penguin édition, 2012. (Cité en page 25.)

[Pichon 1999] Ronan Pichon et Pascale Sébillot. *Différencier les sens des mots à l'aide du thème et du contexte de leurs occurrences : une expérience.* In TALN 1999 - Actes de la 6ème conférence sur le Traitement Automatique des Langues Naturelles, Cargèse, 1999. (Cité en page 78.)

[Pichon 2000] Ronan Pichon et Pascale Sébillot. *From corpus to lexicon : from contexts to semantic features.* Practical applications in language corpora (PALC'99), vol. 1, pages 375–389, 2000. (Cité en page 78.)

[Pincemin 1999a] Bénédicte Pincemin. *Diffusion ciblée automatique d'informations : conception et mise en oeuvre d'une linguistique textuelle pour la caractérisation des destinataires et des documents.* PhD thesis, Université Paris IV, 1999. (Cité en pages 73 et 74.)

[Pincemin 1999b] Bénédicte Pincemin. *Sémantique interprétative et analyses automatiques de textes : que deviennent les sèmes ?* Sémiotiques, vol. 17, pages 71–120, 1999. (Cité en page 81.)

[Pinski 1976] Gabriel Pinski et Francis Narin. *Citation influence for journal aggregates of scientific publications : theory, with application to the literature of physics.* Information processing & management, vol. 12, no. 5, pages 297–312, 1976. (Cité en pages 27 et 33.)

[Pirkola 2010] Ari Pirkola et Tuomas Talvensaari. *Addressing the limited scope problem of focused crawling using a result merging approach.* In SAC'10 - Proceedings of the 2010 ACM symposium on applied computing, pages 1735–1740, Lausanne, 2010. (Non cité.)

[Pons-Porrata 2003] Aurora Pons-Porrata, Rafael Berlanga-Llavori et José Ruiz-Shulcloper. *Building a hierarchy of events and topics for newspaper digital libraries.* In Proceedings of the 25th European conference on IR research (ECIR 2003), pages 588–596, Pisa, 2003. (Non cité.)

[Pons 2007] Pascal Pons. *Détection de communautés dans les grands graphes de terrain.* PhD thesis, Université Paris 7 - Denis Diderot, 2007. (Non cité.)

[Ponte 1997] Jay Ponte et Bruce Croft. *Text segmentation by topic.* In Proceedings of the 1rst European conference on research and advanced technology for digital libraries, 1997. (Non cité.)

[Porter 1980] Martin Porter. *An algorithm for suffix stripping.* Program, vol. 14, no. 3, pages 130–137, Décembre 1980. (Cité en pages 41 et 70.)

[Porter 2009] Mason Porter, Jukka-Pekka Onnela et Peter Mucha. *Communities in networks.* Notices of the American mathematical society, vol. 56, no. 9, pages 1082–1097, 2009. (Cité en page 53.)

[Portner 2009] Paul Portner. Modality. Oxford University Press, 2009. (Cité en page 112.)

[Poulard 2008a] Fabien Poulard. *Analyse quantitative et qualitative de citations extraites dâun corpus journalistique.* In Actes de RECITAL'08, Avignon, 2008. (Cité en pages 102 et 103.)

[Poulard 2008b] Fabien Poulard, Thierry Waszak, Nicolas Hernandez et Patrice Bellot. *Repérage de citations, classification des styles de discours rapporté et identification des constituants citationnels en écrits journalistiques.* In TALN 2008 - Actes de la 15ème conférence sur le traitement automatique des langues naturelles, Avignon, 2008. (Non cité.)

[Poulard 2009] Fabien Poulard, Stergos Afantenos et Nicolas Hernandez. *Nouvelles considérations pour la détection de réutilisation de texte.* In TALN 2009 - Actes de la 16ème conférence sur le traitement automatique des langues naturelles, Senlis, 2009. (Non cité.)

[Prime-Claverie 2002] Camille Prime-Claverie, Michel Beigbeder et Thierry Lafouge. *Clusterisation du web en vue d'extraction de corpus homogènes.* In INFORSID 2002 - Actes du 20ème congrès INFORSID, pages 229–242, Nantes, 2002. (Non cité.)

[Qi 2006] Xiaoguang Qi et Brian D. Davison. *Knowing a web page by the company it keeps.* In CIKM'06 - Proceedings of the 15th ACM international conference on information and knowledge management, page 228, Arlington, Novembre 2006. ACM Press. (Cité en page 42.)

