Trajectoire de l'information dans les médias sociaux

Charles Huyghues-Despointes

N° d'ordre NNT : 2020LYSE2088

THESE de DOCTORAT DE L'UNIVERSITÉ DE LYON

Opérée au sein de

L'UNIVERSITÉ LUMIÈRE LYON 2

École Doctorale : ED 512

Informatique et Mathématiques

Discipline : Informatique

Soutenue publiquement le 29 septembre 2020, par :

Charles HUYGHUES-DESPOINTES

Trajectoire de l'information dans les médias sociaux.

Devant le jury composé de :

Pascal PONCELET, Professeur des universités, Université de Montpellier, Président

Juan Manuel TORRES-MORENO, Maître de conférences HDR, Université d'Avignon, Rapporteur

Clémence MAGNIEN, Directrice de recherche, CNRS, Examinatrice

Camille ROTH, Chargée de recherche, Centre Max Bloch, Examinatrice

Leila KHOUAS, Expert, Examinatrice

Sabine LOUDCHER, Professeure des universités, Université Lumière Lyon 2, Directrice de thèse

Julien VELCIN, Professeur des universités, Université Lumière Lyon 2, Co-Directeur de thèse

Contrat de diffusion

Thèse présentée pour obtenir le grade de

Docteur de l'Université Lumière Lyon 2

École Doctorale Informatique et Mathématiques (ED 512)

Laboratoire ERIC (EA 3083)

Discipline : Informatique

Trajectoire de l'information dans les médias sociaux

Charles HUYGHUES-DESPOINTES

Présentée et soutenue publiquement le 29 Septembre 2020 devant un jury composé de :

Pascal Poncelet, Professeur des Universités, Université de Montpellier	Rapporteur
Juan-Manuel Torres-Moreno, Maître de conférences, Université d'Avignon	Rapporteur
Clémence Magnien, Directrice de recherche CNRS, Sorbonne Université	Examinatrice
Camille Roth, Chargé de recherche CNRS, Centre Marc Bloch	Examinateur
Sabine Loudcher, Professeur des Universités, Université Lyon 2	Co-directrice
Julien Velcin, Professeur des Universités, Université Lyon 2	Co-directeur
Leila Khouas Ingénieure recherche Bertin IT, Montpellier	Invitée

Abstract

The work presented in this thesis, made in collaboration with the company *Bertin IT*, aims to study the way information spreads in social media text corpora.

We present the information Trajectory model, a way to depict information propagation when the pieces of information mutate alongside their propagation. Information propagation and information mutation are generally dissociated in the literature. The information Trajectory model makes the joint representation of these two phenomenons possible.

We describe a two-step method for computing an approximation of the information Trajectory. First, we approximate the propagation structure by computing coherent document chains. Then, we describe a way to exploit this structure in order to extract, characterize and label the propagating information pieces. We run two evaluation campains which experimentally show that our method is relevant.

We also describe our works for using the information Trajectory. From the computed Trajectory approximations, we can efficiently navigate in consequent corpora. We can detect and analyze subtlety between two similar information pieces with different propagation histories, and we can discover low-signal information inside the corpus.

keywords : information Trajectory, information extraction, information mutation, information propagation, social media analysis.

Résumé

Les travaux présentés dans cette thèse, réalisés en collaboration avec l'entreprise *Bertin IT*, ont pour objectif d'étudier la manière dont l'information chemine dans des corpus de documents tirés des médias sociaux.

Nous présentons le modèle de la Trajectoire de l'information, une manière de représenter la propagation d'informations qui mutent en même temps qu'elles se propagent dans des corpus de documents textuels. Dans la littérature, la question de la propagation est généralement dissociée de la question de la mutation de l'information. Le modèle de la Trajectoire permet de représenter les deux phénomènes conjointement ce qui est tout à fait innovant.

Nous détaillons une méthode pour approcher la Trajectoire de l'information en deux temps. Nous commençons par estimer sa structure de propagation à l'aide de la notion de chaînes cohérentes. Ensuite, nous décrivons une méthode pour extraire de cette structure les différentes informations qui se propagent ainsi que pour les nommer. Nous démontrons expérimentalement la pertinence de chaque méthode à l'aide de campagnes d'évaluation par des experts. Nous présentons également nos travaux pour exploiter les objets que nous construisons. Les trajectoires que nous calculons permettent par exemple de naviguer efficacement dans de grand corpus de documents, de détecter et d'analyser la nuance entre deux propagations d'informations similaires, et de découvrir des informations qui correspondent à des signaux faibles potentiels.

Mots-clefs : Trajectoire de l'information, extraction d'informations, mutation de l'information, propagation de l'information, analyse des médias sociaux.

Remerciements

À Liesse et à Aimée

Je tiens à remercier en premier lieu Sabine Loudcher, Julien Velcin et Leila Khouas pour m'avoir accompagné tout au long de cette thèse. Vos conseils, votre rigueur scientifique et votre gentillesse m'ont permis de la mener à bout et de surmonter les moments les plus difficiles. Je tiens également à remercier Pascal Poncelet et Juan-Manuel Torres-Moreno en tant que rapporteurs, et Clémence Magnien et Camille Roth en tant qu'examinateurs, pour avoir accepté d'évaluer ces travaux.

Je tiens à remercier l'ensemble de l'équipe *R&D* de *Bertin IT* basée à Montpellier pour m'avoir accueilli dans une ambiance de travail avec beaucoup de sérieux mais aussi de camaraderie, de bonne humeur, de chocolatines, et de babyfoot.

J'ai également travaillé au laboratoire ERIC et je tiens à remercier tous ses membres pour m'avoir reçu avec bonne humeur, ne manquant jamais de faire de ma venue un prétexte pour organiser un moment convivial en dehors des heures de travail.

Je remercie grandement les gens qui m'ont soutenu durant ces quatre années : mon père, ma mère, ma sœur, mon beau-frère, mes nièces, mes oncles et tantes et cousines, mais également mes amis, Benjamin, Delphine, Odin, Laetitia et Pauline.

Le doctorat est également une fin d'étude et je tiens à remercier ceux qui m'ont donné le goût d'aller au bout. Merci à Sebastien Dumortier de m'avoir fait faire mes premiers programmes informatiques. Merci à Bin-Minh Bui-Xuan et toute l'équipe APR du LIP6 pour m'avoir donné le goût de la recherche et la capacité à croire en moi.

Table des matières

Chapitre 1

Introduction

1.1 Contexte général

L'avènement d'Internet et la démocratisation du *World Wide Web*, ont permis une augmentation dans la publication de contenus et leur accessibilité. Chaque année, la quantité publiée quotidiennement d'articles, de vidéos, de contenus audio ne fait que croître. En 2016, un quotidien comme le *New York Times* publiait en moyenne 240 articles par jour, hors dépêches d'agences de presse, sur son site Web [1] là ou sa version papier ne comporte qu'une dizaine d'articles par édition. Ce chiffre, rapporté à l'ensemble des acteurs de l'information comme les journaux de presse ou les chaînes de télévision et de radio, autrement appelés **les médias**, témoigne de l'explosion des contenus publiés quotidiennement. De plus, le *Web* permet également à tous ses utilisateurs de publier du contenu, que ce soit à l'aide de sites individuels comme les blogs, ou des sites communautaires comme les forums ou les réseaux sociaux, ce qui forme un nouveau panel de médias numériques. Il est ainsi devenu courant pour de nombreux utilisateurs de lire quotidiennement du contenu via Internet, et également d'en publier ou de réagir à d'autres contenus publiés. Par exemple, en 2020, plus de 500 millions de messages sont échangés sur Twitter chaque jour [2].

1. Selon Robinson Meyer, journaliste pour *The Atlantic* : `https://www.theatlantic.com/technology/archive/2016/05/how-many-stories-do-newspapers-publish-per-day/483845/`

2. Selon le site *Internet live stats* `https://www.internetlivestats.com/twitter-statistics/`

Plusieurs millions d'utilisateurs, que ce soit leur métier ou non, écrivent et partagent chaque jour les informations pour lesquelles ils ont un intérêt, une opinion. Il s'agit d'une mine d'informations importante pour quiconque souhaite mesurer l'opinion des utilisateurs, comme un journaliste, se tenir en alerte sur des pratiques émergentes dans un secteur particulier, comme un ingénieur, ou communiquer efficacement le lancement d'un produit, comme un responsable marketing. L'analogie de la mine est particulièrement adéquate, les informations pertinentes pour un analyste étant enfouies parmi les nombreuses informations présentes dans chaque contenu, contenus eux-mêmes éparpillés dans un flux important de publications. Aussi, pour extraire et traiter l'information efficacement et rapidement, il est nécessaire d'utiliser et donc de concevoir des outils automatisés.

Cet afflux quotidien de contenus constitue un corpus dynamique conséquent des différents échanges d'information existants entre les utilisateurs du réseau. Il s'agit donc d'un témoignage important de la manière dont les informations les plus diverses sont échangées entre les internautes et en particulier de la manière dont elles sont reprises, recontextualisées ou récupérées par différents acteurs. Ainsi, les informations mutent, sont discutées, marquent les sujets et le vocabulaire utilisé par les utilisateurs du *Web*. Depuis son origine à aujourd'hui, un même évènement a pu être transmis et analysé selon de multiples angles, chacun ayant à son tour évolué, si bien que l'événement possède une véritable histoire de son traitement médiatique, composée de multiples ramifications. Un contenu, quant à lui, se compose de multiples informations, d'importance variable et mises en ordre par son ou ses auteurs à la lumière des informations préalables dont ils disposent. Le *Web* peut ainsi être vu comme un gigantesque métier à tisser l'information, produisant chaque jour une masse importante de contenus dont les motifs sont construits à l'aide des divers fils d'informations jusqu'alors disponibles. Le sujet central de cette thèse est de savoir comment détricoter ces motifs dans le cas particulier des contenus textuels. Il est question d'une part de retrouver l'histoire de la propagation des différentes informations d'un corpus, d'autre part, d'identifier les informations elles-mêmes. La

connaissance des cheminements pris par l'information, sa **trajectoire**, serait une matière originale pour analyser la manière dont elle se propage et pour analyser la manière dont elle évolue et se nuance sémantiquement.

1.2 Intérêt Industriel

Cette thèse s'est déroulée dans un contexte industriel (CIFRE [3]) au sein de l'entreprise Bertin IT [4] qui est, entre autres, spécialisée dans l'édition logicielle pour la veille stratégique. La thèse s'est déroulée au sein de l'équipe développant spécifiquement la plateforme de veille Web *AMI Enterprise Intelligence*, ou simplement *AMIEI* [5].

Le métier de la veille stratégique consiste à se tenir informé de l'évolution sur certains domaines d'importance capitale pour son organisation. Le chargé de veille délivre ainsi des rapports sur les nouvelles technologies, l'évolution du droit, ou de l'opinion, permettant aux différents décideurs de l'organisation qui l'emploie d'avoir un champ de vision à jour et éclairé. Il est donc crucial pour un chargé de veille de tirer parti de la grande quantité de documents publiés sur Internet chaque jour. La veille concurrentielle par exemple s'intéresse aux nouveaux produits des entreprises concurrentes, leur stratégie marketing, et l'opinion que les clients ont de ces produits. Elle s'intéresse également aux démissions et aux recrutements, l'important étant de ne pas manquer une opportunité ou une alerte.

La solution *AMIEI* est une plateforme collaborative permettant aux chargés de veille d'une entreprise d'effectuer, en collaboration avec les autres employés, l'entièreté du processus de veille. Le processus de veille complet de la plateforme est fourni en Figure 1.1. Nous détaillons les phases de collecte des informations, de capitalisation et d'analyse des documents.

3. Conventions Industrielles de Formation pour la REcherche.
4. `https://www.bertin-it.com/`.
5. Description du produit : `https://www.bertin-it.com/intelligence-numerique/solution-veille-strategique-intelligence-competitive/`

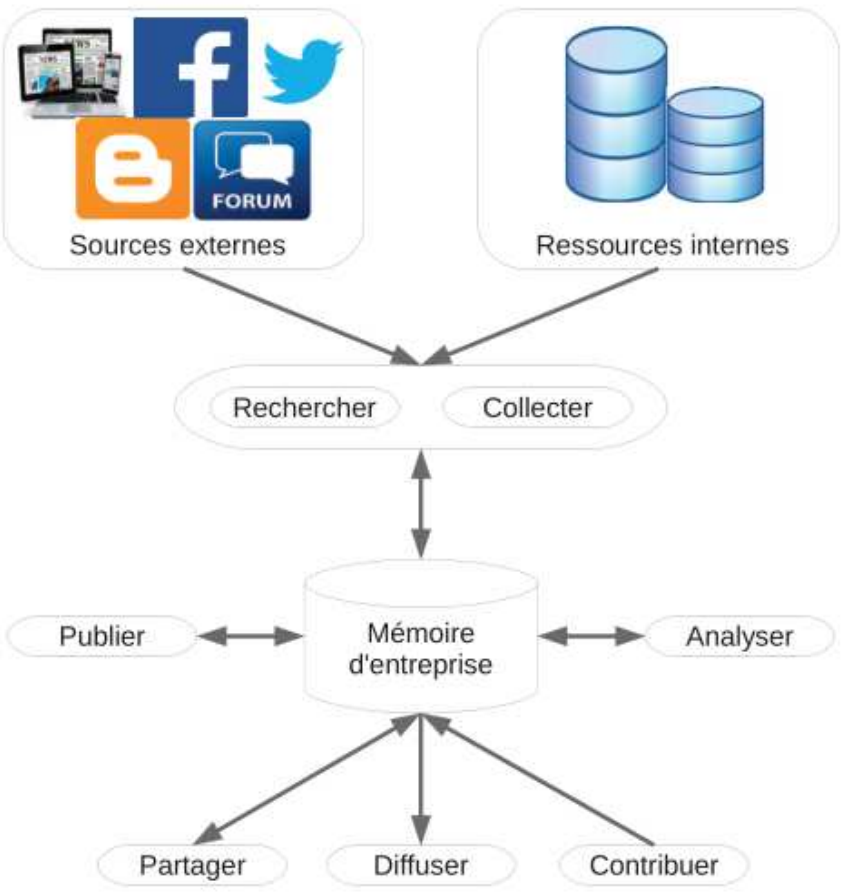


FIGURE 1.1 – Vision d'ensemble du processus de veille d'*AMIEI*.

Collecte des informations

La phase de collecte consiste à récupérer l'ensemble des contenus potentiellement pertinents pour le chargé de veille. La collecte peut se faire automatiquement, à l'aide d'un ensemble de règles de recherche et de sources de contenus déterminées et affinées par le chargé de veille. Elle peut également se faire manuellement pour collecter des documents internes, ou trouvés à la volée lors de la navigation sur Internet des collaborateurs de la plateforme. La collecte automatique se fait de manière récurrente, à une certaine fréquence adaptée à la fréquence de publications des sources.

Capitalisation

La capitalisation est la gestion de l'ensemble des contenus collectés. Il s'agit de la « Mémoire d'entreprise » qui constitue la base de données centrale de la veille. Cette

base est entretenue par les utilisateurs de la plateforme qui contribuent en croisant les différents contenus, en les commentant, les classant, les annotant, etc. Au fur et à mesure des collectes et de l'entretien de la mémoire d'entreprise, elle devient un véritable lieu de partage de connaissances et d'informations.

Analyse

Le but de l'analyse est d'exploiter et explorer les données acquises pour en extraire de l'information nouvelle qui servira au chargé de veille. Par exemple, pour découvrir l'évolution des tendances sur certains sujets, on peut extraire des entités nommées (personnes, lieux, organisations, concepts, ...), ainsi que les réseaux sous-jacents à certains sujets sur les réseaux sociaux, etc. L'analyse permet non seulement au chargé de veille de mieux comprendre ses corpus, mais également de fournir une visualisation synthétique aux décideurs.

C'est dans ce contexte que prennent place les travaux de cette thèse. Il s'agit de fournir aux chargés de veille un outil d'analyse leur permettant de suivre les cheminements de l'information dans les corpus capitalisés. D'une part, ces cheminements ont pour vocation d'aider le chargé de veille à naviguer au sein du corpus, en suivant la mutation de certaines informations. D'autre part, ils pourraient aider à détecter les nuances et les sujets mineurs qui co-existent, potentiellement dans des lieux restreints du corpus, avec le sujet de la veille. Ceci permet au chargé de veille d'avoir une lecture nouvelle et éclairée de l'ensemble de document qu'il a collecté.

1.3 Problématique et contributions

Les informations se propagent entre les individus. Ce faisant, les informations changent de contextes, sont réinterprétées, répétées, imitées, nous disons simplement qu'elles mutent. L'étude conjointe de la propagation de l'information et de sa mutation est peu voire non traitée dans l'état de l'art. La raison tient dans une expérience de pensée bien connue de l'ensemble des phénomènes qui s'inscrivent dans le temps : le bateau de Thésée. Il s'agit d'imaginer un bateau, dont on remplacerait une pièce

tous les jours, si bien qu'au bout d'un certain temps, toutes les planches seraient différentes de sa version originale. S'agit-il toujours du bateau de Thésée ? Une information mutant sans cesse n'est ainsi plus l'information originelle. Aussi, quel sens donner à la propagation d'un objet qui n'existe plus ? Les travaux portant sur la propagation de l'information admettent ainsi, en tout cas dans un premier temps, que l'information est suffisamment stable en différents lieux pour pouvoir discuter de sa propagation. La problématique centrale de cette thèse est l'étude conjointe du phénomène de propagation et du phénomène de mutation de l'information dans un corpus de documents textuels.

Pour répondre à ce problème, nous proposons un nouveau modèle de la propagation de l'information, que nous nommons le modèle de la *Trajectoire de l'information*. Les informations n'existent plus ponctuellement mais comme une structure distincte qui s'étend dans le temps, et qui se situe le long de séquences de documents. Nous appelons de telles séquences de document des chaînes de propagation.

Pour extraire les différentes informations qui se propagent au sein d'un corpus de documents, nous proposons de procéder en deux étapes. Tout d'abord, il est nécessaire de retrouver les séquences de documents susceptibles de contenir des informations qui se sont propagées, ce que nous appelons la structure de la Trajectoire. Dans un second temps, il s'agit d'extraire de ces séquences de documents les différentes informations qui s'y propagent, ce que nous appelons la sémantique de la Trajectoire.

Pour retrouver la structure de la Trajectoire, nous présentons une approche, basée sur un critère de cohérence d'une séquence de documents, qui permet de retrouver de telles séquences. nous montrons la pertinence de notre approche à l'aide d'une campagne d'évaluation par des annotateurs humains. Notre approche, ainsi que la problématique de la thèse et le modèle de la Trajectoire font l'objet de deux publications [HuyghuesDespointes, 2018 ; HuyghuesDespointes, 2019].

Pour retrouver la sémantique de la Trajectoire, nous proposons une méthode permettant l'identification et la caractérisation des principales informations susceptibles de se propager le long des séquences de document. Nous présentons également

diverses approches permettant d'étiqueter ces informations qui tirent parti de la Trajectoire. Nous montrons également la pertinence de notre travail à l'aide d'une campagne d'évaluation par des annotateurs humains.

Enfin, nous proposons plusieurs exploitations des informations contenues dans la Trajectoire de l'information. Nous avons construit une maquette permettant l'exploitation et la navigation dans les différentes chaînes permettant de découvrir les nuances entre des informations similaires, mais également de détecter des sous-corpus particulièrement intéressants pour certains sujets mineurs du corpus.

Nous listons ainsi les différentes contributions de cette thèse :

1. Mise en lumière du problème de l'étude conjointe du phénomène de propagation et du phénomène de mutation de l'information dans un corpus de documents textuels (Chapitre 2).
2. Modèle de la *Trajectoire de l'information* (Chapitre 3).
3. Méthode d'approximation de la structure de la Trajectoire à l'aide d'un critère de cohérence (Chapitre 3).
4. Méthode d'extraction et de caractérisation des informations de la Trajectoire (Chapitre 4).
5. Outil de visualisation, d'exploration et d'exploitation de la Trajectoire (Chapitre 5).

1.4 Structure du manuscrit

Hormis ce chapitre introductif et le chapitre de conclusion, ce manuscrit est composé de quatre chapitres. Le Chapitre 2 commence par introduire précisément le cadre de la propagation de l'information dans les médias sociaux. Il présente d'abord ce que nous nommons médias sociaux, ainsi que les méthodes d'analyse de textes et de graphes qui seront utilisées dans le manuscrit. La suite du chapitre 2 présente un état de l'art sur l'étude de la propagation de l'information. Ces différentes considérations étant en place, le chapitre se conclut par la présentation de notre problématique

sur l'analyse conjointe de la mutation et de la propagation de l'information. On y trouvera également le positionnement nous ayant conduit au modèle de la Trajectoire de l'information.

Le Chapitre 3 commence par exposer le formalisme du modèle de la Trajectoire de l'information. Nous proposons ensuite une première approche pour calculer les chaînes de document le long desquelles de l'information est susceptible de se propager, à l'aide de la notion de chaînes cohérentes. En partant d'une approche simple, nous proposons plusieurs améliorations de l'approche permettant de la rendre applicable sur un corpus conséquent en un temps raisonnable. Le chapitre se conclut sur la présentation d'une campagne d'évaluation des chaînes que nous calculons, nous montrons que les chaînes calculées capturent bien l'intuition humaine d'une chaîne le long de laquelle une information est susceptible de s'être propagée. L'approche de calcul proposée au chapitre 3 est discutée plus en détail en annexe, où on trouvera une version asynchrone de la méthode (Annexe A), une réflexion sur la manière de représenter efficacement en mémoire un ensemble de chaînes (Annexe B), et une étude quantitative des chaînes fortuitement cohérentes ainsi que des différents paramètres de l'approche (Annexe C).

Là où le chapitre 3 propose de calculer les chaînes de documents le long desquelles de l'information est susceptible de s'être propagée, le chapitre 4 présente notre approche pour extraire de telles chaînes les différentes informations. Pour ce faire, nous détaillons une approche en plusieurs étapes : nous commençons par étendre les techniques de description des documents aux chaînes de documents. Après la description d'une chaîne nous, proposons une méthode d'identification des principales informations de la chaîne. Nous proposons ensuite une manière de décrire ces principales informations avant de proposer plusieurs méthodes d'étiquetage de ces informations pour les rendre plus lisibles pour un utilisateur humain. Le chapitre 4 se termine par la présentation d'une campagne d'évaluation de notre approche. À partir des résultats, nous concluons qu'il est possible d'extraire des informations notables des différentes chaînes.

Le chapitre 5 présente l'exploitation des chaînes et des informations extraites dans les chapitres 3 et 4. Il commence par présenter une maquette logicielle permettant d'explorer les chaînes indépendamment les unes des autres, mais également les chaînes dans leur ensemble (la trajectoire). La maquette permet également d'explorer des groupements de chaînes qui contiennent des informations très similaires, ce que nous nommons des *récits*. Le chapitre 5 présente ensuite une étude de corpus mettant en lumière la capacité de notre approche à extraire des informations diverses et marginales d'un corpus, qu'il serait délicat de trouver autrement. Enfin le chapitre présente les principales exploitations que nous envisageons pour la trajectoire.

Le manuscrit se conclue sur les différents travaux réalisés durant ces quatre années de travail et les différentes perspectives qui en découlent.

Chapitre 2

Propagation de l'information et positionnement

Résumé. Ce chapitre est l'occasion de présenter les différents travaux et méthodes existantes autour de la propagation d'informations dans les médias sociaux. La première partie présente la notion d'information, le contexte des médias sociaux et pourquoi l'information s'y propage. Nous travaillons essentiellement sur des documents textuels. Aussi, nous présentons les différentes techniques que nous utilisons pour analyser le texte. La question de la propagation est également une histoire de structure, et nous présentons succintement, pour fixer du vocabulaire, le formalisme des graphes. Dans la deuxième partie, nous dressons un état de l'art sur la question de la propagation des informations. Enfin, nous discutons en dernière partie de l'importance du phénomène de mutation de l'information pour l'étude de sa propagation. Ce qui nous amènera à aborder le phénomène de propagation autrement en posant un nouveau cadre pour son étude.

TABLE 2.1 – Notations utilisées

Notation	Signification
D	Corpus de documents.
d	Document du corpus.
f	Descripteur.
$w_{f,d}$	Importance du descripteur f pour le document d.
F	Ensemble des différents descripteurs des documents du corpus.
V	Ensemble des sommets d'un graphe
E	Ensemble des arcs d'un graphe

2.1 Introduction aux médias sociaux et à l'analyse de texte

L'être humain est un animal social. Nous vivons pour la plupart en société où nous échangeons des informations à un rythme soutenu par divers moyens, du bouche-à-oreille en passant par les écrits, les journaux, la radio ou encore la télévision. Chacun de ces moyens est soutenu par une technologie, que ce soit la parole, l'écriture, l'imprimerie, les ondes hertziennes, etc. L'Internet, littéralement l'interconnexion des terminaux informatiques, est la plus récente de ces technologies. Elle a permis l'émergence de nouveaux médias [1] connectés, permettant au Web d'être, entre autres, une grande plateforme d'échange d'informations. Nous appelons cet ensemble médiatique **les médias sociaux**.

La structure d'Internet et des médias sociaux nous donne l'accès à de grandes masses de données de communication que nous pouvons indexer, enregistrer et traiter à l'aide de machines. Nous parlerons dans un premier temps des caractéristiques principales de cet espace médiatique, avant de lister dans un second temps les différents outils dont nous nous servons pour son analyse dans le cadre de cette thèse. Ces outils sont issus de deux domaines : le traitement automatique de la langue pour l'analyse textuelle et la théorie des graphes pour la modélisation par graphe.

1. Le terme média a plusieurs orthographes différentes, de sa version latine medium/media, à sa version invariable média, nous avons choisi d'utiliser média au singulier et médias au pluriel, comme recommandé dans Le Petit Larousse et le dictionnaire de l'Académie Française.

2.1.1 Définitions

L'information

L'acte de communication a généralement un sens. C'est-à-dire qu'il sert à exprimer la pensée de celui qui l'initie. Cette expression se fait par l'entremêlement de *signes linguistiques*. Par définition, un signe linguistique est une unité d'expression du langage. Charles Sanders Peirce [Peirce, 1991] propose de définir formellement le signe comme un triplet constitué du **représentamen**, il s'agit de la représentation du signe, de l'**objet**, il s'agit cette fois du concept du signe existant au niveau de la pensée, et enfin de l'**interprétant**, qui sert à passer du représentamen à l'objet.

Exemple 1. "Parlons du chien Médor. Le chien est un labrador."
Dans cet exemple, concentrons-nous sur les deux occurrences du mot chien. Dans les deux cas, le représentamen est le mot "chien", et l'objet est le chien auquel je pense, à savoir Médor. Dans la première occurrence, l'interprétant du mot chien sera plutôt le concept de chien. Son rôle est de définir une caractéristique de Médor : être un chien. Dans la seconde occurrence, l'interprétant, vu le contexte, est directement Médor. Peirce appelle cette capacité de changer d'interprétant en fonction du contexte la **sémiose** [Peirce, 1991].

Dans le contexte de la communication, on parle d'**informations** pour les messages qui sont véhiculés. Une information est un cas particulier de signe linguistique qui peut produire un effet dépendant de celui qui l'interprète. Elle vérifie la chaîne suivante :
Pensée (de l'auteur) -> Signe -> Communication -> Receveur -> Effet.
Par exemple, j'écris "Il va faire froid demain" pour signifier ma connaissance de la météo de demain et je le publie sur Internet (sur un média social). Quelqu'un consulte mon message et pense le lendemain à se vêtir d'un pull. Ce que j'ai signifié est donc une information. Il faut noter qu'une information peut être interprétée différemment par le receveur et par l'émetteur, ceux-ci évoluent dans des contextes différents et ont donc un **interprétant** différent.

Exemple 2. Exemple d'information : le fait

L'expérience humaine du monde est par essence phénoménale. Autrement dit, il se passe des *choses* en des lieux et à des moments précis : Il pleut en Ariège. J'ouvre ma fenêtre. Un nouveau président est élu. Ces choses qui surviennent sont appelées des **évènements**. La constatation d'un évènement est appelé un **fait**. Le fait est un cas particulier d'information, dont l'objet est un évènement.

Le média classique

Le média, dans son acception large, est **le support** des représentamens. Il sert à les transmettre depuis leurs auteurs jusqu'à ceux qui vont les interpréter. L'étymologie du mot vient du latin *medium*, signifiant le milieu, l'intermédiaire. L'air, les tablettes d'argile ou les gestes sont donc des médias puisqu'ils servent à porter la parole, l'écrit ou la communication gestuelle.

Il y a une grande quantité de médias différents et plus ou moins sophistiqués. Dans leur utilisation usuelle, l'expression de média, et l'adjectif médiatique se réfèrent à des structures capables de transmettre à un large public. On les qualifie de **médias de masse**. Le média de masse **classique** oppose les receveurs, qui sont ceux qui reçoivent et consomment le média à la chaîne de production de celui-ci.

La chaîne de production est composée du producteur du média, qui est la personne qui possède l'objet produit et d'un ou plusieurs auteurs. Dans le cas d'un livre il y a généralement un auteur, mais dans le cas d'un journal il y en a vraisemblablement plusieurs. Enfin, il y a un ou plusieurs éditeurs, dont le rôle est de produire l'objet qui sera reçu.

Le média classique positionne ainsi plusieurs personnes intermédiaires entre les auteurs et les receveurs et implique un processus relativement long. L'auteur n'est pas libre de publier par lui-même, ce qu'il souhaite et à n'importe quel moment où il le désire.

Exemple 3. Exemple d'activité médiatique : le journalisme

Le **journalisme** est l'activité qui consiste à informer un large public des événements

importants *actuels*. L'activité se découpe en plusieurs étapes :

1. La collecte des faits. Le journaliste cherche des témoignages portant sur différents événements récents. Il sélectionne ceux qu'il estime ayant un intérêt particulier. Lorsqu'un fait parle d'un événement récent, on parle d'un fait d'actualité, ou simplement d'une actualité ou d'une nouvelle. Le terme anglais consacré est *News*.
2. La vérification des faits. Le journaliste détermine si l'événement d'un fait a bien eu lieu ou non. Par exemple, en trouvant plusieurs sources différentes qui relatent des faits similaires.
3. La diffusion de ces faits. Les faits sont rédigés et éventuellement agrémentés de commentaires et d'une mise en contexte. Cet objet diffusé est appelé **un journal**[2].

Le journalisme est une activité professionnelle. Les trois étapes sont soumises à un code de déontologie. Le journalisme utilise la plupart des médias de masse. On trouve ainsi des journaux dans la presse, à la radio, à la télévision ou sur Internet.

Internet et les médias sociaux

Internet est un réseau informatique mondial qui permet la communication via un ensemble de protocoles standardisés. Au 30 juin 2019, on estime que 58% de la population mondiale y est connectée[3]. Sur la base d'Internet s'est construit un ensemble d'applications comme le courrier électronique, la transmission de fichiers, ou le World Wide Web. C'est ce dernier qui permet de publier et consulter des pages sur des sites, à l'aide d'un navigateur. Les sites ont pour vocation d'être des espaces de publication, des médias, à disposition de tous. Là où le média classique encadre la publication, le Web permet à tous les utilisateurs de mettre du contenu à disposition de tous les autres. Benjamin Bayart, alors président du French Data Network[4] traduit cette propriété ainsi : « L'imprimerie a permis au peuple de lire,

2. Le terme journal est polysémique en français. On parle ici d'un contenu édité, relayant des faits d'actualité.
3. `https://internetworldstats.com/stats.htm`
4. Plus ancien fournisseur d'accès Internet français encore en activité.

Internet va lui permettre d'écrire. » [5].

Les **médias sociaux** sont constitués par l'ensemble des sites Web permettant la consultation de contenus informatifs. Il en existe différentes formes, selon la manière dont on y publie :

— Les **forums** sont des espaces de discussion où les messages s'agencent en fils de discussion. Les utilisateurs enregistrés peuvent répondre et créer de nouveaux fils.

— Un **blog** est un site tenu par une (voire plusieurs) personne(s) sur lequel elle publie régulièrement des articles, qu'on appelle des billets. La presse utilise ce format pour publier son contenu papier, mais elle propose aussi du contenu Web exclusif.

— Un **Wiki** est un site collaboratif, dont le but est d'associer l'effort de ses utilisateurs pour créer une base de connaissances. L'exemple le plus célèbre est Wikipedia [6], une encyclopédie collaborative.

— Un **réseau social** est un site sur lequel il est possible de publier du contenu et de suivre les utilisateurs dont le contenu nous intéresse.

Cet ensemble forme un écosystème médiatique, dans lequel les utilisateurs ont la capacité de réagir librement. Une grande partie du contenu publié sur le Web est textuel, mais il y a également du son, des images, des vidéos, et le texte peut lui-même être enrichi (mise en gras ou en italique de certains passages, présences d'hyperliens). La plupart des contenus sont constitués d'une combinaison de ces différents formats, qu'on appelle alors des contenus **multimédias**.

Phénomène de propagation de l'information

Le média est un support de communication. La communication est l'acte de transmettre des informations d'un agent à un ou plusieurs autres. Ces agents, nouvellement informés, peuvent à leur tour, s'ils le souhaitent, transmettre les informations reçues. Le **phénomène de propagation de l'information** est l'événement observable de

5. Conférence tenue à Rennes, le 25 septembre 2009
6. `https://www.wikipedia.fr`

la circulation d'informations entre agents.

Un exemple type du phénomène de propagation de l'information est la **rumeur** : Il s'agit d'un récit (dont la véracité est à confirmer ou infirmer), qui se diffuse de personne à personne à une grande échelle. Ainsi, deux personnes peuvent connaître une même rumeur, sans y avoir été exposées par la même source.

Au-delà de l'exemple de la rumeur, toutes les informations sont susceptibles de se propager. Une fois exposée à une information, une personne peut alors, consciemment ou non, la mémoriser et décider de la relayer dans le futur. La manière dont on sélectionne et mémorise les informations auxquelles nous sommes exposés, et notre volonté de les transmettre est variable selon la personne et le type de l'information. Par exemple, la stratégie de la publicité en général, et du marketing viral en particulier, est justement d'inciter les récepteurs à mémoriser et à relayer une information commerciale.

Le modèle d'information de Charles Sanders Peirce, précédemment présenté, la détaille en trois parties : représentamen, objet et interprétant. Le média, quel qu'il soit, ne permet que la transmission du représentamen. Ainsi, le récepteur d'une information a la tâche de reconstruire l'objet et l'interprétant. L'information perçue est ainsi réinterprétée, modifiée, mutée. On parle du phénomène de **mutation de l'information**. Une partie de l'explication tient à la subjectivité des humains : Ils transforment l'information, consciemment ou non, par le prisme de leurs référents socioculturels [Blackmore, 2000]. Ainsi, même pour une tâche de transmission simple, *le téléphone arabe*, il est difficile de s'extraire du phénomène de mutation.

Le téléphone arabe est un jeu, qui se joue à plusieurs, dont l'objectif est de passer un message sans le déformer entre tous les joueurs. Disposés en file, chaque joueur, du premier au dernier, passe le message à son voisin sans que les autres ne l'entendent. À la fin, le dernier joueur énonce le message qu'il a entendu à voix haute, ce dernier a généralement dévié du message original.

Les médias sociaux sont propices à l'étude du phénomène de propagation de l'information. D'abord par leur nature, ils permettent à tous les utilisateurs de

s'exprimer publiquement et donc de participer au phénomène. Ensuite, par leur support, l'expression publique des utilisateurs étant déjà enregistrée dans un format numérique. Il est ainsi plus simplement disponibles que d'autres médias qu'il serait nécessaire d'enregistrer d'une part et de retranscrire d'autre part pour les étudier.

Collecte et problème des données lacunaires

Pour étudier dans son ensemble le phénomène de propagation de l'information, il faudrait enregistrer l'ensemble des conversations entre une quantité d'agents qui ne communiquent qu'en vase clos, sans aucun contact extérieur. Dans un but applicatif réel, ce dispositif n'est pas réaliste. Les gens sont largement interconnectés, et la plupart des conversations sont hors des radars dont nous disposons (les enregistrer serait, de toute façon, illégal). Nous avons choisi comme cadre les médias sociaux. Il s'agit de la plus grande base de contenus produits par des humains qui soit potentiellement accessible. Comme tous les utilisateurs sont susceptibles de publier, les médias sociaux facilitent le phénomène de propagation de l'information.

Cependant, accéder à l'intégralité du Web n'est pas aisé. En effet, le contenu du Web change constamment. De plus, selon les objectifs que l'on se fixe, tous les contenus ne sont pas pertinents. Ainsi, si on s'intéresse à l'actualité française, il ne semble pas judicieux d'analyser l'intégralité des photos de chats du Web. En général, il suffit de **collecter** un ensemble de contenus pertinents pour l'application que l'on vise. Comment collecter cet ensemble à partir d'un grand corpus comme le Web, est un champ de recherche à part entière : la *Recherche d'information*(RI)[BaezaYatesRibeiroNeto, 1999]. Il est difficile de réaliser des collectes automatiques sans générer du contenu non pertinent. L'élimination de ce bruit nécessite souvent une validation humaine coûteuse. Par ailleurs, il peut arriver aussi que des contenus pertinents échappent à la collecte. Il s'agit du problème des **contenus manquants**.

Les contenus collectés sur le Web ne sont pas toujours de bonne qualité. En effet, certains contenus collectés présentent des erreurs d'encodage ou des défauts d'extraction aboutissant à un résultat imparfait. On parle notamment du problème

de **nettoyage des données**.

Par ailleurs, des données utiles peuvent être indisponibles, par exemple beaucoup de contenus sont publiés anonymement, sans date spécifiée de publication. Ce problème, couplé aux deux problèmes précédents forment le problème global des **données lacunaires**.

Nous travaillons dans un cadre de veille documentaire sur le Web. C'est-à-dire que nous collectons des ensembles documentaires qui deviennent notre base de travail. Nous sommes dans un contexte partiel. Nous n'avons ni tous les documents, ni la connaissance de tous les acteurs de la propagation, en sus la complexité de la langue et les mutations de l'information rendent la différentiation des informations difficile. Nous ne mentionnerons plus ce problème des données lacunaires, mais il s'inscrit en filigrane dans toute cette thèse.

Cette section a été l'occasion de définir ce qu'est l'information, sa propagation dans les médias, et plus particulièrement nous nous sommes intéressés aux médias sociaux et à leurs mécanismes de fonctionnement. Travailler dans ce cadre implique en pratique de travailler avec des données lacunaires. Même si toutes les considérations de propagation d'information dans les médias sociaux s'appliquent à tous les types de contenus, **nous nous intéressons principalement au contenu textuel**. Il s'agit d'un des contenus les plus présents du Web, et d'un des moyens d'expression le plus accessible. La majorité des utilisateurs d'Internet sait lire, écrire et sont munis d'un clavier. La section suivante traite ainsi des techniques d'analyse de textes utilisées dans le reste du manuscrit.

2.1.2 Représentation textuelle de documents et méthodes d'analyse

La matière première de notre étude est constituée de corpus de documents textuels. Ces documents comportent un certain nombre d'attributs, tels qu'une liste d'auteurs, une provenance (sur Internet cela peut par exemple être une URL), une date de publication, un titre, et un contenu. Le titre et le contenu sont généralement le

cœur du document. Ce sont tous les deux des textes, et c'est principalement ce que nous utiliserons pour étudier nos corpus. Si représenter informatiquement le texte comme une suite de caractères est aujourd'hui bien maîtrisé et normé, l'analyse et la compréhension automatique de textes rédigés en langue naturelle est un champ d'étude vaste et actif. Cette section introduit les méthodes usuelles de représentation et de comparaison de textes dont nous allons nous servir. Nous présentons aussi succinctement le problème du résumé automatique de texte ainsi que la méthode que nous utiliserons en partie 4.

Modéliser les textes

Un **texte** est une succession de symboles respectant les règles (la grammaire) d'une ou plusieurs langues. Étymologiquement, le mot texte provient du latin *textus*, le tissu, et de sa forme verbale *texo*, tisser. Il y a dans cette étymologie l'idée que le texte est un entremêlement complexe d'informations constituant un discours, une expression concrète de la pensée de son auteur. Un texte est ainsi caractérisé par :

— Un ensemble de mots. Ce sont les briques élémentaires à partir desquelles le texte est construit. Chaque mot a un, voire plusieurs sens, selon la manière dont il est utilisé.

— Une mise en séquence des mots. Il s'agit de l'ordre dans lequel on agence les mots pour constituer le texte.

— Une validité grammaticale. Les mots et leur assemblage doivent vérifier l'ensemble des règles de la grammaire de la langue dont ils sont issus pour constituer un énoncé compréhensible.

Pour étudier automatiquement un texte, il est possible d'exploiter chacune de ces caractéristiques. La première étape de l'étude automatique d'un texte consiste généralement à décrire les éléments de base qui le composent, qu'on nomme des descripteurs :

Définition 1. On appelle un **descripteur** un motif extrait d'un texte, dont on peut compter l'occurrence.

Les mots sont les descripteurs les plus élémentaires du texte. On peut également utiliser des mots consécutifs d'une taille fixe comme descripteurs, on parle alors de n**-grammes**.

Exemple 4. Le loup est dans la bergerie.
bi-grammes : Le loup, loup est, est dans, dans la, la bergerie.
tri-grammes : Le loup est, loup est dans, est dans la, dans la bergerie.

La quantité de n**-grammes** potentiels augmente avec n. Ainsi, la fréquence d'un 5-gramme contenant le mot « chat » est moindre que la fréquence du mot « chat » lui-même. Pour comparer des documents, on considère généralement des n-grammes de taille 2, 3 et 4. Les cas où deux documents partagent un n-gramme, pour $n > 4$ sont considérés trop rares, et garder les différents n-grammes, nombreux, en mémoire est coûteux. [7]

Les **mots** et les **n-grammes** sont simples à construire et ne nécessitent pas de connaissances grammaticales *a priori* sur la langue, c'est pourquoi il s'agira des principaux descripteurs utilisés dans le reste de ce manuscrit. Notons qu'il existe des descripteurs plus élaborés auxquels il convient de faire appel pour décrire plus finement nos textes. Parmi lesquels on retrouve :

— Les **lemmes**. Ce sont les formes non fléchies des mots, qui sont usuellement celles qu'on trouve dans un dictionnaire. Ainsi les mots « aimerait », « aimant » et « aimé » ont tous le même lemme « aimer ».

— Les **radicaux**. Ils sont obtenus à partir des mots auxquels on ôte leurs suffixes et préfixes : « adjugera » devient « ad-jug-era » dont on ne conserve que le radical « jug ».

— Les **syntagmes**. Ce sont les groupes de mots qui forment une brique grammaticale au sein du texte. Par exemple, pour la phrase « Il a une belle table en bois de chêne », on a les syntagmes suivants : « une belle table en bois de chêne », « une belle table », « bois de chêne ».

7. Il s'agit d'une instance du phénomène intitulé le fléau de la dimension. Plus on considère un espace de dimension élevé, ici la taille des n-grammes, plus les données sont isolées et éparses.

— Les **collocations**. Ce sont des mots qui apparaissent fréquemment ensembles, comme « passer son tour », « au fur et à mesure ».
— Les **entités nommées**. Il s'agit des références à des entités existantes, comme des personnes, des lieux, des entreprises, des monuments ou des œuvres. Selon le contexte, plusieurs locutions peuvent faire référence à la même entité, comme « Mona Lisa » et « La Joconde ».

Un ensemble de descripteurs forme une première représentation d'un texte. À partir de celle-ci, il est possible de calculer une similarité textuelle entre deux documents en utilisant des notions ensemblistes, sans tenir compte de la fréquence des descripteurs.

L'indice de Jaccard calcule le nombre de descripteurs communs aux documents d_1 et d_2 relativement à leur total de descripteurs.

$$sim_{jacc}(d_1, d_2) = \frac{|w_{d_1} \cap w_{d_2}|}{|w_{d_1} \cup w_{d_2}|}$$

L'indice de Jaccard a la bonne propriété d'être nul si les deux documents n'ont aucun descripteur commun, et de valoir un lorsqu'ils ont exactement les mêmes descripteurs. Imaginons un long texte, que l'on veut comparer à un de ses paragraphes. On peut vouloir que la similarité entre le texte et un de ses paragraphes soit élevé. Dans ce cas l'indice de Jaccard ne convient pas, un texte long est susceptible d'avoir bien plus de descripteurs qu'un seul de ses paragraphes.

Le coefficient de recouvrement pallie ce problème, en ne considérant le nombre de descripteurs communs relativement qu'à un seul des deux documents :

$$sim_{overlap}(d_1, d_2) = \frac{|w_{d_1} \cap w_{d_2}|}{|w_{d_1}|} \tag{2.1}$$

Dans le cas où d_1 est un paragraphe de d_2, on obtient bien :

$$sim_{overlap}(d_1, d_2) = 1, \text{ et} \tag{2.2}$$

$$sim_{overlap}(d_2, d_1) < 1. \tag{2.3}$$

Il peut paraître étrange que des mots comme « le », « à » ou « une » aient la même importance pour décrire le texte que d'autres comme « Paul », « marqua », ou « assemblée ». Une première approche consiste à enlever, *a priori*, les mots à faible valeur descriptive générale, qu'on appelle les **mots outils** ou *stop-words* en anglais. Une seconde approche consiste à pondérer l'importance de chaque descripteur pour ce texte. Dans chaque cas, il est nécessaire d'injecter de la connaissance extérieure au texte. Une telle connaissance peut venir de notre corpus de documents. Le reste de cette section présente deux méthodes permettant d'effectuer une telle pondération à l'aide d'un corpus.

Sac de mots

Le modèle sac de mots considère l'union de tous les descripteurs issus de la totalité du corpus, qu'on appelle le **vocabulaire**. On le note F (pour *Features* en anglais). Le corpus est noté D. Pour chaque document $d \in D$ et chaque descripteur $f \in F$, on a un poids $w_{f,d}$ qui représente l'importance de f pour d. En indexant les éléments du vocabulaire, on a le vecteur poids qui est une représentation du contenu du document :

$$w_d = [w_{f_1,d}, w_{f_2,d}, \ldots, w_{f_{|F|},d}].$$

Une fonction de pondération simple pour le modèle sac de mots est le **TF** (Term Frenquency), qui correspond au nombre d'occurrences du descripteur dans le document.

Lorsqu'on ne considère que la présence ou l'absence du descripteur, on parle de

BTF pour Binary Term Frequency :

$$BTF(f,d) = \begin{cases} 1, & \text{si } TF(f,d) > 0. \\ 0, & \text{sinon.} \end{cases} \tag{2.4}$$

Le modèle sac de mots pondéré par le BTF est équivalent à notre représentation simple du document. Il est maintenant possible de compter la fréquence des descripteurs dans le corpus. On la note **DF** pour Document Frequency.

$$DF(f,D) = \frac{|\{d \in D / BTF(f,d) = 1\}|}{|D|}.$$

Plus un descripteur est fréquent au sein du corpus, moins il est spécifique. Il est moins efficace pour différencier un document dans lequel il apparaît. Le cas extrême est un descripteur que tous les documents possèdent. En prenant pour exemple un corpus sur l'imagerie, on s'attend à ce que le mot « image » soit très fréquent. Il a une valeur descriptive très basse dans ce contexte. Partant de cette idée, on construit la famille de pondération **TF-IDF** (Term Frequency - Inverse Document Frequency).

$$TF\text{-}IDF(f,d,D) = TF(f,d) \cdot IDF(f,D).$$

$$\text{Avec } IDF(f,D) = log(\frac{1}{DF(f,d)}).$$

Si on comprend l'inverse de DF comme une mesure de rareté, son passage au logarithme est une heuristique classique [Jones, 1972], dont l'efficacité est constatée mais l'interprétation théorique ardue [Robertson, 2004].

Le TF-IDF est une famille de pondération car il existe de nombreuses normalisations du TF (dont le BTF). Celui-ci est sensible à la longueur du document dans sa forme standard. Dans le reste du manuscrit nous utilisons le TF-IDF standard lorsqu'il est mentionné, sauf mention contraire explicite.

Une contrainte technique du modèle sac de mots est qu'il utilise des vecteurs de la taille du vocabulaire, qui est généralement de grande taille. Cependant, un

document particulier n'utilise qu'une portion, généralement faible, du vocabulaire. Ces vecteurs sont essentiellement creux, c'est-à-dire avec de nombreuses valeurs à zéro. Cette propriété en permet une représentation efficace.

Pour comparer des documents représentés par des vecteurs réels, on utilise traditionnellement [SinghalInc, 2001] la **similarité cosinus**, définie comme suit :

$$sim_{cos}(d_1, d_2) = \frac{w_{d_1} \,.\, w_{d_2}}{||w_{d_1}|| \,.\, ||w_{d_2}||}$$

Géométriquement, elle correspond au cosinus de l'angle entre les deux vecteurs w_{d_1} et w_{d_2}. Deux vecteurs *colinéaires* sont alors parfaitement similaires, tandis que deux vecteurs *orthogonaux* ont une similarité nulle. La similarité cosinus est composée uniquement de produits scalaires, simples et rapides à calculer pour des vecteurs creux de grande dimension.

Le modèle sac de mots utilise essentiellement les descripteurs sans se soucier de la manière dont ils sont agencés dans le texte (leur ordre). Un descripteur comme le n-gramme permet d'exploiter le fait que certaines séquences de mots ont une fréquence distinctive. Autrement dit, un mot apparaît plus fréquemment dans certaines séquences de mots que dans d'autre. Un défaut du n-gramme comme descripteur est cependant qu'il conduit à des représentations des documents par des vecteurs creux et de grande taille. Le modèle suivant tire également parti du contexte des mots, mais il permet de construire des vecteurs denses de représentation du sens des descripteurs d'abord, et des documents ensuite.

Modèle distributionnel

Le modèle distributionnel s'appuie sur l'hypothèse linguistique proposée par Zellig Harris [Harris, 1954] en 1954 :

Hypothèse 1. Le sens d'un mot est caractérisé par la distribution des contextes dans lequel il apparaît.

Cela signifie que deux mots utilisés dans des contextes similaires tendent à avoir

un sens similaire.

Exemple 5. Contexte similaire pour un mot rare :

— Le chien mange dans la gamelle.
— Le chat mange dans la gamelle.
— Le tarsier mange dans la gamelle.

« chien », « chat » et « tarsier » ont, dans l'exemple 5, tous les trois le même contexte. Même sans savoir ce qu'est un tarsier (il s'agit d'un petit primate), on devine qu'il s'agit d'un animal. Le domaine des sciences cognitives s'intéresse à ce phénomène [McdonaldRamscar, 2001 ; Yarlett, 2008] pour expliquer notre capacité à interpréter des mots rares ou inconnus.

Le contexte d'un mot pourrait comprendre l'entièreté du document dans lequel il apparaît. Pour des raisons combinatoires analogues à celles des n-grammes (fléau de la dimension), avoir des contextes trop conséquents réduit les chances d'avoir deux mots qui partagent un même contexte. Ce faisant, pouvoir utiliser l'hypothèse distributionnelle pour comparer le sens de mots devient moins fréquent. On considère donc généralement un contexte court autour des mots, qu'on appelle la *fenêtre*.

Tout comme on a pu représenter un document à l'aide de la fréquence de ses descripteurs dans le corpus, on a une représentation des mots à l'aide de la fréquence des contextes dans lesquels ils apparaissent dans le corpus. En notant $f \in F$ le mot, k la taille de fenêtre, et $w_{ctx,f}$ la fréquence du contexte ctx pour le mot f dans le corpus on a la représentation suivante de f :

$$w_f = [w_{ctx_1,f}, w_{ctx_2,f}, \ldots, w_{ctx_{|F|^k},f}].$$

Cette représentation est illustrée en Figure 2.1. On y constate que le mot f apparaît dans 3 contextes différents, ab_cd, ij_kl et mn_op avec des fréquences différentes. Le mot f n'est pas présent dans certains contextes comme qr_st.

On remarque que ce vecteur a $|F|^k$ composantes, qui reste un nombre élevé pour

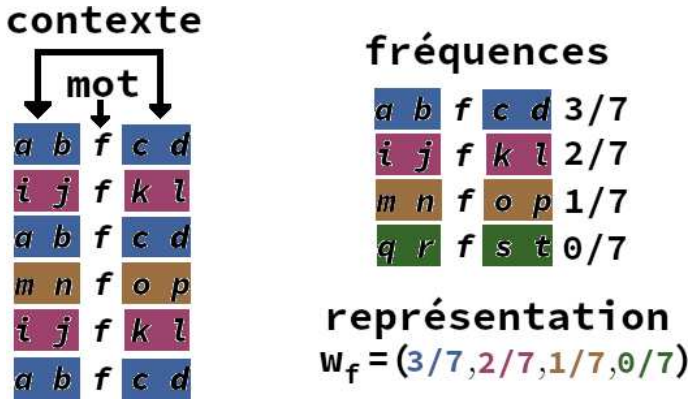


FIGURE 2.1 – Représentation distributionnelle d'un mot f. À gauche sont listées les occurrences du mot f dans le corpus avec un contexte cerné par une fenêtre de taille 2, soit deux mots avants et deux mots après f. À droite la fréquence des différents contextes du corpus pour f et sa représentation.

un vocabulaire conséquent et une taille de fenêtre non nulle. Pour pallier ce problème, on utilise une méthode de réduction de dimension, qu'on appelle **plongement de mots** (word embedding, en anglais). L'idée est d'utiliser un espace réel de dimension fixe dans lequel on représente les mots par des points. Pour placer les points dans cet espace, on suit l'idée selon laquelle deux mots qui apparaissent dans des contextes similaires doivent être représentés par des points proches. La représentation devient alors :

$$w_f = [w_{1,f}, w_{2,f}, \ldots, w_{l,f}]$$

. où les $w_{i,f}$ sont les coordonnées de la représentation du mot f dans un espace de dimension l fixé au préalable.

Il existe plusieurs méthodes de plongement de mots, parmi lesquelles l'analyse sémantique latente [Deerwester, 1990], le clustering de Brown [Brown, 1992], GloVe [Pennington, 2014] et Word2Vec [Mikolov, 2013]. Quoc Le et Tomas Mikolov proposent une extension de Word2Vec pour plonger non seulement les mots, mais également les documents dans le même espace, intitulée Doc2Vec [LeMikolov, 2014].

Doc2Vec et TF-IDF forment les méthodes classiques de représentation de document que nous utilisons dans le reste du manuscrit. Nous décrivons maintenant le fonctionnement de Doc2Vec, sans entrer dans tous les détails de la méthode.

L'idée derrière Doc2Vec (et Word2Vec) est d'entrainer un classifieur [8] à prédire le prochain mot, connaissant les vecteurs des l mots précédants (la fenêtre) et du document [9]. Une fois le modèle entraîné sur le corpus, le vecteur appris pour chaque document devient sa représentation.

Les deux représentations vectorielles (TF-IDF, Doc2Vec), ainsi que la représentation ensembliste, permettent, comme vu précédemment, de comparer les documents.

Résumer les textes

Le résumé automatique de texte est une tâche de traitement automatique des langues qui consiste, à partir d'un texte, à produire un autre texte, plus court, qui reprend l'essentiel du premier. Nous utiliserons principalement une méthode experte brevetée par Go Albert intitulée GMIL.

GMIL est une technique d'extraction. Elle extrait des morceaux du document original pour former un résumé. Il existe une autre famille de techniques pour le résumé automatique, dites d'abstraction. Celles-ci consistent à générer un résumé original, non contenu dans le texte. Pour un contexte plus général sur les tâches et les différentes approches de résumé automatique de texte, l'article [TorresMoreno, 2014] offre une vision large de la question. Les méthodes les plus récentes sont étudiées dans l'article [GambhirGupta, 2017].

La méthode GMIL fonctionne comme suit. Elle commence par extraire les syntagmes nominaux du texte à l'aide d'un ensemble de règles spécifiques à la langue de celui-ci. Ces syntagmes sont alors pondérés selon de multiples paramètres, tels que leur position dans le texte, la fréquence des mots qu'ils contiennent et optionnellement leur proximité d'une requête textuelle. Les syntagmes de plus fort poids, mis bout-à-bout, forment alors le résumé du texte.

La modélisation, la comparaison et le résumé automatique sont les outils princi-

8. l'article de Le et Mikolov, ainsi que les implémentations utilisées dans la thèse, utilisent un réseau de neurones avec une couche cachée comme classifieur.

9. C'est la version PV-DM (Distributed Memory version of Paragraph Vector) de Doc2Vec, le papier originel propose également une autre méthode PV-DBOW, qui prédit une fenêtre à partir du vecteur document, mais dont les performances sont moindres.

paux, liés au texte, nécessaires pour les méthodes présentées plus loin. Si le texte est le constituant principal d'un document, dans le contexte des médias sociaux les connexions entre les documents sont d'importance égale. Nous présentons maintenant les documents dans leur relation les uns avec les autres. Pour ce faire nous commençons par introduire le formalisme des graphes, qui est le cadre théorique usuel pour étudier la structure d'objets en relation les uns avec les autres.

2.1.3 Représentation sous forme de graphes

Les graphes sont une abstraction mathématique d'un ensemble d'objets en relation les uns avec les autres. Ils peuvent servir à de nombreux niveaux : pour l'analyse des textes, l'adjacence des mots (quel mot est à droite de quel autre) forme un graphe. Les phrases du texte sont des *chemins* particuliers de ce graphe. Au niveau des documents, il est possible de définir plusieurs relations. La référence par exemple, consiste à pointer au sein d'un document vers un autre document, l'ensemble de ces pointeurs forme le graphe de références. Dans des domaines différents, les réseaux d'eaux, réseaux électriques, ou le réseau social forment tous des graphes. Cette section introduit le vocabulaire relatif aux graphes utilisé par la suite.

Définition 2. Graphe
Soit V un ensemble fini non vide et $E \subseteq V \times V$. On appelle **graphe** le couple $G = (V, E)$. On appelle de plus V l'ensemble des sommets et E l'ensemble des arcs.

On dit de plus que u est un **parent** de v, et v est un **fils** de u, si $(u, v) \in E$.

Les graphes peuvent être orientés (directed en anglais) ou non selon que les arcs ont un sens précis où sont réciproques ($(a, b) = (b, a)$). Nous considérons, sauf mention contraire, que tous les graphes sont orientés.

Définition 3. Arcs consécutifs
Soit $G = (V, E)$ un graphe, $e_1 = (a, b), e_2 = (c, d) \in E$ deux arcs. On dit que e_1 et e_2 sont **consécutifs** si et seulement si $b = c$.

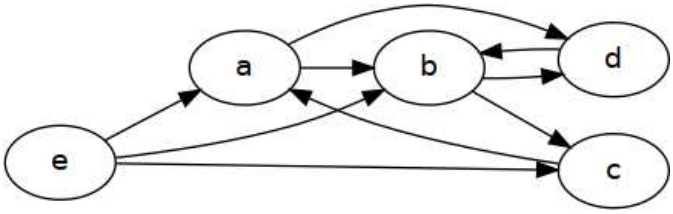


FIGURE 2.2 – Exemple de graphe. Les sommets sont a, b, c, d et e. Les arcs sont représentés par des flèches.

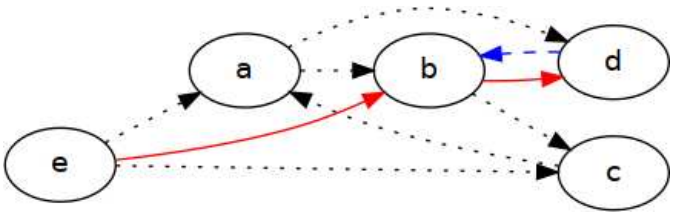


FIGURE 2.3 – Les arcs (e, b) et (b, d) sont consécutifs. Les arcs (e, b) et (d, b) ne le sont pas.

Définition 4. Chemin

Soit $G = (V, E)$ un graphe. On dit qu'une suite d'arcs $c = (e_1, \ldots, e_n)$ est un **chemin** si tous les arcs sont consécutifs :

$$\forall 1 \leq i < n, e_i \text{ et } e_{i+1} \text{ sont consécutifs.}$$

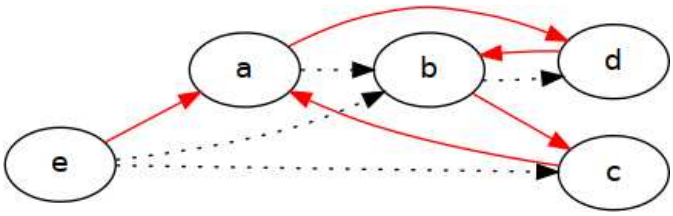


FIGURE 2.4 – $((e, a), (a, d), (d, b), (b, c), (c, a))$ est un chemin du graphe. Comme simplification, on le nomme *eadbca*. *edbc* n'est pas un chemin du graphe, car il n'y a pas d'arcs reliant e et d.

La notion intuitive du chemin est un parcours du graphe en suivant les arcs. Lorsqu'un chemin ne revient jamais sur ses pas, c'est-à-dire qu'il ne passe pas deux fois par le même sommet, on parle de **chemin simple**. Dans la Figure 2.4, *eadbca* n'est pas un chemin simple car il passe deux fois par a. *eadbc*, qui ne fait pas la dernière étape passant de c à a est simple. On appelle un **cycle** un chemin qui commence et fini sur le même sommet. *adbca* est un cycle.

Définition 5. Graphe acyclique (DAG)

On dit qu'un graphe $G = (V, E)$ est un **graphe acyclique** si et seulement si, il ne contient que des chemins simples. De manière équivalente, le graphe est acyclique si, et seulement si, il ne contient aucun cycle. Usuellement on parle de graphe acyclique en utilisant l'acronyme **DAG** pour **D**irected **A**cyclic **G**raph.

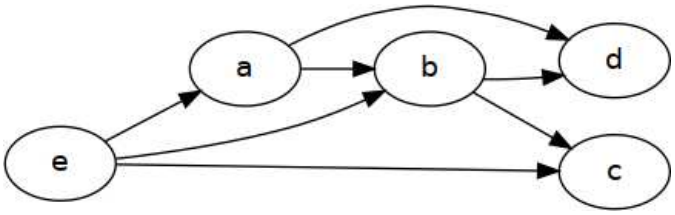


FIGURE 2.5 – Exemple de graphe acyclique. Il s'agit du graphe en Figure 2.2 auquel on a retiré les arcs (d, b) et (c, a).

Les DAGs sont liés à la description des phénomènes qui respectent un *ordre*. Par exemple les événements qui s'écrivent selon une temporalité, comme la généalogie d'une population ou les citations entre documents, est décrit par un DAG. Les hiérarchies sont un autre exemple de structures décrites par des DAGs.

Définition 6. Degré d'un sommet

Soit $G = (V, E)$ un graphe et $v \in V$ un sommet. On appelle :

Degré entrant de v, noté $deg_{in}(v)$ la quantité d'arcs finissant en v.

$$deg_{in}(v) = |\{(a, b) \in E / b = v\}|$$

Degré sortant de v, noté $deg_{out}(v)$ la quantité d'arcs commençant en v.

$$deg_{out}(v) = |\{(a, b) \in E / a = v\}|$$

Degré total de v, noté $deg(v)$ la quantité d'arcs commençant ou finissant en v.

$$deg(v) = deg_{in}(v) + deg_{out}(v)$$

Définition 7. Arborescence

On dit qu'un graphe $G = (V, E)$ est une **arborescence** si, et seulement si, il existe un sommet u, nommé racine, à partir duquel part un chemin unique vers chacun des autres sommets de G.

On nomme **feuilles** les sommets d'une arborescence de degré sortant nul, et **sommets internes** les sommets qui ne sont ni des feuilles ni la racine.

On appelle un ensemble d'arborescences une **forêt**.

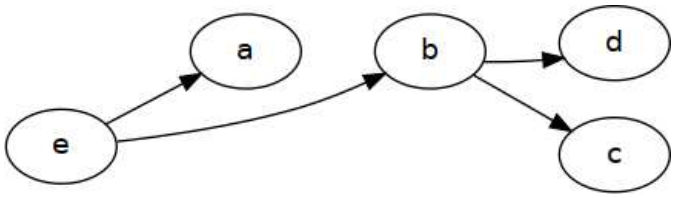


FIGURE 2.6 – Exemple d'arborescence. La racine est E. Il existe un et un seul chemin de E vers tous les autres sommets de l'arborescence. A, C et D sont des feuilles, B est un sommet interne.

Les arborescences sont des DAGs particuliers où les sommets ont tous un degré entrant valant 1, à l'exception de la racine. La Figure 2.6 est un exemple d'arborescence. Ils peuvent représenter des phénomènes comme les affluents d'un fleuve, le branchement d'arbres, mais également des catégorisations strictes comme les taxonomies (cf Figure 2.7), ou les systèmes de fichiers (cf Figure 2.8).

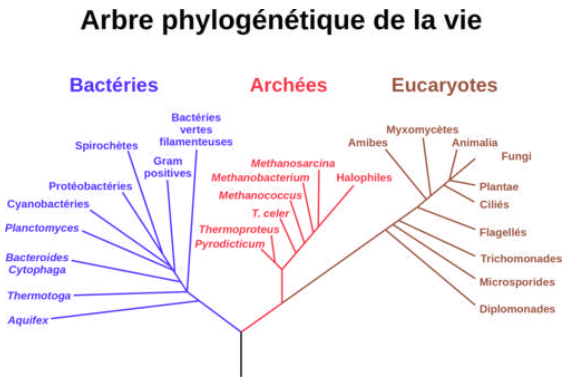


FIGURE 2.7 – Arbre phylogénétique hypothétique de tous les organismes vivants. L'arbre est basé sur des séquences de l'ARNr 16S. À l'origine proposé par Carl Woese [Woese, 1990], il montre l'histoire évolutive des trois domaines du vivant (bactéries, archaea et eucaryotes).

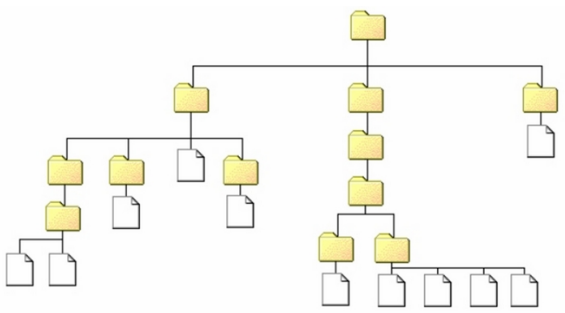


FIGURE 2.8 – Arborescence de fichiers dans des dossiers. Les fichiers sont des feuilles, les dossiers des nœuds internes.

Le formalisme des graphes permet de faire des analogies entre plusieurs phénomènes qui se structurent de la même manière et ainsi leur appliquer les mêmes analyses. En particulier, les réseaux, les communications et les diffusions sont des phénomènes qui ont une structure de graphe, tout comme les médias sociaux.

Modélisation des médias sociaux à l'aide de graphes

On distingue généralement trois familles d'objets dans les médias sociaux.

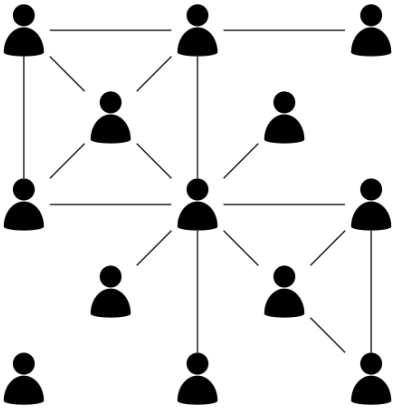
— Les personnes, ou les utilisateurs.

— Les médias, sur Internet en particulier ce sont des sites Web.

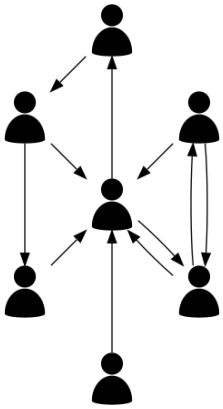
— Les documents, ils sont rédigés par des personnes, publiés sur des médias, consultés par des personnes.

Le **réseau social**, ou graphe social, décrit la manière dont les personnes sont connectées entre elles, du fait qu'elles se connaissent, et sont amenées plus ou moins fréquemment à s'échanger des informations. Dans sa forme la plus simple, le réseau social est donc un graphe où les sommets sont des individus et les arcs relient ceux qui "se connaissent", comme en Figure 2.9a. Par extension, les médias qui permettent de connecter les gens qui se connaissent sont également appelés des réseaux sociaux. C'est le cas de Facebook, par exemple, où on se connecte à ses amis, ou de Linkedin, où on se connecte à ses collègues. Ces deux exemples sont *réciproques* : pour que Bob soit connecté à Alice, il est nécessaire qu'Alice soit aussi connectée à Bob. C'est le cas des gens que l'on rencontre et qu'on fréquente. Il est aussi possible de connaître quelqu'un, sans que celui-ci nous connaisse en retour. Je connais par exemple le

présentateur du journal télévisé, sans que celui-ci ne me connaisse. Il existe donc des connections asymétriques dans le réseau social, ce genre de réseau est donc orienté comme en Figure 2.9b. Sur Twitter ou Instagram pour exemples, le réseau est asymétrique. Chaque utilisateur déclare les utilisateurs pour lesquels il a un intérêt, il est alors notifié de leurs publications.



(a) Deux personnes sont liées si elles se connaissent mutuellement.



(b) Il y a un arc orienté de A vers B si A connaît B.

FIGURE 2.9 – Deux approches du réseau social. (a) Réseau social réciproque. (b) Réseau social non réciproque.

Les documents ont aussi une structure de graphe qui les relie entre eux : le **graphe de citation**. Lorsqu'un document cite explicitement un autre, par exemple via une bibliographie, par l'utilisation d'un lien hypertexte (une URL), ou via un mécanisme de relais (le retweet sur Twitter, le partage sur Facebook), il met en évidence une de ses sources. *A priori*, le graphe de citation est acyclique : un document ne cite généralement que des documents qui sont antérieurs à sa publication. La Figure 2.10 présente un graphe de citation que nous avons extrait à partir d'un ensemble de documents au sujet du mouvement Nuit Debout de 2016.

Un média social se définit par la capacité de ses utilisateurs à produire et consulter des documents, son fonctionnement peut donc aussi être vu comme un graphe entre les documents et les utilisateurs, schématisé en Figure 2.11. Les deux interactions principales sont la publication et la consultation. La publication est certainement la plus visible, parce qu'elle laisse une trace évidente qui est le document. La consultation, quant à elle, peut être enregistrée par le média. Par exemple, une réponse à un

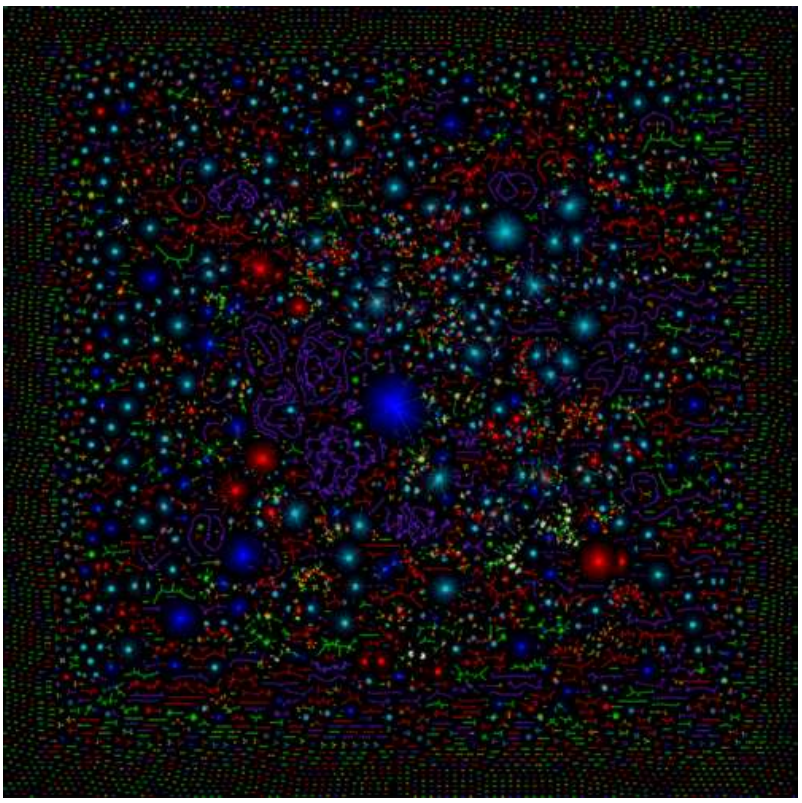


FIGURE 2.10 – Graphe de citations d'un ensemble d'articles collectés par la plateforme AMIEI sur le mouvement Nuit Debout. La couleur des sommets change selon que le document est issu de blogs (vert), de sites de presses (rouge) ou de Twitter (bleu). La couleur d'un arc $A \leftarrow B$ dépend de la provenance de A et B.

tweet, un commentaire en bas d'un article, ou un *like* sur une publication Facebook présupposent tous la consultation du document auxquels ils se rattachent. Cependant ces mécanismes sont facultatifs, la consultation ne laisse généralement pas de trace évidente pour un observateur tiers. L'état de l'art présente certains travaux qui, à partir des publications et du contenu des documents, cherchent à estimer le graphe ayant permis la propagation d'information, et donc, en particulier, les consultations effectives qui ont conduit à une autre publication. Ces travaux sont présentés en section 2.2.3.

Les médias sociaux constituent le principal contexte observable dans lequel se propage l'information. Chaque heure, des quantités importantes de documents sont produites sur différentes plateformes. Après avoir donné un cadre précis pour définir ce que sont les informations, les documents, les médias et pourquoi l'information se propage, cette section a introduit les différents pré-requis techniques à l'analyse des

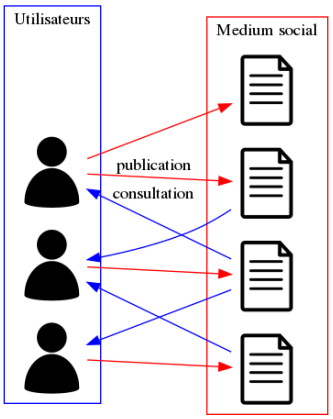


FIGURE 2.11 – Schéma de fonctionnement d'un médium social. Des utilisateurs publient des documents, d'autres les consultent.

documents textuels et à l'étude de la structure des réseaux sociaux. La section suivante présente l'état de l'art de la recherche portant sur la question de la propagation d'information dans les médias sociaux. Enfin, la dernière section du chapitre définit et justifie la recherche de la Trajectoire de l'information, comme une approche innovante pour aborder le phénomène de propagation.

2.2 État de l'art de la propagation d'Information

À la fin du XIXe siècle, les anthropogéographes allemands Leo Frobenius et Friedrich Ratzel constatent et référencent des phénomènes de diffusion des nouveautés entre différents groupes humains [Frobenius, 1898]. Le sociologue français Gabriel de Tarde, pour les expliquer, construit une théorie où la société repose sur l'acte d'imitation (Les lois de l'imitation) [Tarde, 1993]. Selon cette théorie, toute nouveauté (selon ses termes, toute invention), est le résultat de l'imitation, peut-être, très faible, d'inventions antérieures. Il dit ainsi "Toute similitude sociale a l'imitation pour cause.", *Les lois de l'imitation*, [Tarde, 1993], p. 40. Sur la base de ces travaux, se développe le champ d'étude de la diffusion des innovations, synthétisé dans les années 60, puis porté par Everett Rogers [Rogers, 2010]. Il constate, statistiquement, que l'adoption de nouveautés au sein d'une population suit, au cours du temps, un profil en **S**, à la manière d'une fonction logistique. La modélisation mathématique des

phénomènes de diffusion se poursuit, avec notamment le modèle de diffusion de Frank Bass [Bass, 1969] et les modèles issus de l'épidémiologie [Hethcote, 2000] qui auront une influence importante en sciences de gestion et en marketing, ce qu'on appelera le marketing viral.

Avec l'apparition du Web dans les années 90 et la démocratisation des ordinateurs personnels, un nouveau terrain d'étude de la propagation sociale se développe : les échanges sont traçables, parfois formatés. Les contenus sont produits par plusieurs millions d'entités autonomes, dont des chercheurs, des grands groupes privés, des commerçants, des agences gouvernementales, des associations, et aussi des individus. Le Web en lui-même est un réseau social. Jon Kleinberg propose de l'étudier comme tel, ainsi tous les nœuds du réseau, ce sont ici les sites Web, n'ont pas la même importance vis-à-vis de la propagation. Certains sites sont très consultés, tandis que d'autres le sont beaucoup moins. Ces sites très consultés ont donc un rôle important pour la diffusion des informations. Ce sont aussi les sites que des utilisateurs sont susceptibles de rechercher. C'est sur ce constat que Jon Kleinberg, d'abord [Kleinberg, 1999], Sergey Brin et Larry Page (les fondateurs de Google), par la suite [BrinPage, 1998], proposent une méthode de calcul de cette importance des sites (respectivement HITS et PageRank). Une autre application de cette importance concerne directement le marketing : quels nœuds doivent présenter une information pour que celle-ci se diffuse le mieux ? L'introduction classique de ce problème réside dans le travail de Kempe, Kleinberg et Tardos [Kempe, 2003] et est toujours au cœur de travaux récents [Li, 2018 ; Du, 2013 ; GomezRodriguez, 2016]. Une approche descriptive de ce phénomène est donnée par Leskovec, Adamic et Huberman [Leskovec, 2007]. De par sa richesse, il existe de nombreuses manières d'étudier et d'exploiter le contenu du Web. Nous nous concentrerons dans la suite de cet état de l'art sur quelques questions qui nous semblent essentielles dans le cadre de la propagation de l'information. Un lecteur qui voudrait une vision d'ensemble de l'exploitation du Web social peut se tourner vers [Zafarani, 2014].

Parmi les nombreux travaux gravitant autour du thème de la propagation de

l'information, nous présentons un état de l'art que nous rangeons selon les cinq axes de recherche suivants : détecter les informations et suivre leurs mutations, modéliser leur propagation, en inférer le support de diffusion, remonter à la source primaire d'une information, et enfin, proposer une vision synthétique du phénomène de propagation.

2.2.1 Suivre l'information et ses mutations

Dans les médias sociaux, les publications sont des documents composés d'informations et constitués par des individus. Or, l'information se propage entre les individus. On s'attend ainsi à ce que certaines informations d'un document consulté se retrouvent dans une ou plusieurs publications ultérieures du fait de la propagation. Ces informations auront peut-être une forme ou un contexte légèrement différents, elles auront mutées. Nous regroupons ici les travaux qui s'intéressent à la détection des informations propagées. Nous distinguons en particulier les quatre tâches suivantes :

— L'identification des informations propagées.
— La localisation des informations propagées.
— La détection de la mutation.
— L'analyse de la mutation.

Identification et localisation des informations propagées

L'identification et la localisation des informations d'un corpus sont deux problèmes fondamentaux, ce sont généralement des prérequis pour toute analyse du corpus. L'identification automatique repose principalement sur un *a priori*, un modèle, de ce qui constitue ou indique une information. Certains de ces modèles sont détaillés dans la section 2.1.2 traitant la représentation textuelle des documents. Parmi les informations du corpus, certaines se propagent. La question de l'identification de ces informations qui se propagent doit être traitée conjointement avec une modélisation de la propagation. Par exemple, Yang et Zha [YangZha, 2013] proposent d'inférer le support de diffusion, problème sur lequel nous reviendrons, et d'identifier les

informations en même temps. La localisation automatique d'une information, même connue, peut être ardue. Il s'agit d'une question centrale en *Recherche d'Information* (RI), dont les principes peuvent être trouvés dans [Manning, 2008].

Même lorsque les informations ne sont pas identifiées, dès lors qu'on est capable de comparer les documents, il est possible de partitionner le corpus en plusieurs sous-corpus de documents similaires. On parle alors de *clustering*. Un sous-corpus de documents similaires est susceptible de contenir un ensemble d'informations similaires, qu'on appelle le thème (ou thématique). Les approches classiques de *clustering* sont présentées dans [Zafarani, 2014] [10]. Un autre angle consiste à former ces thèmes en premier lieu, on parle alors de *Topic Model*, l'algorithme le plus connu est certainement LDA [Blei, 2003] et sa version dynamique [BleiLafferty, 2006]. Un état de l'art extensif sur le sujet se trouve dans la thèse de Mohamed Dermouche [Dermouche, 2015].

Détection et analyse de la mutation

La mutation peut se reconnaître de deux manières. Soit on détermine exactement le mécanisme de mutation, et on peut constater les objets qui le respectent. Soit on ne détermine pas ce mécanisme, et on s'appuie sur des indice, des traces, des similarités, qui laissent supposer l'existence d'une mutation.

Dans certains cas, on connaît le mécanisme de mutation. C'est le cas de la plupart des fautes de frappe selon Frederick Damerau qui sont soit une insertion, une suppression ou une substitution d'une lettre, soit une transposition de deux lettres[Damerau, 1964]. Il s'agit également de ce que les généticiens appellent une mutation ponctuelle d'ADN ou d'ARN, et ils construisent des systèmes performants pour les détecter [Caporaso, 2007]. Dans le contexte des corpus de document, Leskovec Backstrom et Kleinberg se sont intéressés à l'analyse de la citation [Leskovec, 2009]. La citation est plus simple à détecter explicitement, parce qu'elle se trouve généralement entre guillemets. Ils définissent la mutation d'une citation par le troncage : il s'agit

10. Plus précisément dans les sections 5.5.1 et 6.1 du livre, en section 6.2 pour les approches dynamiques

de la même citation à laquelle il manque certains mots. Le suivi des citations et leurs mutations permet aux auteurs d'étudier la dynamique de propagation des citations au sein de la presse et des blogs à large échelle (avec plus de 90 millions de documents). Dans le sillon de ces travaux, d'autres chercheurs ont proposé des améliorations techniques à l'extraction de citations [Snowsill, 2011], et se sont intéressés aux analyses qu'on peut en déduire. [Cheng, 2016] et [Zhang, 2017] s'intéressent particulièrement au caractère cyclique des informations dans les médias. [MyersLeskovec, 2012] et [Zarezade, 2017] s'intéressent aux interactions entre les différentes informations qui circulent : certaines peuvent entrer en conflit et se partager l'espace médiatique, voire très peu se propager, tandis que d'autres, au contraire, s'entraident dans le phénomène de propagation.

Lorsqu'on ne connaît pas le mécanisme de propagation, on peut s'appuyer sur un calcul des similarités. L'inconvénient par rapport à la première méthode est qu'il s'agit d'une supposition de mutation, la relation n'est plus explicite, comme ça peut être le cas pour une citation ou un retweet. Il s'agit de l'approche standard en détection de réutilisation de texte, les principales mesures de similarité classiques sont listées dans [Bär, 2012] et nous en avons présentées certaines en section 2.1.2. Il existe également des méthodes à base de fonction de hachage, comme décrit dans [SeoCroft, 2008] et [Alvi, 2014]. À notre connaissance, hormis le travail de Shahaf et Guestrin [11] cherchant à connecter deux documents quelconques dans un corpus [ShahafGuestrin, 2010], il n'y a pas de travaux analysant la mutation dans les médias sociaux sous cet angle.

2.2.2 Modèles de propagation

Comment se propage l'information ? L'étude des phénomènes de propagation couvre un large champ d'objets de nature diverse, dont l'information. Aussi, il existe de nombreux modèles, dont plusieurs ont été appliqués à l'information. Certains sont issus de l'épidémiologie, d'autres de la physique (par exemple [Bourigault, 2014] utilise

11. Ce travail est présenté plus en détail dans la section 2.2.5.

des équations de la chaleur), du marketing ou des sciences sociales. Nous distinguons deux grandes familles de modèles de propagation. Les modèles **structurés** sont les modèles qui prennent en compte les interactions entre les différents acteurs de la propagation, ils s'opposent aux modèles **non structurés** qui ne prennent pas en compte ces interactions. Dans les modèles structurés, on distingue les **modèles centrés receveur** (*receiver-centric*), il s'agit des modèles où chaque nœud décide de participer à la propagation en fonction de ses voisins. À l'opposé, dans les **modèles centrés émetteur** (*sender-centric*), chaque nœud qui participe à la propagation décide de faire participer ses voisins.

Les modèles non structurés sont généralement utilisés pour évaluer le nombre de personnes impactées et la vitesse du phénomène de propagation. Goel, Anderson, Hofman et Watts suggèrent que la structure du réseau de propagation a un impact négligeable sur ces mesures [Goel, 2016]. Les modèles non structurés les plus utilisés sont le modèle de Frank Bass [Bass, 1969], et les modèles d'épidémie [12] **SI** (Sain→Infecté), **SIR** (Sain→Infecté→Résistant) et **SIS** (Sain→Infecté→Sain). Ils sont documentés dans [Zafarani, 2014] [13] et [Jackson, 2010] [14]. Comme nous nous intéressons particulièrement à la structure de la propagation d'information, nous traitons les modèles structurés plus en détails que les modèles non structurés.

Modèles centrés receveurs, le comportement grégaire

Le comportement grégaire est un mécanisme de prise de décision issu de la psychologie des foules. L'idée du comportement grégaire est qu'un individu effectue ses choix en fonction du comportement des autres individus qu'il observe. L'individu a tendance à valoriser les choix faits par un grand nombre d'individus. L'exemple classique est le choix entre deux restaurants [Banerjee, 1992] pour déjeuner, l'un est rempli, l'autre est vide. Choisir le restaurant sur le critère de sa fréquentation est un cas type de comportement grégaire.

12. Les modèles d'épidémie ont aussi une variante structurée. Elles sont présentées dans [Jackson, 2010]. Ce sont des modèles centrés émetteur.
13. Au Chapitre 7.
14. Au chapitre 7, section 1 et 2

L'avis ou les choix de ses voisins influencent la prise de décision d'un individu, modéliser ce fait conduit à des modèles centrés receveurs. Le contexte du modèle centré receveur est le suivant :

— Chaque acteur [15] a le choix de participer à la propagation ou non.
— Chaque acteur a un voisinage visible d'acteurs, dont il observe les choix.
— Selon les choix de son voisinage, et son intention *a priori* de participer, chaque acteur prend sa décision.

Il existe de multiples familles de modèles qui partent de ce postulat. La première est celle des **modèles par seuils**, d'abord proposée par Mark Granovetter [Granovetter, 1978]. L'idée est que chaque acteur possède un seuil de participation. Lorsque son nombre de voisins participant dépasse le seuil, l'acteur décide de participer. Eli Berger [Berger, 2001] et Duncan Watts [Watts, 2002] étudient et constatent que les modèles de seuil peuvent donner lieu à des participations massives et exceptionnelles, sous des conditions initiales simples. Le modèle de seuil est un des modèles les plus utilisés pour étudier la propagation d'informations. Il a été utilisé pour étudier la propagation au sein des blogs [Gruhl, 2004] et des réseaux sociaux [Xiong, 2012 ; Lagnier, 2013]. On peut également s'en servir pour trouver les acteurs qui assurent la meilleure propagation [Nematzadeh, 2014].

La deuxième famille, plus générale, étudie la participation à la propagation de chaque acteur comme un jeu, auquel l'acteur doit décider de sa meilleure stratégie. Il s'agit des **modèles issus de la théorie des jeux**. Lawrence Blume d'une part [Blume, 1993] et Glenn Ellison d'autre part [Ellison, 1993], formalisent et étudient un cadre très général des interactions entre des joueurs dans un graphe de voisinage. Plutôt utilisés en économie, ces modèles sont aussi appliqués pour la propagation d'information [Hajibagheri, 2013 ; Jiang, 2014 ; MontanariSaberi, 2010] ou la détection de communautés [Alvari, 2014].

15. Lorsqu'un ensemble d'acteurs effectue des choix, il s'agit d'une modélisation basée sur des agents. Une introduction à l'utilisation de cette catégorie de modèles pour l'étude de phénomènes sociaux peut être trouvée dans [MacyWiller, 2002].

Modèles centrés émetteur : modèles de cascades

Les modèles centrés émetteur partent de l'idée que les acteurs décident d'exposer leurs voisins au phénomène de propagation. Par exemple, dans un réseau social comme Facebook ou Twitter, l'utilisateur du réseau publie du contenu. Ce contenu est présenté à tous ses amis (respectivement, sur Twitter, ses abonnés). Il n'est pas nécessaire qu'une masse critique d'amis d'un utilisateur propage pour que cet utilisateur décide de la relayer à son tour. Une personne suffit. Mark Granovetter suggère à ce propos que l'information nouvelle passe essentiellement par des *liens faibles* [Granovetter, 1973]. Dans les réseaux sociaux, les gens sont fortement connectés lorsqu'ils ont de nombreux contacts communs. Un ensemble de personnes fortement connectées les unes aux autres forment une *communauté*. *A contrario*, deux personnes connectées, mais ayant peu ou pas de contacts communs sont faiblement connectés. Ainsi les liens faibles agissent comme des ponts informationnels entre les différentes communautés. Goldenberg, Libai et Muller constatent l'importance des liens faibles dans leurs travaux sur le bouche-à-oreille [Goldenberg, 2001]. Selon Stanley Milgram, les liens faibles dans les réseaux sociaux sont disposés de telle sorte que la distance entre deux personnes quelconques est extrêmement courte. Il s'agit de l'*hypothèse du petit monde* [Milgram, 1967], elle a été constatée pour le graphe de Facebook (avec une distance moyenne de 4,74 entre deux personnes [Backstrom, 2012]) et celui de Twitter (distance de 3,88 [Bakhshandeh, 2011]).

Les modèles d'épidémie sont un cas type de modèles centrés émetteur. Il suffit d'être au contact d'une seule personne infectée pour avoir un risque d'être à son tour infecté. Dans le cadre de la propagation d'information, lorsqu'on a plusieurs informations qui se propagent, on parle de **modèles de cascades d'informations** [16], formalisme introduit par Kempe, Kleinberg et Tardos [Kempe, 2003]. Le principe du modèle de cascades est le suivant :

— Certains acteurs connaissent une ou plusieurs des informations qu'on suit.

16. La cascade désigne la propagation de proche en proche d'une information. En ce sens, on peut également observer le phénomène de cascade avec des modèles centrés receveur. La notion de **modèles de cascades** dont on parle ici est introduite par Kempe, Kleinberg et Tardos [Kempe, 2003], en opposition aux modèles de seuils.

— Selon les informations qu'il connaît, et le temps depuis lequel il les connaît, un acteur peut partager ponctuellement une ou plusieurs informations à un ou plusieurs de ses voisins.
— Au bout d'un certain temps, on arrête le modèle.

Le modèle original de Kempe et al., ainsi que son inspiration, le modèle de Goldenberg et al. [Goldenberg, 2001], sont discrets, de cascades indépendantes, sans infections multiples et à occasion unique.

Le modèle est dit **discret** lorsqu'il fait intervenir un temps discret. Il se décrit étape par étape, à la manière d'un automate. Il s'agit des premiers modèles de cascades proposés, comme ceux présentés dans [GomezRodriguez, 2010] et [Bakshy, 2012]. À l'inverse, se sont développés des modèles **continus**, où les propagations entre acteurs sont modélisées par des *processus ponctuels* [DaleyVereJones, 2007], c'est le cas des modèles présentés dans [Gomez Rodriguez, 2013 ; Bourigault, 2014 ; YangZha, 2013 ; Louzada Pinto, 2016 ; Zhao, 2015 ; Du, 2013].

Un modèle est dit à cascades indépendantes si les informations n'agissent pas les unes sur les autres. Les travaux sur le suivi des informations montrent que celles-ci, au contraire, interagissent et modifient le comportement final [Zarezade, 2017 ; MyersLeskovec, 2012]. Certains travaux proposent d'intégrer les documents au modèle pour capturer ce phénomène [Guille, 2014 ; Wang, 2012].

Le modèle est dit sans infections multiples si un acteur ne peut pas être soumis à une information qu'il connaît déjà. Wang, Ermon et Hopcroft soulignent le fait que pouvoir rafraîchir la date où un acteur a été soumis à une information améliore la modélisation pour certains problèmes [Wang, 2012].

Enfin, le modèle est à occasion unique si chaque acteur a une et une seule occasion de partager chaque information à ses voisins. Zhang, Zao et Xu montrent que les informations se propagent généralement plusieurs fois (typiquement deux), avec des écarts temporels significatifs [Zhang, 2017]. Cheng, Adamic, Kleinberg et Leskovec arrivent à des conclusions similaires [Cheng, 2016] et proposent de prédire la répétition des cascades.

Nous avons évoqué les modèles de la propagation d'information. Ils sont régis et classés essentiellement selon deux mécanismes : d'une part, ceux qui expliquent la propagation au travers du comportement du plus grand nombre (le comportement grégaire) et ceux qui l'expliquent par l'exposition a de multiples signaux provenant de diverses sources. Les deux phénomènes de propagation existent, et mènent à des résultats différents [Sela, 2016]. Il convient généralement d'étudier les deux [Kempe, 2003]. Les modèles structurés partent d'un réseau sur lequel s'effectue la propagation. Dans certaines situations, ce réseau est inconnu. Le retrouver est l'objet de la section suivante.

2.2.3 Inférence du support de propagation

L'information se propage d'individu à individu au travers de canaux de communication. La structure composée des individus et des canaux de communication impliqués dans la propagation d'un ensemble d'informations constitue ce qu'on appelle son **support de propagation**. Ce support est généralement inconnu. On peut estimer la structure à partir du réseau social des individus, mais celui-ci n'est pas toujours connu non plus. Même lorsqu'il est connu, il est *a priori* incomplet vis-à-vis de la propagation. Dans un média social clos, où l'information ne peut se propager qu'entre utilisateurs connectés, l'information peut se propager en dehors du média, via le bouche-à-oreille, la radio, la télévision, etc. De plus, le réseau social fournit des liens qui ne sont pas nécessairement utilisés par la propagation. La connexion sociale n'implique pas la retransmission de toutes les informations.

Il est donc nécessaire, pour étudier la structure de la propagation, d'en retrouver le support. Il s'agit du problème d'inférence du support de propagation. Le problème se décline en deux types, selon qu'on connaît déjà une partie du support ou non. Lorsqu'on connaît déjà une partie du support, on peut s'en servir pour prédire les parties manquantes, à la manière d'un puzzle en partie rempli. Lorsqu'on ne connaît pas le support, on peut tirer parti des indices laissés par le phénomène de propagation, on parle de **traces de propagation**.

Connaissance d'un support partiel : Prédiction de nœuds et liens manquants

La prédiction de nœuds ou de liens manquants se définit dans un cadre plus général que l'étude des réseaux sociaux. Il émerge lorsqu'on a un graphe dont on sait, ou a minima on suspecte, qu'il en manque des parties. Pour retrouver ces éléments manquants, il est nécessaire de faire des hypothèses concernant la forme du graphe, il en existe de deux types.

Les *hypothèses structurelles* décrivent une structure idéale du graphe. L'écart entre le graphe réel et le graphe idéal correspond aux éléments manquants. Par exemple, l'hypothèse du petit monde de Milgram [Milgram, 1967] implique une structure de graphe avec des chemins les plus courts entre deux nœuds de petite taille. Clauset et al. [Clauset, 2008] proposent une méthode pour retrouver les liens manquants dans des graphes qui respectent une forme de hiérarchie : dans les réseaux sociaux il s'agit de communautés imbriquées (votre cercle d'amis, les habitants de votre ville, ceux de votre pays, etc.). Leskovec et al. proposent le modèle des graphes de Kronecker [Leskovec, 2010] qui permet de générer des graphes contraints par certaines hypothèses de structure. Kim et Leskovec proposent une méthode d'application de ce modèle pour retrouver les éléments manquants d'un réseau [KimLeskovec, 2011].

Les *hypothèses de nature* décrivent comment les nœuds doivent se lier du fait de leurs propriétés intrinsèques. Par exemple, en reprenant le comportement grégaire, le constat de deux individus faisant la même succession de choix indique qu'ils sont probablement liés d'une certaine manière. Lorsque des nœuds sont similaires, parce qu'ils fréquentent les mêmes endroits, ont les mêmes intérêts, écrivent sur les même sujets, on s'attend à les voir liés dans le graphe. On utilise généralement à la fois les deux types d'hypothèses, plusieurs travaux montrent que cela améliore les résultats [MenonElkan, 2011 ; AdarAdamic, 2005 ; Wang, 2015b]. Adar et Adamic [AdarAdamic, 2005] utilisent déjà les deux hypothèses dans leur étude de la propagation des URLs sur la blogosphère.

Pour un tour d'horizon plus poussé sur le problème de la prédiction de nœuds et

liens manquants, Lü et Zhou présentent les travaux dans le cadre général [LüZhou, 2011]. Wang et al. décrivent les principales méthodes utilisées dans le cas des réseaux sociaux [Wang, 2015b].

Cas du support inconnu *a priori*

Il arrive régulièrement que le support de propagation d'une information soit inconnu et inaccessible. Cependant, on peut être capable de détecter la présence de différentes informations relayées par différentes personnes. On appelle **trace de propagation** la donnée des trois éléments suivants :

— L'information détectée.

— L'émetteur de l'information. Il peut s'agir d'un document, un auteur, etc.

— La date d'émission de l'information.