[Qi 2007] Xiaoguang Qi, Lan Nie et Brian D. Davison. *Measuring similarity to detect qualified links.* In AIRWeb'07 - Proceedings of the 3rd international workshop on adversarial information retrieval on the web, page 49, Banff, Mai 2007. ACM Press. (Non cité.)

[Qi 2008] Xiaoguang Qi et Brian D. Davison. *Classifiers without borders : incorporating fielded text from neighboring web pages.* In SIGIR'08 - Proceedings of the 31st annual international ACM SIGIR conference on research and development in information retrieval, page 643, Singapore, Juillet 2008. ACM Press. (Cité en page 42.)

[Qi 2009] Xiaoguang Qi et Brian D. Davison. *Web page classification : features and algorithms.* ACM computing surveys, vol. 41, no. 2, 2009. (Cité en page 27.)

[Radicchi 2004] Filippo Radicchi, Claudio Castellano, Federico Cecconi, Vittorio Loreto et Domenico Parisi. *Defining and identifying communities in networks.* Proceedings of the national academy of sciences, vol. 101, no. 9, 2004. (Non cité.)

[Raghavan 2007] Usha Nandini Raghavan, Reka Albert et Soundar Kumara. *Near linear time algorithm to detect community structures in large-scale networks.* Physical review E, vol. 76, no. 3, 2007. (Cité en pages 55 et 57.)

[Ramage 2010] Daniel Ramage, Susan Dumais et Daniel Liebling. *Characterizing microblogs with topic models.* In ICWSM 2010 - Proceedings of the 4th international AAAI conference on weblogs and social media, Washington DC, 2010. (Non cité.)

[Rastier 1987] François Rastier. *Représentation du contenu lexical et formalismes de l'intelligence artificielle.* Langages, vol. 22, no. 87, pages 79–102, 1987. (Cité en page 82.)

[Rastier 1990] François Rastier. *La triade sémiotique, le trivium et la sémantique linguistique.* Nouveaux actes sémiotiques, vol. 9, pages 5–39, 1990. (Non cité.)

[Rastier 1996] François Rastier. *La sémantique des textes - Concepts et applications.* Journal of linguistics, vol. 16, pages 15–37, 1996. (Cité en page 74.)

[Rastier 1997a] François Rastier. *Défigements sémantiques en contexte.* In La locution, entre langue et usages, pages 305–329. ENS, 1997. (Non cité.)

[Rastier 1997b] François Rastier, Marc Cavazza et Anne Abeillé. Sémantique pour l'analyse. Dunod, 1997. (Non cité.)

[Rastier 2001a] François Rastier. Arts et sciences du texte. Presses Universitaires de France, 2001. (Non cité.)

[Rastier 2001b] François Rastier. Sémantique et recherches cognitives. Presses Universitaires de France, 2001. (Non cité.)

[Rastier 2002] François Rastier. *Enjeux épistémologiques de la linguistique de corpus.* In Actes des deuxièmes journées de linguistique de corpus, Lorient, 2002. (Non cité.)

[Reddy 2001] Krishna Reddy et Masaru Kitsuregawa. *An approach to relate the web communities through bipartite graphs.* In WISE'01 - Proceedings of the 2nd international conference on web information systems engineering, Kyoto, 2001. (Non cité.)

[Reutenauer 2010] Coralie Reutenauer, Evelyne Jacquey, Michelle Lecolle et Mathieu Valette. *Sémème au macroscope : genèse et variation sémiques d'une unité lexicale*. In JADT 2010 - Proceedings of the 10th international conference on the statistical analysis of textual data, Rome, 2010. (Non cité.)

[Richardson 2001] Matt Richardson et Pedro Domingos. *The intelligent surfer : probabilistic combination of link and content information in PageRank*. NIPS, pages 1441–1448, 2001. (Cité en pages 35 et 36.)

[Rizoiu 2010] Marian-Andrei Rizoiu, Julien Velcin et Jean-Hugues Chauchat. *Regrouper les données textuelles et nommer les groupes à l'aide des classes recouvrantes*. In EGC 2010 - Actes de la 10ème conférence extraction et gestion des connaissances, pages 561–572, Hammamet, 2010. (Cité en page 73.)

[Rogers 2003] Everett Rogers. Diffusion of innovations, 5th Edition. Free Press, 2003. (Cité en page 26.)

[Romaine 1992] Suzanne Romaine. *The evolution of linguistic complexity in pidgin and creole languages*. The evolution of human languages, pages 213–238, 1992. (Cité en page 78.)