Muni d'un ensemble de traces, et d'un modèle de propagation, il devient possible, via des méthodes d'optimisation, d'estimer la structure expliquant le mieux les traces de propagation. Ce problème est démontré NP-difficile [Daneshmand, 2014], mais de nombreuses heuristiques ont vu le jour durant la dernière décennie. Les premières, à notre connaissance, sont issues des travaux de Manuel Gomez-Rodriguez (et al.) et sont nommés *NetInf* [GomezRodriguez, 2010] et *NetRate* [GomezRodriguez, 2011]. Elles sont basées sur le modèle de cascades indépendantes de Kempe et al. [Kempe, 2003]. L'idée est la suivante : la probabilité d'existence d'un lien entre deux nœuds, de u vers v, dans le support de propagation, dépend de la fréquence à laquelle une information relayée par u est à son tour relayée par v, dans une certaine fenêtre de temps. La principale différence entre *NetInf* et *NetRate* est que le modèle de propagation est supposé discret pour *NetInf* et continu pour *NetRate*. *NetRate* fournit généralement de meilleurs résultats que *NetInf*, aussi les travaux d'inférence du support suivants se sont focalisés sur une modélisation continue de la propagation, à l'aide de processus ponctuels [YangZha, 2013 ; Zarezade, 2017 ; Zhao, 2015 ; Louzada Pinto, 2016]. Tout comme pour le cas de la prédiction de liens manquants, la connaissance de la nature des nœuds ou une notion de similarité

entre ceux-ci sont autant d'*a priori* sur les liens qui devraient améliorer les résultats de l'inférence. [Wang, 2012 ; Louzada Pinto, 2016 ; Guille, 2014 ; Bourigault, 2014] proposent tous cette intégration. En particulier, [Wang, 2012] propose de considérer la possibilité d'infections multiples, par une même information, d'un nœud. Louzada Pinto [Louzada Pinto, 2016] et Adrien Guille [Guille, 2014] proposent des modèles spécifiques aux informations fournies par Twitter. Bourigault et al. proposent un modèle original de diffusion basé sur l'équation de la chaleur [Bourigault, 2016].

Le support de propagation est essentiellement évolutif. Les gens se connectent et se déconnectent au cours du temps, les chemins de propagation, et donc leur support, changent tout autant. Certains travaux cherchent ainsi à inférer le support évolutif de la propagation, notamment *InfoPath* de Gomez-Rodriguez et al. [Gomez Rodriguez, 2013], *PARAFAC* de Shen et al. [Shen, 2017] et *Dyference* de Ghalebi et al. [Ghalebi, 2018].

Enfin, il est également possible d'utiliser les traces de propagation lorsqu'une partie du graphe est connu. Sundareisan et al. [Sundareisan, 2015] proposent ainsi de s'en servir pour retrouver dans le même temps des nœuds manquants et les sources primaires de la propagation. Wang et al. [Wang, 2015c] proposent un apprentissage du graphe par transfert en utilisant un graphe connu pour d'autres phénomènes de propagation.

L'inférence du support de propagation peut utiliser de nombreuses techniques, selon ce qui est su, supposé et la nature du support souhaité. Il s'agit néanmoins d'une étape importante. La connaissance du réseau de propagation permet de calculer les influenceurs [Kempe, 2003 ; Li, 2018] pour propager efficacement de nouveaux produits ou mieux comprendre et suivre l'information. Gonzalo Mateos et al. proposent un état de l'art à jour de la problématique d'inférence [Mateos, 2019].

Dans la section suivante, nous discutons d'un problème similaire en apparence : retrouver la source primaire d'une information.

2.2.4 Trouver la source primaire

Lorsqu'un nouveau virus se propage, il peut être critique, pour mieux analyser son comportement et ses effets, de retrouver le patient zéro : le premier infecté. De la même manière, lorsqu'une information se propage, son premier émetteur, sa **source primaire**, est une donnée souhaitable. L'autorité d'une source permet, par exemple, de décider ou non de croire en une information. Par une approche rationnelle, on aura plus tendance à croire le travail d'un journaliste ou d'un chercheur que les propos d'un inconnu. Or, d'après les travaux de Berger et Milkman [BergerMilkman, 2012], nous avons surtout tendance à propager, et par extension juger crédible, des informations excitantes sur un plan émotionnel. Ce biais peut être exploité pour nous amener à croire en des faits, même lorsque ceux-ci n'ont pas eu lieu. Il s'agit du champ d'étude des fausses nouvelles (*fake news*). Nous vous renvoyons vers les travaux d'Alcott et Gentzkow [AllcottGentzkow, 2017] sur l'emploi des fausses nouvelles dans la campagne présidentielle américaine de 2016 pour de plus amples informations. La source primaire d'une information est donc un élément de contexte décisif pour le travail de vérification des faits (*Fact checking*). Tout particulièrement à une époque où, via les médias sociaux, des millions d'acteurs témoignent, discutent et relayent chaque jour plusieurs faits, vérifiés ou non.

Le problème consistant à retrouver la source primaire d'une information se pose dans le contexte suivant : la propagation d'une information a eu lieu. Elle est partie d'une, ou plusieurs sources primaires. Cependant, les sources primaires sont indifférenciées des autres acteurs de la propagation. Selon la quantité d'information connue sur la propagation, plusieurs méthodes ont été proposées pour trouver les sources primaires.

Lorsqu'on connaît l'intégralité des participants à la propagation, et le réseau social, Shah et Zaman [ShahZaman, 2010] proposent une méthode basée sur une mesure de centralité : le centre de rumeur. Intuitivement, il s'agit du nœud du graphe duquel partent le plus de chemins de propagation. Luo et al. l'étendent pour le cas des sources primaires multiples [Luo, 2013]. Prakash et al. [Prakash, 2012] proposent

d'estimer la probabilité qu'un nœud soit la source primaire sachant le support de propagation résultant à l'aide de techniques d'analyse spectrale. Fioriti et Chinnici [Fioriti, 2014] vont plus loin. Leur hypothèse est que plus un nœud est proche de la source primaire, plus son impact sur la connectivité du support est grand. En pratique, dans le cadre de la propagation d'informations, l'intégralité du réseau ou des participants sont loin d'être connus.

Plusieurs travaux se placent ainsi dans des contextes lacunaires, où ils n'ont qu'une connaissance partielle du réseau des participants, et de la manière dont ils se connectent. Lokhov et al. [Lokhov, 2014], proposent de tirer parti de la date à laquelle chaque acteur a participé à la propagation. Sur un média social, par exemple, on sait à quel moment un utilisateur a publié du contenu. Farajtabar et al. [Farajtabar, 2015] proposent d'ailleurs de retrouver les sources primaires à partir de traces de propagations partielles et des techniques d'inférence du support de propagation de Daneshmand et al. [Daneshmand, 2014]. Brockmann et Helbing proposent une méthode générale de résolution du problème, à l'aide d'un concept de front d'onde de propagation [BrockmannHelbing, 2013]. Cependant, sa complexité la rend difficile à appliquer à des réseaux de grande taille.

Il existe d'autres approches à la recherche de sources primaires, notamment à l'aide de nœuds qui ont un rôle d'observateur, comme présentés dans les travaux de Pinto et al. [Pinto, 2012]. Elles sont plutôt utilisées dans le cadre de la détection de pannes ou d'attaque de réseaux (pollution d'un réseau d'eau, coupure d'un réseau électrique...). Une vision synthétique et comparée des méthodes de détection des sources primaires est donnée par Jiang et al. [Jiang, 2016].

Dans le problème suivant, il s'agit de produire un récapitulatif intelligible et fidèle des principales informations qui se sont propagées dans un corpus.

2.2.5 Résumer le corpus

Le phénomène de propagation des informations est inobservable dans son détail. Il n'est pas possible, en général, d'affirmer que la publication d'Alice, contenant

l'information a, a pour conséquence une publication de Bob, contenant l'information a', mutation de a. Les sections précédentes montrent cependant que plusieurs méthodes ont été mises en œuvres pour dégager des indicateurs caractéristiques du phénomène de propagation. Les méthodes de suivi de l'information (section 2.2.1) permettent de dégager les principales informations, d'en suivre le volume, ou d'étudier leurs cooccurrences. L'inférence du support de propagation (section 2.2.3) dégage la structure sous-jacente à la propagation : les canaux reliant ses différents acteurs. Cette section présente un problème à la croisée du suivi des informations et de la structure de la propagation. Dans un corpus de documents, comment se mettent en séquence les principales informations du corpus au cours du temps ? Il ne s'agit pas de capter les mutations fines et locales des informations, mais de reconstruire des **histoires**.

La narratologue Mieke Bal analyse la narration selon trois couches connectées [BalVan Boheemen, 2009]. Le **texte** correspond au support de la narration, cela peut être un texte écrit, lu, une bande dessinée, un jeu vidéo, etc. L'**histoire** est la mise en séquence d'événements qui donne lieu à une fabula. La **fabula** est l'effet sémantique, l'interprétation de la narration. Il est possible de mettre en séquences différents événements qui n'ont aucun lien les uns avec les autres. Ils ne produisent aucun sens au lecteur, aucune fabula. Cette séquence ne constitue donc pas une histoire.

Une partie des informations d'un corpus est relative à des événements, ce sont des faits. Certains faits s'organisent en séquences et produisent du sens, ce sont des histoires. Par exemple, la Figure 2.12 est une séquence de titres d'articles du Huffington Post. À la lecture des titres, on peut inférer une séquence d'événements :

1. Ryan Lochte et trois autres nageurs américains déclarent s'être fait voler à Rio.
2. Le juge brésilien n'est pas convaincu et les empêche de rentrer.
3. Ils sont finalement accusés de faux témoignage et de vandalisme.
4. Ils iront en commission de discipline.

Ryan Lochte Robbed At Gunpoint In Rio With 3 Other U.S. Swimmers	Sun, 14 Aug 2016 14:36:19 GMT
Brazil Judge Wants Answers From Ryan Lochte About Alleged Robbery	Wed, 17 Aug 2016 16:33:00 GMT
U.S. Swimmers Allegedly Robbed Along With Lochte In Rio Pulled From Flights Home	Wed, 17 Aug 2016 21:42:45 GMT
A British Olympian Was Allegedly Robbed At Gunpoint In Rio	Thu, 18 Aug 2016 11:05:04 GMT
U.S. Swimmers Should Be Charged For False Testimony, Vandalism: Brazil Police	Thu, 18 Aug 2016 17:36:48 GMT
U.S. Swimmers To Face Disciplinary Commission Over False Rio Robbery Claims	Fri, 19 Aug 2016 18:03:24 GMT

FIGURE 2.12 – Chaîne de titres du Huffington Post à propos d'une affaire de nageurs américains aux Jeux Olympiques de Rio. La chaîne a été extraite automatiquement en utilisant les méthodes développées dans le Chapitre 3.

Un corpus de documents, *a fortiori* des documents parlant de faits d'actualité, contient probablement de nombreuses histoires de la sorte. Plusieurs travaux sont consacrés à l'extraction des principales histoires d'un corpus. Il s'agit à la fois d'un problème de détection d'un type particulier d'informations : les faits, et de l'inférence d'une structure : la succession des faits.

Détecter les principaux faits d'actualité

Contrairement au concept général d'information, le fait d'actualité relate un événement du monde réel récent. Un tel événement a un lieu, une date proche de la date de publication des documents qui en parlent. De fait, les faits d'actualité importants ont une distribution particulière dans leur traitement médiatique. Leskovec, Backstrom et Kleinberg le constatent et proposent un modèle du fonctionnement des médias pour l'expliquer [Leskovec, 2009]. Allan, Papka et Lavrenko [Allan, 1998] sont les premiers, à notre connaissance, à poser le problème de la détection des faits d'actualité dans un flux de documents. Ils notent le rapprochement avec le problème similaire de l'extraction de thématiques, avec la composante supplémentaire de la nouveauté, qui décrit un profil en fonction du temps particulier. Aussi, une approche

classique est d'utiliser les méthodes d'extraction de thématiques dynamiques, comme la version dynamique de LDA proposée par Blei et Lafferty [BleiLafferty, 2006]. Diao et Jiang proposent une version utilisant des processus de restaurant chinois au lieu des processus de Dirichlet de LDA [DiaoJiang, 2014]. Mele et Crestani fournissent des méthodes pour traiter le problème des documents hétérogènes, qui ne proviennent pas des mêmes médias, et ont des longueurs très différentes [MeleCrestani, 2017]. En particulier, plusieurs travaux s'intéressent à la détection des faits d'actualités sur des média où les messages sont courts type Twitter [Long, 2011 ; Metzler, 2012 ; Xie, 2016].

[Long, 2011], [Metzler, 2012], [Xie, 2016].

Résumer le corpus via des histoires

Les corpus sont de plus en plus grands. En particulier, le flux journalier d'articles de presse, de billets de blogs, et de publications sur les réseaux sociaux est devenu conséquent. La détection des actualités permet d'avoir un aperçu des principaux événements de la journée, mais comment se connectent-ils avec ceux qui sont survenus les jours précédents ? Il s'agit de l'axe de recherche de Dafna Shahaf. Elle propose, avec Carlos Guestrin, une méthode pour construire des successions cohérentes de documents qui connectent deux documents quelconques d'un corpus [ShahafGuestrin, 2010]. Sur cette base, ils proposent l'idée de la carte de métro (metromap) de l'information [Shahaf, 2012] : un ensemble d'histoires, composées de documents, expliquant au mieux le corpus, construit comme un problème d'optimisation. Une illustration de carte de métro est donnée en Figure 2.13. En parallèle, Yan et al. proposent également de construire des histoires expliquant au mieux le corpus [Yan, 2011], cette fois à l'aide de techniques de résumés de texte.

L'idée est adaptée de différentes manières. Tran et al. développent une méthode tirant parti des titres d'articles pour s'affranchir de la complexité de la détection de faits [Tran, 2015]. Vossen et al. cherchent à déterminer les relations entre les différentes histoires du résumé [Vossen, 2015]. Liu et al. proposent de construire des

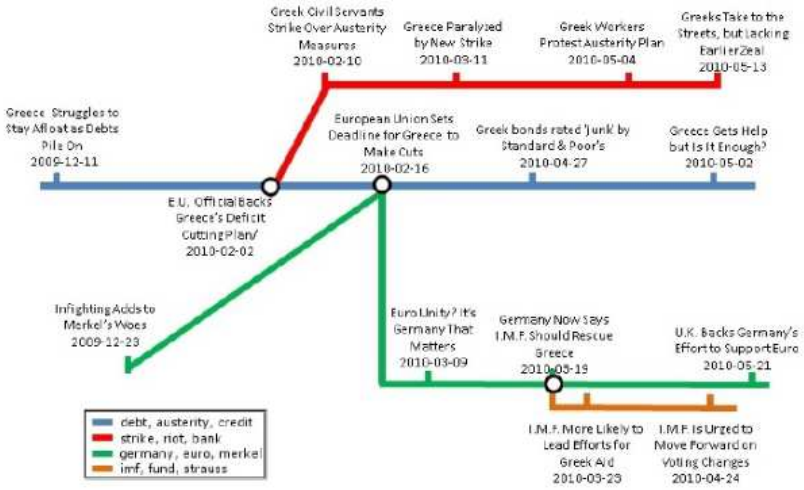


FIGURE 2.13 – Carte de métro construite par [Shahaf, 2012] au sujet de la crise de la dette grecque.

arbres d'histoires, mieux adaptés à une approche en ligne du problème, contrairement aux approches par optimisation qui peuvent restructurer les histoires à chaque ajout de documents [Liu, 2017]. Wang et al. suggèrent qu'exploiter les commentaires sous les articles améliore les résultats [Wang, 2015a].

Enfin, plusieurs autres modèles sont proposés. Notamment, Deyu Zhou propose des méthodes à base de processus de restaurant chinois [Zhou, 2015 ; Zhou, 2016] et de réseaux de neurones [Zhou, 2018]. Wang et al. proposent une approche par factorisation de matrices, qui leur permet également de prendre en compte les images pour leur génération d'histoires [Wang, 2016b]. Un état de l'art à jour des problèmes liés à l'extraction et l'explication de corpus par des histoires peut être trouvé dans le travail de Hou et al [Hou, 2019].

Nous avons vu, dans cette section, différents travaux autour de la propagation d'informations dans les médias sociaux. Nous n'avons pas couvert tous les champs d'étude qui s'y rapportent. Par exemple, nous faisons que mentionner l'existence des travaux sur le marketing viral ou l'analyse des fausses nouvelles. Cet état de l'art s'articule autour des questions qui nous ont semblées fondamentales pour la description de la propagation. Le premier axe exposé traite des informations en elles-mêmes, il s'agit de la facette sémantique de la propagation. Nous y avons vu

que l'identification et la localisation des informations sont des tâches compliquées, en particulier lorsque le phénomène de mutation est considéré. La seconde facette de la propagation est structurelle, il s'agit de déterminer la manière dont les informations se meuvent. Dans le deuxième axe de recherche il s'agit d'expliquer et de modéliser ce mouvement. Les principaux modèles exposés considèrent certaines propriétés de l'information dont on étudie la propagation (sa viralité), sa localisation dans la population d'individus à un instant donné, la manière dont les gens sont connectés (le réseau social). De ces données, les modélisations proposent une évolution de la propagation. Ainsi, les informations et le réseau social des individus sont supposés connus. Or ce réseau social, et plus particulièrement sa partie impliquée dans la propagation que nous appelons le support, n'est généralement connu que partiellement, ou est inconnu. Retrouver ce support est la problématique du troisième axe exposé.

En s'aidant d'une connaissance des informations et des mécanismes de propagation, c'est-à-dire les axes précédents, il devient possible d'inférer le support de la propagation. Ces trois axes couvrent ainsi la partie fondamentale de l'étude sur la propagation d'information dans les média sociaux : Ils englobent les sujets de la propagation, ce sont les relations entre les individus, les objets de la propagation, ce sont les informations, et la manière dont ils se coordonnent, ce sont les mécanismes de la propagation. Ces travaux forment un socle solide de connaissances sur le phénomène de propagation des informations. L'état actuel de l'art repose néanmoins sur notre capacité à identifier et localiser l'information.

Le quatrième axe présente le travaux autour du problème du retour à la source primaire d'une information. L'existence de ce problème souligne les limites de notre connaissance sur les aspects structurel et sémantique de la propagation. Si nous étions capables d'identifier et de localiser précisément l'information dont on souhaite retrouver la source, il s'agirait de choisir sa plus vieille occurrence. De la même manière, la connaissance du support de propagation renseigne sur les voies que l'information emprunte, mais elle n'en explicite pas les cheminements. Nous exposons justement dans le dernier axe les travaux sur la détection d'histoires. Ces travaux

soulignent l'existence de cheminements logiques entre les faits et leur traitement dans les média sociaux. Il s'agit, en un sens, de fournir les grands axes du phénomène de propagation et d'en résumer les étapes clefs et la manière dont elles se mettent en séquence. Ces deux derniers axes témoignent de l'idée qu'une information est changeante lors de sa propagation, qu'il est difficile de la suivre, et qu'elle s'inscrit dans une continuité, une forme d'histoire de la propagation.

Ces idées ne correspondent cependant pas à la manière actuelle de modéliser le phénomène de propagation. Nous pensons qu'il est intéressant de décrire le phénomène de propagation sous un angle différent, qui considère la notion de mutation. Une telle modélisation s'intègrerait dès lors dans ces cinq axes de recherche : elle redéfinirait l'information et expliciterait sa mutation d'une part. Elle contiendrait aussi des informations ou des indices nouveaux de structures dont on pourrait tirer parti pour étudier le support de propagation ou développer de nouvelles techniques pour retrouver les sources primaires des informations ou extraire des histoires. C'est dans cette perspective que nous exposons dans la section suivante notre problématique et nos motivations pour une nouvelle manière d'étudier la propagation.

2.3 Problématique et Positionnement

Ce chapitre a été l'occasion de présenter le concept d'information dans le cadre des médias sociaux. L'information est constitutive de la communication entre les individus, et nous avons noté qu'elle possédait les propriétés suivantes :

1. *Existence du phénomène de propagation.* L'information se propage, d'individu à individu, au travers des médias. On peut observer les effets de la propagation, en constatant que certaines informations similaires apparaissent en plusieurs lieux, même si on ne connaît pas son déroulé exact.

2. *Viralité de l'information.* Toutes les informations ne se propagent pas de la même manière. Certaines atteignent une grande quantité de personnes, d'autres restent dans un cercle très restreint, certaines se propagent autour du monde

en une journée (comme l'événement d'une catastrophe naturelle), d'autres se propagent pendant des années (comme certains travaux de recherche). La viralité peut-être intrinsèque à l'information, mais elle dépend aussi de ceux qui la relaient (il s'agit du problème des influenceurs).

3. *Phénomène de mutation.* L'information mute au cours de la propagation. Elle change de formulation, ou, a minima, de contexte, et par conséquent, elle change de sens.

L'existence du phénomène de propagation et la viralité de l'information sont largement étudiés dans l'état de l'art, ils sont d'ailleurs fondamentaux dans la plupart des axes de recherche présentés. Le phénomène de mutation y est notablement moins présent, ou, a minima, moins central.

Il y a une raison à cela : dans son acception la plus pure, le phénomène de mutation implique qu'il n'existe pas en deux lieux distincts la même information. Cela va, en surface, à l'encontre de l'idée de la propagation (et donc de la viralité). Il est donc plus naturel de définir une information comme un élément qui ne change pas durant la propagation, *a fortiori* lorsqu'on veut étudier ou tirer parti de la viralité de l'information. Rendre central le phénomène de mutation dans la modélisation de la propagation semble, d'une part, contre-intuitif, et contre-productif d'autre part. On peut néanmoins légitimement se demander quelles sont les conséquences de la mutation intégrée à la modélisation de la propagation de l'information.

La propagation s'effectuant d'individu à individu, il est naturel de se demander ce qui pousse un individu à propager une information. Pour l'expliquer, l'état de l'art tourne autour de deux mécanismes individuels généraux. On distingue le comportement grégaire d'une part, lorsque l'individu propage l'information par stratégie vis-à-vis de son entourage et de sa position sociale. D'autre part, on distingue Le mécanisme d'excitation, lorsque l'individu propage l'information selon l'effet que celle-ci a eu sur lui. La théorie de Gabriel de Tarde fusionne ces mécanismes en un seul, qu'il estime fondamental : l'imitation. Dans cette théorie, tout acte social est imitation. En particulier, la propagation d'une information est le résultat d'une

succession d'imitations [17]. Il est notable de constater que l'idée de l'imitation est intimement liée à l'idée de la mutation. L'imitation est la reconstitution approximative d'un fait, et même si l'imitation est parfaite, le contexte dans lequel elle est produite diffère du fait original. Cela correspond à notre concept de mutation : le fait est altéré, mais reste similaire à son origine. Il y a cependant une différence entre l'imitation et la mutation : l'imitation est intentionnelle, c'est un acte qui provient d'une volonté consciente ou inconsciente, et c'est en ce sens que l'entend de Tarde, tandis que la mutation est phénoménale, il s'agit du constat du résultat d'un processus qu'il n'est pas nécessaire d'expliquer. Une vision de la propagation comme une succession d'imitations soulève deux questions :

1. Imaginons qu'une information i se propage d'un individu M à un individu N. Cette information étant présente chez M, il n'y a que trois cas possibles :
   - Il s'agit d'une nouveauté, aucune information similaire connue ne la précède. M en est la source primaire. C'est le cas, par exemple, quand M est le premier à témoigner sur un fait nouveau [18].
   - M imite cette information d'un unique individu L.
   - M tire des informations similaires de multiples individus $L, L', L'', \ldots$.

   Est-il possible de déterminer dans lequel de ces cas s'inscrit cette propagation de M à N ?
2. En particulier, dans le troisième cas, où l'information de M provient de multiples sources $L, L', L'', \ldots$, il n'est pas dit que l'information que M transmet à N respecte indistinctement toutes les sources. M a vraisemblablement synthétisé les informations similaires reçues de $L, L', L'', \ldots$ pour les transmettre à N. Cette synthèse met vraisemblablement plus en avant certaines versions vis-à-vis des autres. Une explication vient directement de la viralité des différentes versions. Par exemple L est une figure d'autorité aux yeux de M, tandis que L'', en plus d'être un inconnu, tient des propos qui vont à l'encontre des

17. " Toute similitude sociale a l'imitation pour cause. " *Les lois de l'imitation*, [Tarde, 1993], p. 40

18. Cependant, M n'est pas la toute première personne à fournir un témoignage. Il y a donc, dans son message, une information sur la forme du discours, qui, elle, est une imitation.

convictions de M. Ainsi, M accorderait plus d'importance aux nuances de L qu'à celles de L''. Peut-on capturer cette filiation dans les nuances ?

Ces deux questions forment essentiellement la question de la généalogie de l'information. La première est relativement similaire aux travaux cherchant le support de propagation d'une information : on constate plusieurs informations similaires émises par différents individus, quel est le réseau sous-jacent ayant conduit à ce résultat ? La seconde question, si on y répond par l'affirmative, a des répercussions étonnantes. Une information a maintenant une provenance, elle est le résultat d'une ou plusieurs séquences d'informations qui se sont propagées et ont muté. Il s'agit, en quelque sorte, de son arbre généalogique, ou de son histoire de propagation. La Figure 2.14 illustre un exemple d'une telle structure. L'ensemble de ces provenances, pour toutes les différentes informations, constitue une représentation alternative de la propagation. Dans les modèles usuels, une information se propage d'un individu à un individu, on représente donc la propagation de cette information comme un graphe sur les individus. Ici, la localisation de l'information n'est plus l'individu, mais le document qui la contient. On représente donc la propagation comme un ensemble de chemins dans l'espace des documents, comme illustré en Figure 2.15. Ce qui nous amène à poser la problématique centrale de cette thèse :

Problématique. Étant donné un corpus de documents textuels, et considérant l'existence du phénomène de propagation de l'information, la viralité de l'information et le phénomène de mutation de celle-ci, est-il possible de reconstituer les chemins de propagation des informations au sein du corpus ? Quels avantages peut-on tirer d'une telle représentation ?

Le problème exposé ici est relativement différent de ceux présentés dans l'état de l'art, il partage évidemment certaines interrogations. Nous mettons en parallèle les questions que soulèvent la problématique et les différents travaux précédemment évoqués. Dans un premier temps, nous passons en revue les travaux qui exploitent la notion de mutation. Ensuite, nous soulevons la difficulté pour définir ce qui se propage dans un contexte de mutation et ses implications. Une implication particulière est

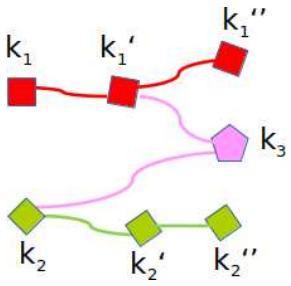


FIGURE 2.14 – Généalogie simple de plusieurs informations. k_1'' provient de la lignée $k_1 - k_1'$, k_2'' provient de la lignée $k_2 - k_2'$. k_3 provient de deux informations de lignées différentes.

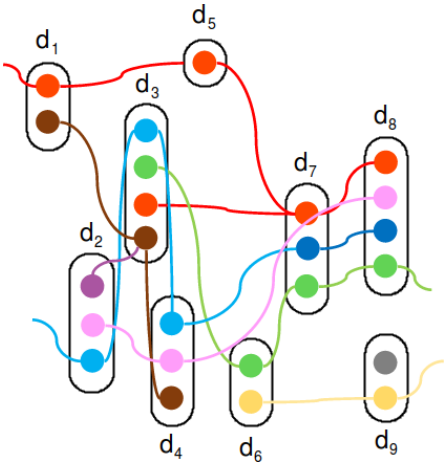


FIGURE 2.15 – Chemins de propagation de l'information dans un corpus de 9 documents. Les documents sont représentés par les contenants aux bords arrondis, les informations comme les pastilles contenues dans les documents. Les fils qui traversent les pastilles représentent les séquences de mutation.

d'étudier la propagation sur les documents, là où elle est généralement étudiée sur les auteurs. Enfin, la structure de chemins n'est pas sans rappeler les travaux sur l'extraction d'histoires, nous montrons qu'il s'agit bien de deux problèmes distincts. Toutes ces réflexions nous ont menés à aborder le problème sous l'angle d'une structure que nous appelons la **Trajectoire de l'information**, et dont la présentation clôture ce chapitre.

La mutation dans l'état de l'art

Nous déclarions en introduction que le phénomène de mutation était moins fondamental dans l'état de l'art, mais il serait faux de dire qu'il n'est pas considéré.

La mutation est d'ailleurs la justification centrale du problème de la détection et du suivi des informations : si l'information ne mutait ni dans sa forme, ni dans son sens, il serait plus aisé de la détecter ou de la suivre. C'est cette propriété qui motive les travaux sur la citation : sa forme change peu, ou de manière contrôlée, par exemple par troncature, et son intention est, *a priori*, de retranscrire tel quel le message de son auteur, sans en altérer le sens.

La mutation est donc toujours reconnue comme un problème endémique du travail sur l'information dans les corpus de textes. Dans la définition que nous donnons de la mutation, il s'agit de la transformation d'une information, en une autre, qui lui est similaire mais légèrement différente. Dans une vision plus algébrique, il y a, entre une information et sa mutation, d'un côté leur différence, de l'autre, un substrat commun, ce qui en fait, par essence, une mutation et non une information strictement différente. C'est par la définition et la recherche de ce substrat commun qu'il devient possible d'étudier la fréquence, la répartition et la viralité de celui-ci. L'extraction de thématique ou le clustering de documents sont, par exemple, deux techniques utilisées pour modéliser ce substrat commun. Pour prolonger l'exemple, toutes les techniques d'inférences du support par traces de propagation présentées font la supposition que les informations tracées auront une certaine redondance, qu'elles apparaîtront identiques en plusieurs endroits. Dans la pratique, ils utilisent donc des méthodes qui extraient ces substrats communs. Le phénomène de mutation est ainsi considéré : les informations sont différentes d'un texte à l'autre. Mais il s'agit d'un problème, résolu par la détection de ce substrat commun. Il s'agit en premier lieu [19] d'étudier une composante immuable de l'information, et non sa fluctuation.

Cette remarque ne signifie pas que la mutation n'est pas étudiée dans certains travaux, comme [Leskovec, 2009 ; Zarezade, 2017 ; Snowsill, 2011], mais que son étude a lieu dans un second temps, ou en complément d'une étude de la composante immuable de l'information. Or la nuance, la composante mutante, est peut-être source

19. Dans l'état de l'art sur la détection et le suivi des informations, nous avions relevé un champ d'étude en particulier qui considère la mutation de manière générale et non contrainte. Il s'agit de la détection de réutilisation de texte. Ici encore, le but final est de détecter une partie qui ne change pas.

d'informations précieuses. Wang et al., dans leurs travaux d'inférence du support de propagation [Wang, 2012], proposent, par exemple, d'ajouter à des méthodes, une notion de similarité entre les nœuds obtenus par la similarité des textes et constatent que cela améliore la performance de leur inférence. Les travaux d'extraction d'histoires se basent notablement sur l'idée que certaines séquences de documents sont meilleures que d'autres, et tirent parti, via la similarité entre les documents, des nuances pour sélectionner les meilleurs chemins.

Dans la lignée de ces travaux, et dans la perspective d'exploiter les détails de la composante mutante de l'information, nous nous demandons à quoi ressemblerait une modélisation non résumée de l'histoire de la propagation qui tienne compte, en premier lieu, du phénomène de mutation.

Savoir ce qui se propage

Nous souhaitons prendre en compte le phénomène de mutation pour étudier la propagation. Cela soulève une question clef, qui est de savoir, de pouvoir nommer, ce qui se propage d'un individu à un autre. Nous ne pouvons lister toutes les informations contenues dans un corpus, chaque document ayant son propre ensemble d'informations, et chaque information étant différente des autres. Lister les informations contenues dans un seul document est une tâche compliquée, il y en a, *a priori*, beaucoup, et la manière d'interpréter le document est subjective. Cependant, dans le cas où des informations similaires apparaissent en de nombreux lieux, cela indique l'existence d'une propagation qui induit la répartition constatée de ces informations. Ce qui se traduit ainsi :

Hypothèse 2. Un motif récurrent dans un corpus est signe d'une propagation d'informations.

Des connaissances sur les informations propagées se déduisent ainsi des connaissances sur la propagation. Mais l'inverse est également vrai : à partir de connaissances sur la propagation, on peut déduire des connaissances sur les informations propagées. Une manière de l'illustrer est de considérer deux documents A et B. Savoir que A et

B contiennent tous les deux un marqueur i, représentant un ensemble d'informations similaires issues d'un vaste corpus, conforte l'hypothèse que le phénomène de propagation lie A et B. À l'inverse, savoir que des informations communes, sans savoir exactement lesquelles, se sont propagées et ont atteint A et B permet de cibler la détection d'informations communes sur A et B, et d'en dégager des spécificités.

Il semble donc possible de déduire ce qui se propage à l'aide de la structure de la propagation. Si nous connaissons, de surcroît, toute une histoire d'une propagation, cela devient d'autant plus simple. Il est difficile de lister les informations contenues dans un document A par manque de contexte. Si on sait que des informations se sont propagées de A à B, l'ajout de contexte permet de mieux cerner ce qui se propage. Si on sait maintenant que des informations ont suivi un chemin $A \rightarrow B \rightarrow C \rightarrow D$, le contexte grandit d'autant. Et si on sait par ailleurs que des informations ont suivi un autre chemin $A \rightarrow B \rightarrow C \rightarrow E$, il devient possible de comparer la différence de ce qui se propage sur les deux chemins. Ce niveau de détail est rendu possible d'une part, par l'absence de détection préalable des informations, d'autre part, par une vision séquentielle, de l'information propagée.

Positionnement

Les travaux de l'état de l'art portant sur la propagation de l'information se basent sur l'hypothèse 2. Leur approche se déroule, schématiquement, en deux étapes :

1. La détection des informations du corpus.
2. La déduction, à partir, entre autres, des informations détectées, de propriétés relatives au phénomène de propagation. C'est ici qu'intervient l'hypothèse 2.

Or, nous venons de voir que si la connaissance préalable des informations nous renseigne sur la structure de propagation, à l'inverse la connaissance préalable de la structure de propagation nous renseigne sur les informations qui se propagent, et ce d'une manière détaillée et localisée. Notre approche, et cette thèse, se déroulent donc ainsi :

1. Détection de la structure de propagation sur le corpus, sans calcul *a priori* de

ce qui se propage (Chapitre 3).

2. Détection détaillée et localisée de ce qui se propage (Chapitre 4).
3. Déductions de propriétés relatives au phénomène de propagation (Chapitre 5).

Structure de propagation sur les documents

La première question est maintenant de savoir comment détecter la structure de propagation sur le corpus sans définir une notion d'information globale. L'idée principale est de s'appuyer sur des comparaisons locales des documents. Deux documents qui sont comparativement différents, ou qui, au contraire, ont un certain nombre de points communs sont plus ou moins similaires. Cela nous amène à considérer une version plus assouplie de l'Hypothèse 2 :

Hypothèse 3. Si deux documents sont suffisamment similaires, ils sont liés par un phénomène de propagation [20].

La similarité est ainsi un indicateur de la quantité d'informations similaires, et donc de la propagation. Ses variations d'un document à l'autre peuvent être un phare nous guidant sur les sentiers des mutations probables. Mais la similarité a de multiples inconvénients pour ce que nous souhaitons faire. Ce n'est d'abord pas un indicateur direct de la propagation d'un document vers un autre. Il suffirait, sinon, de construire le graphe de similarité pour obtenir un support de la propagation. Le fait que deux documents A et B soient similaires peut s'expliquer par plusieurs structures, dont :

— Une propagation directe $A \to B$,

— une propagation indirecte $A \to D_1 \to \ldots \to D_n \to B$,

— ou d'un ancêtre commun $C \to \ldots \to A$; $C \to \ldots \to B$.

Ceci nous incite à penser la propagation comme des cheminements, et non des événements de propagation d'individu à individu. La même raison motive la notion

20. Il s'agit, encore une fois, d'un cas particulier du postulat de Gabriel de Tarde : " Toute similitude sociale a l'imitation pour cause. " dans *Les lois de l'imitation*, [Tarde, 1993], p. 40

de cascade de propagation dans l'état de l'art, une cascade de propagation correspond à l'ensemble des chemins empruntés par une information. Ce modèle des chemins soulève une autre interrogation : comment faire les jonctions ? La similarité n'est *a priori* pas transitive : le fait que A soit similaire à B, et B à C n'implique en rien que A sera similaire à C. Il s'agit du même argument que mettent en avant les travaux d'extraction d'histoires. On peut l'illustrer ainsi :

Exemple 6. Dans les trois phrases suivantes, A et B ont des similarités, B et C aussi. Mais A et C n'ont rien à voir.

— **A :** J'adore les hamsters.

— **B :** J'adore le café.

— **C :** L'abus de café est mauvais pour la santé.

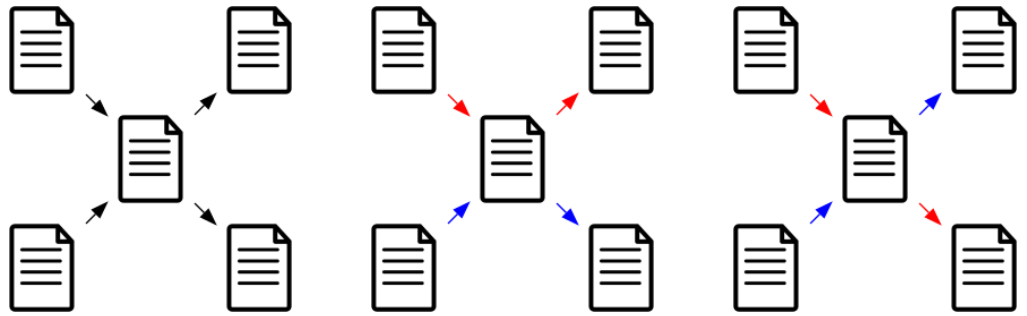
Enfin, la multiplicité des informations fait que de nombreux chemins de propagation différents peuvent relier deux documents, si bien que ceux-ci peuvent sembler très similaires et pourtant être le résultat de multiples chemins indirects. La similarité n'est donc pas un indicateur parfait. Cela explique peut-être pourquoi son utilisation a principalement été comme aide secondaire dans certains travaux d'inférence du support, ou dans les travaux d'extraction d'histoires. On peut d'ailleurs remarquer que les travaux de l'état de l'art sur l'inférence du support de propagation traitent, à notre connaissance, exclusivement d'un support sur les auteurs et non sur les documents comme c'est le cas ici. La similarité est néanmoins porteuse d'indices de structure. Nous montrons dans les chapitres 3 et 4 son utilité pour déterminer des structures cohérentes et les manières de pallier ses apparentes faiblesses.

Histoires et chemins de propagation

Parmi les différents axes de recherche présentés dans l'état de l'art, l'extraction d'histoires est certainement celui qui, conceptuellement, se rapproche le plus de notre problématique. Structure de séquences de documents, non-détection *a priori* des informations, et utilisation de la similarité sont autant de points que nous avons argumentés et qui sont présents dans les travaux d'extraction d'histoires. Il y a

cependant une différence fondamentale d'intention. L'extraction d'histoires cherche à synthétiser un corpus, à l'aide de séquences de documents, de manière à en dégager les grandes lignes factuelles. Pour ce faire, ils considèrent leur objectif comme un problème d'optimisation soumis à plusieurs contraintes, comme la cohérence narrative, la couverture du corpus, ou la non-redondance d'histoires similaires et la détection de leurs intersections. Il s'agit donc d'un résumé narratif du phénomène de propagation. De notre côté, il s'agit, dans un premier temps, de décrire la structure du phénomène de propagation dans le corpus. En ce sens, notre travail est plus proche des questions d'inférence du support de propagation. L'exhaustivité des chemins de propagation est un but, la cohérence recherchée n'est pas narrative mais descriptive du phénomène de mutation. Le travail de cette thèse est, ainsi, à la croisée de ces deux axes de recherche en proposant une manière de structurer la propagation en histoires, sur les documents, et de manière non résumée.

On peut cependant s'interroger sur l'utilisation de séquences de documents pour représenter la généalogie de la propagation. Dans le cadre de l'extraction d'histoires, il semble naturel qu'une narration soit séquentielle. Une généalogie est plus fréquemment représentée comme une arborescence, par exemple l'arbre généalogique d'une famille, ou l'arbre phylogénétique des espèces. Dans le modèle classique des cascades indépendantes [Kempe, 2003], la cascade d'une information est d'ailleurs un arbre. Il y a cependant deux raisons à l'utilisation d'une représentation en séquence. La première consiste à dire que des informations peuvent partir d'un même document et se propager jusqu'à un autre par plusieurs chemins différents. Ceci n'est pas représentable en utilisant un seul arbre. Dans les modèles de cascades, ils proposent de représenter ce phénomène en superposant un arbre par information qui se propage, mais comment déterminer la bonne dissociation de ces arbres sans connaître, *a priori*, les informations comme c'est notre cas ? D'ailleurs un arbre ne suffit pas non plus, car une même information peut partir d'un document et se décliner le long de plusieurs chemins différents qui finissent tous leur course au même endroit. C'est pourquoi certains travaux sur les cascades proposent l'utilisation d'un DAG



(a) (b) (c)

FIGURE 2.16 – Un DAG (a) peut se décomposer en plusieurs chemins, par exemple (b) et (c). Or nous voulons savoir d'où les informations contenues dans les deux documents de droite proviennent (du document en haut à gauche, de celui en bas à gauche, des deux, d'aucun des deux).

plutôt que d'un arbre comme [Wang, 2012]. Or le DAG souffre d'un problème majeur pour notre approche, illustré en Figure 2.16. Il n'est pas décomposable de manière unique en chemins, il n'est donc pas une structure suffisamment claire pour parler précisément de la généalogie de l'information. La séquence est donc le cas d'arbre le plus simple et le moins ambigu pour parler de la généalogie de la propagation. La seconde raison est que la séquence a un sens en isolation : il s'agit d'un chemin de propagation d'une ou plusieurs informations, il devient possible de l'étudier en tant que tel en dehors du reste du corpus. Elle a aussi un sens direct, celui d'une filiation, elle est donc plus simple à comprendre et à juger par un évaluateur humain qu'une structure plus complexe où l'évaluateur devrait juger de liens indirects non évidents.

Conclusion

Ce chapitre a été l'occasion de donner le cadre de ce travail : La propagation de l'information dans les médias sociaux. Dans un premier temps, nous avons défini la notion d'information, celle de médias et particulièrement de médias sociaux et nous avons vu que ceux-ci entraînent un phénomène de propagation de l'information au travers de la publication de documents. Il est possible de collecter ces documents, de manière, certes, toujours partielle, mais ils permettent de nombreuses analyses sur des quantités conséquentes de données. Cela nous a permis de voir les principales méthodes d'exploitation et d'analyses de textes, ainsi que les notions de théorie

des graphes, qui constituent le socle méthodique premier du travail des prochains chapitres. Dans un deuxième temps, nous avons détaillé un état de l'art sur la question de la propagation de l'information, que nous avons découpé selon cinq axes qui ont tous trait avec notre sujet. Nous avons d'abord étudié les méthodes de détection et de suivi de l'information, en plus d'être un préliminaire essentiel pour d'autres axes, ces méthodes permettent d'avoir une vue d'ensemble en dégageant des indicateurs chiffrés sur le phénomène de propagation. Ensuite nous avons vu les différentes approches de modélisation du phénomène de propagation avant de présenter les axes qui s'intéressent à la structure de la propagation. L'inférence du support de propagation cherche, à partir des indices laissés par les documents à retrouver le graphe de communication sous-jacent ayant permis la propagation. La recherche de la source primaire cherche à remonter à la source probable d'une information dans un milieu lacunaire. Enfin les travaux d'extraction d'histoires cherchent, à partir des documents, à reconstituer et mettre en lien les faits marquants d'un corpus sous la forme de séquences chronologiques.

A la lumière de cet état de l'art, nous avons dégagé que le phénomène de mutation de l'information n'était traité qu'en complément ou dans un second temps dans ces différents travaux. Après avoir donné une raison pragmatique à ce choix (il s'agit de dégager plus simplement les grandes lignes du phénomène de propagation), nous avons souligné que mettre la mutation au premier plan permettrait une compréhension plus fine et un modèle plus détaillé du phénomène de propagation en s'attaquant à la généalogie des informations. Cela dit, ce changement de paradigme entraîne son lot de questions, il n'est plus si simple de définir ce qui se propage exactement, il devient nécessaire de procéder localement, avec des comparaisons entre les documents. La structure induite devient linéaire et non plus binaire comme dans les modèles classiques où l'information passe d'un individu à un autre. Maintenant celle-ci chemine de documents en documents et se nuance en mutant. Les travaux d'extraction d'histoires considèrent aussi des séquences de documents, cependant ils ne le font pas pour étudier la structure de la propagation directement, mais pour en fournir un

résumé narratif. Nous proposons ainsi de représenter le phénomène de propagation en un ensemble de séquences de documents le long desquels une ou plusieurs informations se propagent, ensemble que nous appelons **la Trajectoire de l'information**. Nous consacrons le reste de ce manuscrit à sa définition, son calcul, et son exploitation. Nous montrons d'une part que notre méthode permet de construire des ensembles de séquences cohérentes. D'autre part nous montrons que nous sommes capables d'identifier précisément l'information qui se propage le long de ces séquences. Ce modèle permet également une manière innovante de visualiser et de naviguer dans un corpus, que nous présentons au chapitre 5 avec plusieurs applications que nous prévoyons et d'axes d'approfondissement du modèle.

Chapitre 3

Trajectoire de l'information et sa généalogie

Résumé. Dans ce chapitre, nous présentons la notion de Trajectoire de l'information. Il s'agit de l'ensemble des chemins empruntés par les informations lors de leur propagation. Nous commençons par exposer la question du calcul de ces chemins dans un corpus de documents. Nous posons en particulier la question de ce calcul sans connaissance *a priori* des informations qui s'y propagent. Après une section sur la formalisation du problème, nous proposons une méthode de calcul de chaînes cohérentes efficace. Nous présentons ensuite la campagne d'évaluation menée pour estimer les performances de notre méthode.

3.1 Introduction

Le problème de l'identification de la Trajectoire de l'information dans les médias sociaux consiste à retrouver les fils d'information qui lient les documents entre eux. Un document contient un ensemble d'informations que nous ne connaissons pas *a priori* (Figure 3.1). Du fait du phénomène de propagation, certaines de ces informations se propagent de documents en documents. On peut s'intéresser au support de propagation, qui peut être représenté par un graphe comme illustré en Figure 3.2. Nous voulons suivre le cheminement des informations. Par définition,

TABLE 3.1 – Notations utilisées

Notation	Signification
D	Corpus de documents
d_1, d_2	Documents du corpus
$date(d)$	Date de publication du document d
K	Ensemble des informations.
k_1, k_2	informations
$I(d)$	Ensemble des informations de d
sem_K	Similarité sémantique entre les informations
$P_{i,j}$	Ensemble des mutations entre d_i et d_j
$c = d_i d_j \dots d_k$	Chaîne de document
T	Trajectoire de l'information
T_D	Trajectoire de l'information projetée sur D
$T_{coh}(D)$	Trajectoire cohérente sur D
coh	Fonction de cohérence
γ	Seuil de cohérence
γ^*	Seuil de faible cohérence

FIGURE 3.1 – Un document contient de multiples informations, représentées par des pastilles colorées.

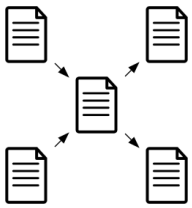
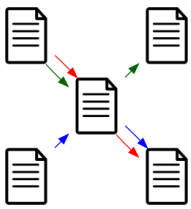
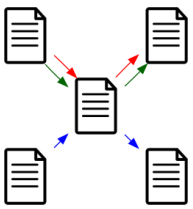


FIGURE 3.2 – Support de propagation entre des documents. Il y a un lien entre deux documents quand il y a de l'information qui s'est propagée de l'un à l'autre.

celles-ci se meuvent sur le support de propagation. Mais le support de propagation seul n'est pas suffisant pour déterminer la Trajectoire, les informations peuvent y circuler de plusieurs manières (cf. Figure 3.3). Nous proposons dans ce chapitre de suivre



(a) L'information rouge passe en bas.



(b) Ici elle passe en haut

FIGURE 3.3 – Sur un même support, les informations, identifiées par la couleur des arêtes, peuvent se déplacer de plusieurs manières.

l'information à l'aide d'une nouvelle structure, que nous appelons la **Trajectoire de**

l'information. Il s'agit de l'ensemble des successions de documents le long desquelles de l'information s'est propagée. On appelle chacune de ces successions une chaîne de propagation. Un exemple est donné en Figure 3.4. La Trajectoire fournit une nouvelle représentation du phénomène de propagation pouvant servir de base pour différentes études et analyses sur un corpus de documents. Nous en présenterons plusieurs dans le chapitre 5. En particulier, connaître la Trajectoire aide à identifier l'information qui se propage. Pour cette raison, nous souhaitons retrouver la Trajectoire avant même de détecter et identifier les différentes informations qui se propagent dans le corpus. Nous distinguerons *la Trajectoire de l'information*, écrit avec une majuscule, qui est une propriété particulière du phénomène de propagation de l'information, au même titre du support de la propagation. À l'inverse, pour parler de la structure qui correspond à un ensemble de successions de documents, qu'elle corresponde à une propagation ou non, nous parlerons de *trajectoire*, écrit avec une minuscule. La Trajectoire de l'information est une trajectoire particulière. Nous définissons précisément les notions de Trajectoire et de trajectoire dans le formalisme.

Nous proposons dans ce chapitre une méthode pour construire une trajectoire approchant la Trajectoire de l'information en utilisant uniquement une similarité entre deux documents. Le chapitre 4 est ainsi dédié à trouver les informations qui se propagent le long des chaînes d'information.

Le chapitre est structuré comme suit. Nous commençons par donner un formalisme des différents objets que nous manipulons, puis nous exposons le problème du calcul de la Trajectoire de l'information ainsi que la méthode que nous proposons pour l'approcher. Nous présentons ensuite la campagne d'évaluation que nous avons menée pour vérifier la pertinence des trajectoires ainsi construites ainsi que ses résultats.

3.2 Formalisme, notion de trajectoire

Dans cette section, nous présentons le formalisme au travers duquel nous exprimons nos problèmes. Nous commençons par rappeler plusieurs notions vues au chapitre précédent. Ensuite nous proposons un formalisme du phénomène de propa-

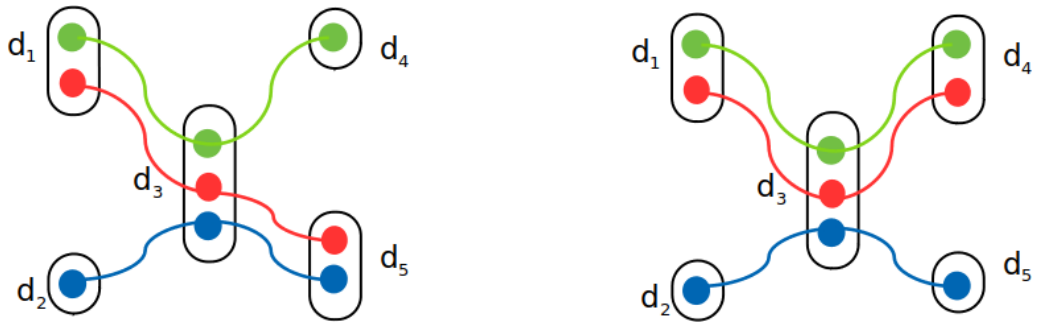


FIGURE 3.4 – La Trajectoire des informations selon les deux propagations vues en Figure 3.3, chaque ligne représente une chaîne de propagation.

gation qui conduit à la notion de trajectoire et sa représentation sur un corpus.

Rappels divers

Un **texte** est une succession de symboles respectant les règles (la grammaire) d'une ou plusieurs langues.

On appelle **document** un texte publié, c'est-à-dire accessible à la lecture. Il est en général constitué d'un corps (texte), d'un titre (texte), d'un ou plusieurs auteurs, d'une date et d'un lieu de publication.

On appelle **corpus** un ensemble d'au moins deux documents. Le corpus de documents est l'objet premier de notre étude. Dans la suite, nous supposons que nous avons à notre disposition un corpus de documents que l'on notera D. Dans nos applications, nous pouvons classer chaque document selon qu'il provienne d'un blog, d'un site d'actualité ou du site de microblogging Twitter. Il s'agit de ce qu'on appelle des médias sociaux.

Les **médias sociaux** sont des plateformes sur lesquelles il est possible de consulter et de publier des documents. Nous utilisons des corpus de documents Web issus de médias sociaux. Des exemples de médias sociaux sont les sites d'actualité, les blogs et. Dans la suite, nous travaillons exclusivement sur des corpus de documents provenant de médias sociaux.

Les documents issus de médias sociaux ont généralement un sens, c'est-à-dire qu'ils expriment la pensée de ceux qui les ont écrits. Cette expression se fait par l'entremêlement d'**informations**. L'ensemble des informations du corpus sera noté

K. L'ensemble des informations contenues dans un document d se note $I(d) = \{k_0, k_1, \ldots\}$.

Il existe une notion de **similarité sémantique** sur K. Des phrases comme "Il va pleuvoir." et "Il pourrait pleuvoir" ne véhiculent pas exactement les même informations, pourtant celles-ci restent proches en sens. On pose $sem_K : K \times K \to [0, 1]$ une fonction de similarité sur K. 0 signifie que les deux informations n'ont rien à voir, 1 signifie que les deux informations sont exactement les mêmes. Toute valeur entre les deux qualifie un dégrée de similitude entre les deux informations.

3.2.1 La Trajectoire de l'information

Un document d_i du corpus D contient de multiples informations. De part la nature des médias sociaux, ce document peut être consulté par quelqu'un qui, à son tour et sous l'influence de d_i, publiera un document d_j contenant certaines informations sémantiquement très proches de celles de d_i. Lorsque cet effet est réalisé, on parle de **propagation d'information** et la transformation de l'information originale en une autre est appelée **mutation de l'information**. Formellement, on note la mutation comme la paire d'information (k_0, k_1) avec $k_0 \in I(d_i)$ et $k_1 \in I(d_j)$. On note P_{d_i,d_j} l'ensemble des mutations entre d_i et d_j.

Une mutation vérifie une condition de similarité forte, pour un $\epsilon > 0$ dont la petitesse est à définir, on a :

$$(k_0, k_1) \in P_{d_i,d_j} \implies sem_K(k_0, k_1) \geq 1 - \epsilon$$

Comme P_{d_i,d_j} décrit exactement toutes les mutations entre d_i et d_j, il décrit toute la propagation d'information entre ces deux documents. On dit alors que P_{d_i,d_j} est un **événement de propagation**. Un événement de propagation peut se représenter comme un graphe, comme illustré dans la Figure 3.5.

Cependant, un ensemble d'événements de propagation fournit une description riche du processus de mutation, qui est plus précise que la simple information de

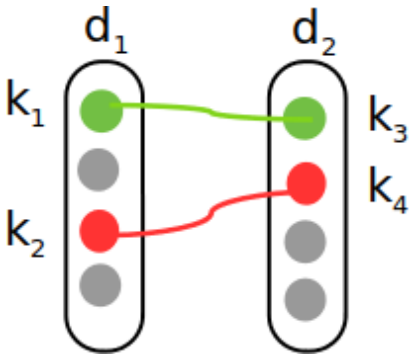


FIGURE 3.5 – Représentation de l'événement de propagation $P_{d_1,d_2} = \{(k_1, k_3), (k_2, k_4)\}$.

propagation d'un document à un autre. Chaque information étant unique, il devient possible de retisser le cheminement de l'information. La Figure 3.6 souligne la mise en évidence d'un chemin à partir de deux événements de propagations.

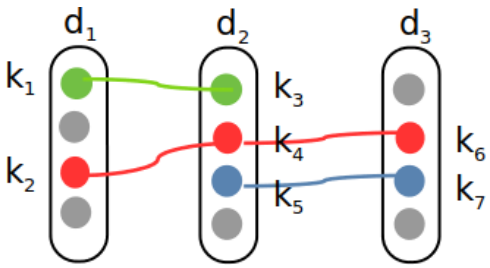
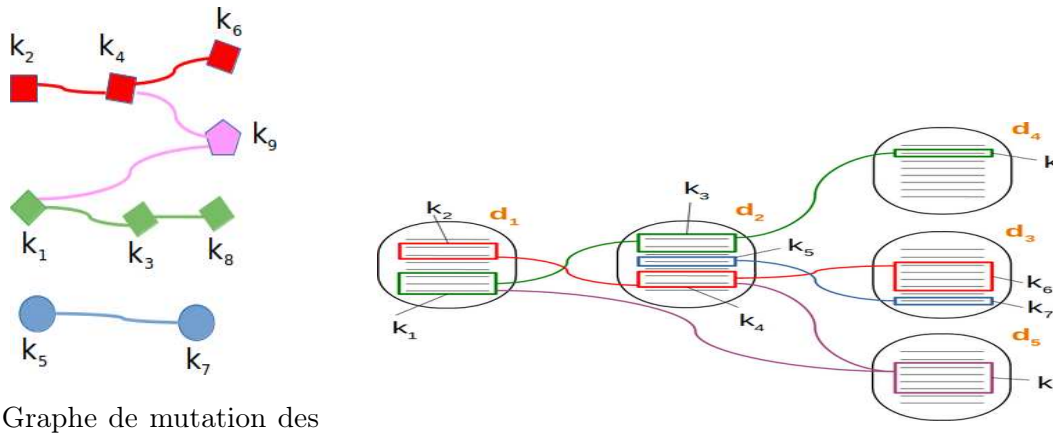


FIGURE 3.6 – Représentation de deux événements de propagation : $P_{d_1,d_2} = \{(k_1, k_3), (k_2, k_4)\}$ et $P_{d_2,d_3} = \{(k_4, k_6), (k_5, k_7)\}$. On observe la mise en évidence du chemin de propagation $(k_2, k_4)(k_4, k_6)$, plus simplement noté (k_2, k_4, k_6).

Ainsi, l'ensemble des événements de propagation synthétise deux structures : le graphe de mutation des informations d'une part, le cheminement de propagation de ces informations au sein des documents d'autre part. Ces deux représentations sont données en Figure 3.7b et 3.7a. On peut voir ainsi que l'information se propage le long de chemins. Ces chemins peuvent être transcrits sur les documents, c'est ce qu'on appelle une **chaîne de propagation**.

Définition 8. On appelle une suite d'au moins deux documents ordonnée chronologiquement (par date de publication) $d_0, d_1, \ldots, d_n$, une **chaîne** et on la note $c = d_0 d_1 \ldots d_n$. On note $Chains(D)$ l'ensemble des chaînes sur un corpus D.

Soit P un ensemble d'événements de propagation, et $c = d_0 d_1 \ldots d_n$ une chaîne.



(a) Graphe de mutation des informations.

(b) Propagation des informations dans les documents.

FIGURE 3.7 – Une représentation de la propagation des informations dans les documents. D'un côté la manière dont l'information mute (a), de l'autre côté, la structure que cela engendre à travers les documents (b).

On dit que c est une **chaîne de propagation** si, et seulement si :

$$\forall i \in \{1, 2, \ldots, n-1\}, \exists\ p_i, p_{i+1} \in P / \begin{cases} p_i = P_{d_{i-1}, d_i}, \\ p_{i+1} = P_{d_i, d_{i+1}}, \\ \exists (k, k') \in p_i \wedge \exists (k', k'') \in p_{i+1}. \end{cases}$$

Dans la suite de cet section nous ferons particulièrement attention à ne pas confondre les deux termes chaîne et chaîne de propagation. Une chaîne est une succession ordonnée dans le temps de document, tandis qu'une chaîne de propagation est une chaîne particulière. Nous introduisons le vocabulaire et les opérateurs sur les chaînes comme suit :

Définition 9. Soient $d, d' \in D$, on dit que $d < d'$ si et seulement si $date(d) < date(d')$.

Soient $c_{ai} = d_a \ldots d_i$ et $c_{xz} = d_x \ldots d_z$ deux chaînes du corpus. On dit que $c_{ai} < c_{xz}$ si et seulement si $d_i < d_x$.

Si $c_{ai} < c_{xz}$, on peut définir l'opération de concaténation $\cdot$ comme suit : $c_{ai} \cdot c_{xz} = d_a \ldots d_i d_x \ldots d_z$. $c_{ai} \cdot c_{xz}$ est aussi une chaîne.

On dit qu'une chaîne c_1 est une **sous-chaîne** d'une autre chaîne c_0, que l'on note $c_1 \subset c_0$, si une des assertions suivantes est vraie :

— $\exists c' < c_1 / c' \cdot c_1 = c_0$

— $\exists c' > c_1 / c_1 \cdot c' = c_0$

— $\exists c' < c_1 < c'' / c' \cdot c_1 \cdot c'' = c_0$

On dit aussi que c_0 est une **sur-chaîne** de c_1, que l'on note $c_0 \supset c_1$.

Proposition 1. Soit c une chaîne de propagation, alors toute sous-chaîne $c' \subset c$ est une chaîne de propagation.

La réciproque est généralement fausse. Toutes les sous-chaînes de c peuvent être des chaînes de propagation sans que c le soit nécessairement.

Étant donné un ensemble d'événements de propagation dont le support est un corpus de document D. On peut définir l'ensemble des chaînes de propagation de D. On appelle cet ensemble la **Trajectoire de la propagation** qu'on note $T(D)$. On appelle une **trajectoire** un ensemble de chaînes de documents. $T(D)$ est une trajectoire particulière, qui correspond à l'ensemble des événements de propagation.

Définition 10. Source et Destination

Soit T une trajectoire pour un corpus D. On peut définir les deux ensembles suivants :

— $Src = \{u \in D / \exists c \in T, c \texttt{ commence en } u\}$

— $Dst = \{v \in D / \exists c \in T, c \texttt{ finit en } v\}$

On dit que Src est l'ensemble source de T et D son ensemble destination.

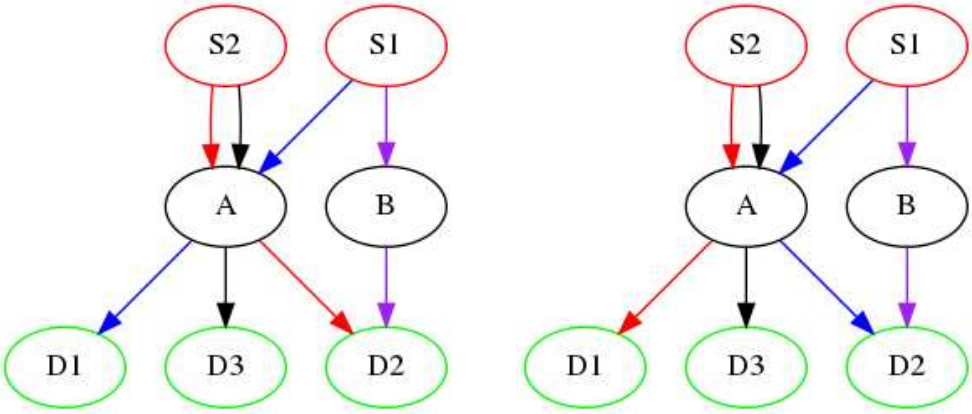


FIGURE 3.8 – Exemples de trajectoires constituée de quatre chemins (rouge, bleu, noir et violet). Les deux trajectoires sont différentes mais ont la même source et la même destination.

Définition 11. Support minimal

Soit G un graphe sur D et T une trajectoire sur D.

On dit que T est dans G si toute chaîne de T est constituée d'arcs de G. Dans ce cas, on dit que G est un *support* de T. On dit que G est le *support minimal* de T si G est inclus dans tout autre support de T.

Nous pouvons toujours construire le support minimal d'une trajectoire donnée. L'inverse est généralement faux (cf Figure 3.9), le support minimal d'une trajectoire ne suffit pas pour identifier cette trajectoire. Il existe un support simple à calculer pour toute trajectoire. Il s'agit du graphe temporel sur le corpus D. Dans ce graphe, il y a un arc entre un document d_i et un document d_j si et seulement si $date(d_i) < date(d_j)$. Ce graphe encode la relation temporelle des chaînes. On l'appelle le **support temporel** du corpus D.

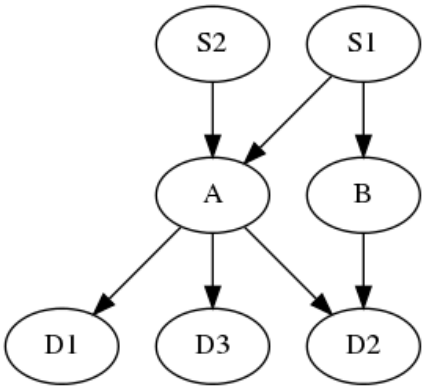


FIGURE 3.9 – Support minimal des trajectoires présentées en Figure 3.8. Les deux trajectoires ont le même support minimal bien qu'elles soient différentes. Le support minimal d'une trajectoire n'est pas une donnée suffisante pour déterminer cette trajectoire.

Définition 12. Trajectoire arborescente

Soit T une trajectoire, Src son ensemble source, Dst son ensemble destination et $Card$ l'opérateur de cardinalité usuel sur les ensembles.

On dit que T est une *trajectoire arborescente* si, et seulement si :

— $Card(Src) = 1$

— Si c_1 et $c_2 \in T$ passent par un même sommet A alors jusqu'à A c_1 et c_2 utilisent les mêmes arcs.

Remarque. Le support minimal d'une trajectoire arborescente est un arbre dans le sens où on l'entend en théorie des graphes. De plus, pour un arbre donné, il existe une et une seule trajectoire ayant cet arbre comme support minimal et c'est une trajectoire arborescente.

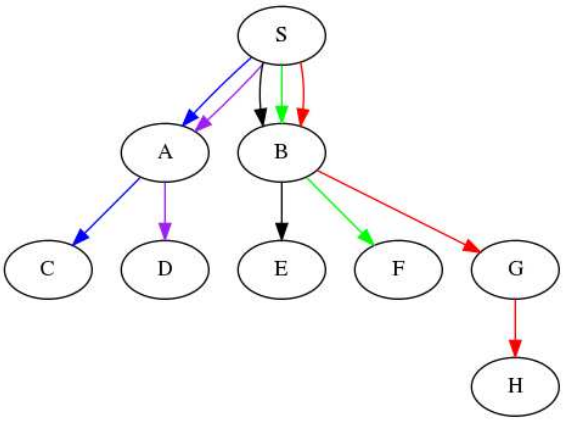
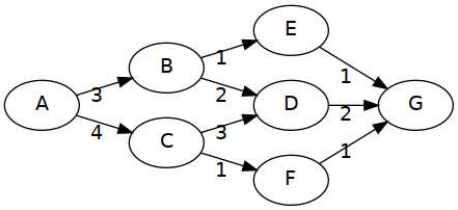
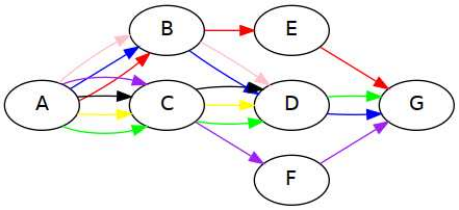


FIGURE 3.10 – Exemple de trajectoire arborescente.

Du point de vue de la trajectoire, l'information se comporte dans les documents comme des pièces lors de transactions. L'exemple en Figure 3.11 s'intéresse aux mouvements de pièces lors de transactions d'argent. Le graphe de transaction présente les sommes échangées, tandis que la trajectoire des pièces spécifie la manière dont elles se sont réparties au cours du temps.



(a) Mouvement de pièces entre transactions, sur les arcs sont notés le nombre de pièces échangées.



(b) Même mouvement de pièces, l'historique de chaque pièce du système est représenté par une chaîne.

FIGURE 3.11 – La notion de trajectoire apparaît lorsqu'on veut capturer le déplacement d'éléments précis dans des flots. Il pourrait aussi s'agir de voitures dans les rues d'une ville.

La Trajectoire de l'information est une structure riche permettant d'expliciter les chemins empruntés par les informations lors de leur propagation. Il reste à savoir comment la calculer et si ce calcul présente un intérêt. Dans le reste de ce chapitre,

nous proposons des méthodes pour approcher cette trajectoire.

3.3 Calcul de la Trajectoire de l'information

Pour parvenir à calculer les successions de documents le long desquelles l'information se propage, la Trajectoire de l'information, il convient de soulever un certain nombre de questions. Certaines sont techniques : par exemple, comment calculer cette Trajectoire de manière performante et peu complexe. D'autres sont plus fondamentales : comment extraire cette trajectoire dans un contexte où il manque des données ? Même sans cela, est-il possible de dévoiler le phénomène de propagation sans aucune connaissance préalable d'une structure reliant les documents ?

Trajectoire sur un corpus de document

Concernant les données manquantes, on peut distinguer deux sortes de problèmes. Soit le contenu de certains documents est partiel ou faux, soit il manque des documents entiers dans le corpus pour bien considérer la propagation. Le premier problème est plus pernicieux qu'il n'y paraît. S'il peut s'agir d'un problème de collecte (texte incomplet ou mal nettoyé, date manquante, auteur mal extrait, ...), il peut aussi s'agir d'une erreur humaine (mauvaise date, fautes d'orthographe, ...) ou d'une ambiguïté (un même nom d'auteur peut être porté par deux personnes distinctes). Obtenir un corpus correctement constitué reste la préoccupation principale dans les métiers de veille, aussi il existe un certain nombre de bonnes pratiques. Le second problème de données manquantes, les documents absents du corpus, constitue un vrai verrou pour étudier le phénomène de propagation. Celui-ci s'effectue sur un grand réseau de communication que nous ne savons pas capturer en entier : bouche-à-oreille, TV, radio, livres, journaux, Internet, ... Si on note T la Trajectoire de l'information relative à toute la communication humaine, il y a deux manières de définir T_D, sa restriction au corpus de documents D :

1. On considère uniquement les chaînes de T qui sont des chaînes valides sur D.

2. On modifie les chaînes de T pour leur enlever tous les documents extérieurs à D. Une chaîne $d_1\delta_1 d_2\delta_2\delta_3\delta_4 d_3$ de T où les $d_i \in D$ et les $\delta_j \notin D$ correspondra à une chaîne $d_1 d_2 d_3$ dans T_D. Dans cette situation, le lien d'une chaîne entre deux documents de D symbolise une sortie arbitrairement longue (voire de longueur nulle) du corpus. Dans cette situation, les chaînes de T_D ne sont plus nécessairement des chaînes réelles (de T), mais seulement des parties observables de chaînes réelles.

La différence entre ces deux définitions de T_D se situe au niveau du sens que l'on attribue aux maillons (deux documents successifs) des chaînes. Dans la première, le maillon d'une chaîne représente une propagation directe. Dans la seconde cette propagation est potentiellement indirecte. Il est possible de travailler dans le premier cas lorsque les données désignent explicitement des maillons fiables. Les références sur Wikipedia, les URLs sur Internet en général, les *quote tweets* sur Twitter, ou la bibliographie dans les articles universitaires sont des exemples de situation où l'auteur fournit de lui-même ses influences dans son document : on parle alors de citation explicite. De manière générale, ce genre de connections évidentes peut être relativement rare, aussi nous nous plaçons dans le second cas, celui où le maillon signifie une propagation potentiellement indirecte.

Données exploitables

Pour comprendre le phénomène de propagation, qui est *a priori* inconnu et dont les seules traces sont les documents du corpus, on peut distinguer plusieurs approches, selon les données dont on dispose sur les documents, qui peuvent être complémentaires :

1. Modéliser le phénomène sur la base de connaissances fiables sur la propagation entre deux documents. Il s'agit, lorsque c'est possible, d'obtenir des relations évidentes de propagation. Par exemple une URL ou une citation explicite sont des exemples fiables de propagation (cf. section 2.2.1).

2. Modéliser le phénomène de propagation. Il s'agit de proposer une repré-

sentation mathématique, *a priori*, imitant au mieux le comportement de la propagation, et donc de la forme de la trajectoire, puis d'en estimer les paramètres (cf. section 2.2.2).

3. Exploiter les données à la recherche d'indices de propagation. Il s'agit de chercher des tournures de phrases, des champs lexicaux, des similarités sémantiques claires qui augmentent le crédit accordé à l'existence d'une chaîne passant par tels ou tels maillons (cf. section 2.2.1).

Comme dit précédemment, il n'y a aucune garantie *a priori* d'avoir assez de connaissance sûre, la première approche générale est donc à écarter dans un premier temps. Pour la deuxième, il est nécessaire d'avoir une vérité terrain, a minima partielle, pour estimer les paramètres du modèle. Aussi, c'est la troisième méthode que nous approfondissons dans ce chapitre, dans le but d'identifier des chaînes de propagation et re-construire la Trajectoire de l'information.

L'exploitation d'indices nous amènera à la définition d'une notion de **cohérence** de chaîne, permettant de mesurer sa pertinence en tant que chaîne de propagation. Dans un premier temps, nous détaillons la construction de cette notion de cohérence. Puis, nous proposons une méthode pour construire les chaînes cohérentes du corpus.

3.3.1 Notion de chaîne cohérente de documents

On dit d'une chaîne de documents qu'elle est cohérente, lorsqu'elle semble être une chaîne de propagation pour un observateur. Cet observateur attribue un crédit élevé au fait que cette chaîne soit dans la Trajectoire. Une chaîne de propagation n'est pas nécessairement cohérente. La cohérence est un critère subjectif, un observateur peut ne pas repérer une filiation entre des documents par ailleurs très différents. Inversement, une chaîne cohérente n'est pas nécessairement une chaîne de propagation. Par exemple, deux documents peuvent s'inspirer d'une même source et un observateur pourrait juger, au vu de leurs similitudes, que le plus récent s'inspire du plus ancien. Nous faisons l'hypothèse qu'une grande partie des chaînes de propagation sont cohérentes, et que l'ensemble des chaînes de propagation cohérentes constitue un pan

intéressant de la propagation.

Hypothèse 4 (Cohérence des chaînes de propagation)**.** Une partie « conséquente » des chaînes de propagation sont des chaînes cohérentes.

Nous modélisons la cohérence d'une chaîne c par une mesure, $coh(c)$, dont la valeur se situe entre 0 (cohérence nulle) et 1 (cohérence sans équivoque). Pour bien définir une chaîne cohérente, on fixe un seuil de cohérence γ, et on dit que c est une chaîne cohérente si et seulement si :

$$coh(c) > \gamma. \tag{3.1}$$

On appelle le couple (coh, γ) le **critère de cohérence**. Il y a plusieurs manières de construire la fonction coh. Par exemple, elle peut être construite via l'annotation des chaînes par des experts humains. Elle peut être construite par la quantité de vocabulaire commun aux documents de la chaîne. Nous pouvons appréhender intuitivement le sens et l'objectif de la fonction coh, mais il reste à déterminer la manière de la calculer.

Nous considérons que la cohérence d'une chaîne se détermine à la lecture (ordonnée) des documents de la chaîne et du corpus.

$$\forall c = d_1 \dots d_k, coh(c, D) = f(d_1, \dots, d_k, D). \tag{3.2}$$

Pour alléger les notations, nous postulons l'existence du corpus D comme un contexte général disponible et notons simplement $coh(c)$ la cohérence de la chaîne c. De plus, nous émettons l'hypothèse suivante :

Hypothèse 5. La cohérence d'une chaîne $c = d_1 \dots d_k$, $coh(c, D)$, est fonction d'une similarité sémantique entre les documents :

$$coh(c) = f(\{sim_{sem}(d_i, d_j, D), i \neq j \in \{1, \dots, k\}\}). \tag{3.3}$$

L'idée derrière cette hypothèse est que, le long d'une chaîne de propagation, les documents partagent une information stable perceptible à travers la similarité sémantique entre documents. La manière dont on calcule la similarité sémantique sim_{sem}, et la manière dont on la combine le long de la chaîne, vont avoir un impact sur la forme des chaînes cohérentes. Nous proposons plusieurs manières de construire la similarité et la cohérence.

Construction de la fonction de cohérence

On souhaite estimer la cohérence d'une chaîne de documents. L'hypothèse 5 fournit une piste de réflexion pour construire une estimation de la cohérence. Pour ce faire, il est d'abord nécessaire d'avoir une fonction de similarité sémantique entre les documents, les similarités entre les documents constitutifs de la chaîne fournissent ainsi les paramètres de base permettant le calcul de la cohérence. Il reste à déterminer les paires de documents qu'il est intéressant de comparer pour estimer la cohérence, il n'est pas évident qu'ils soient tous nécessaires, où que certaines paires ne jouent pas un rôle plus important que les autres. Nous appelons **fonction de sélection**, une fonction qui, à partir d'une chaîne, détermine les paires de documents intéressantes pour estimer la cohérence de la chaîne. Enfin, il faut déterminer comment agréger cet ensemble de similarité entre différents documents pour fournir une valeur de cohérence. Nous appelons **fonction de combinaison** une fonction qui agrège les différentes similarités en une valeur de cohérence. La donnée des trois fonctions, la similarité sémantique, la sélection et la combinaison fournit une approche générale pour construire une estimation de la cohérence, que nous donnons en Algorithme 1. Une telle approche est d'ailleurs naturellement utilisée dans la proposition de Shahaf et Guestrin pour estimer la cohérence d'une chaîne [ShahafGuestrin, 2010]. Nous présentons maintenant quelques fonctions candidates pour estimer la cohérence.

Fonction de similarité sémantique

Nous avons déjà discuté certaines fonctions de **similarité sémantique** en partie 2.1.2. En utilisant une représentation vectorielle des documents, comme le TF-IDF ou

Doc2Vec, on peut calculer une similarité cosinus. Celle-ci produit une valeur absolue comprise entre 0, lorsque les représentations sont orthogonales, et 1 lorsqu'elles sont colinéaires. Un autre exemple, la similarité proposée par Shahaf et Guestrin, définie la similarité entre deux documents vis-à-vis de chaque descripteur du corpus. En toute généralité, la similarité sémantique peut être multicritère, c'est-à-dire que certains aspects des documents peuvent se ressembler tandis que d'autres non. La signature de la fonction de similarité générale est un vecteur réel, de taille n :

$$sim_{sem} \ : \ D \times D \to \mathbb{R}^n.$$

En notant $w_1, \ldots, w_{|F|}$ les différents descripteur du corpus de document D, la similarité de Shahaf et Guestrin s'écrit ainsi :

$$sim_{Shahaf}(d_i, d_j) = [sim_{Shahaf}(d_i, d_j|w_1), \ldots, sim_{Shahaf}(d_i, d_j|w_{|F|})] \tag{3.4}$$

Où la similarité selon un descripteur $sim_{Shahaf}(d_i, d_j|w_k)$ est calculée selon l'influence que possède ce descripteur dans la connectivité entre d_i et d_j dans un graphe biparti entre les documents et les descripteurs. Pour plus de détails nous vous renvoyons à l'article original [ShahafGuestrin, 2010].

Fonction de sélection

La fonction de sélection, notée *select*, choisit, à partir d'une chaîne, la séquence de paires de documents à considérer pour calculer la cohérence. On en considère principalement deux :

— $select_{all}$ choisit l'intégralité des paires de documents apparaissant dans la chaîne. Tous les documents de la chaîne doivent se ressembler deux à deux.

$$select_{all}(d_1 d_2 \ldots d_k) = \{(d_i, d_j), 1 \leq i \neq j \leq k\}. \tag{3.5}$$

— $select_{succ}$ choisit uniquement les paires consécutives de documents de la chaîne. Seuls les maillons assurent la cohérence de la chaîne. C'est la fonction de

Algorithme 1 : Procédure générale de construction de $coh(c)$.

Entrées : sim, une fonction de similarité.
$select$, une fonction de sélection.
$combine$, une fonction de combinaison.
c une chaîne

Sortie : $coh(c)$

```
Paires ← select(c)
Sims ← ∅
for (d_i, d_j) ∈ Paires do
    Sims ← Sims ∪ {sim(d_i, d_j)}
end
Coh = combine(Sims)
Retourner Coh
```

sélection choisie par Shahaf et Guestrin.

$$select_{succ}(d_1 d_2 d_3 \dots d_{k-1} d_k) = \{(d_i, d_{i+1}), 1 \leq i < k\}. \tag{3.6}$$

Fonction de combinaison

Enfin, **la fonction de combinaison** prend en entrée les paires de documents choisies par *select* et leur similarité et les combinent pour fournir un score de cohérence. La procédure générale pour calculer cette fonction de cohérence est donnée dans l'algorithme 1.

Voici quelques exemples de fonctions de combinaison usuelles :

— La moyenne géométrique, définie ainsi :

$$combine_{geo}([v_1, \dots, v_k]) = (\prod_{v_i \neq 0} v_i)^{\frac{1}{k}}.$$

où le produit de plusieurs vecteurs est défini pour chaque composante i [1] comme :

$$[u \cdot v \cdot \ldots \cdot w]_i = [u]_i \times [v]_i \times \ldots \times [w]_i.$$

1. Il s'agit du produit d'Hadamard, appliqué ici à des vecteurs.

— La moyenne standard, ou arithmétique, définie ainsi :

$$combine_{avg}([v_1, \ldots, v_k]) = \frac{\sum v_i}{k}.$$

— Le minimum, défini pour chaque composante i comme :

$$[combine_{min}([u, v \ldots, w])]_i = min([u]_i, [v]_i, \ldots, [w]_i).$$

Lorsque Sim est réduit à un ensemble de scalaires, il s'agit des notions habituelles de moyenne géométrique, arithmétique et du minimum. Dans le cas où la similarité est multicritère, comme c'est le cas pour la similarité de Shahaf et Guestrin, les vecteurs sont de dimension supérieure à 1, ces fonctions de combinaison fournissent la notion habituelle de moyenne pour chaque critère. Il reste à combiner les différents critères pour obtenir une unique valeur de cohérence, pour cela on peut utiliser une autre fonction de combinaison. D'ailleurs, la fonction de combinaison de Shahaf et Guestrin est posée comme un problème d'optimisation sur l'ensemble des descripteurs qui consiste en la succession d'un maximum et d'un minimum sur les similarités. La description complète de leur approche sort du cadre de cette section. Néanmoins, nous comparons dans les expérimentations leur méthode avec des estimations de la cohérence basées sur les fonctions décrites ici.

Cette famille de fonctions de cohérence, construite *a priori*, fournit une base de travail. Dans un premier temps, elle permet de construire des premières trajectoires à l'aide de l'approche présentée dans la prochaine section. Dans un second temps, elle fournit des approches de base auxquelles se comparer pour construire ou apprendre des mesures de cohérences plus pertinentes.

3.3.2 Algorithme de construction : Idée générale

Nous proposons maintenant de construire l'ensemble des chaînes cohérentes. En partant d'une idée simple : tester toutes les chaînes possibles. Nous construisons, étape par étape, une méthode efficace pour construire toutes les chaînes cohérentes. Par la

Algorithme 2 : Procédure générale de calcul de $T_{coh}(D)$.

Entrées : D, le corpus de documents.
coh et γ, qui forment le critère de cohérence.
Sortie : $T_{coh}(D)$

1 $C \leftarrow Calculer\text{-}Candidats(D)$
2 $T \leftarrow Filtrer\text{-}Candidats(C, coh, \gamma)$
3 **Retourner** T

suite, nous constaterons que les chaînes qui finissent en un certain document sont essentiellement des prolongements de chaînes qui finissent en des documents publiés antérieurement. En poursuivant ce raisonnement, on remarquera que les chaînes cohérentes de taille 2 ont un rôle primordial pour construire les chaînes cohérentes potentielles. Enfin, nous proposons une heuristique pour gérer les situations où le nombre de chaînes cohérentes est trop important.

Idée générale

Nous séparons le problème de construction en deux étapes. Dans un premier temps, il s'agit de construire un ensemble de chaînes potentiellement cohérentes, que nous appelons l'ensemble **Candidats**. Ensuite, il s'agit de tester les chaînes candidates et de ne conserver que les chaînes qui satisfont le critère de cohérence. La procédure générale est donnée dans l'Algorithme 2.

Il est crucial, d'autant plus dans le cadre d'une veille documentaire, de construire notre ensemble dans un temps raisonnable. Le temps d'exécution de cette approche se décompose en une somme de deux termes paramétrés : t_{calcul}, le temps de calcul, et $t_{filtrage}$, le temps de filtrage.

$$Temps(\text{Algorithme 2}, D, coh, \gamma) = t_{calcul}(D) + t_{filtrage}(C, coh, \gamma). \tag{3.7}$$

La procédure de filtrage est donnée dans l'Algorithme 3. Le temps de calcul du filtrage se calcule suivant la formule :

$$t_{filtrage}(C, coh, \gamma) = \sum_{c \in C} (Temps(coh(c)) + O(1)). \tag{3.8}$$

Algorithme 3 : Procédure de filtrage des candidats *Filtrer-Candidats*(C, coh, γ).

Entrées : C, l'ensemble candidat
coh et γ, qui forment le critère de cohérence
Sortie : *Filtrer-Candidats*(C, coh, γ)

```
Resultats ← ∅
for c ∈ C do
    if coh(c) > γ then
        Resultats ← Resultats ∪ {c}
    end
end
Retourner Resultats
```

Avec, représentées par $O(1)$, les différentes comparaisons et affectations durant le filtrage. Le temps de filtrage est ainsi lié au le cardinal de l'ensemble candidat :

$$t_{filtrage}(C, coh, \gamma) = O(|C|) + \sum_{c \in C} Temps(coh(c)). \tag{3.9}$$

La procédure de filtrage est relativement directe et incompressible. Le problème de cette approche est donc de soumettre le moins de candidats possible tout en s'assurant de construire le plus de chaînes cohérentes.

Calcul naïf de l'ensemble candidat

La manière la plus simple de construire l'ensemble candidat est de lister toutes les chaînes possibles sur le corpus. Dans le cas où tous les documents ont une date distincte, une chaîne est analogue à un sous-ensemble de D contenant au moins 2 éléments, on peut définir l'ensemble candidat de la sorte, avec $\mathcal{P}(D)$ l'ensemble des sous-ensembles de D :

$$Candidats(D) = Chains(D) = \mathcal{P}(D) \setminus (\{\emptyset\} \cup \{\{d\}, d \in D\}). \tag{3.10}$$

On connaît le nombre de sous-ensembles d'un ensemble, $|\mathcal{P}(D)| = 2^{|D|}$. La taille maximale de l'ensemble candidat est donc :

$$|Candidats(D)| = 2^{|D|} - |D| - 1. \tag{3.11}$$

La taille de cet ensemble candidat le proscrit pour tout corpus d'une taille conséquente. Dès 20 documents publiés à des dates différentes, il faudrait soumettre plus d'un millions de chaînes au filtrage. Nous proposons maintenant un succession de deux améliorations de cette méthode pour réduire la taille de l'ensemble candidat et pour le calculer efficacement même pour des corpus de plusieurs milliers de documents.

3.3.3 Première amélioration : Hypothèse de l'hérédité de la cohérence

Toutes les chaînes de propagation ne sont pas indépendantes les unes des autres. En particulier, être une chaîne de propagation implique que toutes ses sous-chaînes soient également des chaînes de propagation (cf. proposition 1). On aimerait, assez naturellement, que la même propriété s'étende aux chaînes cohérentes :

Hypothèse 6. Soit c une chaîne **cohérente**, alors toute sous-chaîne $c' \subset c$ est une chaîne cohérente.

En fait, l'hypothèse 6 pose un gros problème en allant à l'encontre de notre motivation première : la chaîne produit un contexte plus riche que ses parties. Une chaîne peut donc être plus cohérente que ses sous-chaînes. En particulier, une chaîne c peut être cohérente au sens de notre critère, $coh(c) > \gamma$, tandis qu'une de ses sous-chaînes c' n'est pas cohérente.

Pour atténuer ce problème, nous proposons de tolérer, dans un premier temps, des chaînes de cohérence moindre. On introduit un second seuil, $0 < \gamma^* \ll \gamma$, et on dit qu'une chaîne c est **faiblement cohérente** si $coh(c) > \gamma^*$. On admet, à la place de l'hypothèse 6, l'hypothèse suivante :

Hypothèse 7. Soit c une chaîne faiblement cohérente, alors toute sous-chaîne $c' \subset c$ est une chaîne faiblement cohérente.

L'idée est que les sous-chaînes d'une chaîne cohérente ne sont pas totalement incohérentes, elles peuvent simplement être d'une cohérence plus faible. On va pouvoir tirer parti de cette hypothèse pour calculer les chaînes cohérentes, en calculant au

Algorithme 4 : Procédure de calcul de l'ensemble des chaînes cohérentes $T_{(coh,\gamma)}(D)$.

Entrées : D, le corpus de documents.
coh et γ, qui forment le critère de cohérence.
$T_{(coh,\gamma^*)}$, une procédure qui calcule les chaînes faiblement cohérentes.
Sortie : $T_{(coh,\gamma)}(D)$

$C \leftarrow T_{(coh,\gamma^*)}(D)$ `// Utilise l'Algorithme 2`
$T \leftarrow Filtrer\text{-}Candidats(C, coh, \gamma)$
Retourner T

préalable les chaînes faiblement cohérentes. L'approche est décrite dans l'Algorithme 4, qui remplace et utilise l'Algorithme 2. L'idée est d'utiliser un algorithme qui tire parti de l'hypothèse 7 pour construire rapidement l'ensemble des chaînes faiblement cohérentes. Puis de simplement filtrer cet ensemble. Cependant, la même critique de l'hypothèse 6 peut être appliquée à l'hypothèse 7. Nous acceptons que certaines chaînes faiblement cohérentes nous échappent, par construction. Nous étudions les différents paramètres de notre méthode de calcul, dont l'impact du choix de γ et de γ^* dans l'annexe C.

Nous disposons maintenant de la procédure générale du calcul des chaînes cohérentes donnée par l'algorithme 4. Celui-ci utilise l'algorithme 2 dont on a vu à la section précédente qu'il était, pour le moment, inutilisable pour des corpus de taille moyenne. Il s'agit maintenant de trouver un remplaçant à l'algorithme 2. Comme celui-ci est utilisé pour le calcul des chaînes faiblement cohérentes, nous décidons , par simplicité, dans la suite de cette section, de confondre les notions de cohérence et cohérence faible, γ^* se notera simplement γ et on confondra également l'hypothèse 6 et l'hypothèse 7. Cependant, il faut garder à l'esprit qu'il y aura un post-traitement sur la trajectoire qu'on calcule afin de trouver les chaînes cohérentes pour le seuil γ réel.

Calcul incrémental : un document après l'autre

L'hypothèse 6 permet d'utiliser une approche du problème étape par étape, en construisant la trajectoire document après document. L'idée est que le processus

de calcul de trajectoire est essentiellement incrémental. Pour un corpus auquel on adjoint un nouveau document, plus récent que tous les documents précédents, la trajectoire cohérente sur ce corpus est une mise à jour de la trajectoire cohérente sur ce corpus sans ce nouveau document. Plus formellement, on note $t_1 < t_2 < \ldots < t_N$ l'ensemble des dates de publication des documents d'un corpus (certains documents peuvent être publiés à la même date). On introduit la notion de corpus antérieur à une date :

Définition 13 (Corpus antérieur)**.** Soit D un corpus de documents et t une date, on appelle **corpus antérieur** à t, noté $D_{<t}$, l'ensemble des documents de D publiés avant t :

$$D_{<t} = \{d \in D / date(d) < t\}$$

Comme pour une récurrence, on suppose que $T_{coh}(D_{<t_n})$ est d'ores et déjà calculée. Il s'agit maintenant de définir les chaînes candidates pour $T_{coh}(D_{<t_{n+1}})$. Une première option est de définir l'ensemble $Candidats(d)$ comme suit :

$$Candidats(d) = \{c \cdot d / c \in T_{coh}(D_{<date(d)})\} \cup \{d'd / d' \in D_{<date(d)}\} \tag{3.12}$$

Il s'agit de l'ensemble des chaînes finissant en un document publié avant d, continuées en d, auxquelles on ajoute les nouvelles chaînes de tailles 2 possibles. Une représentation pour 4 documents est disponible en Figure 3.12. Il suffit ensuite de filtrer les candidats qui vérifient le critère de cohérence :

$$T_{coh}(D_{<date(d)} \cup \{d\}) = T_{coh}(D_{date(d)}) \cup \{c \in Candidats(d) / coh(c) > \gamma\} \tag{3.13}$$

Ainsi, on peut résoudre le problème en utilisant une approche ascendante. En partant du document le plus ancien, et en allant au plus récent, on peut calculer toutes les chaînes finissant en chaque document. La procédure générale est donnée dans l'Algorithme 5, qui remplace l'Algorithme 2.

Il y a une boucle supplémentaire sur les temps de publication. Son but est de

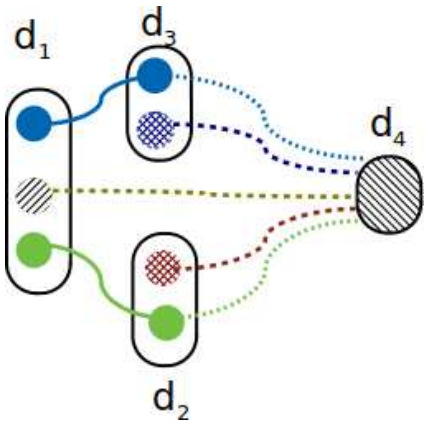


FIGURE 3.12 – Liste des chaînes candidates pour d_4, représentées en pointillées. Les chaînes pleines (d_1d_3, d_1d_2) sont déjà connues. La chaîne $d_1d_2d_3d_4$ n'est pas considérée car elle n'a pas été calculée précédemment.

Algorithme 5 : Procédure **incrémentale** de calcul de $T_{coh}(D)$.

Entrées : D, le corpus de document
$\{t_1, \ldots, t_n\}$, les dates de publications des documents
coh et γ, qui forment le critère de cohérence
Sortie : $T_{coh}(D)$, ensemble des chaînes cohérentes sur D

```
1  T = ∅
2  for t ∈ {t_1, ..., t_n}, par date de publication croissante do
3  |  U = ∅
4  |  for d ∈ D/date(d) = t do
5  |  |  C ← Calculer-Candidats(d, T)
6  |  |  F ← Filtrer-Candidats(C, coh, γ)
7  |  |  U ← U ∪ F
8  |  end
9  |  T ← T ∪ U
10 end
11 Retourner T
```

prendre en considération les documents qui ont la même date de publication, mais au final chaque document est traité une seule fois. Le calcul s'effectue, comme auparavant, en deux temps : le calcul des candidats, puis le filtrage des chaînes cohérentes. On retrouve la structure générale de l'Algorithme 2, mais cette fois le calcul et le filtrage se font au niveau de chaque document et non plus au niveau du corpus.

Dans ce cas de figure, l'ensemble candidat pour le document d correspond à toutes les chaînes cohérentes existantes avant d. Il s'agit de la trajectoire cohérente

sur le corpus tronqué avant d. En notant $D_{<d}$ l'ensemble $D_{<date(d)}$ on a :

$$|Candidats(d)| = |T_{coh}(D_{<d})| + |D_{<d}|. \tag{3.14}$$

Le second terme correspond aux nouvelles chaînes de taille 2. Le nombre total de chaînes candidates soumises au filtrage est donc :

$$qte(Candidats(D)) = \sum_{d \in D} |Candidats(d)|. \tag{3.15}$$

$$qte(Candidats(D)) = (\sum_{d \in D} |T_{coh}(D_{<d})|) + (\sum_{d \in D} |D_{<d}|). \tag{3.16}$$

Dans le pire cas, tous les documents ont des dates de publication différentes et sont parfaitement ordonnés, ce qui donne :

$$qte(Candidats(D)) = (\sum_{d \in D} |T_{coh}(D_{<d})|) + \frac{(|D|-1)(|D|-2)}{2}. \tag{3.17}$$

Un intérêt de cette approche est qu'elle est facilement mise à jour. Lors d'une collecte de documents nouveaux, la mise à jour de l'ensemble de chaînes cohérentes est immédiat.

On peut se demander dans quelles proportions évolue l'équation 3.17 vis-à-vis d'une augmentation de la taille du corpus D. Si on ajoute un document d_{nouv} à D plus récent que tout autre document de D, on obtient :

$$qte(Candidats(D \cup \{d_{nouv}\})) - qte(Candidats(D)) = |T_{coh}(D)| + |D|. \tag{3.18}$$

Ajouter un document d_{nouv} implique de vérifier la cohérence de toutes les chaînes existantes poursuivies en d_{nouv} et de toutes les chaînes de taille 2 de la forme $d_i d_{nouv}$, pour $d_i < d_{nouv}$. À cette étape, on fait du travail en double. Pour une chaîne $c = d_1 \dots d_k$ on calcule à la fois :

— la cohérence de $c' = d_k d_{nouv}$,

— la cohérence de $c'' = c \cdot d_{nouv} = d_1 \dots d_{k-1} d_k d_{nouv} = d_1 \dots d_{k-1} \cdot c'$.

Étant donné que c' est une sous-chaîne de c'', si c' est incohérente, on peut éviter de calculer inutilement la cohérence de c''. Lors de l'ajout d'un document, il est intéressant de d'abord vérifier les chaînes cohérentes de taille 2 de la forme $d_k d_{nouv}$. Si $d_k d_{nouv}$ n'est pas cohérente, toutes les chaînes de la forme $d_i \dots d_k d_{nouv}$ ne le sont pas. L'ensemble des chaînes cohérentes de taille 2 a donc un rôle particulier, permettant d'améliorer notre approche.