[Rosell 2005] Magnus Rosell et Sumithra Velupillai. *The impact of phrases in Document clustering for Swedish*. In NODALIDA 2005 - Proceedings of the 15th nordic conference of computational linguistics, Joensuu, 2005. (Non cité.)

[Rosell 2009] Magnus Rosell. *Introduction to information retrieval and text clustering*. Rapport technique, 2009. (Non cité.)

[Rossignol 2005] Mathias Rossignol. *Acquisition sur corpus d'informations lexicales fondées sur la sémantique différentielle*. PhD thesis, Université Rennes 1, 2005. (Cité en pages 74, 78 et 86.)

[Saitou 1987] Naruya Saitou et Masatoshi Nei. *The neighbor-joining method : a new method for reconstructing phylogenetic trees*. Molecular biology and evolution, vol. 4, no. 4, pages 406–425, 1987. (Cité en page 107.)

[Salem 1988] André Salem. *Approches du temps lexical*. Mots, vol. 17, pages 105–143, 1988. (Non cité.)

[Salton 1975] Gerard Salton, Andrew Wong et Chung-Shu Yang. *A vector space model for automatic indexing*. Communications of the ACM, vol. 18, no. 11, pages 613–620, 1975. (Cité en pages 69 et 75.)

[Sankaranarayanan 2009] Jagan Sankaranarayanan, Hanan Samet, Benjamin Teitler, Michael Lieberman et Jon Sperling. *TwitterStand : news in tweets*. In GIS'09 - Proceedings of the 17th ACM SIGSPATIAL international conference on advances in geographic information systems, pages 42–51, Seattle, 2009. (Non cité.)

[Saussure 1916] Ferdinand Saussure. Cours de linguistique générale. Payot, Paris, 1916. (Cité en pages 71 et 80.)

[Savoy 1993] Jacques Savoy. *Stemming of French words based on grammatical categories*. Journal of the American society for information science, vol. 44, no. 1, pages 1–9, 1993. (Cité en page 70.)

[Savoy 2000] Jacques Savoy et Justin Picard. *Recherche documentaire sur le Web : les hyperliens sont-ils vraiment utiles ?* In JADT 2000 - Proceedings of the 5th international conference on the statistical analysis of textual data, Lausanne, 2000. (Cité en page 51.)

[Sayyadi 2009] Hassan Sayyadi, Matthew Hurst et Alexey Maykov. *Event detection and tracking in social streams.* In ICWSM 2009 - Proceedings of the 3rd international AAAI conference on weblogs and social media, San Jose, 2009. (Cité en page 98.)

[Schmid 1994] Helmut Schmid. *Probabilistic part-of-speech tagging using decision trees.* In Proceedings of the international conference on new methods in language processing, Manchester, 1994. (Non cité.)

[Schneidermann 2004] Daniel Schneidermann, Pascal Froissart et Guillaume Soulez. *Rumeurs et emballements - Comment les décrire, comment leur résister ?* Médiamorphoses, 2004. (Non cité.)

[Schutze 1998] Hinrich Schutze. *Automatic word sense discrimination.* Computational linguistics, vol. 24, no. 1, pages 97–123, 1998. (Non cité.)

[Seglen 1992] Per Seglen. *The skewness of science.* Journal of the American society for information science, vol. 43, no. 9, pages 628–638, 1992. (Non cité.)

[Simmons 2011] Matthew Simmons, Lada Adamic et Eytan Adar. *Memes online : extracted, subtracted, injected, and recollected.* In ICWSM 2011 - Proceedings of the 5th international AAAI conference on weblogs and social media, Barcelona, 2011. (Cité en pages 101, 105 et 109.)

[Slodzian 2000] Monique Slodzian. *WordNet : what about its linguistic relevancy ?* In Proceedings of the 12th EKAW international conference (EKAW 2000), Juan-les-Pins, 2000. (Cité en page 76.)

[Small 1973] Henry Small. *Co-citation in the scientific literature : a new measure of the relationship between two documents.* Journal of the American society for information science, vol. 24, no. 4, pages 265–269, 1973. (Cité en page 27.)

[Steinhaus 1957] Hugo Steinhaus. *Sur la division des corps matériels en parties.* Bulletin de l'académie polonaise des sciences, vol. 3, no. 4, pages 801–804, 1957. (Cité en page 53.)

[Stern 1902] William Stern. *Zur Psychologie der Aussage. Experimentelle Untersuchungen über Erinnerungstreue.* Zeitschrift für die gesamte Strafechtswissenschaft, vol. XXII, no. 2, 1902. (Cité en page 16.)