3.3.4 Seconde amélioration : exploitation des chaînes de taille 2

Dans l'approche précédente, on propose de calculer, pour chaque document d, les chaînes potentiellement cohérentes finissant en d. Il s'agit de l'ensemble candidat donné à l'équation 3.12. On vient de voir que lorsqu'on ajoute un nouveau document d_{nouv}, il n'est pas pertinent de considérer toutes les chaînes précédemment calculées pour déterminer les chaînes cohérentes finissant en d_{nouv}. Il est possible d'en élaguer certaines en se basant sur les documents qui forment une chaîne de taille 2 cohérente avec d_{nouv}. Nous appelons support de cohérence le graphe reliant les documents tels que deux documents liés forment une chaîne de taille 2 cohérente :

Définition 14 (Support de cohérence)**.** On dit que $G = (D, E)$ est le **support de cohérence** du critère de cohérence (coh, γ) si et seulement si :

$$E = \{(d, d') \in D \times D / coh(dd') > \gamma\}.$$

Le calcul du support de cohérence est direct, la procédure est donnée dans l'Algorithme 6. Un exemple de support est donné en Figure 3.13.

À l'aide du support de cohérence, il devient possible de traiter le calcul des chaînes cohérentes plus efficacement, nous proposons une approche par programmation dynamique.

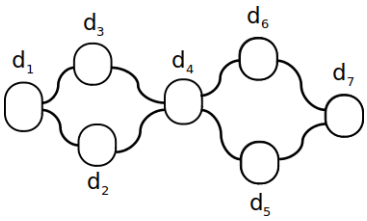


FIGURE 3.13 – Exemple de support de cohérence pour un corpus de 7 documents. La présence d'une arête (d_i, d_j) marque la cohérence de la chaîne $d_i d_j$.

Algorithme 6 : Procédure de calcul du support de cohérence $G(D, coh, \gamma)$.

Entrées : D, le corpus de document
coh et γ, qui forment le critère de cohérence
Sortie : $G(D, coh, \gamma)$, le support de cohérence.

$E = \emptyset$
for $d \in D$, *par date de publication croissante* **do**
  for $d' \in D_{<d}$ **do**
    if $coh(d'd) > \gamma$ **then**
      $E \leftarrow E \cup \{(d', d)\}$
    end
  end
end
Retourner $G = (D, E)$

Approche par programmation dynamique

La programmation dynamique permet la résolution de problèmes à partir de la résolution d'instances plus petites du même problème (on parle de sous-problèmes). Le calcul de l'ensemble des chaînes cohérentes finissant en un certain document d est un problème de ce type.

Définition 15 (Chaînes finissant en un document)**.** Pour un document d, on appelle $FinitEn(d)$ l'ensemble des chaînes cohérentes **finissant** en d. Une chaîne finit en d si et seulement si elle est la concaténation (notée avec l'opérateur $\cdot$) d'une chaîne $d'd$ avec, au plus, une chaîne de $FinitEn(d')$.

$$c \in FinitEn(d) \iff \begin{cases} \exists d' \in D, c = d'd \\ \exists d' \in D, c' \in FinitEn(d') / c = c' \cdot d \end{cases} . \tag{3.19}$$

Le premier cas $c = d'd$ est déjà connu à l'aide du support de cohérence. Il

Algorithme 7 : Procédure de calcul de $FinitEn(d, coh, \gamma, Pred(d), Finis(d))$.

Entrées : d, le document.
coh et γ, qui forment le critère de cohérence.
$Pred(d)$, les documents précédant d dans le support de cohérence.
$Finis(d) = \{FinitEn(d')/d' \in Pred(d)\}$, les ensembles $FinitEn(d')$ précédents.
Sortie : $FinitEn(d, coh, \gamma, Pred(d), Finis(d))$

$P \leftarrow \{d'd, d' \in Pred(d)\}$
$C \leftarrow Calculer\text{-}Candidats(d, Finis)$
$F \leftarrow Filtrer\text{-}Candidats(C, coh, \gamma)$
Retourner $F \cup P$

reste à déterminer les chaînes de $FinitEn(d)$ du second cas. Pour ce faire, comme précédemment, on construit un ensemble candidat :

$$Candidats(d) = \bigcup_{d' \in Pred(d)} \{c \cdot d, c \in FinitEn(d')\}. \tag{3.20}$$

$Pred(d)$ correspond aux documents qui précédent d dans le support de cohérence $G = (D, E)$:

$$Pred(d) = \{d' \in D/(d', d) \in E\} \tag{3.21}$$

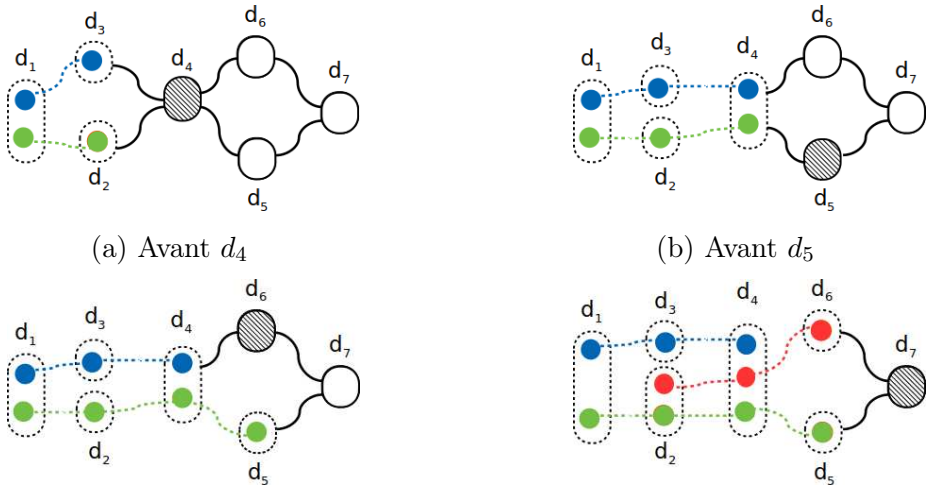
Le support de cohérence correspond au graphe de dépendance de la structure $FinitEn(d)$. C'est-à-dire que pour calculer l'ensemble $Candidats(d)$ il faut avoir résolu les sous-problèmes de calcul des $FinitEn(d')$ pour $d' \in Pred(d)$. La procédure de calcul des ensembles $FinitEn(d)$ est donnée en Algorithme 7. La procédure complète par support de cohérence est donnée en Algorithme 8, elle remplace la procédure donnée en Algorithme 5. Quelques étapes successives du calcul sont illustrées en Figure 3.14.

Par rapport à l'approche du calcul incrémental présentée dans la sous-section précédente (l'Algorithme 5), il y a une quantité moindre de chaînes à filtrer pour l'approche exploitant le support de cohérence. Pour les différencier, nous notons $(qte(Candidats(D))$ la quantité de chaînes filtrées lors du calcul incrémental (donnée dans l'Équation 3.17), et nous notons $qte^*(Candidats(D))$ la même quantité pour

Algorithme 8 : Procédure par **support de cohérence** de calcul de $T_{coh}(D, coh, \gamma)$.

Entrées : D, le corpus de documents.
coh et γ, qui forment le critère de cohérence.
Sortie : $T_{coh}(D)$, ensemble des chaînes cohérentes sur D

$T \leftarrow \emptyset$
`/* Calcul du support de cohérence */`
$G \leftarrow G(D, coh, \gamma)$
`/* Calcul des` $FinitEn$ `*/`
$Finis \leftarrow newHashTable()$
for $d \in D$, *par date de publication croissante* **do**
  $F \leftarrow \emptyset$
  $P \leftarrow Pred(d, G)$ `/* prédécesseurs de` d `dans` G `*/`
  for $d' \in P$ **do**
    $F \leftarrow F \cup Finis(d')$
  end
  $Finis(d) \leftarrow FinitEn(d, coh, \gamma, P, F)$
  $T \leftarrow T \cup Finis(d)$
end
Retourner T



(a) Avant d_4 (b) Avant d_5

(c) Avant d_6 (d) Avant d_7

FIGURE 3.14 – État de l'algorithme avant de traiter d_4 jusqu'au traitement de d_6. Les documents déjà traités sont en pointillés. Les arêtes pleines représentent le support de cohérence. Les arêtes en pointillés représentent des chaînes (ainsi que leurs sous-chaînes). Au traitement de d_6, la sous-chaîne d_2d_4 est prolongée créant une nouvelle chaîne (en rouge).

l'approche exploitant le support de cohérence. Cette dernière vaut ainsi :

$$\begin{aligned} qte^*(Candidats(D)) &= (\sum_{d \in D} |Candidats(d)|) + O(|D|^2), \\ qte^*(Candidats(D)) &= (\sum_{d \in D} \sum_{(d',d) \in E} |FinitEn(d')|) + O(|D|^2). \end{aligned} \tag{3.22}$$

Le second terme correspond aux filtrages via le support de cohérence, il est donné par le second terme dans l'équation 3.17. On peut comparer la quantité de chaînes filtrées entre l'approche incrémentale et l'approche tenant compte du support de cohérence :

$$\begin{aligned}
&qte(Candidats(D)) - qte^*(Candidats(D)) \\
&= \sum_{d\in D} |T_{coh}(D_{<d})|) - (\sum_{d\in D} \sum_{(d',d)\in E} |FinitEn(d')|), \\
&= \sum_{d\in D} (|T_{coh}(D_{<d})| - \sum_{(d',d)\in E} |FinitEn(d')|), \\
&= \sum_{d\in D} (\sum_{d'\in D_{<d}} |FinitEn(d')) - \sum_{(d',d)\in E} |FinitEn(d')|, \\
&= \sum_{d\in D} \sum_{\substack{d'\in D_{<d} \\ d'\notin Pred(d)}} |FinitEn(d')|.
\end{aligned} \tag{3.23}$$

L'approche qui calcule d'abord le support de cohérence effectue au pire autant de filtrages que son homologue. Pour peu qu'au moins une des chaînes de taille 2 ne soit pas cohérente, cette approche est plus efficace, tout en restant simple à mettre à jour lors d'ajouts de nouveaux documents.

3.3.5 Heuristique pour le contrôle de la quantité de chaînes

Nous avons abouti à une approche simple et qui nous semble efficace pour le calcul des chaînes cohérentes. Cependant, dans le pire des cas, si beaucoup de chaînes sont cohérentes, il peut devenir nécessaire d'avoir à gérer une quantité exponentielle de chaînes (par rapport au nombre de documents, voir l'équation 3.11 pour la quantité exacte). Il y a donc, *a priori*, des cas où l'algorithme nécessitera un temps et un espace mémoire élevés. Pour pallier cette incertitude, il peut être intéressant d'ajouter une contrainte heuristique à l'approche, qui garantit une quantité de chaînes raisonnable.

Nous retrouvons ici les considérations des travaux d'extraction d'histoires présentés dans l'état de l'art en section 2.2.5. Il s'agit de calculer un ensemble de séquences cohérentes de documents ou d'événements de taille "raisonnable", qui optimise

certaines contraintes de diversité et de connectivité. Ici, on cherche à conserver un maximum de telles séquences, mais il est parfois nécessaire, pour des raisons pratiques, de filtrer les chaînes et nous aimerions a minima conserver celles qui offrent une certaine forme de diversité.

Pour commencer nous revenons sur l'approche développée dans la section précédente. Nous avons soulignés que commencer par calculer les chaînes cohérentes de tailles 2, ce que nous appelons le support de cohérence, fournissait une parcours des chaînes candidates plus efficace. Par le même argument, nous aurions pu calculer conséquemment les chaînes cohérentes de taille 3 à partir des chaînes de taille 2, et généraliser au calcul des chaînes de taille n à partir des chaînes de taille $n-1$. Nous ne procédons pas ainsi pour deux raisons, la première est que cette approche ne permet pas facilement d'ajouter un nouveau document et de mettre à jour la trajectoire. La seconde est une raison heuristique. Il est possible d'avoir une quantité exponentielle de chaînes cohérentes, ce que nous voulons contraindre.

La contrainte naturelle pour une récurrence sur la taille des chaînes est de contraindre la taille maximale des chaînes. Par exemple, cela consiste à ne pas construire des chaînes plus grandes qu'une certaine taille maximale constante l_{limit}. On sait qu'il y a au plus $\binom{|D|}{l}$ chaînes de taille l, la quantité maximale de chaînes est ainsi bornée par $O(|D|^{l_{limit}})$.

Cette heuristique pose plusieurs problèmes. D'abord il n'y a pas de raisons de préférer les chaînes courtes aux longues. Au contraire, nous supposons que les chaînes longues peuvent être plus cohérentes que leurs parties. Ensuite, d'un point de vue pratique, cette heuristique peut éliminer des chaînes même lorsqu'elles sont déjà peu nombreuses. Nous souhaitons ne brider que les corpus qui contiennent un nombre de chaînes cohérentes trop important et obtenir le résultat exact dans les autres situations.

Dans les expérimentations présentées en Annexe C.5, nous identifions que l'explosion combinatoire du nombre de chaînes cohérentes intervient dans des jeux réels à cause de certains documents, capables de se connecter et de s'intégrer dans de

nombreuses chaînes. De la même manière qu'un mot qui apparaît dans de nombreux documents du corpus est peu discriminant [2], un document présent dans la plupart des chaînes a peu d'impact sur la diversité des chaînes obtenues. Aussi, nous avons choisi de limiter le nombre de chaînes qui peuvent finir en un certain document d par une quantité fixe, notée q_{limit}. Cette heuristique s'intègre directement à l'approche que nous avons construit jusqu'ici. Pour une valeur adéquate, l'heuristique atténue le rôle des documents connecteurs sans nuire à la diversité des chaînes calculées. De cette manière la quantité totale de chaînes est bornée par $q_{limit} \times |D|$.

La complexité temporelle de cette heuristique peut être calculée. En notant $G = (D, E)$ le support de cohérence, les étapes principales du calcul de $FinitEn(d)$ sont :

1. Accumulation des candidats. Il y a au plus $n = q_{limit} \times deg_{in}(d)$ chaînes, avec $deg_{in}(d)$ le degré entrant de d dans G.

2. Filtrage des chaînes cohérentes. L'étape coûte au plus $n \times f(l_{max}^2)$. Le coût du calcul de cohérence est borné par $O(f(l_{max}^2))$, avec $l_{max} - 1$ la taille maximale des chaînes cohérentes. Cette borne vient de l'hypothèse 5 selon laquelle la cohérence est fonction des similarités sémantiques entre les documents de la chaîne.

3. Choix (au plus) des k meilleures chaînes cohérentes obtenues après filtrage. Cette étape peut-être réalisée en $O(n)$ à l'aide d'un algorithme de sélection.

Au total, l'approche heuristique nécessite un temps de calcul qui, dans le pire

2. c'est le concept de l'IDF présenté en section 2.1.2.

cas, a la forme suivante :

$$\begin{aligned} Temps(T_{coh}(D)) &= O(|D|) + \sum_{d \in D} Temps(FinitEn(d)) \\ &= O(|D| + \sum_{d \in D} deg_{in}(d) \times k \times (1 + f(l^2_{max}) + 1)) \\ &= O(|D| + (\sum_{d \in D} deg_{in}(d)) \times k \times (2 + f(l^2_{max}))) \\ &= O(|D| + |E| \times k \times (2 + f(l^2_{max}))) \\ Temps(T_{coh}(D)) &= O(|D| + |E| \times k \times f(l^2_{max})) \end{aligned} \tag{3.24}$$

Cette approche heuristique a une complexité similaire à celle d'un parcours de graphe. La taille maximale de chaîne obtenue est rarement élevée dans nos expériences. Mais au besoin, et selon la manière dont on calcule la cohérence, il peut être nécessaire de la limiter.

Dans cette section nous avons introduit la notion de chaîne cohérente comme un angle d'attaque pour approcher la Trajectoire de propagation de l'information. Nous avons développé une approche de calcul de l'ensemble des chaînes cohérentes. Lors de ce développement, nous avons souligné les différentes difficultés inhérentes au calcul de chaîne. En particulier, le nombre de chaînes cohérentes peut être exponentiel, et il convient de prendre certaines dispositions assurant l'arrêt de notre approche en un temps raisonnable. Nous avons passés certaines considérations techniques autour de notre méthode calcul de la trajectoire. Par exemple notre approche se prête presque immédiatement à une implémentation asynchrone qui permet d'améliorer le temps de calcul sur les architectures parallèles, nous en donnons les détails en Annexe A. De la même manière nous n'avons pas parlé des structures de données permettant de représenter et de stocker la trajectoire et les chaînes de documents efficacement. Il existe des manières plus efficace en espace mémoire que l'approche naïve qui représente l'ensemble des chaînes comme une liste de liste de documents. Nous les développons dans l'annexe B. La section qui suit présente les expérimentations que nous avons réalisées pour évaluer la qualité des chaînes obtenues par notre approche.

3.4 Expérimentations

La section précédente introduit et explore la question de la construction d'un ensemble de chaînes cohérentes, noté $T_{coh}(D)$. Nous sommes parvenus à une méthode pour construire cet ensemble d'une manière efficace en temps et en mémoire. Pour ce faire nous nous sommes appuyés sur plusieurs hypothèses, comme l'hérédité de la cohérence : si une chaîne c est cohérente, alors ses sous-chaînes sont *faiblement* cohérentes. Rien ne garantit cette hypothèse, aussi l'ensemble que nous calculons peut ne contenir qu'une fraction des chaînes cohérentes. Aussi, il reste à valider que les chaînes que nous estimons cohérentes correspondent bien à l'intuition humaine de la cohérence. Cela soulève les deux questions suivantes :

1. L'ensemble de chaînes cohérentes calculé par notre approche est-il exhaustif ?
2. Les chaînes calculées sont elles conformes à l'idée que nous nous faisons de la cohérence ?

Plus simplement, nous cherchons à savoir si la construction de $T_{coh}(D)$ est efficace dans son domaine d'application, c'est-à-dire comme approximation de la Trajectoire de l'information dans un corpus tiré des médias sociaux. Une manière supervisée d'appréhender cette efficacité serait de comparer notre approche à l'aide d'une vérité terrain, un corpus de documents pour lequel les chaînes de propagation ou, a minima, les chaînes cohérentes, sont connues exhaustivement à l'aide d'annotations humaines. Or, d'une part, nous n'avons pas une telle vérité terrain à notre disposition, d'autre part, la combinatoire de l'ensemble des chaînes rend extrêmement coûteux sa constitution pour un corpus de documents non trivial. Ainsi, il est compliqué de quantifier l'exhaustivité de l'approche (par exemple, son rappel). Nous menons en Annexe C un début d'analyse quantitative sur la question de l'exhaustivité de l'approche. Il reste cependant possible d'étudier l'exactitude de notre approche : les chaînes fournies par l'approche ont-elles la cohérence qu'on leur réclame ? Pour répondre à cette question, nous avons effectué une campagne d'évaluation des chaînes par des annotateurs humains. Les résultats de cette campagne montrent que notre

approximation de la cohérence, sur les chaînes calculées, paraît corrélée effectivement une partie du jugement humain.

3.4.1 Campagne d'évaluation

Pour pouvoir évaluer la cohérence de l'intégralité de l'approche, on demande à un ensemble de personnes qualifiées pour juger de la cohérence des chaînes. Une personne est qualifiée si elle est capable de comprendre les documents du corpus qu'elle doit étudier. Pour que cette approche soit réalisable, il faut que la quantité par participant de chaînes à analyser soit raisonnable. Nous expliquons d'abord comment nous avons constitué nos jeux de données, avant de présenter le protocole d'évaluation.

Jeux de données

Nous utilisons deux jeux de données pour notre campagne :

1. Le **Citation Network Dataset V1 AMINER** [3], construit par [Tang, 2008]. Il s'agit d'un corpus anglophone constitué principalement de résumés d'articles scientifiques extraits d'ACM ou de DBLP.
2. L'intégralité des articles de presse postés sur la version américaine du Huffington Post, entre le 1er juillet et le 30 novembre 2016. Il s'agit également d'un corpus anglophone.

Chacun de ces jeux de données contient plusieurs dizaines de milliers de documents, ce qui engendre un nombre conséquent de chaînes à évaluer. Il est nécessaire de limiter le nombre de chaînes et l'adapter au nombre d'évaluateurs. La campagne a bénéficié de 4 participants issus du milieu scientifique et parlant couramment l'anglais. Ils étaient en mesure de lire et comprendre les différents articles issus des deux jeux de données qui leurs ont été présentés. Dans la mesure du possible, on veut évaluer l'intégralité des chaînes produites par notre approche. Pour cela, on réduit le nombre de documents des corpus. Pour ce faire, nous avons tiré uniformément 150

3. Le jeu de données entier peut être obtenu à l'adresse : `https://aminer.org/citation`

documents pour chaque corpus. Nous appelons dans la suite nos jeux de données échantillonnés **AMINER** et **HuffPost**.

Pour construire la similarité entre les documents, nous avons considéré l'approche par TFIDF. L'ensemble des termes est enrichi par les n-grammes de taille 2 à 4. La comparaison se fait par un produit cosinus.

Nous avons considéré plusieurs critères de cohérence : la cohérence de similarité minimale coh_{min} et la cohérence par moyenne arithmétique coh_{avg} :

$$coh_{min}(c) = \min_{1 \leq i < j \leq n} sim(d_i, d_j).$$

$$coh_{avg}(c) = \frac{2}{n(n-1)} \sum_{1 \leq i \leq j \leq n} sim(d_i, d_j)$$

Pour le seuil de cohérence, nous avons pris $\gamma = 0,1$, $\gamma = 0,2$ et $\gamma = 0,5$. Au total, cela constitue 6 critères de cohérence différents. Pour chaque critère, et chaque corpus, nous avons calculé un ensemble de chaînes. Certaines chaînes apparaissent pour plusieurs critères, il y a au total 81 chaînes pour AMINER et 149 pour HuffPost.

Plateforme d'évaluation

Pour évaluer les chaînes, nous avons développé une plateforme Web, permettant aux experts de facilement traiter les chaînes. Nous avons réparti les chaînes entre les experts de telle sorte que :

— Chaque chaîne est évaluée par au moins deux experts.

— Deux experts ne sont pas exposés aux mêmes successions de chaînes à évaluer.

Le processus d'évaluation pour une chaîne est le suivant :

1. Lecture du contexte de la chaîne : Un court paragraphe expliquant le jeu de données à partir duquel la chaîne a été extraite.
2. Lecture du premier document de la chaîne. Il s'agit de celui qui a la date de publication la plus ancienne. Les deux premières étapes sont représentées en Figure 3.15.
3. Pour chaque document suivant, du plus ancien au plus récent (Figure 3.16) :

(a) Lecture du nouveau document, qui s'affiche à la suite des documents précédents.

(b) Question : Y a-t-il un lien sémantique entre le document que vous venez de lire et le document que vous avez lu juste avant ? Trois réponses possibles : **lien sémantique fort**, **lien sémantique faible**, **aucun lien**.

(c) Question : Est-il plausible qu'une ou plusieurs informations se soient propagées le long des documents présentés ? Trois réponses possibles : **Oui fortement**, **Oui faiblement**, **Non**.

4. Question optionnelle (Figure 3.17) : Pouvez-vous nommer les informations qui se sont propagées le long de la chaîne (si vous estimez qu'il y en a).

FIGURE 3.15 – Plateforme d'évaluation : Il faut lire le contexte du corpus et le premier document avant de passer à l'étape suivante. Par soucis de taille dans cette Figure et les suivantes, les documents n'ont pas de contenu.

3.4.2 Résultats

La campagne d'évaluation a permis d'obtenir un ensemble de chaînes annotées par différentes personnes. Nous commençons par comparer les réponses des différents évaluateurs sur les chaînes identiques, puis nous nous commentons la répartition des chaînes jugées cohérentes. Enfin, nous nous servons de cette nouvelle vérité terrain pour comparer différentes méthodes d'approximation de la cohérence.

Accords inter-évaluateurs

Décider si deux documents sont liés sémantiquement, où s'il est plausible qu'une ou plusieurs informations se soient propagées le long d'une chaîne, est, *a priori* en partie, subjectif. Deux personnes soumises aux mêmes documents n'aboutiront pas

FIGURE 3.16 – Plateforme d'évaluation : À chaque document à partir du second, il est demandé à l'évaluateur de qualifier le lien avec le document précédent et la chaîne dans son ensemble. Les évaluations et les documents précédents restent disponibles.

FIGURE 3.17 – Plateforme d'évaluation : Lorsque tous les documents de la chaîne sont évalués, l'évaluateur peut, s'il y a lieu, formuler l'information qui lui semble se propager le long de la chaîne.

nécessairement à la même conclusion, selon ce qu'ils savent des sujets abordés dans les documents, et les informations qu'ils jugent importantes ou non. La première question est donc de savoir si les évaluateurs sont d'accord sur la manière dont ils répondent aux questions. Pour ce faire, nous calculons un ratio d'accord inter-évaluateur. La manière dont les chaînes ont été distribuées est telle que chaque chaîne a été soumise à deux évaluateurs différents, et chaque évaluateur a co-évalué autant de chaînes avec tous les autres évaluateurs. Le ratio d'accord inter-évaluateur est le nombre d'évaluations sur lesquels les deux évaluateurs ont répondu la même chose, divisé par le nombre total de chaînes évaluées.

Question 1 : Lien sémantique avec le document précédent

Nous évaluons les résultats de la question du lien sémantique de la chaîne selon deux modalités :

— La **présence** du lien. Le lien est jugé présent s'il est annoté fort ou faible. Sinon, il est jugé absent.

— L'**intensité** du lien. Dans ce cas, on différencie les liens annotés forts de ceux annotés faibles.

Les résultats sont donnés dans la Table 3.2.

TABLE 3.2 – Accord Inter-évaluateur pour la question du lien sémantique avec le document précédent.

Modalité	AMINER	HuffPost
#annotation	81	149
présence	76.6%	79.6%
intensité	68.1%	77.3%

Les experts sont d'accord sur la présence de lien, plus de 75% du temps, et ce sur chaque jeu de données. Cette statistique baisse peu lorsqu'on passe à l'intensité du lien, tout en gardant un accord supérieur à 66%.

Question 2 : Plausibilité de la chaîne

Nous évaluons les résultats de la question de la plausibilité de la chaîne selon deux modalités :

— La **présence** de la plausibilité. La chaîne jugée plausible si elle est annotée forte ou faible. Sinon on juge d'une absente de plausibilité.

— L'**intensité** de la plausibilité. Dans ce cas on différencie les chaînes annotées fortement plausibles de celles annotées faiblement plausibles.

TABLE 3.3 – Accord Inter-évaluateur pour la question de la plausibilité des chaînes.

Modalité	AMINER	HuffPost
#annotation	66	107
présence	80.7%	85.9%
intensité	57.9%	83.7%

De la même manière que pour la question du lien sémantique, on calcule un accord inter-évaluateurs pour la plausibilité. Cependant, on retire les chaînes de taille 2 de cette évaluation. Ces chaînes de taille 2 sont déjà évaluées lors de la question du lien sémantique. Les résultats sont donnés en Table 3.3. On remarque que la présence de plausibilité dépasse 80%, c'est-à-dire que les résultats sont supérieurs à ceux obtenus sur la question du lien sémantique. Cela renforce l'intuition selon laquelle il est plus simple d'aboutir à un consensus lorsqu'on a plus de contexte. Enfin, cela montre que la détection d'une chaîne cohérente par des experts humains est réalisable avec consistance, que ce n'est pas une tâche purement subjective et qu'il existe des indices dans le texte permettant d'effectuer cette décision.

Répartition des évaluations

Pour chaque modalité, présence de lien sémantique ou plausibilité de la chaîne, nous répartissons les évaluations en cinq catégories. Chacune des catégories correspond à la majorité des annotations pour une chaîne. La **Catégorie 1** est le cas où la majorité signale une **intensité forte**. La **Catégorie 2** est celui où la majorité signale une **intensité faible**. La **Catégorie 3** est le cas où les évaluateurs sont d'accord pour dire qu'il y a une **présence**, mais ils sont en désaccord sur l'intensité

TABLE 3.4 – Catégories d'annotation

Catégorie	La majorité des évaluateurs...
1	... ont jugé de la présence avec une intensité forte.
2	... ont jugé de la présence avec une intensité faible.
3	... ont jugé d'une présence mais sont en désaccord sur l'intensité.
4	... ont jugé de l'absence.
5	Les évaluateurs sont en désaccord.

TABLE 3.5 – Distribution des annotations (en %)

Modalité	Jeu de données	Cat. 1	Cat. 2	Cat. 3	Cat. 4	Cat. 5
Lien	AMINER	40,7	23,5	4,9	17,3	13,6
sémantique	HuffPost	18,8	10,7	1,3	63,8	5,4
Plausibilité	AMINER	34,8	19,7	19,7	9,1	16,7
de la chaîne	HuffPost	7,5	4,7	1,9	74,7	11,2

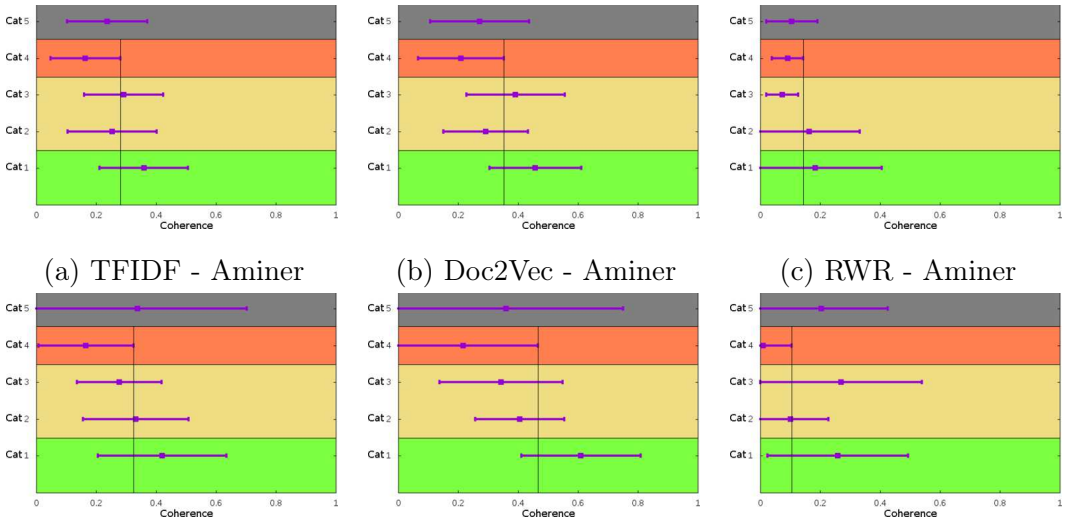
(certains l'ont jugé forte, d'autres faibles). La **Catégorie 4** est le cas où la majorité signale **l'absence**. Enfin, il arrive qu'aucun consensus ne soit trouvé, lorsqu'une moitié des évaluateurs juge la présence et l'autre l'absence. Il s'agit de la **Catégorie 5**. Le détail des catégories est rappelé en Table 3.4. La distribution des annotations par catégorie est donnée en Table 3.5.

Les résultats sur AMINER semblent satisfaisants, avec presque 70% de liens sémantiques présents et 75% de chaînes plausibles (Catégories 1,2 et 3). De l'autre côté, les résultats du HuffPost semblent plus décevants : 64% des liens sont jugés inexistants et 75% des chaînes sont jugées non plausibles. Nous verrons dans la section suivante que ces mauvais résultats sur HuffPost viennent du choix du seuil de cohérence.

Vérité terrain pour la cohérence

Ces annotations forment une vérité terrain qui nous permet de comparer différentes fonctions de cohérence. Une bonne fonction de cohérence devrait séparer clairement toutes les catégories, à l'exception de la catégorie 5, celle-ci étant, par définition, la catégorie des chaînes sur lesquelles les évaluateurs n'ont pas eu de consensus. Nous évaluons des cohérences à partir de trois mesures de similarité :

1. Une similarité cosinus basée sur le $TF\text{-}IDF$.



(a) TFIDF - Aminer (b) Doc2Vec - Aminer (c) RWR - Aminer

(d) TFIDF - HuffPost (e) Doc2Vec - HuffPost (f) RWR - HuffPost

FIGURE 3.18 – Amplitude de cohérence pour différentes fonctions. Une ligne verticale marque l'endroit ou la catégorie des chaînes non plausibles (Cat 4) intersecte les autres.

2. Une similarité cosinus basée sur Doc2Vec. Les vecteurs de Doc2Vec sont appris sur le jeu de données, et sont de dimension 20.

3. Une approche de la similarité par marche aléatoire, proposée dans [Shahaf-Guestrin, 2010], paramétrée avec une probabilité de retour de 99%.

À chaque fois, la cohérence est construite avec les mêmes paramètres :

— Sélection de toutes les paires de documents de la chaîne, fonction $select_{all}$.

— Combinaison et réduction des vecteurs par moyenne arithmétique, fonction $combine_{avg}$.

On considère qu'une fonction de calcul de la cohérence est discriminante s'il y a une intersection faible ou nulle de l'amplitude de cohérence entre les chaînes fortement plausibles et non-plausibles. L'amplitude de cohérence est définie comme la cohérence moyenne de la catégorie plus ou moins deux écarts types. Les amplitudes de cohérences pour les trois fonctions de cohérence sont présentées en Figure 3.18.

Pour chaque jeu de données, la cohérence basée sur Doc2Vec est la plus discriminante des trois. De manière générale, les chaînes fortement et faiblement plausibles ont une cohérence plus élevée que les chaînes non plausibles. Pour le jeu HuffPost, ces résultats expliquent la proportion de mauvaises chaînes observées. Une mauvaise

chaîne typique dans ce jeu de données a une cohérence basée sur le TFIDF inférieure à 0.2. Cela signifie que de telles chaînes proviennent pour la plupart de la trajectoire calculée avec un seuil de cohérence de 0.1 et sont donc filtrables à l'aide d'un seuil plus élevé. Les résultats de cette section nous conforte dans l'hypothèse selon laquelle il est possible de capturer le jugement humain sur la plausibilité des chaînes à l'aide de similarités classiques.

3.5 Conclusion

À la fin du chapitre précédent, nous motivions l'intérêt de prendre en considération le phénomène de mutation dans l'étude de la propagation d'informations. Nous constations que cela impliquait d'étudier la propagation d'une manière relativement différente. Dans ce chapitre, nous proposons un formalisme du phénomène de propagation bâti sur le phénomène de mutation. De ce formalisme, nous dégageons une structure propre du phénomène de propagation : **la Trajectoire de l'information**. Il s'agit de l'ensemble des chemins de documents le long desquels de l'information s'est propagée. La connaissance certaine et entière de cette structure est un mythe. Il semble cependant possible, en exploitant les indices de propagation disséminés dans le corpus, de construire une structure similaire, une trajectoire cohérente, composée des chemins pour lesquels on a pu évaluer une certaine cohérence.

Nous proposons et étudions dans un second temps une méthode pour construire, à partir d'un corpus, une trajectoire cohérente. Il s'agit d'une solution itérative, nécessitant une fonction de cohérence et construisant les chaînes cohérentes à partir des sous-chaînes plus petites. Cette méthode repose sur l'idée que nous sommes capable d'estimer la cohérence d'une chaîne à partir de la similarité entre ses documents constitutifs. À partir d'une campagne d'évaluation des trajectoires calculées selon notre méthode, nous constatons que les annotateurs sont d'accord sur ce qui est ou n'est pas une chaîne cohérente. Nous constatons également que ce que les annotateurs jugent être une chaîne cohérente se corrèle effectivement avec les fonctions de cohérence que nous avons utilisées, malgré leur simplicité, et ce pour plusieurs notions de

similarités différentes. Ces constats sont encourageants. D'abord, ils indiquent qu'il est possible d'estimer (au moins grossièrement) la cohérence à partir des similarités. Plus généralement, ils signifient que nous sommes capables de construire un ensemble de chaînes globalement cohérent.

La question suivante est de savoir si l'ensemble que nous construisons est exhaustif. Répondre à cette question pour un corpus réel, même de petite taille, demande un travail conséquent à cause de la nature combinatoire de l'ensemble des chaînes possibles. Nous proposons une ébauche de réponses à l'aide de chaînes construites dans un cadre synthétique dans l'Annexe C. Nous y montrons que, pour un modèle du bruit gaussien, notre approche capture une quantité non négligeable de l'ensemble des chaînes qui satisfont notre critère de cohérence. Cependant, dans ce modèle, toutes les chaînes sont fortuites. Une chaîne est fortuite si elle satisfait notre critère de cohérence alors qu'elle ne le devrait pas. Dans le modèle gaussien, les critères de cohérence que nous proposons construisent naturellement des chaînes fortuites. Cependant, il est possible, en sélectionnant les bons paramètres, de mitiger la quantité de chaînes fortuites. De plus, les chaînes fortuites calculées ne vérifient pas une hypothèse de notre approche : elles sont, presque toutes, composées de sous-chaînes qui ne sont même pas faiblement cohérentes. À l'inverse, une partie conséquente des chaînes calculées à partir de nos jeux de données réels sont constituées de sous-chaînes au moins faiblement cohérentes.

Ce chapitre constitue ainsi un premier pas satisfaisant pour la construction d'une approximation de la Trajectoire de l'information. Nous disposons maintenant d'une méthode capable de construire des chaînes cohérentes, à partir desquelles nous allons pouvoir, dans les chapitres suivants, étudier la propagation d'information. Nous avons cependant plusieurs idées d'améliorations que nous n'avons pas menées à leur terme. Nous présentons maintenant plusieurs pistes de réflexion sur ce sujet.

Tout d'abord, il y a la question de l'estimation de la cohérence d'une chaîne. Nous avons choisi une construction simple et *a priori* de la cohérence comme une moyenne de similarités. Sa qualité principale est sa vitesse, en adéquation avec la manière dont

se déroule la méthode. Son pendant négatif est sa sensibilité aux chaînes fortuites, et donc le fait que, dans l'absolu, ce soit un mauvais indicateur de cohérence. Le caractère itératif de la méthode mitige ce problème en imposant la faible cohérence à plusieurs sous-chaînes, sans l'imposer à toutes les sous-chaînes. D'une manière générale, nous préférons construire les chaînes cohérentes, même mélangées à des chaînes fortuites, rapidement, quitte à filtrer à l'aide de critères plus fins dans un second temps. Construire, ou apprendre, un critère de cohérence à la fois rapide et efficace reste une amélioration ouverte à proposition.

De la même manière, il serait intéressant de construire l'ensemble des chaînes cohérentes d'une manière radicalement différente. Nous avons déjà soulignés par exemple que les chaînes cohérentes au sein d'un cluster de documents similaires peuvent être nombreuses. Pour s'en prémunir, nous avons considéré une heuristique de limitation du nombre de chaînes par documents. Une autre manière d'y pallier serait de construire un ensemble de chaînes directement maximales, où d'identifier *a priori* les clusters, et d'y calculer des ensembles d'arbres couvrants cohérents pour en sélectionner les optimaux.

Nous considérons maintenant que nous sommes capables de construire une trajectoire cohérente. Bien que nous soyons conscients de la marge de progression importante sur ce seul point, nous souhaitons démontrer dans les chapitres qui suivent qu'il est déjà possible de fournir des analyses pertinentes sur le phénomène de propagation de l'information. La trajectoire est un objet purement structurel, il s'agit d'un ensemble de chaînes. Dans le chapitre suivant, nous abordons ce que nous appelons sa caractérisation. Il s'agit, à partir de la trajectoire, d'extraire pour chaque chaîne la, ou les informations susceptibles de s'y propager. À l'issue du calcul de la trajectoire et de sa caractérisation nous aurions ainsi extrait une représentation cohérente d'une partie du phénomène de propagation.

Chapitre 4

Caractérisation de la trajectoire

Résumé. Dans le chapitre précédent nous avons développé une approche originale permettant de calculer des trajectoires de l'information. Nous obtenons ainsi une approximation de l'historique de propagation de l'information sous forme d'un ensemble de chaînes chronologiques de documents mais sans connaître préalablement l'information qui y circule. Pour permettre l'exploitation de cet objet, il serait utile, voire nécessaire d'identifier l'information propagée dans chacune des chaînes. Dans ce chapitre, nous décrivons notre approche pour identifier et caractériser l'information qui se propage le long des chaînes. Après une réflexion sur les chaînes sémantiquement redondantes et la manière de décrire les chaînes, nous proposons une méthode pour distinguer et représenter les principales informations le long d'une chaîne. Nous proposons également des méthodes d'étiquetage de ces informations. Enfin, nous menons en œuvre une campagne d'évaluation pour estimer la pertinence de notre approche.

4.1 Introduction

La Trajectoire de l'information est l'ensemble des chaînes de documents le long desquelles de l'information s'est propagée. Lors du chapitre précédent nous avons proposé une méthode pour approcher cette Trajectoire, sans avoir à identifier précisément les différentes informations qui se propagent dans le corpus. Nous postulons que

l'information change de contexte, mute, en se propageant d'un document à l'autre, et qu'il n'existe ainsi jamais exactement deux fois la même information dans deux documents distincts. Sous ce postulat, nous sommes arrivés à la conclusion qu'il était difficile d'identifier *a priori* les informations qui se propagent au sein du corpus, mais qu'il semblait possible d'identifier ce qui se propage le long d'une chaîne de propagation (cf. section 2.3, et plus spécifiquement la sous-section *Ce qui se propage*). Nous proposons dans ce chapitre une approche permettant d'identifier et d'étiqueter des unités d'information se propageant le long de chacune des chaînes contenues dans une trajectoire. Le but est de fournir une représentation globale et intuitive de la propagation en vue d'une analyse humaine. Des représentations synthétiques du phénomène de propagation sous forme d'histoires ont été proposées dans la littérature à base de timelines [Wang, 2015a ; Yan, 2011], story forest [Liu, 2017] ou encore de metromap [Shahaf, 2012], un état de l'art détaillé est fourni en section 2.2.5. Notre ambition est de fournir une vision exhaustive et détaillée de la propagation de l'information et de sa mutation.

Le chapitre est structuré ainsi : la section 4.2 présente notre approche pour arriver à une description des différentes informations présentes sur les chaînes d'une trajectoire. Cette approche se décompose en plusieurs étapes. D'abord, nous constatons qu'il existe des sous-chaînes sémantiquement redondantes et qu'il serait utile de les filtrer. Nous proposons ensuite une extension de la notion de $TF\text{-}IDF$ aux chaînes de documents d'une trajectoire pour obtenir une représentation vectorielle pondérée des chaînes. Armé de cette représentation, nous proposons une méthode pour identifier les différentes informations spécifiques à chaque chaîne. Enfin, nous élaborons différentes stratégies permettant d'étiqueter ces informations en vue d'une analyse humaine. Dans la section 4.3, nous présentons les résultats de nos expérimentations sur des corpus de données issus du Web ainsi qu'un protocole d'évaluation mis en œuvre pour juger de la qualité et de la pertinence des résultats obtenus.

d

4.2 Identification de l'information le long d'une chaîne dans une trajectoire

Cette section décrit la méthodologie utilisée pour identifier, caractériser et nommer l'information se propageant le long de chacune des chaînes constituant la trajectoire. Nous commençons par exposer le problème des sous-chaînes sémantiquement redondantes et proposons une manière de les filtrer. Ensuite, nous proposons d'extraire les différentes informations des chaînes de la trajectoire en trois étapes :

1. **Description** des chaînes de la trajectoire à l'aide d'une représentation vectorielle pondérée.
2. **Identification** des différentes unités d'information présentes sur chaque chaîne.
3. **Étiquetage** des unités d'information en vue d'une analyse humaine.

Ces trois étapes sont schématisées en Figure 4.1.

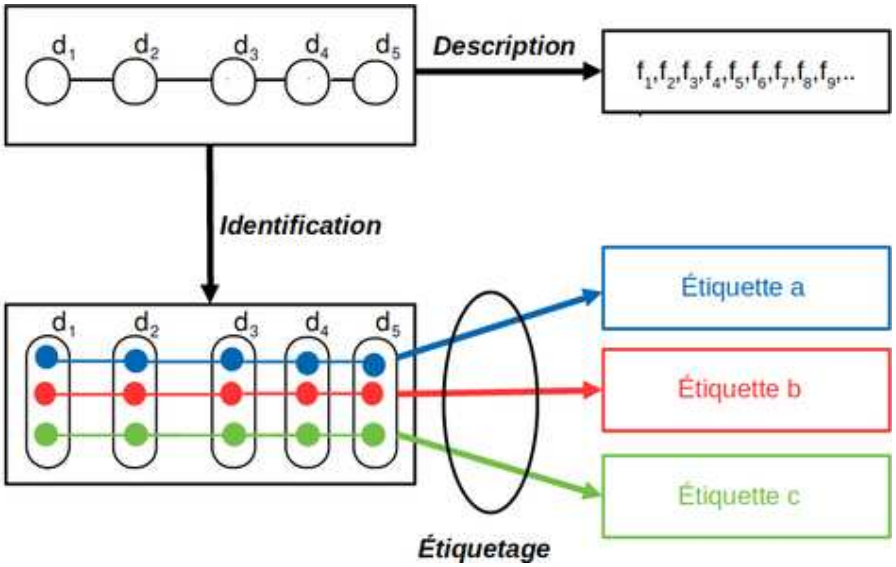


FIGURE 4.1 – Illustration des différentes étapes servant à extraire les informations d'une chaîne. La *description* fournit une représentation vectorielle de la chaîne, tandis que l'*étiquetage* fournit une représentation lisible pour un utilisateur final des informations. L'*identification* discerne les différentes informations de la chaîne.

4.2.1 Filtrage des sous-chaînes redondantes

Le but de cette section est d'extraire les différentes informations qui se propagent le long des chaînes d'une Trajectoire de l'information, notée T_D. Par définition, si c est une chaîne de propagation (c'est-à-dire que $c \in T_D$), alors toute sous-chaîne c_s de c est également une chaîne de propagation : si une information i s'est propagée le long de c, elle s'est en particulier propagée le long de c_s. Autrement dit, en notant $I(c)$ l'ensemble des informations qui se propagent sur la chaîne c, on a :

$$i \in I(c) \implies i \in I(c_s). \tag{4.1}$$

Cela signifie que l'identification de l'information propagée sur c entraîne de fait une connaissance sur certaines informations se propageant sur c_s. Ainsi, identifier l'information se propageant sur c_s lorsqu'on connaît déjà l'information propagée sur c est inutile. Dans l'absolu, on souhaiterait travailler uniquement sur les chaînes présentant un intérêt informationnel. Les sous-chaînes sont ainsi une forme de bruit non souhaitable pour le travail d'identification, en particulier elles peuvent brouiller la fréquence d'un document ou d'une information au sein de la trajectoire du fait de la répétition des fragments d'une chaîne.

Une première approche consisterait à simplement supprimer toutes les chaînes de T_D qui sont sous-chaînes d'une autre, on ne conserverait ainsi que les chaînes maximales. Cependant, rien n'empêche une sous-chaîne de contenir une information qui lui est spécifique comme illustrée en Figure 4.2. Il est donc nécessaire de mettre en place une procédure plus complexe pour déceler les sous-chaînes susceptibles de contenir un contenu informatif propre. Pour ce faire, nous plongeons les chaînes dans un espace sémantique. À l'intérieur de celui-ci, nous construisons un critère de dissimilarité tel que la représentation de la partie spécifique d'une sous-chaîne dans cet espace soit suffisamment différente de la représentation de sa sur-chaîne dans ce même espace. Nous nous servons de ce critère pour écarter les sous-chaînes redondantes tout en conservant les sous-chaînes qui présentent un contenu informatif

spécifique.

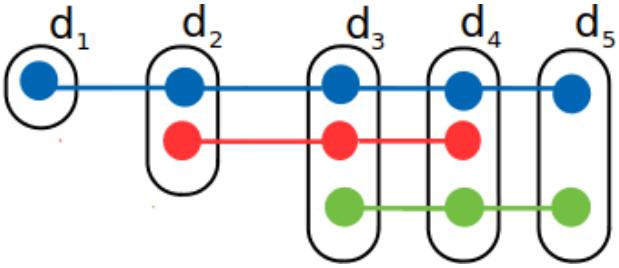


FIGURE 4.2 – Chaque lignée colorée représente une unité d'information. Les sous-chaînes $d_2d_3d_4$ et $d_3d_4d_5$ transportent des informations différentes de leur sur-chaîne $d_1d_2d_3d_4d_5$.

La divergence sémantique spécifique d'une sous-chaîne c_s de c est calculée à l'aide d'une similarité cosinus. En notant c la sur-chaîne, c_s la sous-chaîne, et ϵ un seuil d'admission positif aussi petit que l'on souhaite, le critère pour écarter une sous-chaîne est le suivant :

$$|cos(\vec{c_s}, \vec{c})| \geq \epsilon \tag{4.2}$$

Ainsi, toute sous-chaîne n'ayant pas d'information spécifique par rapport à la sur-chaîne est écartée.

Dans le reste du chapitre, nous considérons que T_D a été filtrée à l'aide de cette approche. Nous proposons maintenant une extension du $TF\text{-}IDF$ pour représenter les éléments de T_D à l'aide de descripteurs textuels.