[Steyvers 2005] Mark Steyvers et Joshua Tenenbaum. *The large-scale structure of semantic networks : statistical analysis and a model of semantic growth.* Cognitive science, vol. 29, pages 41–78, 2005. (Non cité.)

[Stokes 2002] Nicola Stokes, Joe Carthy et Alan Smeaton. *Segmenting broadcast news streams using lexical chains.* In STarting AI Researchers Symposium (STAIRS 2002), pages 145–154, 2002. (Non cité.)

[Swan 2000] Russel Swan et James Allan. *Automatic generation of overview timelines*. In SIGIR'00 - Proceedings of the 23rd annual international ACM SIGIR conference on research and development in information retrieval, pages 49–56, 2000. (Cité en page 98.)

[Sze 2006] Sing-Hoi Sze, Yue Lu et Qingwu Yang. *A polynomial time solvable formulation of multiple sequence alignment*. Journal of computational biology, vol. 13, no. 2, pages 309–319, 2006. (Cité en page 107.)

[Tan 2006] Bin Tan, Xuehua Shen et ChengXiang Zhai. *Mining long-term search history to improve search accuracy*. In KDD'06 - Proceedings of the 12th ACM SIGKDD international conference on knowledge discovery and data mining, pages 718–723, Philadelphia, 2006. (Non cité.)

[Tauveron 2012] Matthias Tauveron. *De la co-occurrence généralisée à la variation du sens lexical*. In La cooccurrence : du fait statistique au fait textuel, Besançon, 2012. (Cité en page 78.)

[Teevan 2011] Jaime Teevan, Daniel Ramage et Merredith Ringel Morris. *#TwitterSearch : a comparison of microblog search and web search*. In WSDM'11 - Proceedings of the 4th ACM international conference on web search and data mining, pages 35–44, Hong Kong, 2011. (Non cité.)

[Thanopoulos 2002] Aristomenis Thanopoulos, Nikos Fakotakis et George Kokkinakis. *Comparative evaluation of collocation extraction metrics*. In LREC 2002 - Proceedings of the 3rd language resources evaluation conference, Athens, 2002. (Non cité.)

[Thompson 1994] Julie Thompson, Desmond Higgins et Toby Gibson. *CLUSTAL W : improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice*. Nucleic acids research, vol. 22, no. 22, pages 4673–4680, 1994. (Cité en page 107.)

[Tournier 1980] Maurice Tournier. *D'où viennent les fréquences de vocabulaire ? La lexicométrie et ses modèles*. Mots, vol. 1, pages 189–209, 1980. (Non cité.)

[Toyoda 2001] Masashi Toyoda et Masaru Kitsuregawa. *Creating a web community chart for navigating related communities*. In HYPERTEXT'01 - Proceedings of the 12th ACM conference on hypertext and hypermedia, pages 103–112, Arhus, 2001. (Non cité.)

[Travers 1969] Jeffrey Travers et Stanley Milgram. *An experimental study of the small world problem*. Sociometry, vol. 32, no. 4, pages 425–443, 1969. (Cité en page 32.)

[Tricot 2006] Christophe Tricot. *Cartographie sémantique, des connaissances à la carte*. PhD thesis, Université de Savoie, 2006. (Non cité.)

[Tsukiyama 1977] Shuji Tsukiyama, Mikio Ide, Hiromu Ariyoshi et Isao Shirakawa. *A new algorithm for generating all the maximal independent sets*. SIAM journal on computing, vol. 6, no. 3, pages 505–517, 1977. (Cité en page 58.)

[Valette 2004] Mathieu Valette. *Sémantique interprétative appliquée à la détection automatique de documents racistes et xénophobes sur internet.* In Approches sémantiques du document numérique - Actes du 7ème colloque international sur le document électronique, pages 215–230, La Rochelle, 2004. (Non cité.)

[Valette 2006] Mathieu Valette, Alexander Estacio-Moreno, Étienne Petitjean et Evelyne Jacquey. *Éléments pour la génération de classes sémantiques à partir de définitions lexicographiques. Pour une approche sémique du sens.* In Actes de TALN2006, pages 357–366, Leuven, Belgique, 2006. (Cité en page 76.)

[Valette 2008] Mathieu Valette et Monique Slodzian. *Sémantique des textes et recherche d'information.* Revue française de linguistique appliquée, vol. 13, no. 1, pages 119–133, 2008. (Non cité.)

[Valette 2009] Mathieu Valette. *Approche textuelle du lexique.* Habilitation, Institut national des langues et civilisations orientales (INALCO), 2009. (Cité en page 76.)

[Valette 2010] Mathieu Valette. *Des textes au concept. Propositions pour une approche textuelle de la conceptualisation.* In IC 2010 - Actes des 21emes journées francophones dâingénierie des connaissances, pages 5–16, Nimes, 2010. (Cité en pages 71, 76 et 95.)

[Van Raan 2004] Anthony Van Raan. *Sleeping beauties in science.* Scientometrics, vol. 59, no. 3, pages 461–466, 2004. (Cité en page 27.)

[Velcin 2007] Julien Velcin et Jean-Gabriel Ganascia. *Topic extraction with AGAPE.* In ADMA'07 - Proceedings of the 3rd international conference on advanced data mining and applications, pages 377–388, Harbin, 2007. (Non cité.)

[Véronis 2003] Jean Véronis. *Cartographie lexicale pour la recherche d'information.* In TALN 2003 - Actes de la 10ème conférence sur le traitement automatique des langues naturelles, Batz-sur-Mer, 2003. (Cité en pages 78 et 83.)

[Véronis 2006] Jean Véronis et Emilie Guimier De Neef. *Le traitement des nouvelles formes de communication écrite.* In Compréhension automatique des langues et intéraction, pages 227–248. 2006. (Non cité.)

[Viprey 1997] Jean-Marie Viprey. *Dynamique du vocabulaire des Fleurs du Mal.* PhD thesis, 1997. (Cité en page 79.)

[Viprey 2006] Jean-Marie Viprey. *Structure non-séquentielle des textes.* Langages, vol. 163, no. 3, pages 71–85, 2006. (Non cité.)

[Waern 2004] Annika Waern. *User involvement in automatic filtering : an experimental study.* User modeling and user-adapted interaction, vol. 14, pages 201–237, 2004. (Non cité.)

[Wang 2009a] Jian Wang et Brian D. Davison. *Counting ancestors to estimate authority.* In Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval - SIGIR '09, page 658, New York, New York, USA, Juillet 2009. ACM Press. (Non cité.)

[Wang 2009b] Jian Wang et Brian D. Davison. *Counting ancestors to estimate authority.* In SIGIR'09 - Proceedings of the 32nd international ACM SIGIR conference on research and development in information retrieval, page 658, Boston, Juillet 2009. ACM Press. (Non cité.)

[Wang 2011] Yujing Wang, Xiaochuan Ni, Jian-Tao Sun, Yunhai Tong et Zheng Chen. *Representing document as dependency graph for document clustering.* In CIKM'11 - Proceedings of the 20th ACM international conference on information and knowledge management, pages 2177–2180, Glasgow, 2011. (Non cité.)

[Wasserman 1994] Stanley Wasserman et Katherine Faust. Social network analysis - Methods and applications. Cambridge University Press, 1994. (Cité en page 26.)

[Watts 1998] Duncan Watts et Steven Strogatz. *Collective dynamics of 'small-world' networks.* Nature, vol. 393, pages 440–442, 1998. (Cité en pages 32 et 77.)

[Watts 2004] Duncan Watts. Six degrees : the science of a connected age. 2004. (Non cité.)

[Watts 2007] Duncan Watts. *Challenging the influentials hypothesis.* WOMMA - Word of mouth marketing association - measuring word of mouth : current thinking on research and measurement of word of mouth marketing, vol. 3, pages 201–211, 2007. (Non cité.)

[Weiss 1996] Ron Weiss, Bienvenido Velez, Mark Sheldon, Chanathip Namprempre, Peter Szilagyi, Andrzej Duda et David Gifford. *HyPursuit : a hierarchical network search engine that exploits content-link hypertext clustering.* In Proceedings of the 7th ACM conference on hypertext, Bethesda, 1996. (Non cité.)

[Weng 2010] Jianshu Weng, Ee-Peng Lim, Jing Jiang et Qi He. *TwitterRank : finding topic-sensitive influential twitterers.* In WSDM'10 - Proceedings of the 3rd ACM international conference on web search and data mining, pages 261–270, New-York City, 2010. (Cité en pages 44 et 45.)

[White 1989] H D White et K W McCain. *Bibliometrics.* Annual review of information science and technology, vol. 24, pages 119–186, 1989. (Cité en page 27.)