4.2.2 Descripteurs textuels d'une chaîne

Différentes techniques peuvent être utilisées pour extraire des descripteurs pondérés d'un document à partir de son contenu textuel (cf. Section 2.1.2). Ce type de technique peut être étendu à une chaîne de documents. Nous proposons deux extensions de la mesure TF-IDF que nous notons CF-IDF et CF-ICF. Ces mesures permettent de pondérer l'importance d'unités textuelles dans chaque chaîne de documents. Une unité textuelle consiste en un mot, lemme ou encore un n-gramme. Une telle unité est appelée descripteur et sera dénotée par f. Nous notons $Desc_c$ l'ensemble des descripteurs pondérés pour une chaîne donnée c.

Définition 16. La mesure **CF** (Chain Frequency), pour un descripteur donné, correspond au nombre de documents de la chaîne contenant ce descripteur divisé par la longueur de la chaîne (nombre de documents de la chaîne). Cet indicateur permet de résumer l'importance du descripteur dans la chaîne.

$$CF(f, c) = \frac{|\{d_i \in c / f \in d_i\}|}{|c|}$$

Définition 17. La mesure **IDF** (Inverse Document Frequency) correspond à la métrique standard utilisée dans le TF-IDF. Intuitivement, cet indicateur permet de mesurer la rareté d'un descripteur au sein d'un corpus de documents D.

$$IDF(f, D) = log(\frac{|D|}{|\{d_i \in D / f \in d_i\}|})$$

Définition 18. La mesure **ICF** (Inverse Chain Frequency) correspond à l'extension de la mesure IDF à l'objet trajectoire. Cet indicateur permet de mesurer la rareté d'un descripteur dans une trajectoire T comme suit :

$$ICF(f, T) = log(\frac{|T|}{|\{c \in T / \exists d_i \in c, f \in d_i\}|})$$

Les mesures **CF-IDF** et **CF-ICF** sont définies de manière similaire à la mesure standard TF-IDF comme suit :

$$CF\text{-}IDF(f, c, D) = CF(f, c) \times IDF(f, D)$$

$$CF\text{-}ICF(f, c, T) = CF(f, c) \times ICF(f, T)$$

À l'aide de ces mesures, nous pouvons ainsi fournir pour chaque chaîne c de T une représentation vectorielle sur l'ensemble d'un vocabulaire F. La représentation ne sera pas la même selon que l'on choisit le CF-IDF ou le CF-ICF. Le premier favorise les descripteurs rares dans le corpus et fournit donc une description de la chaîne qui la discrimine vis-à-vis des documents du corpus, de la même manière que le

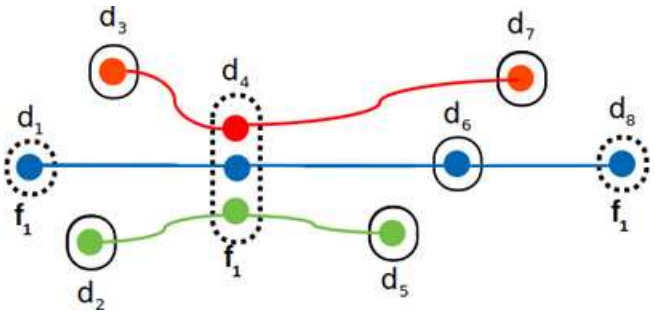


FIGURE 4.3 – Le CF de f_1 dans les chaînes du haut, du milieu est du bas vaut respectivement de 1/3, 3/4 et 1/3. L'IDF de f_1 vaut $log(\frac{8}{3})$ tandis que son ICF vaut $log(\frac{3}{3}) = 0$.

TF-IDF pour les documents. Le CF-ICF favorise les descripteurs rares parmi les différentes chaînes de la trajectoire, il s'agit donc de mettre en valeur les spécificités de cette chaîne vis-à-vis des autres chaînes de l'ensemble. Un exemple de mesures du CF, de l'IDF et de l'ICF pour un descripteur sont données en Figure 4.3, on constate que le descripteur étant présent dans l'intégralité des chaînes, sa valeur discriminante pour représenter une chaîne vis-à-vis des autres (ICF) est nulle. À l'aide de telles représentations, on peut maintenant chercher à détecter les différentes informations se propageant le long d'une chaîne. Dans le reste du chapitre, nous utilisons la représentation CF-IDF. Celle-ci étant indépendante de T, elle nous permet de raisonner sur une chaîne en isolation des autres, de plus elle met en valeur les descripteurs rares du corpus aussi son interprétation est plus claire. Nous donnons au chapitre 5 quelques utilisations de la représentation CF-ICF pour différencier les différentes chaînes de la trajectoire.

4.2.3 Extraction des informations d'une chaîne

Sur chaque chaîne, il se déplace une ou plusieurs informations. L'objectif est maintenant d'extraire ces informations pour une chaîne c donnée. Pour ce faire, nous proposons de construire pour chaque unité d'information i à identifier, un ensemble de descripteurs spécifiques en se basant sur l'ensemble des descripteurs de la chaîne. Nous supposons qu'un descripteur ayant un poids élevé pour la chaîne est susceptible de représenter une unité d'information se propageant dans la chaîne.

Hypothèse 8. Si un descripteur f a un poids élevé pour représenter une chaîne c, alors f a un poids élevé pour représenter une des principales informations se propageant le long de c.

À partir de cette hypothèse on construit une démarche générale pour extraire les différentes informations d'une chaîne :

— Extraire les descripteurs pondérés de la chaîne c.

— Discriminer les principales informations se propageant sur c. Il s'agit de ce que nous appelons *l'identification* des informations.

— *Caractériser* les informations de la chaîne c en associant à chacune les descripteurs les plus représentatifs.

Une illustration de cette démarche d'extraction, augmentée de la phase d'étiquetage, est donnée en Figure 4.4. À partir d'une chaîne, on extrait ses descripteurs. De ces descripteurs, on identifie ceux susceptibles de représenter une information importante de la chaîne. On regroupe ceux qui sont susceptibles de représenter la même information, c'est la caractérisation de l'information. Enfin, cet ensemble descriptif pour chaque information servira, nous en discutons en section suivante, à étiqueter l'information pour lui fournir une description plus lisible pour un utilisateur final humain.

Une première manière d'aborder l'extraction des informations serait d'admettre que de très nombreuses informations circulent sur chaque chaîne. De là, on pourrait utiliser chacun des N meilleurs descripteurs de la chaîne c comme représentant d'une des N principales informations se propageant le long de la chaîne (identification et caractérisation). Cependant, une telle stratégie présente plusieurs limitations. D'abord, un seul descripteur est insuffisant pour décrire précisément une unité d'information. De plus, un ou plusieurs descripteurs peuvent être sémantiquement très similaires et correspondre ainsi à la même unité d'information. Enfin, il est étrange de statuer que toutes les chaînes possèdent la même quantité d'informations importantes qui y circulent. Il semble raisonnable de penser que certaines chaînes ont une et une seule information principale tandis que d'autres en ont plusieurs.

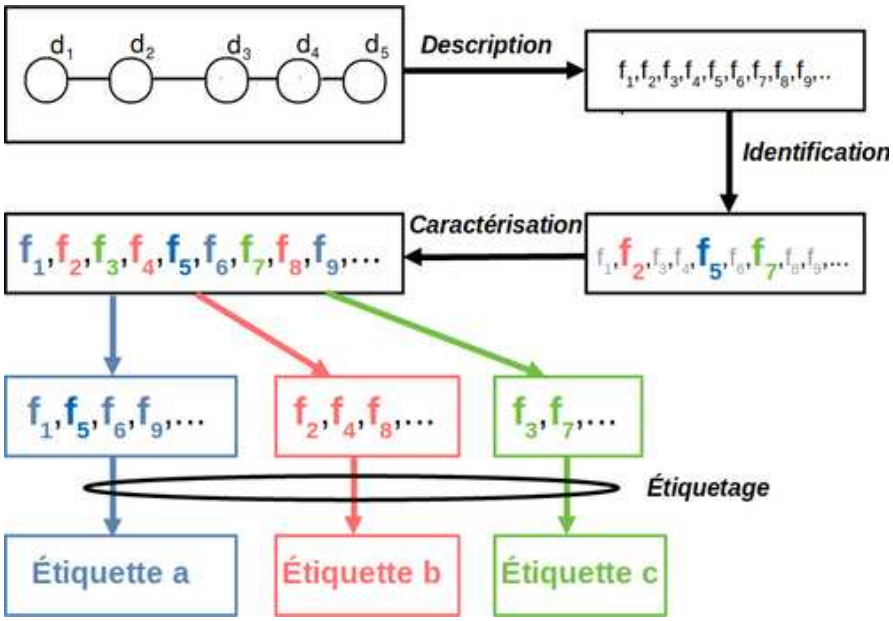


FIGURE 4.4 – Extraction et Étiquetage des informations d'une chaîne. Contrairement à la méthode illustrée en Figure 4.1, on identifie et caractérise les informations à partir de la description de la chaîne.

Nous souhaitons ainsi, pour une chaîne c, détecter de bons descripteurs qui sont susceptibles de représenter des unités d'informations. De tels descripteurs doivent être d'une part suffisamment pertinents pour la chaîne considérée, d'autre part, ces descripteurs devraient se distinguer suffisamment les uns des autres pour représenter les potentielles différentes unités d'informations de la chaîne.

Pour extraire des descripteurs représentatifs et spécifiques à chaque unité d'information, nous proposons d'utiliser un critère MMR (Maximal Marginal Relevance) tel que décrit dans [CarbonellGoldstein, 1998]. La pertinence d'un descripteur f, notée $Rel(f)$, correspond à son poids dans la chaîne. Pour calculer la redondance d'un descripteur f par rapport aux autres descripteurs dans la chaîne, nous avons besoin d'une fonction de similarité sémantique entre descripteurs. Cette dernière, dénotée par sim, pourrait être une simple mesure de cooccurrence ou une mesure plus élaborée telle que celle capturée par un algorithme comme Word2Vec.

La redondance peut être alors définie comme suit :

$$Red(f, Desc) = \max_{f' \in Desc} sim(f, f')$$

Nous définissons par ailleurs la pertinence marginale d'un descripteur comme suit :

$$MR(f, Desc, \alpha) = \alpha \times Rel(f) - (1 - \alpha) \times Red(f, Desc)$$

α étant un nombre réel entre 0 et 1 permettant de pondérer l'importance du critère de redondance. À l'aide de la pertinence marginale nous pouvons détecter un ensemble de descripteurs de la chaîne qui vérifie deux propriétés : d'une part, chacun des descripteurs de l'ensemble est pertinent pour la chaîne. D'autre part, chaque descripteur se distingue suffisamment des autres descripteurs de l'ensemble. Ces descripteurs sont donc à la fois pertinents et différents pour la chaîne, ce qui, en vertu de l'hypothèse 8, en font un ensemble candidat pour l'identification des informations principales de la chaîne.

À partir de cette réflexion, nous avons conçu un algorithme (cf. Algorithme 9) d'extraction de plusieurs ensembles de descripteurs correspondant pour chacun à une information spécifique se propageant le long de la chaîne donnée. Nous appelons un tel ensemble la **signature** de cette information. Cet algorithme se décompose en deux phases : la première identifie les descripteurs importants de la chaîne qui ne sont pas trop similaires les uns avec les autres (à l'aide du critère MMR). Cela constitue la phase d'*identification*, chacun des descripteurs isolé correspond à une information de la chaîne. Cependant, nous avons vu qu'un descripteur seul fournit une représentation ambiguë d'une information et qu'il est préférable d'en avoir plusieurs pour constituer le contexte de l'information. Aussi, la seconde phase traite les descripteurs restants et les affecte à la signature des différentes informations identifiées lors de la première phase. Il s'agit de la phase de *caractérisation*. Ainsi nous obtenons, à partir d'un ensemble de descripteurs représentant une chaîne, plusieurs sous-ensembles représentant les principales informations de la chaîne.

La procédure est illustrée par la Figure 4.4 : trois descripteurs de la chaîne sont sélectionnés lors de la phase d'identification, ils forment les graines pour les signatures des principales informations de la chaîne. Les descripteurs non sélectionnés sont affectés au descripteur-graine dont ils sont les plus proches. Cette seconde

Algorithme 9 : Extraction des signatures d'information d'une chaîne par critère MMR.

Data : Une liste pondérée de descripteurs $Desc$, un paramètre de redondance α, un seuil d'acceptation β
Result : Un ensemble de signatures d'informations $Sign$

$Untreated \leftarrow \emptyset$;
$Labels \leftarrow \emptyset$;
$Sign \leftarrow \emptyset$;
/* Phase d'identification. */
while $Desc \neq \emptyset$ **do**
  /* Extraction du descripteur de pertinence marginale maximale. */
  $f_{max} \leftarrow \underset{f \in Desc}{\arg\max}\ MR(f, Labels, \alpha)$;
  $Labels \leftarrow Labels \cup \{f_{max}\}$;
  $Desc \leftarrow Desc \setminus \{f_{max}\}$;
  /* Filtrage des descripteurs sous le seuil d'acceptation de pertinence marginal. */
  $F \leftarrow \{f \in Desc / MR(s, Labels, \alpha) < \beta\}$;
  $Desc \leftarrow Desc \setminus F$;
  $Untreated \leftarrow Untreated \cup F$;
end
/* Pour chaque descripteur extrait, on crée une signature. */
for $f \in Labels$ **do**
  $Sign_f = \{f\}$;
  $Sign \leftarrow Sign \cup Sign_f$;
end
/* Phase de caractérisation. */
/* On assigne chaque descripteur sous le seuil d'acceptation à une signature. */
for $u \in Untreated$ *par pertinence décroissante* **do**
  $Sign_f \leftarrow \underset{S \in Sign}{\arg\max}\ Red(u, S)$;
  $Sign_f \leftarrow \text{Sign}_f \cup \{u\}$;
end
return $Sign$

phase pourrait être assimilée à l'algorithme de classification K-means, avec une seule itération et ayant pour point de départ les descripteurs-graines calculés lors de la phase d'identification. Il serait donc utile de comparer les résultats obtenus à ceux d'un K-means entier, ou même un K-means initialisé aléatoirement. Nous verrons lors des expérimentations qu'une itération suffit pour obtenir un ensemble de signatures presque identique à celui calculé par un K-means. L'avantage de notre approche est que, contrairement au K-means, elle ne nécessite pas de plonger les descripteurs dans un espace vectoriel, ni de calculer des centroïdes dans cet espace.

Nous fournissons en Figure 4.5 un exemple d'une chaîne de documents issue d'un corpus du New York Times. Une des signatures d'information obtenue après application de notre algorithme inclut les descripteurs suivants :

a) vending machine	**b)** sugary	**c)** school
d) obesity	**e)** epidemic	**f)** snack
g) researchers	**h)** premises	**i)** increase

Ces descripteurs mettent fortement en évidence la corrélation entre les distributeurs de snack dans les écoles et l'obésité infantile. Ils forment un ensemble cohérent de termes qu'on pourrait rapprocher d'une thématique. Pour rappel, une thématique représente et synthétise un ensemble d'informations similaires dans un corpus. En ce sens, la signature est une thématique au niveau d'un corpus qui ne contiendrait que les documents de la chaîne : elle représente des informations similaires le long de la chaîne, ce qui est, potentiellement, une information qui se propage. Notre approche diffère cependant de l'approche des modèles thématiques. Là où les modèles thématiques font des thématiques les catégories générales qui structurent le corpus, la signature représente une des informations d'une chaîne de quelques documents, parmi les plusieurs chaînes de la trajectoire sur un corpus de plusieurs documents. La signature peut ainsi représenter une information très localisée, qui n'existe qu'à des endroits spécifiques d'un petit nombre de documents dans une masse importante, ce qu'un modèle thématique comme LDA aurait du mal à isoler. De plus, là où deux thématiques différentes ont pour vocation de représenter des choses différentes, deux

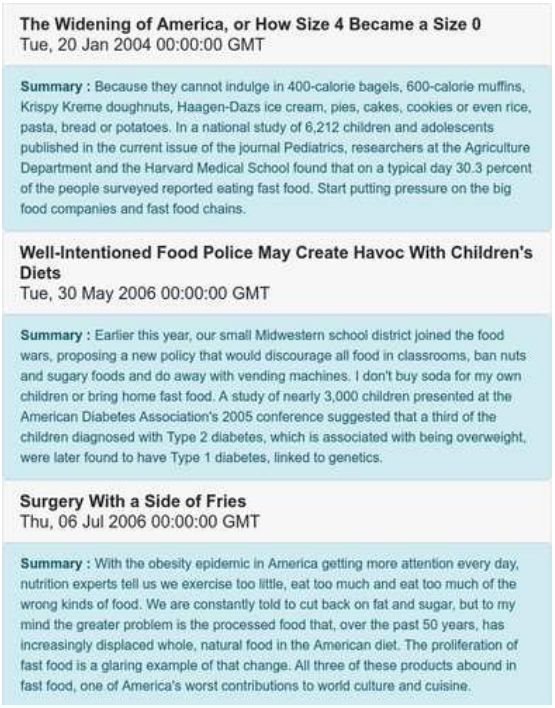


FIGURE 4.5 – Une chaîne de documents au sujet du fastfood aux États-Unis d'Amérique, extraite d'un corpus d'articles du New York Times.

signatures peuvent être similaires tout en étant différentes, ce qui permet d'étudier la nuance entre différentes chaînes.

4.2.4 Étiquetage des chaînes

La signature d'une information (composée de descripteurs pondérés) telle que calculée précédemment fournit une première représentation de l'information circulant le long d'une chaîne. Nous présentons dans cette section différentes stratégies permettant de construire des étiquettes significatives à partir de ces signatures.

Étant donnée une chaîne de documents c, et une signature d'information $Sign_i$ calculée à partir de cette chaîne, nous avons expérimenté les stratégies suivantes pour construire des étiquettes pour l'information :

— **Meilleur titre**. Nous construisons un texte global en concaténant les contenus textuels de tous les documents de la chaîne c. Nous réalisons ensuite une extraction automatique de titre orientée par les mots clés inclus dans la signature $Sign_i$. Cette extraction est basée sur des méthodes usuelles de résumé automatique de texte.

TABLE 4.1 – Différentes stratégies d'étiquetage d'information appliquées à la chaîne fournie en Figure 4.5

Dernier titre	"Surgery With a Side of Fries"
Titre central	"The Widening of America, or Hot Size 4 Became a Size 0"
Meilleur Titre	"Earlier this year, our small Midwestern school district joined the food wars, proposing a new policy that would discourage all food in classrooms, ban nuts and sugary foods and do away with vending machines."
Meilleur résumé	"Nor can they grad a candy bar or down a sugary soda when the snack bug bites. Well-intentioned food police may create havoc with children's diets earlier this year, our small Midwestern school district joined the food wars, proposing a new policy that would discourage all food in classrooms, ban nuts and sugary foods and do away with vending machines. We must have nutritionally literate doctors, but getting fast food out of hospitals will also require the kind of grassroots activism that has removed sugary sodas and candy from vending machines."

— **Meilleur résumé**. Un titre automatique orienté par la signature est construit pour chaque document de la chaîne. Ensuite, les titres ainsi obtenus sont concaténés pour construire un résumé global.

— **Titre central**. Cette stratégie se base sur la structure de la trajectoire T pour identifier un document central. Le processus consiste d'abord à identifier toutes les chaînes c' de T fortement similaires à la signature $Sign_i$. La similarité utilisée est un simple cosinus entre les vecteurs $Sign_i$ et $Desc_{c'}$. Nous désignons ensuite comme document central le document le plus fréquent parmi toutes les chaînes c', ainsi identifiées. Le titre retenu est alors celui du document central.

— **Dernier titre**. Cette stratégie consiste à considérer le titre du dernier document de la chaîne dans la chronologie. L'idée est que puisqu'il s'agit du dernier document de la chaîne, il est le document le plus à même d'être informé des autres. Il s'agit d'une stratégie de base dans le but de la comparer aux autres.

Les résultats obtenus selon ces différentes stratégies sont montrés dans la Table 4.1 pour la chaîne fournie en Figure 4.5.

4.3 Expérimentations

Cette section décrit d'abord la campagne d'évaluation conduite pour étudier nos méthodes d'extraction d'information à partir de la trajectoire. Nous avons utilisé deux corpus de documents et nous avons sollicité 15 experts pour les tâches d'annotation. Dans un deuxième temps, nous commentons les résultats de la campagne pour la tâche d'extraction des informations, et nous comparons les différentes stratégies d'étiquetage proposées. Enfin, nous comparons notre méthode d'extraction d'information basée sur la MMR à une méthode par $k - means$ et nous montrons que, bien que les deux approches semblent similaires, notre approche est plus adaptée à ce problème.

4.3.1 Protocole et corpus de documents

Nous avons utilisé deux corpus de documents. Le premier corpus est en français. Il traite de la coupe du monde de football de 2018 ; en particulier, de la finale entre la France et la Croatie. Le corpus comprend 748 documents issus de différents sites de presse, nous le nommons *mondial*. Le second, en anglais, traite de frites et de fast-food. Il est issu du New York Times entre 1987 et 2007. Il comprend 829 documents et nous le nommons *Fries*.

Nous avons calculé une trajectoire pour chacun des corpus. Après suppression des chaînes redondantes, nous avons extrait des signatures d'information pour chaque chaîne avec en paramétrant le critère MMR avec $\alpha = 0.7$ et $\beta = 0.8$, ce qui fournit un total de 145 chaînes et 214 signatures. Le choix d'$\alpha = 0.7$ est une préconisation de l'article introduisant la MMR [CarbonellGoldstein, 1998] pour extraire les parties les plus importantes de l'information. Nous constatons que cela fournit entre une et cinq signatures par chaînes mais nous n'avons pas étudié formellement la sensibilité de notre méthode aux paramètres α et β.

Concernant le protocole d'évaluation, chaque évaluateur devait évaluer un ensemble de chaînes pour chaque corpus. L'évaluation s'est faite en trois étapes :

1. Chaque chaîne est présentée au participant sous forme d'une liste de docu-

ments accompagnée d'un ensemble de mots clés (une signature d'information). L'évaluateur doit décider si l'ensemble des mots clés correspond à une information potentielle véhiculée par la chaîne (plusieurs choix **oui**, **non** ou **je ne sais pas**). Si la réponse est **oui**, l'évaluateur doit qualifier la correspondance par l'un des qualificatifs suivants : **bien**, **modérément** ou **mal**.

2. L'évaluateur se voit présenter une chaîne de phrases issue de la chaîne de documents et signature présentées à l'étape précédente. Il doit décider si une information potentielle se propage le long de cette chaîne (choix entre **oui**, **non** ou **je ne sais pas**). Si la réponse est **oui**, l'évaluateur doit préciser si l'information qu'il a identifiée sur la chaîne pourrait être décrite par l'ensemble des mots clés de l'étape précédente (qui lui sont de nouveau présentés). Il peut répondre par **oui**, **non** ou **je ne sais pas**.
3. L'évaluateur se voit présenter plusieurs titres calculés à partir de la chaîne et et de la signature des étapes précédentes. Pour chacun des titres, il doit préciser s'il décrit correctement l'information identifiée dans la chaîne. Il peut répondre par **bien**, **modérément** ou **mal**. Il doit par ailleurs choisir le meilleur titre dans la liste proposée.

4.3.2 Accord inter-évaluateurs

Nous avons demandé aux évaluateurs d'évaluer chacun le maximum de chaînes possible. Nous avons obtenu un total de 193 évaluations, dont 68 chaînes évaluées par au moins deux participants. Nous avons calculé l'accord inter-évaluateurs en utilisant le coefficient de Pearson (r). Les résultats sont fournis dans la Table 4.2. L'étape 2 montre le meilleur accord pour les deux corpus évalués et l'étape 3 est la plus controversée. L'accord est globalement meilleur pour le jeu anglophone *Fries* que pour le jeu francophone *mondial*, à l'exception de l'étape 3 dont l'accord est meilleure dans le jeu francophone. Nous considérons ces résultats comme satisfaisants étant donné le caractère très subjectif de la tâche demandée.

L'objectif visé par l'évaluation est de valider ou pas la capacité de notre approche

TABLE 4.2 – Accord inter-évaluateurs

Corpus	mondial	Fries
# chaînes interevaluée	44	24
# évaluations	108	85
r Étape 1 (mots)	0,47	0,66
r Étape 2 (résumé)	0,55	0,75
r Étape 3 (titres)	0,39	0,19

TABLE 4.3 – Distribution des évaluations pour des chaînes de trois documents et plus (%)

Étape	Réponse	mondial	Fries
1 (mots)	Oui / Bien	27,3	32,6
	Oui / Modérément	35,2	24,8
	Oui / Mal	02,3	01,2
	Je ne sais pas	06,8	13,2
	Non	28,4	28,2
2 (résumé)	Oui / Oui	42,2	40,9
	Oui / Je ne sais pas	07,8	09,9
	Oui / Non	21,6	07,1
	Je ne sais pas	05,7	13,7
	Non	22,7	28,4
3 (titres)	Au moins un bon titre	71,8	59,4
	Aucun bon titre	28,2	40,6

à capter la sémantique de l'information se propageant le long d'une chaîne. Notre approche semble fournir de bons descripteurs de l'information (signature) deux fois sur trois pour le corpus français et une fois sur deux pour le corpus Anglais. La distribution des réponses aux différentes étapes est donnée en Table 4.3. Les résultats sont satisfaisants sur les deux jeux de données dès l'étape d'extraction de mots avec 64.8% pour *mondial* et 58.6% pour *Fries*. Les étiquetages par résumé et par titre fournissent des résultats similaires.

Une des stratégies d'étiquetage fournit aussi des résultats satisfaisants, dans des proportions similaires au calcul de signatures. Nous montrons dans la Table 4.4 la répartition des différentes stratégies d'étiquetage. Chaque stratégie est représentée par l'une des réponses possibles à savoir **bien**, **modérément** ou **mal**, ainsi que le cas de figure où le titre issu de la stratégie est élu comme meilleur titre.

La stratégie du **Meilleur titre** est celle qui réunit le moins de mauvaises évalua-

TABLE 4.4 – Évaluation des titres par corpus (%).

Stratégie	Évaluation	mondial	Fries
Titre central	bien	41,9	**30,2**
Titre central	modérément	37,1	26,8
Titre central	mal	21,0	43,0
Titre central	meilleur	20,6	33,4
Meilleur titre	bien	**43,8**	27,9
Meilleur titre	modérément	38,1	41,9
Meilleur titre	mal	**18,1**	**30,2**
Meilleur titre	meilleur	**45,5**	25,1
Dernier titre	bien	31,5	24,4
Dernier titre	modérément	31,4	32,6
Dernier titre	mal	37,1	43,0
Dernier titre	meilleur	33,9	**41,5**

tions avec une moyenne de moins de 25% de mauvaises réponses (**mal**) sur les deux corpus de test.

Les stratégies **Titre central** et **Meilleur titre** semblent fournir des proportions similaires de bonnes réponses (**bien** et **modérément**), ce qui est peut-être surprenant compte tenu de la nature très différente des deux approches. En effet, la première stratégie est basée sur la structure induite par des chaînes similaires tandis que la seconde se base uniquement sur le contenu de la chaîne considérée.

La stratégie **Titre central** est la stratégie la moins élue comme fournissant le meilleur titre. Ce qui peut s'expliquer par le fait qu'elle fournisse en général un titre global, non spécifique à la chaîne considérée. À l'inverse, la stratégie **Dernier titre**, malgré des mauvais résultats en général, fournit régulièrement le **meilleur** titre, et en particulier sur le jeu anglophone. Il s'agit souvent du titre préféré lorsqu'aucun titre n'est jugé bon, ce qui explique ses bons résultats sur le jeu anglophone où 40.6% des évaluations n'ont trouvés aucun des trois titres bon. La sélection du Dernier titre dansc ce cas peut s'expliquer par le fait que, contrairement aux autres, ce titre provient au moins de le chaîne, ce qui lui assure une certaine cohérence.

TABLE 4.5 – ARI entre différentes stratégies de départ de K-means et notre approche basée sur la MMR

Méthode	mondial	Fries
Départ aléatoire	$0,30 \pm 0,04$	$0,33 \pm 0,06$
Même cluster	$0,65 \pm 0,15$	$0,58 \pm 0,14$
Même départ	$0,80 \pm 0,07$	$0,72 \pm 0,12$

4.3.3 Comparaison entre l'approche MMR et une approche K-means

La méthode d'extraction d'information basée sur un critère MMR que nous proposons repose sur l'idée suivante : les descripteurs importants de la chaîne c sont un agrégat des descripteurs importants des informations principales se propageant le long de la chaîne. L'approche peut ainsi se comprendre comme un clustering des descripteurs de la chaîne, où chaque cluster correspond aux descripteurs d'une information. Dès lors, il convient de se demander si notre approche a des avantages vis-à-vis des méthodes classiques de clustering, qui pourraient répondre à la même question.

Nous comparons notre méthode d'extraction à la méthode classique de clustering des K-means. La méthode des K-means est sensible à la manière dont elle est initialisée, aussi nous comparons l'approche par critère MMR avec trois stratégies d'initialisation différentes des K-centrers de clusters. Dans les trois cas, le nombre K de clusters est fixé au nombre de clusters trouvés par notre approche par critère MMR. Pour la première stratégie, on choisit le descripteur de départ de chaque cluster au hasard, on la nomme *départ aléatoire*. Pour la deuxième stratégie, on choisit le descripteur de départ de chaque cluster au hasard parmi les descripteurs de chaque cluster trouvés par l'approche par critère MMR, on la nomme *même cluster*. Enfin, pour la troisième stratégie, on choisit le descripteur de départ de chaque cluster comme étant le descripteur *label* calculé à la première étape de l'algorithme 9. On nomme cette stratégie *même départ*, puisqu'il s'agit des descripteurs de départ utilisés dans la seconde partie de l'algorithme pour construire les signatures. Nous mesurons l'écart entre l'approche par critère MMR et les différentes stratégies d'initialisations

de K-means à l'aide de l'indice de Rand ajusté (ARI) (cf. [Vinh, 2010]). Les résultats portent sur les 427 et 732 chaînes des deux jeux de données *mondial* et *Fries* et sont donnés en Table 4.5. On constate d'abord que les résultats d'un K-means avec départ aléatoire sont particulièrement différents ($ARI < 0.4$) de ceux obtenus par notre approche. En introduisant le biais de MMR au départ, on se retrouve avec des résultats plus similaires. En fait, utiliser exactement les descripteurs obtenus par le critère MMR (stratégie *même départ*) construit des clusters fortement similaires ($ARI > 0.7$). L'approche basée sur un critère MMR que nous exposons se comporte donc dans sa sortie comme un K-means biaisé par le critère MMR. Cependant, notre approche ne nécessite pas de plonger les descripteurs dans un espace vectoriel ni de calculer des centroïdes. Il s'agit d'une alternative plus rapide mais aussi efficace qu'une méthode classique de clustering dans notre cas d'application.

4.4 Conclusion

La trajectoire obtenue peut être le support et le point de départ pour plusieurs applications liées à l'étude et l'analyse de la propagation d'information. Néanmoins, la notion de chaîne qui constitue un bon indicateur de la structure de propagation, ne fournit pas précisément une description de l'information qui s'y propage. Nous avons proposé dans ce chapitre une approche globale permettant d'extraire et de caractériser l'information qui se propage le long d'une chaîne donnée.

Nous avons aussi conduit une campagne d'évaluation humaine pour valider l'approche proposée. Cette campagne a permis de démontrer que notre approche fournissait des descripteurs de chaîne pertinents quant à une information potentielle se propageant le long de la chaîne.

De plus, une description textuelle satisfaisante de type titre ou résumé de l'information, pouvait être obtenue sur la base de ces descripteurs.

A notre connaissance, la problématique qu'on s'est posée, à savoir la caractérisation d'une chaîne de propagation, est tout à fait innovante et originale. Les techniques que nous avons développées pourront servir de référence pour des travaux

futurs. Par exemple, il serait intéressant d'étudier d'autres pistes pour l'extraction de descripteurs de chaînes tels que l'extension d'approches de type Word2Vec aux chaînes de documents ou étudier des stratégies plus sophistiquées d'étiquetage de chaîne basées sur une combinaison du contenu et de la structure de la chaîne dans la trajectoire.

Dans ces travaux, nous avons focalisé l'attention sur la partie immuable de l'information. Nous fournissons tout de même l'extraction des résumés : chaque document de la chaîne est résumé par une phrase, orientée par la signature d'une information. Cela donne une succession ordonnée de phrases étiquetant cette idée de mutation le long de la chaîne. Dans le chapitre suivant, nous présenterons la notion de récit, qui est le regroupement de plusieurs signatures très similaires dans le but de fournir une représentation des différentes informations du corpus, mais aussi pour comparer les nuances qui existent entre plusieurs chaînes similaires. Dans le chapitre suivant nous nous intéressons ainsi à les nuances entre différentes chaînes qui possèdent une information similaire, ces nuances sont ainsi des témoins d'une mutation, que nous n'exhibons pas. Nous laissons en suspend la question de la caractérisation précise de la mutation d'une information le long de la chaîne, le couple chaîne-signature est dans ce travail notre représentation atomique d'une information. Nous envisageons quelques pistes pour la caractérisation de la mutation le long d'une chaîne : une première consiste à creuser l'idée d'activation des descripteurs développée par Shahaf [ShahafGuestrin, 2010], permettant de quantifier l'importance d'un descripteur pour la connectivité sémantique entre deux documents et à l'étendre à une chaîne. Une seconde, plus directe, serait d'étudier les signatures non seulement d'une chaîne mais aussi de toutes ses sous-chaînes pour en comparer les variations. Une prémisse de cette idée est brièvement présentée au chapitre 5 : il s'agit de l'attachement, qui étudie la fluctuation de la cohérence d'une chaîne et de ses sous-chaînes. Lorsqu'une information mute "trop", elle commence à diverger de son sens originel, et sa cohérence baisse mécaniquement. Ainsi l'attachement est une indication de l'intensité potentielle de la mutation le long d'une chaîne. Ce constat de divergence

lors de la mutation pose également la question de la reconstitution de l'histoire de la mutation, nos chaînes ne capturent pas les divergences trop importantes, et il serait intéressant de retrouver ces points de divergence entre les chaînes.

Chapitre 5

Explorer et exploiter la trajectoire

Résumé. Les chapitres précédents se sont attachés à définir, motiver et calculer la Trajectoire de l'information, ainsi qu'à extraire et caractériser l'information qui s'y propage. Ce chapitre est l'occasion de discuter nos différentes réflexions sur l'exploitation des objets que nous calculons et la manière dont on peut les intégrer dans un cadre applicatif convenant à un analyste. Nous commençons par présenter différentes approches pour visualiser et explorer le contenu d'une trajectoire à l'aide d'une maquette technique que nous avons réalisés. Ensuite nous présentons une étude de corpus dirigée par l'étude de la trajectoire. Enfin, nous discutons plusieurs idées d'application de la trajectoire.

Dans les chapitres 3 et 4 nous avons proposés des méthodes pour calculer des trajectoires, des ensemble de chaînes de documents le long desquelles de l'information est susceptible de se propager, et pour identifier et caractériser ces informations. À ce stade, nous sommes donc en mesure de construire une description à la fois structurelle et sémantique des informations qui se propagent dans un corpus. La suite logique de notre travail est d'exploiter cette description pour étudier la propagation d'information dans les corpus de documents.

Nous présenterons dans ce chapitre différentes manières d'exploiter l'outillage des trajectoires de l'information. Un prérequis pour l'usage et l'exploitation de la trajectoire est sa **visualisation**. Il s'agit de développer des approches d'affichage

et d'interaction permettant d'explorer la structure d'un ensemble de chaînes. Nous avons expérimenté plusieurs types de représentation en vue de développer des outils intuitifs de navigation dans la trajectoire. Ceci a conduit à la réalisation d'un outil que nous décrivons en première partie. Ensuite, nous proposons une étude de corpus assistée par les outils développés pour illustrer la pertinence de la trajectoire. Enfin, nous discutons les différentes pistes d'exploitation futures de la trajectoire que nous considérons.

5.1 Visualiser la trajectoire

Une trajectoire est un ensemble de chaînes de documents. Chaque document contient un texte plus ou moins long et peut apparaître dans plusieurs chaînes. De plus, chaque chaîne contient une ou plusieurs informations. Lorsque la trajectoire contient seulement une poignée de chaînes, il est possible de l'explorer manuellement, mais, dès lors qu'elle contient plusieurs dizaines, centaines ou milliers de chaînes, la trajectoire devient alors un objet complexe pour lequel il n'est pas évident d'avoir une représentation intuitive. Or, pour mieux comprendre et analyser la trajectoire, il serait intéressant de la voir. La visualisation de la trajectoire poursuit trois objectifs :

— Fournir une représentation macroscopique et intuitive de la trajectoire. Il s'agit de présenter la structure entière des différentes chaînes de manière lisible et digeste.

— Fournir des points d'entrée pour l'analyse et la compréhension de la propagation d'une information donnée.

— Fournir un arbre de navigation multidimensionnelle pour l'exploration du corpus en général.

Pour répondre à ces objectifs, nous avons expérimenté plusieurs représentations et modes d'interaction. Nous avons tout assemblé dans une maquette sous forme d'une interface Web [1] (cf. Figure 5.1). Celle-ci permet à un utilisateur de sélectionner un jeu de données ainsi qu'une trajectoire pré-calculée à partir de ce jeu de données

1. Une démonstration de la maquette est disponible à l'adresse `http://lks.charleshd.com/`

et de naviguer dans la trajectoire selon différents niveaux de granularité.

Cette section est structurée comme suit : après une présentation générale de la maquette, nous présentons les parties qui permettent la visualisation de chaque chaîne, puis celles qui permettent l'exploration de la trajectoire dans sa globalité. Enfin, nous présentons les *Story*, une manière d'explorer des regroupements de chaînes qui contiennent des information similaires.

FIGURE 5.1 – Page d'accueil de la maquette de visualisation de trajectoire

5.1.1 Présentation générale de la maquette

L'outil de visualisation propose plusieurs points d'entrée accessibles depuis plusieurs onglets, affichés dans la barre de navigation :

— *Accueil*. Page d'accueil fournissant le détail de fonctionnement de l'outil.

— *Configuration*. Tableau de bord permettant de calculer des trajectoires.

— *Jeux* Cet onglet permet la sélection d'une trajectoire déjà calculée pour l'analyser.

— *Visualisation*. Cet onglet permet de visualiser la trajectoire sélectionnée. Il est présenté en section 5.1.3.

— *Chaîne*. Cet onglet permet d'afficher la séquence de documents constituant la chaîne sélectionnée. Il est présenté en section 5.1.2.

— *Chaînes semblables*. Cet onglet permet de visualiser le graphe des chaînes similaires. Il est présenté en section 5.1.2.

— *Story*. Cet onglet propose les principaux *récits* issus de l'analyse de la trajectoire. La notion de récit, ainsi que l'onglet, sont présentés en section 5.1.4.

— *Timeline*. Cet onglet permet de visualiser et explorer un récit sélectionné. Il est présenté en section 5.1.4.

L'outil fonctionne essentiellement par sélection d'objets à visualiser. L'objet principal est la Trajectoire, aussi faut-il pouvoir la calculer et la sélectionner.

Calculer et sélectionner une trajectoire

Nous avons expérimenté de multiples manières de calculer la trajectoire : il y a plusieurs corpus différents, plusieurs manière de calculer la similarité entre les documents, plusieurs manières d'estimer la cohérence, la faible cohérence, ses seuils, etc. Cela constitue un nombre de paramètres important. Pour s'y repérer facilement, nous avons construit un tableau de bord pour choisir ces paramètres et lancer les calculs de trajectoire. Il s'agit de l'onglet Configuration, dont on peut voir une illustration en Figure 5.2. Il contient de nombreux paramètres regroupés en section dont la description exacte est donnée dans la maquette. Il permet entre autre de choisir un corpus de document, une requête pour restreindre ce corpus aux documents contenant certains mots et filtrer les documents sur leur nombre de caractères, choisir et paramétrer la similarité entre les documents ainsi que la cohérence, et enfin déterminer les paramètres de calcul de la trajectoire cohérente.

De l'autre côté, il est pratique de pouvoir simplement voir les paramètres et sélectionner les différentes trajectoires qui ont été calculées. Pour cela nous avons construit l'onglet *Jeux* illustré en Figure 5.3. Les trajectoires y sont triées selon leur chronologie, les plus récentes en premier. Pour chacune des trajectoires plusieurs informations sont fournies, notamment : Le nom jeu de données dont est issue la trajectoire, la requête utilisée pour restreindre la taille du jeu de données, le nombre de documents dans ce jeu de données après filtrage de la requête, ainsi que la quantité de chaînes non redondantes obtenues dans la trajectoire calculée. Cet onglet permet de sélectionner la trajectoire pour pouvoir l'analyser.

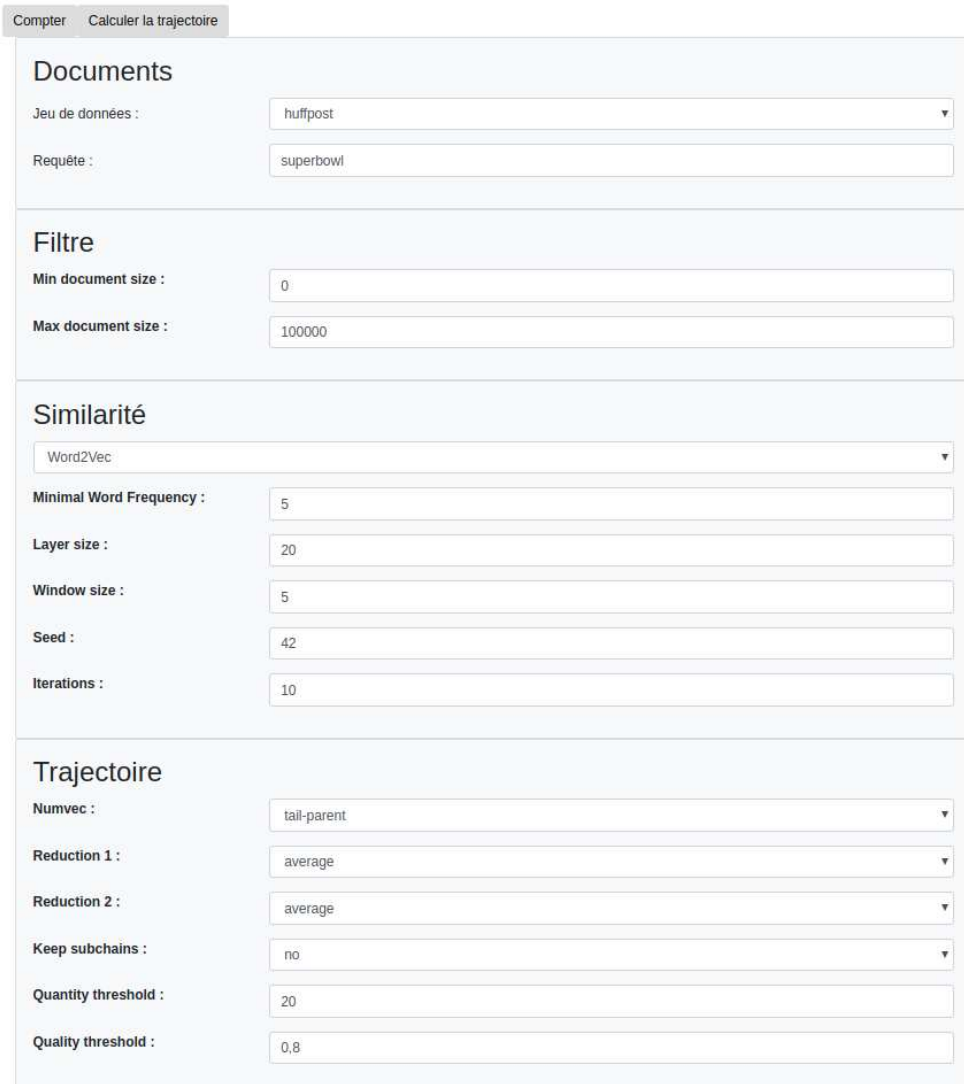


FIGURE 5.2 – Onglet *Configuration* de la maquette.

Cette partie technique de la maquette permet de construire et de gérer différentes trajectoires sur différents jeux de données. Nous présentons maintenant trois manières d'explorer et visualiser la trajectoire selon qu'on veut observer des chaînes en particulier, la trajectoire dans sa globalité, ou des ensemble de chaînes sémantiquement similaires.

5.1.2 Explorer les chaînes

La trajectoire est composée de chaînes, aussi la visualisation de chacune des différentes chaînes et des documents qui les composent est une première étape nécessaire à l'analyse de la trajectoire. Comme il s'agit d'une brique essentielle, on

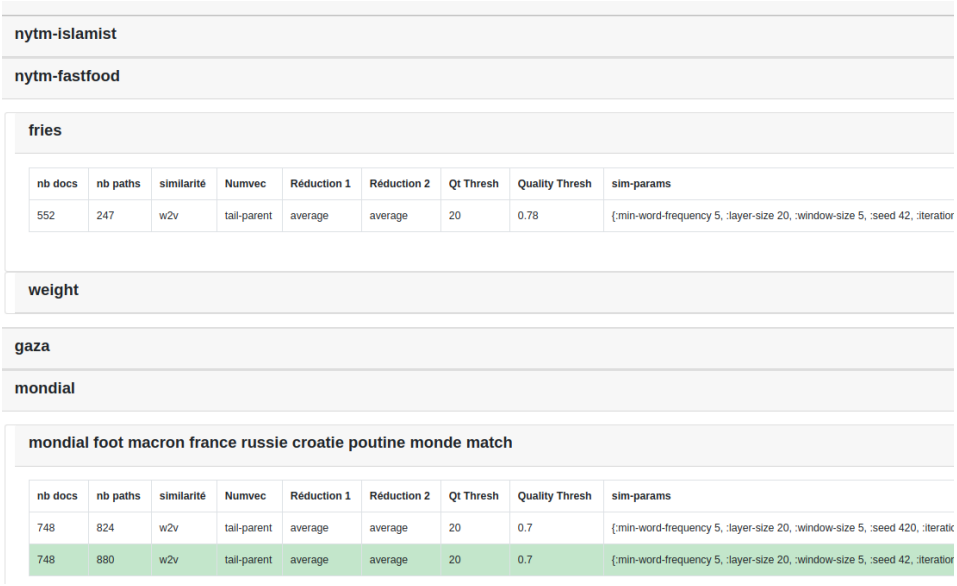


FIGURE 5.3 – Onglet *Jeux* de la maquette.

retrouve la capacité à visualiser une chaîne sélectionnée dans la plupart des onglets de visualisation de l'outil. Il est possible, dans l'onglet *Visualisation* d'avoir une vue de la liste des chaînes de la trajectoire. Il s'agit de la représentation **Chaîne** de la trajectoire. Un exemple de l'onglet *Visualisation* avec la représentation **Chaînes** est donné en Figure 5.4 avec les différentes parties des informations encadrées. Nous présentons maintenant les différents encadrés.

Représentation Chaînes (encadré 1)

Les différentes chaînes de la trajectoire sont présentées en succession de gauche à droite, dans un ordre décroissant de cohérence (la plus cohérente est à gauche). Il est possible de survoler avec la souris les différentes trajectoires, ou d'en sélectionner une en cliquant dessus. Dans les deux cas, cela fait apparaître les encadrés 2 à 5 qui décrivent la chaîne survolée ou sélectionnée. Les deux autres représentations, **Graphe** et **Date** servent à obtenir une visualisation macroscopique de la trajectoire et seront décrites dans la section suivante.

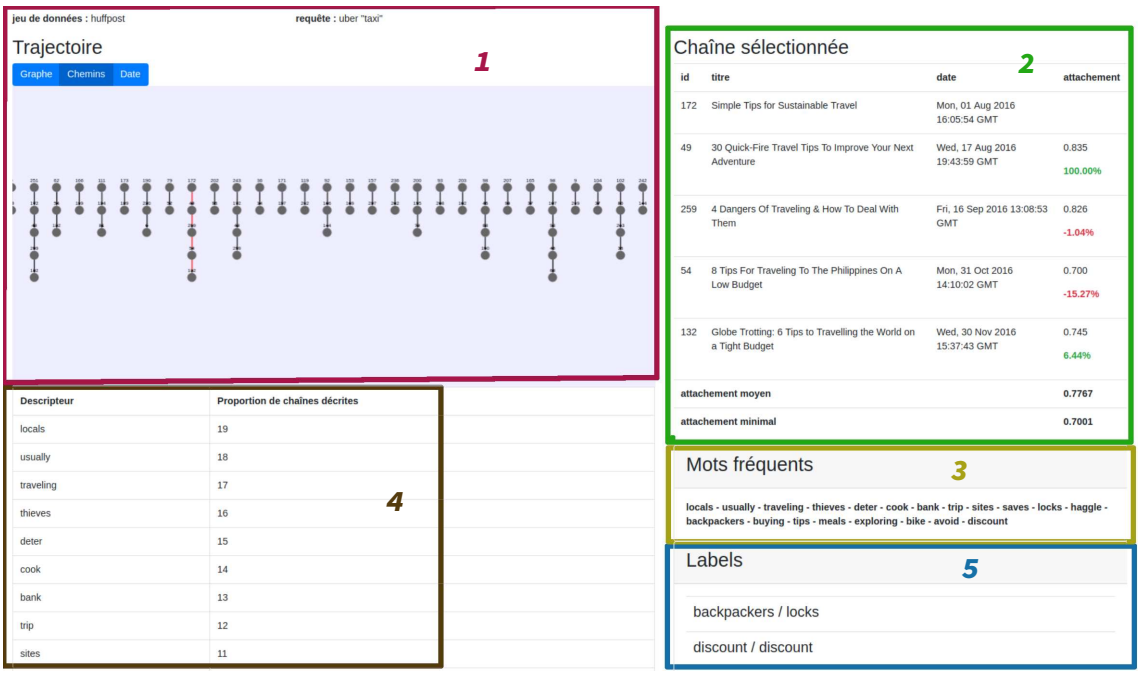


FIGURE 5.4 – Illustration de la représentation **Chaînes** de la trajectoire et une chaîne sélectionnée (en rouge dans l'encadré 1). Cette représentation est accessible depuis l'onglet *Visualisation*, elle se nomme **Chemins** dans la maquette, plus évocateur pour un analyste que le concept de chaîne.

Chaîne sélectionnée (encadré 2)

Il s'agit d'une représentation séquentielle des différents documents de la chaîne de haut en bas, avec en haut le document le plus ancien de la chaîne et en bas le plus récent. Pour chaque document, on affiche son titre et sa date. En cliquant sur la ligne, on a également accès au contenu du document (cf. Figure 5.5). Certains mots du contenu ont un fond coloré. Il s'agit des descripteurs des différentes informations extraites de la chaîne à l'aide de la méthode décrite au chapitre 4.

La dernière information fournie pour chaque document est son attachement. Il s'agit de la cohérence de la sous-chaîne qui part du document le plus ancien jusqu'au document actuel. La notion d'attachement est reliée à la méthode que nous proposons pour calculer une trajectoire (cf. Chapitre 3), il s'agit à chaque étape de voir comment l'ajout d'un document influe sur la cohérence globale de la chaîne. Sous l'attachement, en vert lorsqu'elle est positive ou en rouge lorsqu'elle est négative, est affichée la variation de l'attachement. Dans l'exemple affiché en Figure 5.4, on voit que les documents 259 et 54 font diminuer l'attachement, tandis que le document

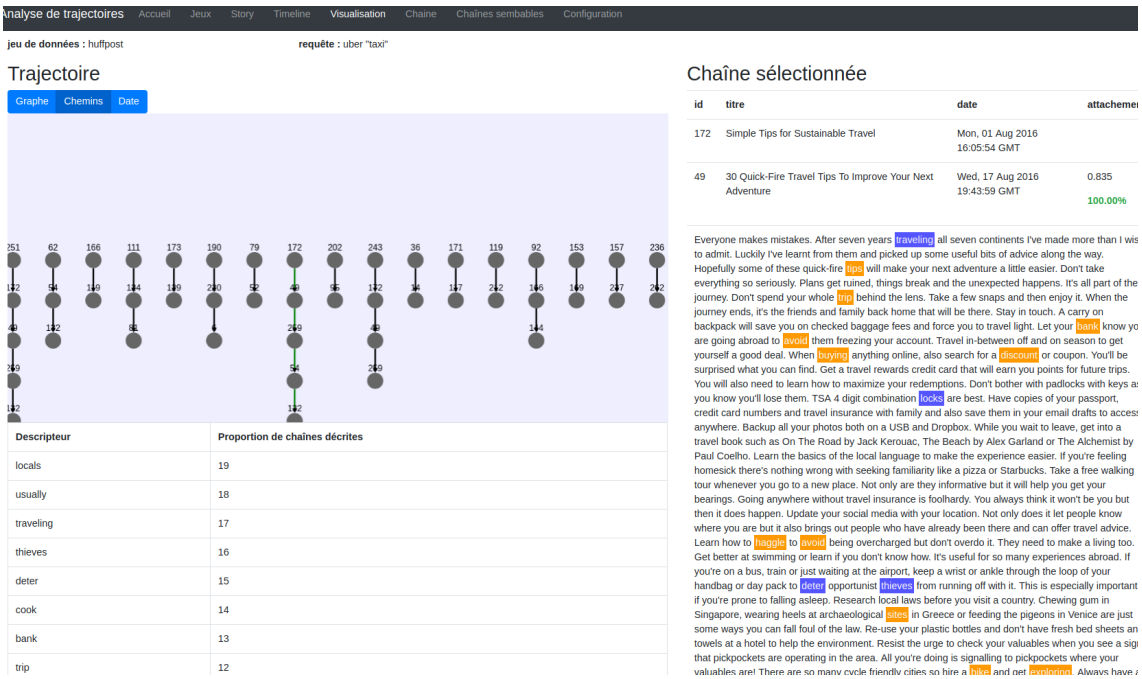


FIGURE 5.5 – Avec un clic sur une ligne de la chaîne sélectionnée, on affiche le contenu du document. Les mots en surbrillance correspondent aux descripteurs des différentes informations extraites de la chaîne. Ici, il y en a deux représentés par les couleurs bleu et jaune.

132 l'augmente. Nous interprétons la variation de l'attachement comme un critère de stabilité de la chaîne : s'il varie peu, la chaîne est stable, si au contraire il varie grandement, certains maillons de la chaînes sont plus faibles que d'autres. Une étude plus approfondie du rôle de l'attachement est une des pistes de réflexion future de notre travail.

Mots Fréquents (encadrés 3 et 4)

Les mots fréquents sont les mots dont le CF-IDF (extension du TF-IDF aux chaînes, cf. Chapitre 4) est le plus élevé pour la chaîne sélectionnée. Par défaut, la plateforme affiche les 20 premiers. Ils ne sont pas classés par ordre de CF-IDF, mais par la quantité de chaînes dans la trajectoire pour lesquels ils sont mots fréquents. Cette quantité est donnée dans l'encadré 4. Les mots fréquents sont donc les mots les plus représentatifs de la chaîne.

Labels (encadré 5)

Ces mots fréquents se retrouvent directement dans l'extraction des informations de la chaîne, procédure détaillée au chapitre 4. Ces informations sont données dans l'encadré 5, sous le nom de *labels*. Il s'agit d'une liste de paire de descripteurs, chaque paire correspond à une information extraite. Le premier est le descripteur-graine de MMR maximale pour cette information (cf. Chapitre 4), il s'agit donc du descripteur le plus important pour cette information. Le second est le descripteur le plus similaire à tous les autres descripteurs pour la signature de cette information. Il est le mot le plus représentatif de la signature et il ne s'agit pas toujours du descripteur-graine. Dans l'exemple de la Figure 5.4, il y a le cas où il s'agit du même mot *« discount / discount »*, et le cas où il s'agit de deux mots distincts *« backpackers / locks »*. En survolant chaque ligne avec la souris, les autres mots décrivant l'information extraite apparaissent. En cliquant sur la ligne, s'affiche une représentation de l'information extraite sous forme d'une séquence de phrases. Chacune des phrases est extraite dans l'ordre des documents de la chaîne, une illustration pour l'exemple courant est donnée en Figure 5.6. Cette représentation permet de se faire une idée du fil directeur de l'information dans chaque document.

La totalité de ces indicateurs permet ainsi de visualiser les différentes chaînes de la trajectoire et fournit un socle pour à la fois analyser la chaîne et avoir un retour sur la sortie de nos différentes méthodes. S'il s'agit ici d'une visualisation chaîne par chaîne de la trajectoire, les possibilités d'analyse ne s'arrêtent pas à ce niveau là. Dans la section suivante, nous explorons la trajectoire comme une structure dans son ensemble.

5.1.3 Explorer la trajectoire

La trajectoire est composée de chaînes elles-mêmes constituées de documents. Il existe ainsi plusieurs manières de visualiser la structure de la trajectoire. On peut représenter la manière dont les chaînes lient les documents, il s'agit de la visualisation la plus intuitive de la trajectoire, et la représentation utilisée dans les chapitres

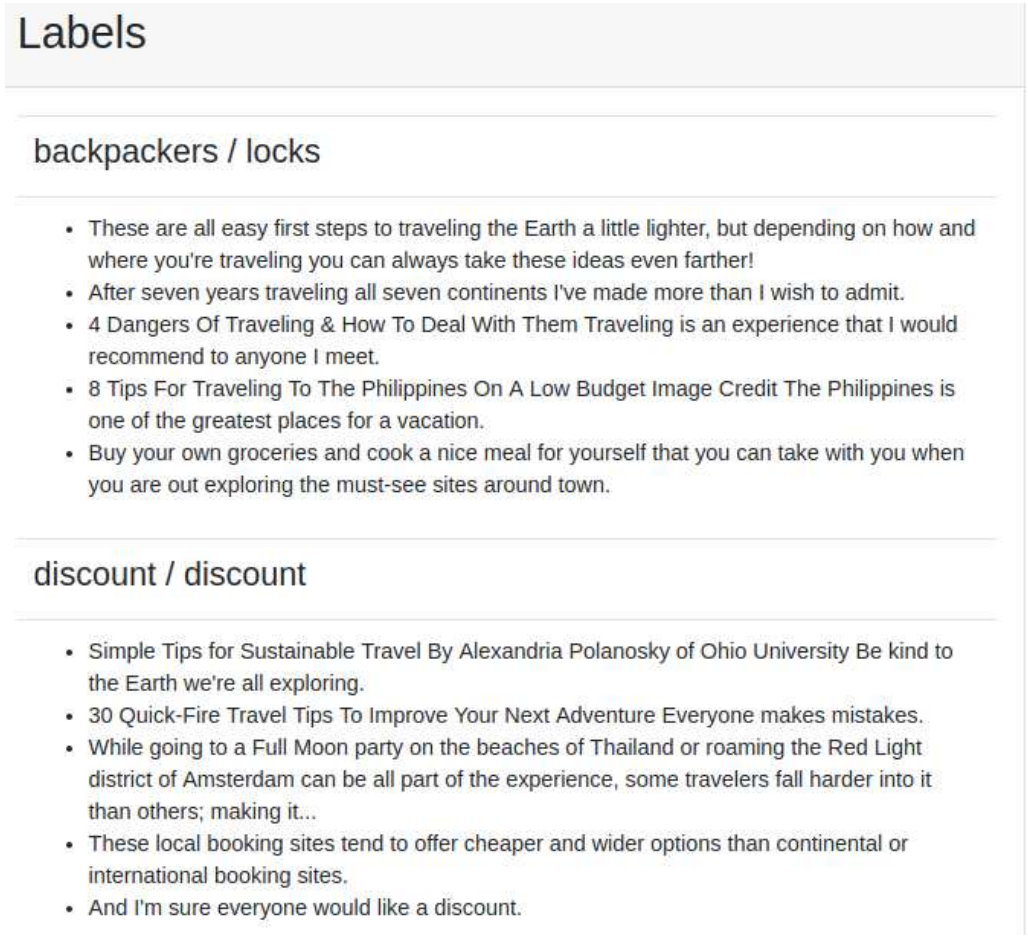


FIGURE 5.6 – Les différents labels de la chaîne sélectionnée en Figure 5.4. Chaque label est représenté par une séquence de phrases extraites des documents de la chaîne.

précédents pour l'illustrer. On peut également afficher la manière dont les chaînes interagissent les unes avec les autres. Nous présentons les deux manières de visualiser la trajectoire.

La trajectoire sur les documents

Dans l'onglet *Visualisation*, nous avons présenté la représentation **Chaînes** de la trajectoire qui affiche les différentes chaînes composant la trajectoire dans l'ordre décroissant de cohérence. Les deux autres représentations de cet onglet, **Graphe** et **Date**, affichent les différentes chaînes sur un *multi-graphe*. Un multi-graphe est une extension du concept graphe qui permet d'avoir plusieurs arcs différents entre deux sommets identiques. Ainsi, deux chaînes $d_1d_2d_3$ et $d_1d_2d_4$ peuvent être représentées dans un multi-graphe tel que l'arc d_1d_2 soit présent une fois pour chaque chaîne.

La représentation **Graphe** présente ainsi un multi-graphe sur les documents tel que tout maillon de chaîne $d_1 d_2$ soit présent autant de fois que de chaînes dans lequel il apparaît. Cela permet d'observer, en plus de ce qu'une représentation graphe permet de faire comme les composantes connexes, les zones de fortes ou faibles densité. La Figure 5.7 illustre la représentation **Graphe**. Le placement des documents et des arcs est optimisé par l'algorithme CoSe [Dogrusoz, 2009] qui nous a semblé donner des représentations convenables. On peut observer qu'il y a plusieurs composantes, et des zones plus ou moins denses en quantité de chaînes. De la même manière que pour la représentation **Chaînes**, il est possible de survoler et sélectionner les chaînes qui intéressent l'utilisateur, ce qui lui donne la forme exacte de la chaîne à l'aide d'un changement de couleur (cf Figure 5.8).

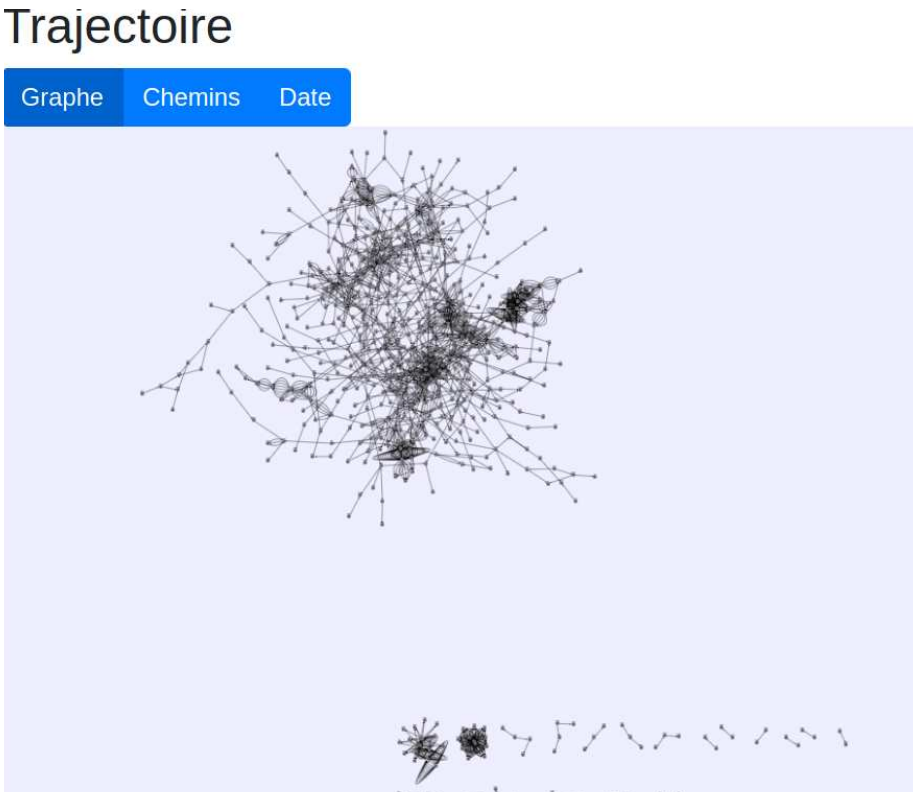


FIGURE 5.7 – Représentation de la trajectoire sous forme d'un multi-graphe. La trajectoire est calculée à partir du jeu de données *mondial*.

La représentation **Date** est très similaire à la représentation **Graphe**, la principale différence réside dans le placement des documents sur le plan. Cette fois, ils respectent la chronologie des documents : les documents de gauche sont plus anciens que les documents à leur droite. Cela permet d'observer le comportement des chaînes et des

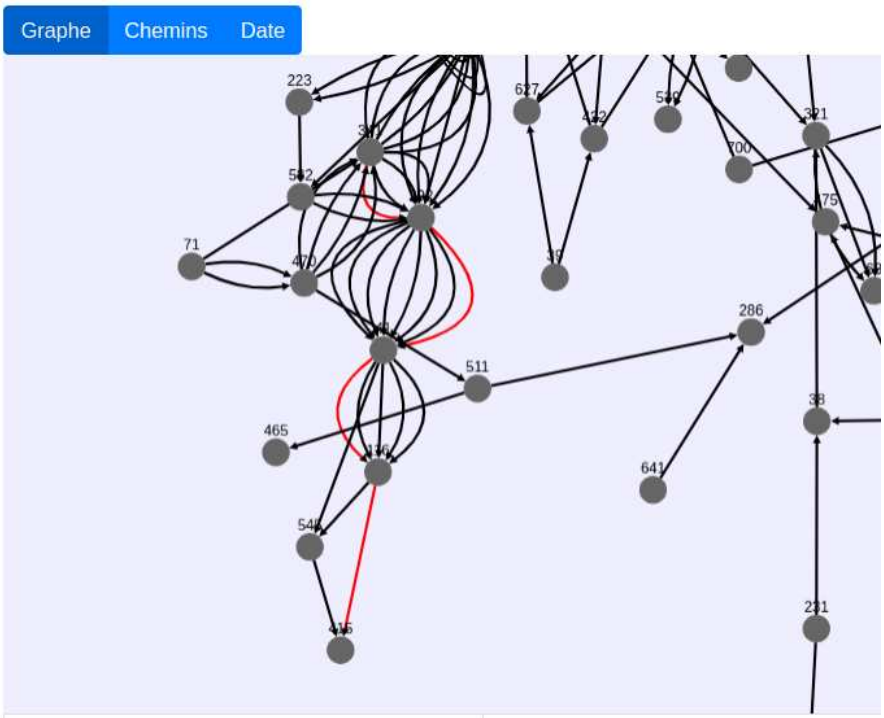


FIGURE 5.8 – Il est aussi possible de sélectionner et visualiser une chaîne dans la représentation **Graphe**. Il peut y avoir plusieurs arcs entre deux documents car chaque arc représente un maillon spécifique d'une chaîne. Cela permet également de rendre compte visuellement si de nombreuses chaînes passent par un document ou non.

documents sur un axe temporel. Une illustration est donnée en Figure 5.9.

Ces représentations permettent d'obtenir une intuition de la forme de la trajectoire sur les documents. On peut aussi observer comment les chaînes de la trajectoire interagissent les unes avec les autres.

Graphes d'interaction des chaînes

Une trajectoire peut contenir un nombre très important de chaînes dont certaines très semblables. Certaines sont semblables parce qu'elles convoient des informations qui sont sémantiquement similaires, d'autres sont semblables parce qu'elles partagent une quantité importante de documents. Il est à noter que ces idées sont indépendantes de celle de la similarité entre les chaînes, deux chaînes peuvent être proches sémantiquement sans partager un seul document, et deux chaînes peuvent partager plusieurs documents mais représenter des informations qui ne sont pas similaires. Ces

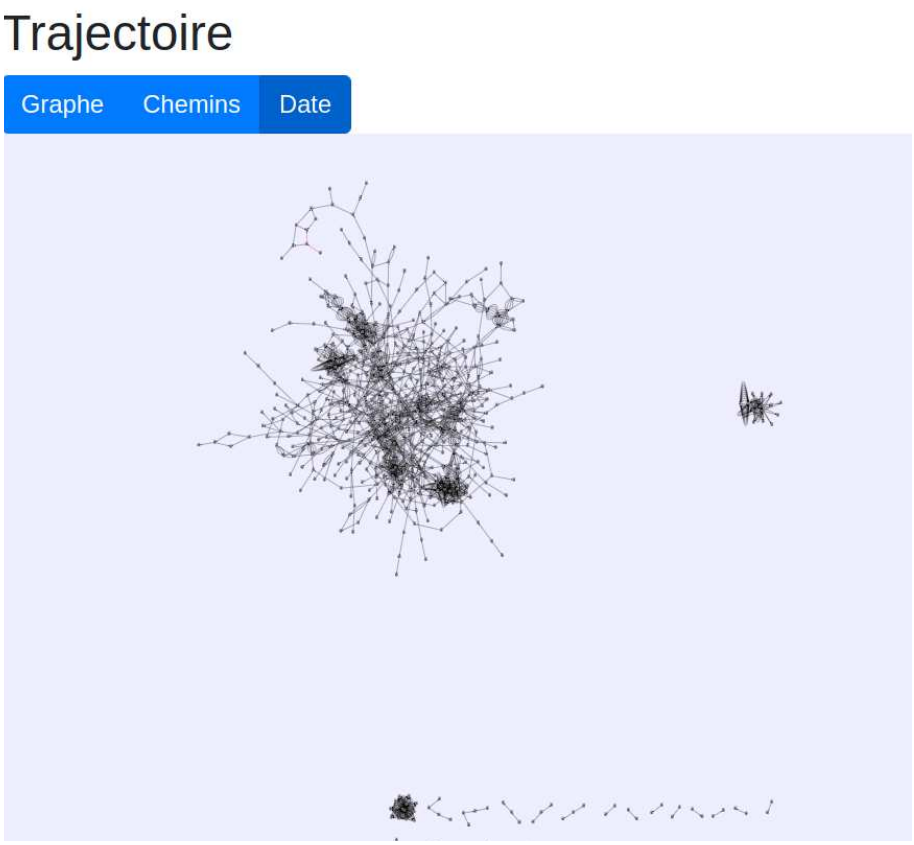


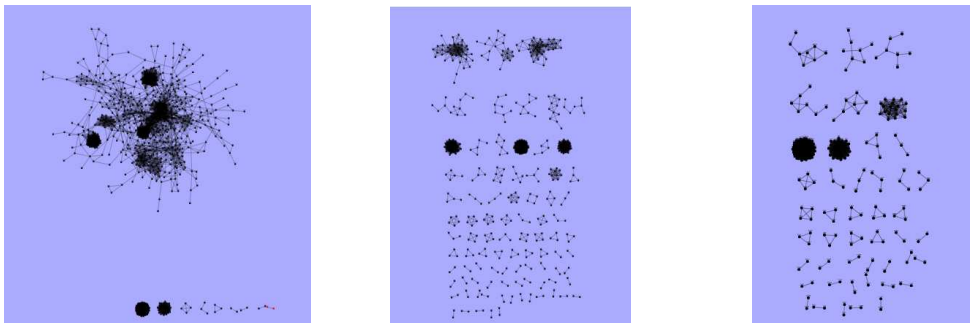
FIGURE 5.9 – Représentation de la trajectoire sous forme d'un multi-graphe en respectant la chronologie des documents. La trajectoire est calculée à partir du jeu de données *mondial*.

notions de similarité sémantique et structurelle peuvent être intéressantes d'abord pour mieux comprendre la structure de la trajectoire, ensuite elles peuvent être intéressantes pour définir une notion de redondance sémantique et de redondance structurelle, qu'on pourrait vouloir minimiser à des fins, entre autres, de résumé de trajectoire.

Le problème pourrait se formuler de la manière suivante : on a un ensemble de chaînes T qui contient de nombreuses chaînes. On voudrait le réduire en un ensemble T' contenant moins de chaînes tout en minimisant la perte d'informations. Il s'agit d'un problème différent de celui de la suppression des sous-chaînes sémantiquement redondantes évoqué au chapitre 4. Il s'agit ici d'enlever des chaînes différentes, qui ne sont pas nécessairement dans une relation de sous-chaîne l'une par rapport à l'autre, qui parlent de sujet relativement similaires en vue d'avoir une quantité réduite de chaîne à présenter à un analyste, par exemple de l'ordre de la dizaine. Avoir un

nombre de chaîne réduit permet de se focaliser sur les histoires clefs du corpus et d'avoir une vision résumée de la propagation. Cela soulève de nombreuses questions. Notamment pour deux chaînes ayant exactement les mêmes descripteurs : laquelle conserver ? Faut-il les combiner en une seule chaîne ? Si oui, comment ? Faut-il tronquer certaines chaînes ? Plutôt en amont, en aval, ou au milieu ? En vue de répondre à cette problématique et à l'ensemble des questions soulevées, nous avons commencé à explorer la similarité entre chaînes.

Dans cette optique, nous avons commencé à construire un **graphe de similarité** dont les sommets représentent les différentes chaînes de la trajectoire et les arcs entre les sommets indiquent la présence d'une similarité entre les chaînes. Dans l'onglet *Chaînes semblables*, on représente un graphe de similarité sémantique. La similarité est définie ici par le pourcentage de mots fréquents communs aux deux chaînes. Le seuil de similarité est défini par l'utilisateur, ce qui lui permet de visualiser la similarité des chaînes pour différents seuils, comme illustré en Figure 5.10. Pour étudier les chaînes similaires, l'outil de visualisation permet également de sélectionner une composante connexe du graphe de similarité pour étudier les différentes chaînes qui la composent sur les documents, comme dans l'onglet *Visualisation* (cf. Figure 5.11). On peut ainsi observer la structure des chaînes similaires, la manière dont elles s'intersectent et les descripteurs qu'elles partagent.



(a) Seuil de 10% (b) Seuil de 50% (c) Seuil de 80%

FIGURE 5.10 – Graphe de similarité sémantique pour différents seuils. Le jeu de données utilisé est *mondial*.

La notion de similarité structurelle des chaînes est une piste de réflexion intéressante que nous n'avons pas étudiée en détail et reste à l'état d'ébauche. Elle semble

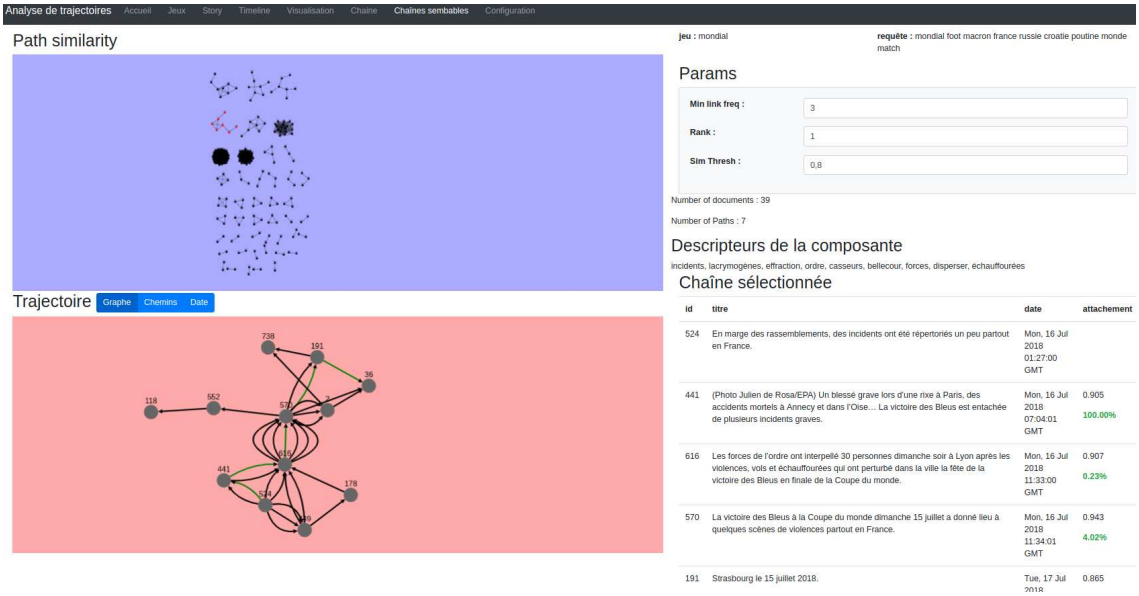


FIGURE 5.11 – L'onglet *Chaînes semblables*. La composante connexe de chaînes sélectionnée dans le premier graphe (points rouges) s'affiche sur les documents, comme dans l'onglet visualisation dans le second graphe. Le jeu de donnée utilisé est *mondial*.

tout de même présenter des possibilités intéressantes dans l'optique d'un résumé du corpus en histoires ou d'un résumé de la trajectoire en elle-même. La notion de similarité sémantique permet tout de même de visualiser et de donner une première analyse sur les chaînes qui utilisent des mots similaires. Cette approche repose sur les descripteurs fréquents des chaînes. Nous proposons maintenant une méthode d'exploration de la trajectoire basée sur les informations extraites des chaînes à l'aide des méthodes détaillées au Chapitre 4.

5.1.4 Explorer les *récits*

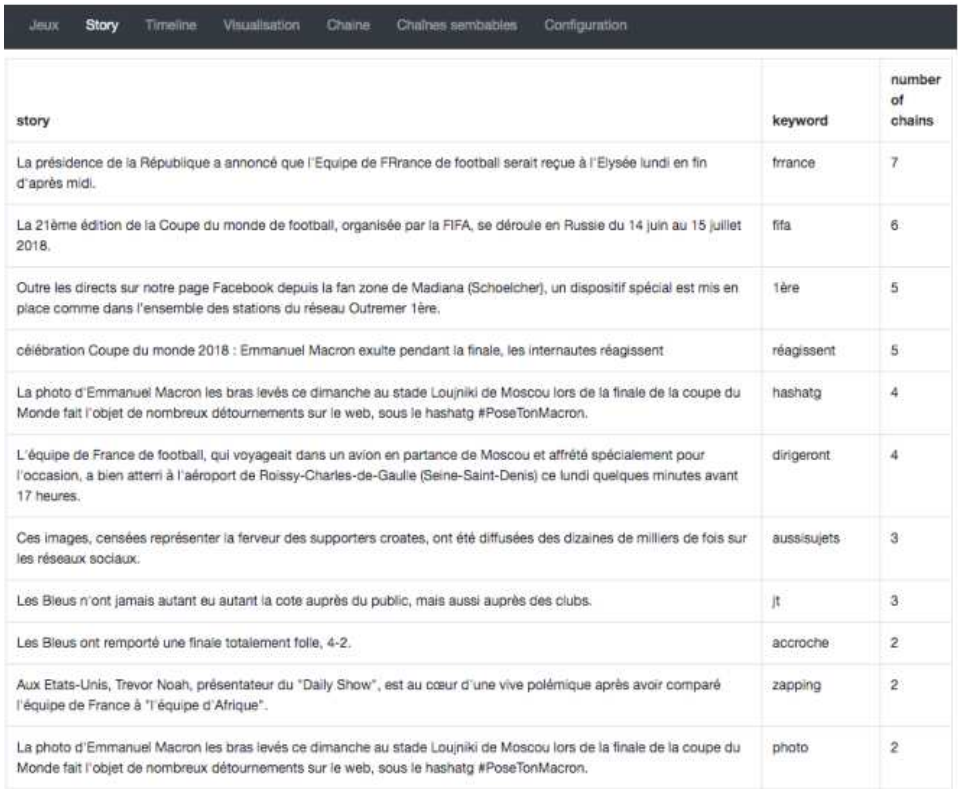
La trajectoire calculée consiste en un ensemble de chaînes de documents. Ces chaînes sont souvent très nombreuses et transportent vraisemblablement, dans beaucoup de cas, une information similaire. Pour en faciliter l'analyse et l'interprétation, il nous a paru intéressant de regrouper ces chaînes par similarité, comme nous venons de le voir. Or le long des chaînes peuvent se propager plusieurs informations différentes et non similaires entre elles. Nous avons présenté lors du Chapitre 4 une méthode pour identifier et étiqueter les différentes informations circulant sur chacune des

chaînes de la trajectoire. Il devient pertinent d'étudier les chaînes le long desquelles circulent une information similaire. Pour ce faire, nous avons construit la notion de *récit*. Un récit est un ensemble de signatures d'information fortement similaires entre elles. Nous classifions ainsi toutes les informations que nous identifions dans les différentes chaînes de la trajectoire en récits. Chaque composante ainsi calculée représente un des *récits* de la trajectoire.

Pour représenter chaque récit, nous utilisons à la fois le descripteur le plus représentatif des différentes informations composant le récit, et nous utilisons la méthode de titrage du **Titre central** présentée lors du Chapitre 4 pour donner un titre au récit. Il s'agit, structurellement, du titre du document le plus central dans l'ensemble des chaînes qui transmettent une des informations composant le récit.

La liste des récits pour la trajectoire sélectionnée est disponible dans l'onglet *Story* de l'outil de visualisation. Pour chaque récit, on a accès à son titre, son descripteur représentatif et le nombre de chaînes contenant une information du récit. La liste des récits pour le jeu de données *mondial*, parlant de la victoire de la France à la coupe du monde de football de 2018, et pour le jeu de données *fries*, contenant les articles du Huffington Post traitant de fastfood, et plus spécifiquement de frites, sont donnés en Figure 5.12 et en Figure 5.13 respectivement. Ces récits correspondent à différents moments forts du mondial, notamment l'annonce de la victoire, la photo montrant la réaction du président français lors du sifflet final de la rencontre, la polémique sur l'origine des joueurs, etc. Ceci représente un bon résumé des principales informations propagées et constitue un point d'entrée permettant d'orienter le veilleur directement vers l'information qui l'intéresse.

Il est possible de sélectionner un des récits disponibles dans l'onglet *Story* pour le visualiser dans l'onglet *Timeline*. L'onglet *Timeline* permet de visualiser le récit sélectionné de manière similaire au graphe de trajectoire de l'onglet *Visualisation*, à la différence près qu'un récit ne constitue qu'un sous ensemble de toute la trajectoire. Par ailleurs, les nœuds du graphe correspondants aux documents sont représentés par leur titre, ce qui ne pouvait être envisagé pour le graphe complet de la trajectoire



Jeux | **Story** | Timeline | Visualisation | Chaine | Chaînes semblables | Configuration

story	keyword	number of chains
La présidence de la République a annoncé que l'Equipe de FRrance de football serait reçue à l'Elysée lundi en fin d'après midi.	frrance	7
La 21ème édition de la Coupe du monde de football, organisée par la FIFA, se déroule en Russie du 14 juin au 15 juillet 2018.	fifa	6
Outre les directs sur notre page Facebook depuis la fan zone de Madiana (Schoelcher), un dispositif spécial est mis en place comme dans l'ensemble des stations du réseau Outremer 1ère.	1ère	5
célébration Coupe du monde 2018 : Emmanuel Macron exulte pendant la finale, les internautes réagissent	réagissent	5
La photo d'Emmanuel Macron les bras levés ce dimanche au stade Loujniki de Moscou lors de la finale de la coupe du Monde fait l'objet de nombreux détournements sur le web, sous le hashatg #PoseTonMacron.	hashatg	4
L'équipe de France de football, qui voyageait dans un avion en partance de Moscou et affrété spécialement pour l'occasion, a bien atterri à l'aéroport de Roissy-Charles-de-Gaulle (Seine-Saint-Denis) ce lundi quelques minutes avant 17 heures.	dirigeront	4
Ces images, censées représenter la ferveur des supporters croates, ont été diffusées des dizaines de milliers de fois sur les réseaux sociaux.	aussisujets	3
Les Bleus n'ont jamais autant eu autant la cote auprès du public, mais aussi auprès des clubs.	jt	3
Les Bleus ont remporté une finale totalement folle, 4-2.	accroche	2
Aux Etats-Unis, Trevor Noah, présentateur du "Daily Show", est au cœur d'une vive polémique après avoir comparé l'équipe de France à "l'équipe d'Afrique".	zapping	2
La photo d'Emmanuel Macron les bras levés ce dimanche au stade Loujniki de Moscou lors de la finale de la coupe du Monde fait l'objet de nombreux détournements sur le web, sous le hashatg #PoseTonMacron.	photo	2

FIGURE 5.12 – Liste des récits pour le jeu de données *mondial*.

Seafood and More at a Croton Falls Spot	mussels	4
The New Boar on the Butchers' Block	syrupy	3
McDonald's, stung by a loss, summons a high-level meeting to rework its marketing.	omnicom	3
From a Punjabi Hand, Lavish Dishes	raita	3
Contemporary American's French Accent	enhanced	2
QUICK BITE/Jersey City; The Skinny on Fatburger	silbert	2
Seafood and More at a Croton Falls Spot	accepted	2
Surgery With a Side of Fries	hospitals	2
A Stir-Fry That Isn't	garnish	2
Not Fast Food. Good Food Fast.	glowing	2
KFC and Royal Philips Electronics surprise their longtime agencies with dismissals.	ddb	2
A Milk Shake and Fries With Your Surgery?	profitable	2
McDonald's New Recipe Lowers Goo For Arteries	saturated	2
'Treasure Island' Flies Into Neurosis	comical	2
A South-of-the-Border Menu, Minus Tacos	primarily	2

FIGURE 5.13 – Liste des récits pour le jeu de données *fries*.

pour des raisons de lisibilité et de performance du rendu. Le graphe affiché peut être manipulé et les chaînes peuvent être explorées à l'aide du descriptif s'affichant dans la partie droite du graphe. Deux exemples de récits sont fournis en Figures 5.14 et

5.15.

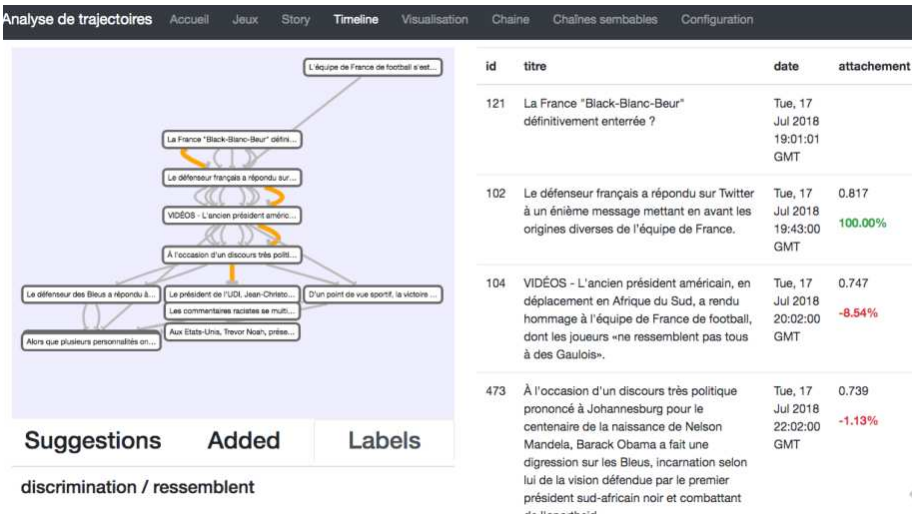


FIGURE 5.14 – Récit sur la discrimination de l'origine des joueurs français.

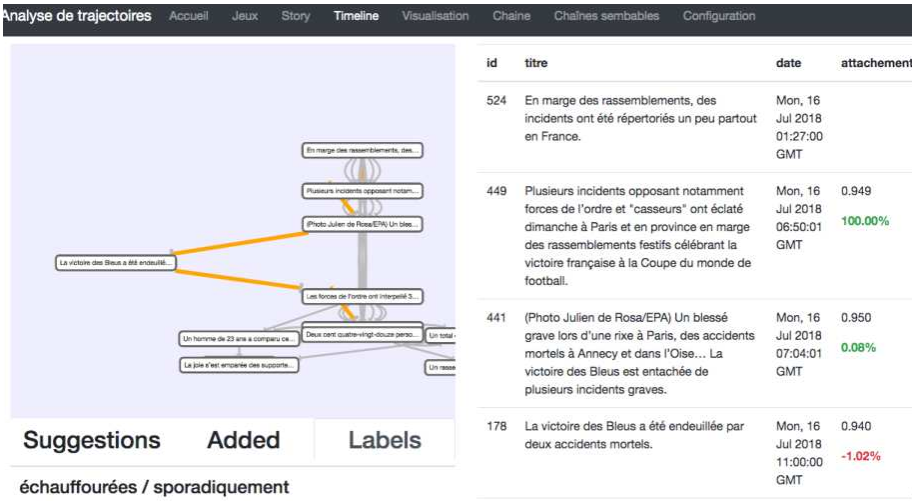


FIGURE 5.15 – Récit sur les échauffourées en marge des célébrations de la victoire française.

Il est également possible d'assembler plusieurs récits différents qui s'intersectent en certains documents. Sous le graphe du récit, se trouvent les trois sections **Suggestions**, **Added** et **Labels**. La section **Labels** est affichée par défaut et présente les différentes informations identifiées pour la chaîne actuellement sélectionnée. La section **Added** présente l'intégralité des objets actuellement affichés dans le graphe, que ce soit des chaînes ou des récits. Il est possible d'afficher en surbrillance les différents éléments qui correspondent à l'objet survolé par la souris, et de supprimer certains objets pour simplifier l'affichage de la Timeline. Une illustration est donnée

en Figure 5.16.

FIGURE 5.16 – Différents objets actuellement présents dans une Timeline. Section **Added** de la Figure 5.15.

En sélectionnant un document de la *Timeline* à l'aide d'un clic souris, on peut voir l'intégralité des autres récits qui passent par ce document dans la section **Suggestions** et les ajouter à la visualisation actuelle (cf. Figure 5.17). Ainsi il est possible de visualiser comment deux récits se superposent et interagissent, comme illustré en Figure 5.18. Ici il s'agit de deux récits similaires, l'un parle des violences en général dans toute la France tandis que l'autre relate un épisode particulièrement violent dans l'agglomération de Grenoble.

L'onglet *Timeline* permet à un utilisateur de l'outil de visualisation de partir d'un récit d'entrée, puis ensuite d'explorer de proche en proche les différentes chaînes pour découvrir les principales informations du corpus à étudier. L'utilisateur est également capable d'enlever les éléments qui lui semblent superflus et est libre de

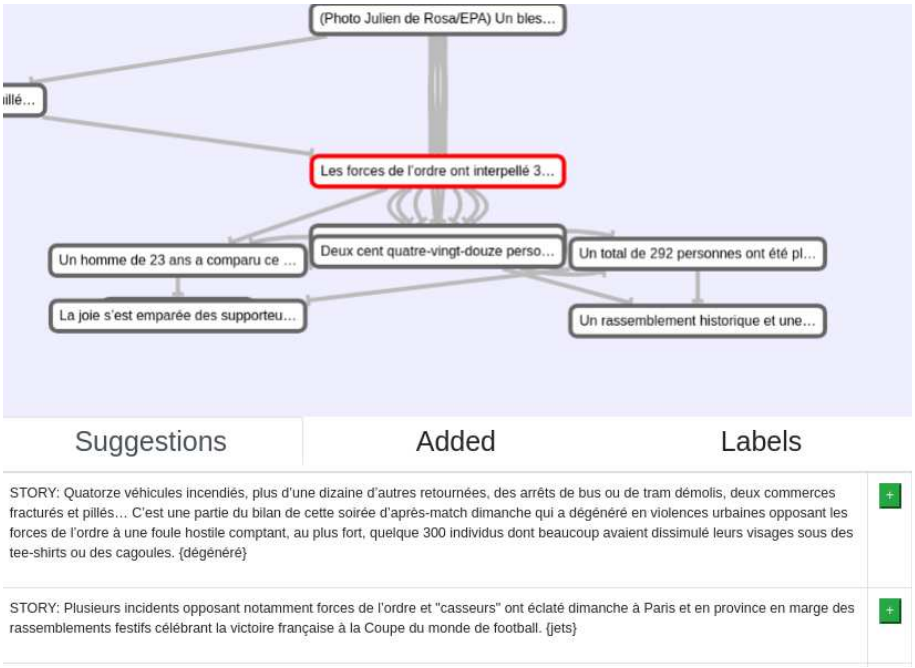


FIGURE 5.17 – Les différents récits supplémentaires passant par le document sélectionné (titre cerclé de rouge). Section **Suggestions** de la Figure 5.15.

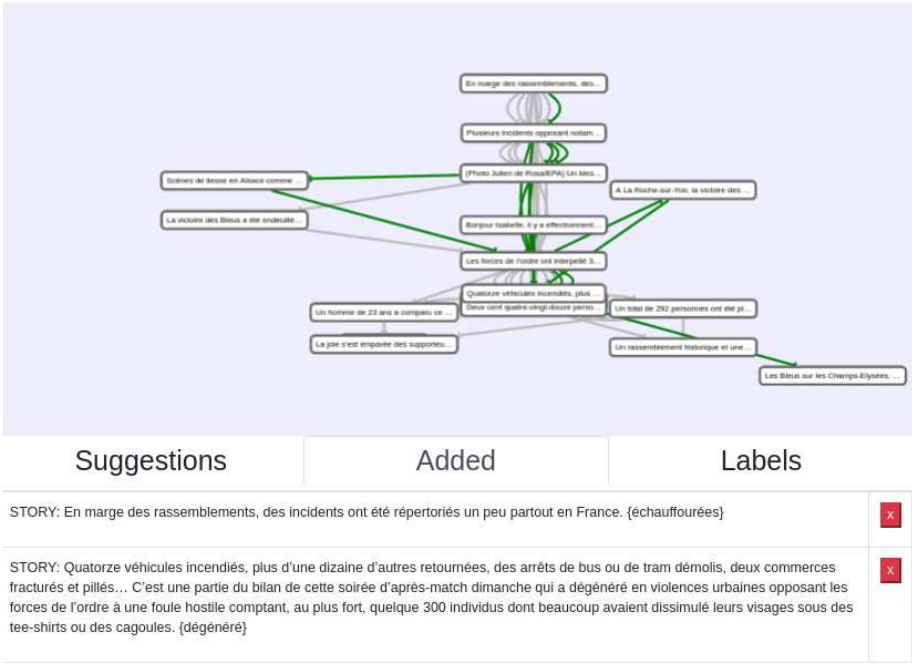


FIGURE 5.18 – Superposition de deux récits du jeu de données *mondial*. Les arcs gris correspondent au récit de violence dans tout le pays, tandis que les arcs verts correspondent à un épisode de violence particulier dans l'agglomération de Grenoble.

construire une visualisation locale des éléments qui l'intéressent et leurs interactions.

La plateforme de visualisation de la trajectoire que nous avons développée nous a permis d'explorer plusieurs représentations et modes d'interaction dans une optique

de test et de validation auprès des veilleurs. Ce travail n'est qu'une première esquisse des possibilités qui méritent d'être approfondies avant d'aboutir à des outils d'analyse intuitifs et utilisables par des veilleurs. Elle nous permet néanmoins d'exhiber l'objet qu'est la trajectoire et d'en présenter ses nombreuses manières d'être étudiées. La visualisation, la navigation et l'exploitation de la trajectoire constituent un sujet nouveau et très ouvert. Dans la section suivante, nous proposons une étude de corpus à l'aide de la trajectoire et des différents outils que nous construisons autour pour illustrer leurs capacités.

5.2 Étude de corpus

Dans cette section, nous nous mettons à la place d'un veilleur qui vient de collecter un corpus conséquent de documents sur un sujet en particulier et qui souhaite en apprendre plus sur ce sujet et sur le contenu du corpus qu'il a constitué. Nous avons mené une étude d'un corpus francophone constitué à l'aide de la plateforme de veille AMIEI. La plateforme permet de collecter des articles de presse et de blogs issus de plusieurs sources internationales et francophones. Le but étant d'illustrer l'utilisation de la trajectoire, nous commençons par présenter, sans rentrer dans les détails, le corpus, puis nous exhibons les différentes informations que nous tirons de la trajectoire sur le sujet d'une part, et sur le corpus d'autre part.

5.2.1 Corpus *mondial*, coupe du monde 2018

Le 10 juillet 2018 à Saint-Pétersbourg, l'équipe de France de football se qualifie pour la finale de la coupe du monde de football face à son homologue belge. Elle rencontrera pour la finale, le 15 juillet à Moscou, l'équipe de Croatie, qualifiée le 11 du même mois après une victoire contre l'équipe anglaise. La finale est remportée par l'équipe de France sur un score final de 4 buts à 2.

Nous avons collecté sur la période allant du 10 au 19 juillet 2018 les documents francophones disponibles via *AMIEI* présentant un ou plusieurs des termes suivants :

« mondial, foot, macron, france, russie, croatie, poutine, monde, match ». Cela constitue un ensemble de 748 documents sur lequel nous voulons travailler.

Nous calculons une trajectoire en utilisant *Doc2Vec* comme mesure de similarité entre les documents. Cette mesure semblait être la mieux adaptée selon les résultats des expérimentations au chapitre 3 (cf. section 3.4.1). Nous paramétrons le critère de faible cohérence à 20%, le critère de cohérence à 70%. Cela nous permet d'obtenir des chaînes fortement similaires en un temps raisonnable. Nous limitons également la quantité de chaînes finissant en un document à 20. Cela permet de clarifier la visualisation au cas où certains documents seraient très connectés et généreraient beaucoup d'arcs sur les graphes. La fonction de cohérence est la moyenne arithmétique de la similarité des différentes paires de documents de la chaîne. Nous filtrons également les sous-chaînes redondantes comme décrit au chapitre 4. Au final, la trajectoire obtenue contient 880 chaînes.