[Widdows 2002] Dominic Widdows et Beate Dorow. *A graph model for unsupervised lexical acquisition.* In COLING'02 - Proceedings of the 19th international conference on computational linguistics, volume 1, pages 1–7, Taipei, Août 2002. Association for Computational Linguistics. (Cité en page 77.)

[Wilson 1983] Patrick Wilson. Second-hand knowledge. An inquiry into cognitive authority. 1983. (Cité en page 21.)

[Wu 2006a] Baoning Wu et Brian D. Davison. *Detecting semantic cloaking on the web.* In Proceedings of the 15th international conference on World Wide Web - WWW '06, page 819, New York, New York, USA, Mai 2006. ACM Press. (Cité en page 27.)

[Wu 2006b] Baoning Wu et Brian D. Davison. *Detecting semantic cloaking on the web*. In WWW'06 - Proceedings of the 15th international conference on world wide web, page 819, Edinburgh, Mai 2006. ACM Press. (Non cité.)

[Wu 2006c] Baoning Wu et Brian D. Davison. *Undue influence : eliminating the impact of link plagiarism on web search rankings*. In SAC'06 - Proceedings of the 2006 ACM symposium on applied computing, page 1099, Dijon, Avril 2006. ACM Press. (Non cité.)

[Yardi 2010] Sarita Yardi, Daniel Romero, Grant Schoenebeck et Danah Boyd. *Detecting spam in a Twitter network*. First monday, vol. 15, no. 1-4, 2010. (Non cité.)

[Yarowsky 1995a] David Yarowsky. *Unsupervised word sense disambiguation rivaling supervised methods*. In ACL'95 - Proceedings of the 33rd annual meeting on association for computational linguistics, pages 189–196, Cambridge, 1995. (Non cité.)

[Yarowsky 1995b] David Yarowsky. *Unsupervised word sense disambiguation rivaling supervised methods*. In Proceedings of the 33rd annual meeting on Association for Computational Linguistics (ACL '95), pages 189–196, 1995. (Non cité.)

[Yen 2005] Luh Yen, Denis Vanvyve, Fabien Wouters, François Fouss, Michel Verleysen et Marco Saerens. *Clustering using a random walk based distance measure*. In ESANN 2005 - Proceedings of the 13th European symposium on artificial neural networks, Bruges, 2005. (Non cité.)

[Zamir 1998] Oren Zamir et Oren Etzioni. *Web document clustering : a feasibility demonstration*. In SIGIR'98 - Proceedings of the 21st annual international ACM SIGIR conference on research and development in information retrieval, pages 46–54, Melbourne, 1998. (Non cité.)

[Zhong 2005] Shi Zhong. *Efficient online spherical k-means clustering*. In IJCNN'05 - Proceedings of the IEEE international joint conference on neural networks, pages 3180–3185, Montreal, 2005. (Cité en page 54.)

Aurélien LAUF

Propagation du buzz sur Internet
Identification, analyse, modélisation et représentation dans un contexte de veille

Résumé

S'inscrivant dans un contexte de veille et d'intelligence d'entreprise sur Internet, l'objectif de cette thèse est d'élaborer des outils et des méthodes permettant d'identifier, analyser, modéliser et représenter le cheminement des buzz sur Internet. Tout buzz a un ou plusieurs points d'origine : les sources primaires. L'information est ensuite relayée par des sources secondaires qui vont accélérer ou non la propagation en fonction de leur degré d'influence. Tout au long du cycle de vie du buzz, le contenu sémantique est amené à évoluer. La compréhension d'un buzz sur Internet passe ainsi par l'analyse de ce qui se dit et la qualification des émetteurs. Nos travaux s'axeront donc autour de deux types d'analyses complémentaires : une analyse topologique des sources (théorie des graphes et des réseaux) et une analyse du contenu textuel (linguistique de corpus).

Buzz, veille, intelligence d'entreprise, internet, graphe, autorité, clustering, cooccurrence, linguistique de corpus

Résumé en anglais

This thesis is in the context of strategic and competitive intelligence. Its goal is to develop tools and methods to identify, analyze, model and represent how buzz spread on the Internet. Any buzz has one or more starting point(s), i.e. primary source(s). The information is then passed on by secondary sources which may speed or slow down its spreading depending on their influence. Throughout the buzz lifecycle, the semantic content can evolve. To understand a buzz on the Internet, one needs to analyze what is said and qualify who speaks. This thesis will focus on two main points : a topological analysis of the sources (graph theory and networks), and an analysis of the textual content (corpus linguistics).

Buzz, competitive intelligence, internet, graph, authority, clustering, cooccurrence, corpus linguistics