Visualisation

En premier lieu, on peut commencer par se faire une idée de la structure générale de la trajectoire. La visualisation est donnée en Figure 5.19. On constate une super composante qui a l'air de couvrir la majorité des documents, trois composantes plus petites qui ont l'air fortement connexes et une dizaine de chaînes isolées ou peu connectées.

On peut commencer par examiner les petites composantes fortement connexes. La plus volumineuse est composée uniquement de dépêches provenant de la section football du site *rmcsport.bfmtv.com*. Si le corps des articles parle effectivement de la coupe du monde de football, ils contiennent tous du contenu générique concernant les *CGU* et la politique de *cookie* du site web (cf. Figure 5.20). Ce fragment de texte qui existe à l'identique dans tous les documents forme des chaînes hautement similaires (à raison) et a saturé la quantité limite de chaînes finissant en un document. Ceci explique que cette composante soit détachée de la super composante. Les deux autres petites composantes ont exactement le même comportement, l'une contient

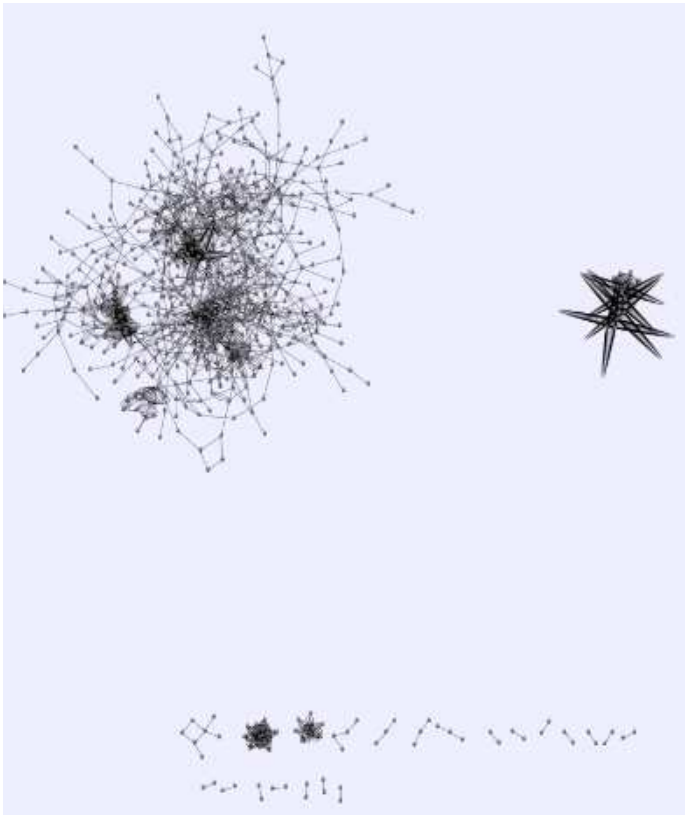


FIGURE 5.19 – La trajectoire calculée pour *mondial*.

des dépêches de *www.tf1.fr*, l'autre de *www.france24.com* toutes deux extraites avec du contenu générique qui a saturé les chaînes y passant. On constate que la trajectoire telle que nous l'avons paramétrée permet rapidement de constater ce genre de problèmes de données génériques non nettoyées en isolant les sous-corpus en question. Il conviendrait ici pour un veilleur de noter le problème et d'effectuer un nettoyage des documents incriminés avant de recalculer la trajectoire. Ici, nous nous contentons d'ignorer ces documents et d'étudier le reste de la trajectoire telle quelle.

URL: *https://rmcsport.bfmtv.com/football/equipe-de-france-griezmann-remet-en-place-courtois-apres-ses-critiques-sur-les-bleus-1488998.html*

pEn poursuivant votre navigation sur ce site, vous acceptez nos CGU et l'utilisation de cookies afin de réaliser des statistiques d'audiences et vous proposer une navigation optimale, la possibilité de partager des contenus sur des réseaux sociaux ainsi que des services et offres adaptés à vos centres d'intérêts. Pour en savoir plus et paramétrer les cookies... Menu Recherche Direct TV Direct Radio Connexion Se connecter Mot de passe oublié Pas encore de compte ? Inscrivez-vous ! Sport Newsletters Sport Inscrivez-vous gratuitement ! Suivez-nous sur "La meilleure Coupe du monde de tous les temps", s'enflamme Infantino pour Russie 2018 Equipe de France: Griezmann remet en place Courtois après ses critiques sur les Bleus 13/07/2018 à 12h25 Antoine Griezmann - AFP × OK L'après-match a été tendu dans les rangs belges mardi après la défaite en demi-finale de Coupe du monde face à la France (1-0).

FIGURE 5.20 – Un des textes de la composante des documents de *rmcsport.bfmtv.com*. L'information propagée identifiée correspond clairement au message générique des CGU du site web.

Parcours des récits

S'il est possible d'explorer les chaînes depuis la vue macroscopique en sélectionnant des chaînes au hasard, il est plus efficace d'aller parcourir la liste des récits calculés pour la trajectoire. Les récits sont des regroupements d'informations très similaires. Pour ce corpus il y a un total de 1296 récits calculés. Les récits qui impliquent le plus de chaînes font tous référence aux composantes de contenu générique que nous venons d'observer. La Figure 5.21 présente deux segments de la liste des récits. La liste des récits nous semble fournir une bon point d'entrée pour observer les différentes informations de la trajectoire++. On apprend qu'il y ainsi 8 chaînes qui parlent de *discrimination*, ou 10 qui parlent d'*échauffourées*. Certaines parlent de la *stratégie* de l'équipe de France. En survolant cette section, on commence à voir apparaître des évènements relatifs au corpus. Il y a eu vraisemblablement des propos discriminant sur l'origine des joueurs de l'équipe de France. Il y a eu également des violences en marge du match. Nous proposons d'étudier plus en détail le récit sur les questions de discrimination pour illustrer un récit contenant plusieurs chaînes, et le récit concernant la stratégie pour illustrer un récit contenant peu de chaînes.

Étude du récit *discrimination*

La visualisation du récit de *discrimination* est donné en Figure 5.22. Elle est composée de 8 chaînes, un exemple de chaîne est donné en Figure 5.23. Toutes les chaînes passent par les deux documents centraux titrés *Le défenseur français a répondu sur Twitter à un énième message mettant en avant les origines diverses de l'équipe de France* et *VIDÉOS - L'ancien président américain, en déplacement en Afrique du Sud, a rendu hommage à l'équipe de France de football, dont les joueurs «ne ressemblent pas tous à des Gaulois»*. Les deux documents centraux sont construits de la même manière. Ils commencent par présenter la prise de parole d'une personnalité sur l'origine des joueurs français. Le premier document présente le défenseur français Benjamin Mendy qui réagit à un tweet. Le second présente l'ancien président américain Barack Obama qui salue la diversité des joueurs de

POLITIQUE - Une nette majorité de Français se disent "optimistes" concernant leur avenir après la victoire des Bleus en finale de la Coupe du monde de football , dont ils estiment qu'elle aura un "impact positif sur la fierté des Français".	matins	10
En marge des rassemblements, des incidents ont été répertoriés un peu partout en France.	échauffourées	10
Le défenseur français a répondu sur Twitter à un énième message mettant en avant les origines diverses de l'équipe de France.	discrimination	8
Foot - Equipe de France	g	8
Le défenseur français a répondu sur Twitter à un énième message mettant en avant les origines diverses de l'équipe de France.	ressemblent	8
Emmanuel Macron célèbre le premier but français dimanche 15 juillet à Moscou.	emmanuel	7
célébration Coupe du monde 2018 : Emmanuel Macron exulte pendant la finale, les internautes réagissent	réagissent	7
32,5 millions d'euros, c'est la prime que va toucher la Fédération française de football (FFF) de la part de la Fédération internationale (Fifa) pour sa victoire dans la Coupe du monde 2018.	2014	7
Après un voyage dans les étoiles, c'est l'heure de l'embarquement pour Paris pour les supporters français qui ont fait le déplacement à Moscou (Russie) pour voir le sacre des Bleus.	aussisujets	7

(a) Exemple de récits de *mondial* contenant plusieurs chaînes.

Les commentaires racistes se multiplient en Pologne.	africains	2
Deux Brésiliens, deux Anglais, deux Croates, un Belge et quatre Français.	tripier	2
32,5 millions d'euros, c'est la prime que va toucher la Fédération française de football (FFF) de la part de la Fédération internationale (Fifa) pour sa victoire dans la Coupe du monde 2018.	enfants	2
célébration Coupe du monde 2018 : Emmanuel Macron exulte pendant la finale, les internautes réagissent	grandioses	2
À la veille de la finale du Mondial 2018 face à la Croatie, dimanche 15 juillet au stade Loujniki de Moscou (17h00 en France, 18h00 locales, en direct sur TF1), le capitaine et gardien de l'Équipe de France Hugo Lloris a répondu aux questions de la presse internationale.	encadrement	2
Thibaut Courtois (Belgique) : question fair-play, le portier belge n'aura pas été le meilleur, avec ses déclarations amères, pour être élégant, après la demi-finale face à la France.	matches	2
Après un mois de compétition acharnée, le Mondial 2018 est terminé et il est désormais l'heure des bilans pour les consultants.	type	2
Mondial russe; supporters de football présents en Russie à savoir	poutine	2
Notre journaliste Fanny Conquy a fait une sélection des images et vidéos qui circulent sur le web.	circulent	2
Adil Rami; équipe de France seront reçus par Emmanuel Macron	elysée	2
Pour se hisser en finale, les Bleus ont laissé le ballon aux Belges et misé sur le talent et leur discipline tactique.	stratégie	2

(b) Exemple de récits de *mondial* contenant deux chaînes.

FIGURE 5.21 – Deux segments de la liste des récits pour *mondial*. Chaque récit est décrit par un titre, un descripteur (il arrive que plusieurs récits aient le même titre), et la quantité de chaînes impliquées dans le récit.

l'équipe française lors d'un discours en Afrique du Sud. Ensuite, chaque document présente une mise en contexte des différentes prises de paroles sur les origines de l'équipe de France. Ainsi, on apprend que de nombreuses personnalités et joueurs se sont exprimés sur la diversité de l'équipe de France, comme le président vénézuélien Nicolas Maduro, le présentateur vedette américain Trevor Noah, le vice président du parlement iranien Ali Motahari, l'ancien ministre français Azouz Begag, et plusieurs joueurs comme Paul Pogba, Antoine Griezmann ou l'entraîneur Didier Deschamps.

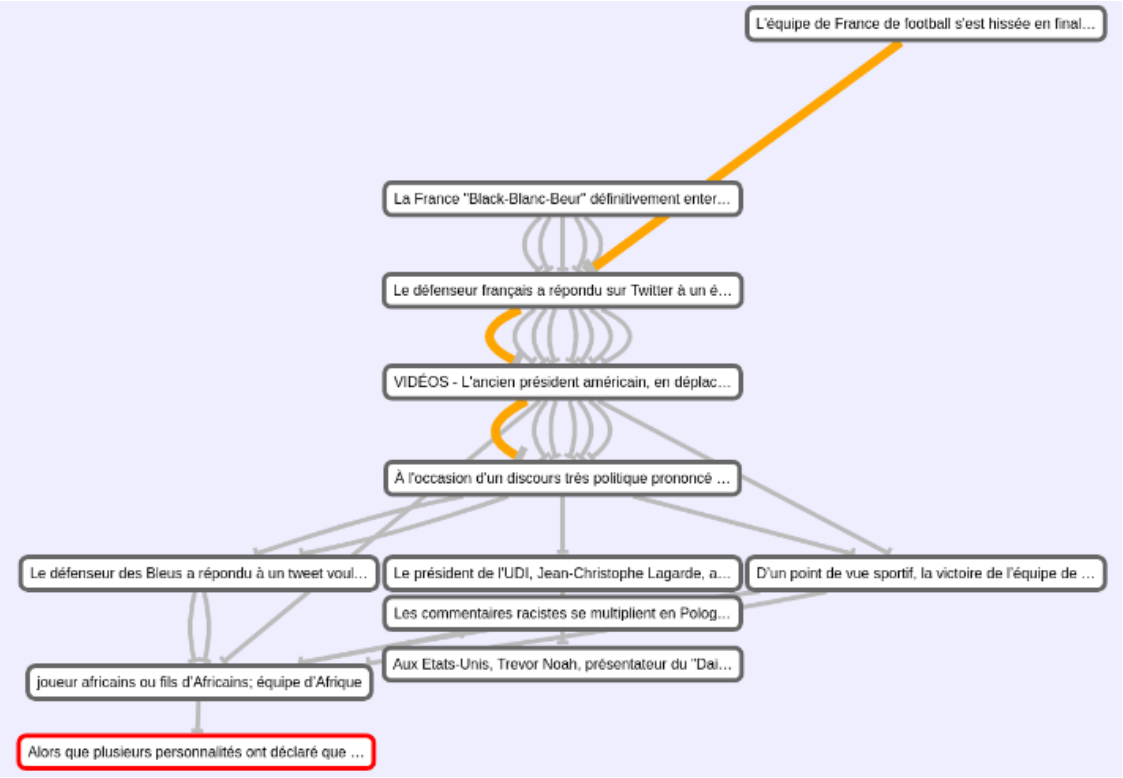


FIGURE 5.22 – Visualisation des chaînes contenues dans le récit *discrimination*.

titre	date
La France "Black-Blanc-Beur" définitivement enterrée ?	Tue, 17 Jul 2018 19:01:01 GMT
Le défenseur français a répondu sur Twitter à un énième message mettant en avant les origines diverses de l'équipe de France.	Tue, 17 Jul 2018 19:43:00 GMT
VIDÉOS - L'ancien président américain, en déplacement en Afrique du Sud, a rendu hommage à l'équipe de France de football, dont les joueurs «ne ressemblent pas tous à des Gaulois».	Tue, 17 Jul 2018 20:02:00 GMT
À l'occasion d'un discours très politique prononcé à Johannesburg pour le centenaire de la naissance de Nelson Mandela, Barack Obama a fait une digression sur les Bleus, incarnation selon lui de la vision défendue par le premier président sud-africain noir et combattant de l'apartheid.	Tue, 17 Jul 2018 22:02:00 GMT
Le défenseur des Bleus a répondu à un tweet voulant vanter la diversité des origines des joueurs de l'équipe de France.	Wed, 18 Jul 2018 00:00:00 GMT

FIGURE 5.23 – Exemple de chaîne du récit *discrimination*, titre et date des documents.

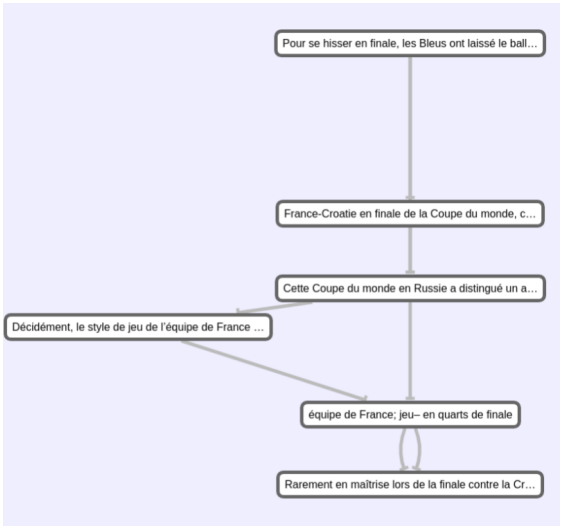
En lisant le document le plus ancien du récit, on peut remonter l'origine de ce phénomène à une publication *Facebook* depuis retirée d'Igor Stimac, ancien joueur international croate, qui affiche un trombinoscope de l'équipe de France où chaque photo est accompagnée du drapeau de leur pays d'origine ou du pays de leurs parents. Cette présentation avait déjà été relayée, une semaine avant Igor Stimac, par David Campese ancien international de rugby sud-africain. L'image a été reprise par le site humoristique *www.sporf.com*. Les autres documents du récit soulignent que la victoire de l'équipe de France a également eu un afflux important de contenus haineux sur les réseaux sociaux.

Ce récit contient donc principalement des articles compilant les discours sur l'origine des joueurs, qu'ils soient critiques de la composition de l'équipe de France ou qu'ils l'encensent. Il s'agit d'un corpus riche pour comprendre un évènement périphérique de la coupe du monde de football 2018. Le corpus est de taille moyenne (douze documents), dont les documents principaux et périphériques sont directement visibles grâce à la structure de chaîne.

Étude du récit *stratégie*

La visualisation du récit de *stratégie* est donnée en Figure 5.24. On constate tout de suite qu'il s'agit d'un récit contenant uniquement deux chaînes, de plus la deuxième chaîne est presque identique à la première avec un document supplémentaire en milieu de chaîne. À la lecture des titres, il semble bien s'agir de documents fournissant une analyse sur la stratégie du jeu de l'équipe de France. La stratégie en football est un domaine technique qui peut être compliqué à appréhender pour un non initié. On peut commencer à regarder l'information extraite et la séquence de phrase qui l'étiquette (cf. Figure 5.25). À l'aide des titres et des phrases extraites, on comprend qu'il s'agit bien d'analyses techniques du jeu de l'équipe de France sans avoir à rentrer dans le corps des documents. Ce sont des articles de fond qui sont plus longs que les articles du récit *discrimination*.

Ce récit contient deux bonnes chaînes cohérentes sur la stratégie du football,



(a) Chaînes du récit *Stratégie*.

titre
Pour se hisser en finale, les Bleus ont laissé le ballon aux Belges et misé sur le talent et leur discipline tactique.
Cette Coupe du monde en Russie a distingué un autre style de football : plus pragmatique et moins axé sur la possession.
Décidément, le style de jeu de l'équipe de France ne plait pas à ses adversaires pendant cette Coupe du monde.
équipe de France; jeu– en quarts de finale
Rarement en maîtrise lors de la finale contre la Croatie, les Bleus ont fait la différence sur des inspirations individuelles.

(b) Quelques titres des documents du récit *Stratégie*.

FIGURE 5.24 – Visualisation des chaînes contenues dans le récit *stratégie*.

stratégie / balle

- Mais en laissant la Belgique faire tourner le ballon et augmenter ses statistiques dans le vide (64 % de possession et 21 centres, mais seulement 9 tirs à 19 et 42 % de duels gagnés par les Diables...
- Je pense que c'est un match qui s'achèvera par un score de 1-1 au terme des prolongations Et qu'il y aura des tirs au but… Dans cet exercice, les Balkaniques ont prouvé leur savoir faire face au...
- Et au final, c'est bien le pays hôte qui s'est qualifié à l'issue de la séance des tirs au but.
-] ce qui en fait un élément crucial de leur jeu», analyse Jorge Valdano.
- Dans cette finale, on a d'abord vu la difficulté de l'équipe de France à gérer un 4-3-3 avec trois milieux

FIGURE 5.25 – Information extraite de la chaîne du récit *stratégie* et la séquence de phrases qui l'étiquette.

mais un veilleur pourrait vouloir extraire un récit plus large sur la question de la stratégie. Pour cela, on peut regarder les récits qui croisent celui-ci. Le document titré *Cette Coupe du monde en Russie a distingué un autre style de football : plus pragmatique et moins axé sur la possession* est structurellement au centre de la chaîne, de plus son titre semble clairement indiquer un commentaire technique, on peut étudier les récits qui passent par ce document. La liste de ces récits est donnée en Figure 5.26. A première vue, ces récits ont également l'air de parler de stratégie, la première raison est qu'ils sont en relation avec un document dont on sait qu'il

est technique et stratégique, la seconde est que leur descripteur correspond à du vocabulaire technique : tirs, dominer, joué, ligue. Le dernier récit a un descripteur non technique (Allemagne) mais son titre prouve qu'il contient du contenu technique. L'ajout de tous ces récits construit une trajectoire sur 24 documents différents constituée de 15 chaînes. La lecture des différents titres montre que les analyses portent principalement sur une stratégie peu orthodoxe de l'équipe de France qui consiste à délaisser la possession de balle.

L'outil d'analyse de récit nous a permis, à partir d'un récit contenant peu de chaînes et portant sur des documents techniques dont nous n'avons pas la maîtrise, de constituer un corpus de taille moyenne conséquent d'analyse stratégique. Un des rôles du veilleur est de pouvoir fournir les documents pertinents aux personnes compétentes pour les analyser. Sous cet angle, le travail sur ce récit, bien que simple, est pertinent.

FIGURE 5.26 – Autres récits passant par le document titré *Cette Coupe du monde en Russie a distingué un autre style de football : plus pragmatique et moins axé sur la possession.*

Conclusion

Cette étude du corpus *mondial* a permis de mettre en évidence plusieurs qualités des outils que nous développons autour de la trajectoire. D'abord, elle nous permet d'avoir une vision macroscopique structurelle rapidement, et ainsi détecter des anomalies dans le corpus. Ensuite, notre approche par récit permet à un utilisateur de **découvrir** les diverses histoires du corpus, même lorsqu'elles lui sont inconnues

comme les échauffourées ou la polémique sur les origines des joueurs, ou lorsqu'elles impliquent une petite quantité précise de documents, comme les articles d'analyse stratégique du corpus. Le récit contient généralement peu de chaînes et permet donc une analyse à taille humaine des corpus, quitte à explorer les documents qui sont proches structurellement. Une limitation est cependant que les récits sont souvent très nombreux, il est ainsi difficile d'avoir une vision exhaustive de ce qui se déroule au sein du corpus.

5.3 Idées d'application de la trajectoire

La trajectoire de l'information est une proposition originale pour aborder la propagation de l'information dans les médias sociaux. Nous avons d'abord soumis et justifié l'idée de la trajectoire, puis nous avons développé une méthode pour calculer des trajectoires ainsi qu'une autre pour identifier et étiqueter les informations qui se propagent le long des différentes chaînes. Nous avons également construit une maquette de visualisation qui permet déjà de faciliter l'analyse d'un corpus. Nous avons illustré la richesse brute de cette structure lors d'une étude de corpus. Dès le départ, la trajectoire a été pensée comme une représentation intermédiaire de la propagation à partir de laquelle on pourrait développer diverses analyses poussées du corpus. Nous présentons maintenant plusieurs analyses de corpus que nous envisageons à partir de la trajectoire. Ces analyses sont essentiellement des réflexions et sont autant d'ouvertures pour prolonger les travaux présentés dans ce manuscrit.

Analyse des mutations de l'information

Dans le chapitre 4, nous identifions les différentes informations présentes sur une chaîne à l'aide d'ensemble de descripteurs que nous appelons signatures des informations. La signature représente l'information à travers toute la chaîne. Il s'agit donc d'une représentation de la partie immuable de cette information le long de la chaîne. Elle ne s'intéresse pas aux spécificités de cette information lorsqu'elle passe d'un document de la chaîne à un autre. En d'autres mots, cette représentation ne

considère pas **la mutation** le long de la chaîne. Il serait intéressant de justement étudier la partie mutante d'une information le long d'une chaîne. Pour ce faire, on pourrait étudier les fluctuations du TF-IDF, ou tout autre mesure d'importance, des descripteurs le long de la chaîne. Ceci formerait une signature de la composante mutante de l'information. De la même manière que pour la composante immuable, réussir à étiqueter la composante mutante est une piste d'analyse de la nuance le long des chaînes.

Dans un second temps, on peut envisager d'utiliser cette fluctuation des informations le long des chaînes pour construire une classification des informations stables, divergentes ou fluctuantes. Au niveau d'un récit, qui est un cluster d'informations similaires, cette classification pourrait permettre de distinguer si les éléments de langage propre à ce récit sont figés, s'ils ont une évolution temporelle, ou encore s'ils varient selon certains regroupement d'auteurs. Ce sujet rejoint l'étude de la manière dont les informations cohabitent le long des chaînes. La fluctuation de certaines informations est-elle corrélée à la fluctuation de certaines autres, et si oui pourquoi ? Myers et Leskovec ont introduit le sujet de la compétition et la coopération des informations dans le cadre de la propagation [MyersLeskovec, 2012], et une étude à travers l'approche des trajectoires serait un angle original.

Lors du positionnement (cf. section 2.3), nous avions souligné le manque de travaux sur l'étude de la mutation dans la propagation d'informations. Il s'agit d'une des motivations pour la construction de la trajectoire, aussi nous la pensons particulièrement adaptée pour ce genre d'études.

Résumé de la propagation d'information

La trajectoire telle que nous la construisons est un ensemble de chaînes conséquent : elle contient *a priori* de nombreuses chaînes pour un corpus de documents non trivial. Par conséquent, elle n'est pas adaptée pour être par elle-même un point d'entrée simple et intuitif d'un corpus pour un veilleur ou un analyste. Avoir une vue claire et simple des différents axes du corpus est un sujet de recherche dont nous avons

déjà discuté dans l'état de l'art (cf. section 2.2.5). La carte de métro est un exemple d'une telle vue (cf. Figure 5.27), elle présente les principaux événements du corpus et la manière dont ils sont reliés. Nous pensons qu'il est possible également de produire une représentation similaire à partir des trajectoires que nous calculons. On peut envisager deux types de représentation, selon que l'on souhaite présenter les documents ou les informations.

La première représentation résumée de la trajectoire consiste à trouver les chaînes qui forment son squelette. Ce sont les chaînes qui représentent au mieux sémantiquement et structurellement l'ensemble des chaînes de la trajectoire. Ce squelette forme une trajectoire réduite, simple à naviguer et explorer pour un veilleur qui souhaiterait voir les principaux documents du corpus et leur articulation.

La seconde représentation résumée présenterait, à la manière d'une carte de métro, les principales informations extraites de la trajectoire et la manière dont elles s'articulent. Il reste à mener la réflexion sur la manière dont on distingue les informations principales des autres, et sur la manière dont on passe d'une structure sur les documents à une structure sur les informations. Dans les deux cas, nous suspectons que la mesure de l'ICF [2] d'un descripteur prend ici son sens, parce qu'on souhaite que les informations fortement indépendantes soient bien représentées.

Motifs des données satellites dans la trajectoire

Pour finir, nous présentons un axe de réflexion légèrement différent des deux précédents qui étaient en droite lignée du travail présenté jusqu'alors. Une chaîne a toujours été définie comme une séquence de documents. Par exemple, la chaîne $d_1d_2d_3$ représente la séquence des documents d_1, d_2 et d_3. Dans ce manuscrit, nous nous sommes principalement intéressés au contenu textuel des documents : leur corps et leur titre. Un document contient pourtant d'autres données, notamment une date de publication et un ou plusieurs auteurs. Dans le cas des médias sociaux, le document provient généralement d'un site dont le domaine est connu : *www.twitter.com*,

2. (*Inverse Chain Frequency*, l'inverse du nombre de chaîne dans lequel le descripteur apparaît, cf. section 4.2)

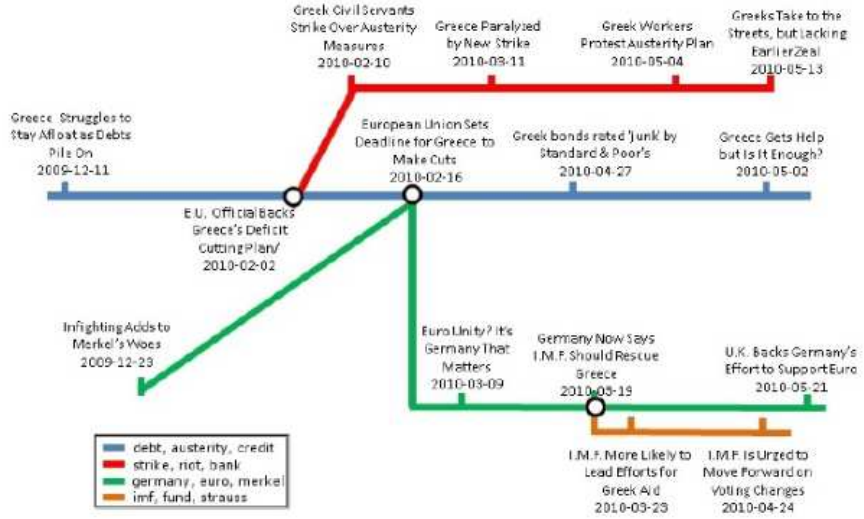


FIGURE 5.27 – Carte de métro construite par [Shahaf, 2012] au sujet de la crise de la dette grecque.

www.lemonde.fr, *etc.* Les chaînes de documents produisent des motifs sur ces données qu'il peut être intéressant d'étudier. Si par exemple tous les documents d'une chaîne sont écrits par la même personne, ou sur le même domaine, ou à des dates régulièrement espacées, il s'agit de données pertinentes pour l'analyse de cette chaîne.

Ainsi l'analyse des motifs induits par les chaînes sur les autres données des documents peut donner lieu à une modélisation de ces motifs et une prévision de leur apparition future. Par exemple, si la séquence d'auteurs A, B, C apparaît fréquemment dans les chaînes de la trajectoire, cela semble indiquer que si la séquence A, B réapparaît à l'avenir, il est crédible d'attendre une publication de C qui constituera une nouvelle séquence A, B, C. Cet aspect bayésien est autant une piste pour prédire les futures chaînes que pour apprendre, à partir d'une trajectoire soigneusement construite sur un corpus d'entraînement, la trajectoire sur un corpus plus important.

Les chapitres précédents se focalisaient sur la raison d'être et la manière de calculer la trajectoire de l'information. Ceci fait, il restait à savoir comment exploiter au mieux le contenu de la trajectoire calculée. Nous avons présenté dans ce chapitre un outil pour visualiser et explorer les différentes chaînes de la trajectoire. Il permet de visualiser la trajectoire dans son ensemble, chacune des chaînes à part, les relations sémantiques entre les chaînes, et permet également de visualiser et naviguer dans

les documents à l'aide des informations identifiées selon les méthodes du Chapitre 4, que nous avons regroupées sous l'appellation de *récits*. Nous avons ensuite illustré la capacité de la trajectoire à orienter un analyste par une étude de corpus. Enfin, nous avons ouvert sur plusieurs pistes applicatives qui suivent naturellement les travaux de ce manuscrit : l'analyse de la mutation le long des chaînes, la représentation synthétique du corpus en résumant la trajectoire, ou encore l'exploitation des motifs exhibés par les chaînes pour construire de meilleures trajectoires ou construire des modèles prédictifs. La trajectoire étant une proposition originale de ce manuscrit, il reste de nombreuses interrogations sur son efficacité pour les divers besoins de la recherche et du développement autour du sujet de la propagation d'informations dans les médias sociaux.

Chapitre 6

Conclusion et perspectives

Au cours de cette thèse, nous avons mené diverses réflexions et proposé une approche originale pour l'étude de la propagation des informations dans les corpus textuels.

Nous avons commencé par constater, lors du chapitre 2, que l'information se propage dans les médias sociaux. Les individus consultent des documents et publient de nouveaux documents en s'inspirant ou en réagissant aux documents précédemment consultés. Lors de cette propagation, l'information mute, changeant de contexte et de sens. En établissant un état de l'art de la recherche autour du phénomène de propagation, nous avons constaté que le phénomène de mutation était rarement traité conjointement à celui de propagation. Nous avons alors mis en lumière l'intérêt d'unifier l'étude du phénomène de propagation et l'étude du phénomène de mutation. Ainsi, en s'attaquant à la généalogie des informations, on pourrait accéder à une compréhension plus fine et un modèle plus détaillé du phénomène de propagation. Cette unification implique en particulier que l'information qui circule ne peut pas être déterminée précisément avant d'en connaître le cheminement.

Pour cela, nous avons proposé dans le chapitre 3 le formalisme de la Trajectoire de l'information, qui consiste en un ensemble de chaînes chronologiques de documents le long desquelles de l'information s'est propagée. Il s'agit de la partie structurelle du phénomène de propagation, c'est-à-dire les chemins que prennent les informations. C'est le long de telles chaînes qu'il devient possible de déterminer l'information qui

se propage. Nous avons également proposé dans ce chapitre une approche permettant de calculer des trajectoires potentielles de l'information. Cette approche repose sur la notion de chaîne cohérente. Une chaîne est cohérente si elle semble être une chaîne de propagation pour un expert humain. En modélisant la cohérence d'une chaîne à partir des similarités qui existent deux à deux entre les documents du corpus, nous parvenons à construire un ensemble de chaînes cohérentes. Nous obtenons ainsi une approximation satisfaisante de l'historique de propagation de l'information. À l'aide d'une campagne d'évaluation, nous concluons qu'il est bien possible de capturer le jugement humain d'une chaîne cohérente et que nous parvenons effectivement à construire un ensemble de chaînes cohérentes.

Ensuite, lors du chapitre 4, nous proposons une approche pour identifier et nommer les différentes informations contenues dans les différentes chaînes d'une trajectoire. Pour ce faire, nous étendons la représentation TF-IDF, classique pour les documents, aux chaînes de documents, ce qui permet d'une part de décrire les chaînes, d'autre part de fournir un point d'entrée pour différencier les principales informations de la chaîne. Nous proposons une approche basée sur la notion de pertinence marginale maximale (MMR) pour identifier les principales informations sur chaque chaîne. Nous proposons également diverses méthodes pour étiqueter ces informations en tirant parti de la structure fournie par la chaîne qui la contient et la trajectoire. L'approche dans sa globalité est également soumise à une campagne d'évaluation qui nous permet de conclure que nous parvenons effectivement à capturer les fils sémantiques probables des chaînes.

À ce stade, nous sommes donc en mesure de construire une description à la fois structurelle et sémantique des informations qui se propagent dans un corpus. Aussi, lors du chapitre 5, nous proposons diverses manières d'exploiter cette description. La toute première passe par la visualisation de la trajectoire. Comme il s'agit d'une structure qu'il est intéressant de traiter autant à l'échelle microscopique (la chaîne elle-même comme un sous-corpus pertinent pour une ou plusieurs informations) qu'à l'échelle macroscopique (la manière dont l'ensemble des chaînes interagissent entre

elles), nous avons construit une maquette Web permettant de tester différentes façons de représenter la trajectoire. Il est ainsi possible d'explorer chaque chaîne indépendamment ou d'observer l'ensemble de celles-ci, avec leurs interactions structurelles ou sémantiques. Ces réflexions nous ont conduit à proposer une représentation supplémentaire au niveau des informations fortement similaires. Celles-ci se regroupent en ce que nous appelons un *récit*, de petits ensembles de chaînes pour lesquels on a identifié une information similaire. Ces récits permettent d'une part de détecter certaines thématiques faibles du corpus, d'autre part d'explorer la manière dont cette thématique a évolué au cours du temps dans le corpus. Pour bien illustrer cette capacité, nous avons procédé à une étude de cas sur un corpus francophone, duquel nous avons pu interactivement construire des sous-corpus traitant d'un sujet particulier dont nous n'avions *a priori* pas connaissance comme le racisme lors de la coupe du monde de football 2018 ou les articles techniques parlant de stratégie. Nous avons également proposé au chapitre 5 plusieurs perspectives d'exploitation de la trajectoire.

L'ensemble du travail de cette thèse fournit ainsi une approche globale, en partant du constat du problème de compatibilité entre le phénomène de mutation et le phénomène de propagation pour aboutir à la proposition du modèle de la Trajectoire de l'information. Ce travail inclut une méthode de calcul humainement évaluée pour sa structure d'une part et sa sémantique de l'autre. Ce travaille intègre l'aspect de l'exploitation pour détecter des informations et fournir des analyses de corpus. Nous avons ainsi constitué durant les quatre années de cette thèse une preuve de concept du modèle de la Trajectoire de l'information. Cette manière originale d'aborder la question de la propagation reste néanmoins à ses débuts, nous avons choisi de commencer par utiliser des approches simples qui pourront servir de référence pour des travaux futurs.

Nous avons ouvert la voie du problème du calcul des chaînes cohérentes en proposant une méthode efficace. Cette méthode pourra être affinée, par exemple en cherchant à régler le problème des chaînes cohérentes fortuites que nous développons

en Annexe C. Même si nous donnons un début de réponse à ce problème, il pourra être pertinent de construire, voire d'apprendre automatiquement, une fonction de cohérence plus sophistiquée afin de capturer des chaînes cohérentes plus subtiles et rejeter un maximum de chaînes fortuites. Nous nous sommes également appuyés sur une hypothèse raisonnable mais forte selon laquelle une chaîne cohérente est composée de sous-chaînes faiblement cohérentes et il sera intéressant de détecter et d'étudier les chaînes cohérentes ne vérifiant pas cette hypothèse. Cela nécessitera de réfléchir à une approche complémentaire pour construire la trajectoire.

Nous avons également ouvert la voie de l'identification et de l'étiquetage d'informations dans le contexte conjoint des phénomènes de mutation et de propagation de l'information. La chaîne de documents est une unité sémantique judicieuse pour laquelle nous avons proposé des méthodes de représentation construites à partir de méthodes classiques de représentation des documents. L'extension et l'étude des différentes méthodes de représentation de l'état de l'art, des documents aux chaînes, pourra être une étape importante vers une compréhension plus fine des chaînes. Nous avons proposé une méthode d'identification de l'information qu'il est possible d'interpréter comme un clustering des descripteurs de la chaîne. Nous avons montré que notre méthode basée sur un critère MMR est aussi efficace, moins contraignante et plus rapide qu'un K-means, une méthode de clustering éprouvée. Une comparaison de notre approche aux autres méthodes de clustering pourrait se révéler utile. Nous proposons également trois méthodes d'étiquetage de l'information extraite qui exploitent la connaissance structurelle supplémentaire fournie par la trajectoire, qu'il sera possible d'approfondir en tirant parti des divers travaux d'étiquetage de l'information.

Enfin, nous avons présenté plusieurs pistes d'exploitation de la trajectoire au chapitre 5. Il reste certainement de nombreuses possibilités pour utiliser cette représentation innovante de la propagation de l'information dans les médias sociaux.

Nous espérons que la lecture de ce manuscrit aura suscité chez vous suffisamment de questions pour qu'à votre tour vous propagiez, selon vos propres idées, une partie

des informations qu'il contient.

Annexes

Annexe A

Extension asynchrone du calcul de la Trajectoire

Dans le chapitre 3, nous construisons, étape après étape, une approche pour le calcul de trajectoires composées de chaînes cohérentes. L'approche finale à laquelle nous aboutissons se déroule de manière séquentielle : les documents du corpus sont traités les uns après les autres, dans l'ordre chronologique selon leur date de publication. Pour chaque document, nous calculons un ensemble de chaînes faiblement cohérentes qui finissent en ce document. La réunion de tous ces ensembles forment, après un filtrage pour ne garder que les chaînes cohérentes, le résultat de l'approche. On peut remarquer que le travail pour chaque document peut, dans une certaine mesure, s'effectuer indépendamment des autres, et donc se paralléliser. Pour effectuer le calcul de l'ensemble des chaînes cohérentes finissant en un document d il est nécessaire de connaître :

— Les documents d_i tels que la chaîne $d_i d$ est faiblement cohérente.

— L'ensemble des chaînes faiblement cohérentes finissant en chaque d_i.

En d'autres termes, pour qu'on puisse effectuer le travail sur le document d, il est nécessaire et suffisant d'avoir effectuer le travail sur les documents d_i, tels que les chaînes $d_i d$ sont faiblement cohérentes. On peut dire que le travail sur d dépend du travail sur un tel document d_i et, par transitivité, que le travail sur d dépend du

travail sur les documents dont dépendent le travail sur d_i. Lorsque les travaux pour deux documents ne sont pas en dépendance transitive l'un vis-à-vis de l'autre, ils sont indépendants, et il est donc possible de les effectuer en parallèle. Cela permet, sur une architecture matérielle adaptée, de réduire le temps de calcul de l'approche.

Il est nécessaire d'adapter la méthode de calcul pour profiter de cette amélioration parallèle. Pour cela nous proposons une version basée sur le modèle des agents initialement proposé par Yoav Shoham [Shoham, 1993]. Un agent est une entité capable d'effectuer séquentiellement une tâche qui lui est demandée. Chaque agent est capable d'envoyer et de recevoir des messages avec les autres agents, ce qui leur permet de communiquer. La réception d'un message est une action bloquante, ce qui signifie que l'agent en attente d'un message n'agira pas tant qu'il ne l'aura pas reçu. Le modèle des agents s'adapte particulièrement bien aux architectures parallèles, chaque agent peut être affecté à une unité de calcul indépendante et les agents ayant fini leur tâche ou en attente peuvent laisser les ressources de calcul aux agents prêts à agir.

Nous modélisons notre approche de la manière suivante :

— Un agent central qui s'occupe de créer le graphe de dépendance des agents, de créer les agents et de les démarrer. Une fois ceci fait il se met en attente pour récolter et réunir le travail des autres agents. Le code de cet agent est donné en Algorithme 10.

— Un agent est créé pour chaque document, cet agent est en attente des agents dont les résultats sont nécessaires pour accomplir son travail. Ce sont ses *agents prédecesseurs*. Lorsqu'il a finit de réaliser son travail, il envoie son résultat à l'agent central et à tous les agents qui dépendent de lui, ce sont ses *agents successeurs*. Le code d'un agent est donné dans l'Algorithme 11.

L'existence de ce modèle d'agent a un léger surcoût algorithmique lié à la gestion des agents actifs et en attente. De manière générale, lorsque le corpus de document est grand et composé de documents divers, il existe de multiples chaînes cohérentes qui n'ont aucun document en commun, ces chaînes peuvent être calculées de manière

indépendante et cela permet une amélioration du temps de calcul, parfois de plusieurs magnitudes selon le graphe de dépendance et l'architecture, ce qui compense ce surcoût du modèle d'agent.

Algorithme 10 : Procédure asynchrone de calcul de $T_{coh}(D, coh, \gamma)$.

Entrées : D, le corpus.
coh et γ, qui forment le critère de cohérence.
Sortie : $T_{coh}(D)$

`// Calcul du support de cohérence.`
$(D, E) \leftarrow G(D, coh, \gamma)$ $T \leftarrow \emptyset$
`// Création d'un agent pour réceptionner toutes les chaînes.`
$Agent(T) \leftarrow Nouvel\text{-}Agent()$
`// Création d'un agent pour chaque document.`
for $d \in D$ **do**
  $Agent(d) \leftarrow Nouvel\text{-}Agent()$
end
`// Pour chaque document, calcul de ses agents prédécesseurs et successeurs, puis mise en route de l'agent.`
for $d \in D$ **do**
  $Pred \leftarrow \{Agent(d') \in D, (d', d) \in E\}$
  $Succ \leftarrow \{Agent(d') \in D, (d, d') \in E\}$
  $Demarrer\text{-}Agent(Agent(d), d, coh, \gamma, Agent(T), Pred, Succ)$
end
`// On attend une réponse pour` $FinitEn(d)$ `de chaque agent document` $Agent(d)$.
for $d \in D$ **do**
  $T \leftarrow T \cup Attendre\text{-}Message(Agent(T))$
end
Retourner T

Algorithme 11 : Procédure interne des agents pour calculer $FinitEn(d, coh, \gamma)$.

Entrées : $Agent(d)$, l'agent actuel.
d, le document actuel.
coh et γ, qui forment le critère de cohérence.
$Agent(T)$, l'agent source à qui envoyer les résultats.
$Pred$, les agents dont le travail est un pré-requis.
$Succ$, les agents à prévenir par la suite.

Sortie : $FinitEn(d)$ est calculé puis envoyé aux agents qui en ont besoin.

$Finis \leftarrow \emptyset$
`// On attend que chaque prédécesseur termine et communique ses chaînes.`
for $p \in Pred$ **do**
  $Finis \leftarrow Finis \cup Attendre\text{-}Message(p)$
end
`// On calcule les chaînes qui finissent ici.`
$F \leftarrow FinitEn(d, coh, \gamma, Pred(d), Finis)$ `// On communique les chaînes aux successeurs.`
for $s \in Succ$ **do**
  $Envoyer\text{-}Message(Agent(d), s, F)$
end
`// On communique les chaînes à la procédure principale.`
$Envoyer\text{-}Message(Agent(d), Agent(T), F)$ `// On libère les ressources de l'agent.`
$Finir(Agent(d))$

Annexe B

Compression mémoire de la Trajectoire

Dans le chapitre 3, nous construisons une approche pour le calcul d'un ensemble de chaînes cohérentes. Ce qui guide cette construction est essentiellement une approche temporelle : nous souhaitons construire cet ensemble en un temps raisonnable. Nous ne discutons cependant pas la question de l'espace mémoire nécessaire pour stocker un ensemble de chaîne quelconque ou une trajectoire, et *a fortiori* la question d'une représentation adaptée à notre méthode de calcul.

Dans cette annexe, nous nous intéressons à deux méthodes pour représenter une trajectoire. Nous commençons par discuter de la représentation naïve d'une trajectoire et comment on peut tirer partie des propriétés de la trajectoire pour compresser cette représentation, à l'aide de ce que nous appelons les chaînes maximales. Enfin, nous proposons une représentation de la trajectoire adaptée à la méthode de calcul construite au chapitre 3. Selon les opérations qu'on souhaite appliquer aux chaînes, certaines représentations sont plus adaptées que d'autres. Aussi aux chapitres 4 et 5, on utilisera principalement une représentation par chaînes maximales, qui permet de parcourir simplement les différentes chaînes, sans considérer les sous-chaînes redondantes.

B.1 Représentation naïve et chaînes maximales

Pour représenter une chaîne de taille l, la solution naïve est de la représenter par un vecteur ou une liste chaînée de taille l, où chaque élément est un pointeur vers le document correspondant. Ce vecteur occupe un espace linéaire vis-à-vis de la chaîne : $Espace(d_1 \dots d_l) = O(l) \geq O(2)$. Ainsi, ce que nous appelons la **représentation naïve** de la trajectoire occupe la somme de l'espace occupé par ses chaînes :

$$Espace(T) = \sum_{c \in T} Espace(c) \geq O(2 \times |T|). \tag{B.1}$$

Cette représentation a l'avantage d'être exacte et direct, toutes les chaînes de la trajectoire sont accessibles à l'instant sans calcul supplémentaire. Les chaînes peuvent être stockées dans un vecteur ou un dictionnaire avec un index qui convient à l'application qu'on souhaite en faire. Par exemple, l'accès à l'ensemble chaînes qui finissent en un document d peut se faire en temps constant à l'aide d'une table de hachage.

Il est cependant possible de compresser cette représentation. L'idée est de tirer parti de la proposition 1 :

Proposition 1. Soit c une chaîne de propagation, alors toute sous-chaîne $c' \subset c$ est une chaîne de propagation.

Il y a une duplication d'informations lorsqu'on stocke à la fois une chaîne et une de ses sous-chaînes. Il n'est donc pas nécessaire de stocker toutes les chaînes pour représenter une trajectoire qui vérifie la proposition 1.

Définition 19 (Chaîne maximale)**.** Soit T une trajectoire, et $c \in T$ une chaîne. On dit que c est une **chaîne maximale** si et seulement si aucune autre chaîne de T ne contient c :

$$\forall c' \neq c \in T, c \not\subset c'.$$

On note $T_{max} \subset T$ l'ensemble des chaînes maximales de T.

Il est simple de récupérer T depuis T_{max}. Il suffit d'ajouter pour chaque chaîne maximale ses sous-chaînes afin d'obtenir T. La question est maintenant de savoir quelle quantité de chaînes sont maximales au sein d'une trajectoire. On peut donner une borne maximale et minimale du nombre de chaînes maximales dans T à partir de la connaissance du support minimal $G = (D, E)$ de T.

Pour trouver le minimum, on peut utiliser la construction suivante : Soit un document d. Ce document est traversé par une certaine quantité de chaînes, pour assurer la couverture de G, il est nécessaire que chaque arc arrivant en d soit couvert par au moins une chaîne, et que chaque arc partant en d soit couvert par au moins une chaîne. Puisqu'il est possible de prolonger les chaînes arrivant en d, la seule situation nécessitant d'augmenter le nombre de chaînes est lorsqu'il y a plus d'arcs sortants que de chaînes entrantes en d, si bien que chaque arc n'est couvert que par une et une seule chaîne. Le nombre minimal de chaînes maximales sur un graphe G est ainsi :

$$min_G|T_{max}| = sum_{\substack{d\in D \\ deg_{out}(d)>deg_{in}(d)}} deg_{out}(d) - deg_{in}(d). \tag{B.2}$$

Pour trouver le maximum, l'idée est similaire. Parmi toutes les chaînes qui arrivent au document d, si d a des successeurs, il faut prolonger **toutes** les chaînes à chaque successeur. Ainsi, par récurrence toutes les chaînes arrivent jusqu'aux documents de degré sortant nul. On peut compter le nombre de chaînes maximales finissant en un document d à l'aide de la fonction que nous notons ϕ. Par convention $\phi(d)$ vaut 1 pour un document de degré entrant nul, parce qu'il est expliqué par la chaîne dégénérée d'un seul document d. La définition de ϕ et le calcul du nombre maximal de chaînes maximales sont ainsi :

$$\begin{aligned} \forall d \in D, \phi(d) &= \begin{cases} 1 & \text{si } deg_{in}(d) = 0, \\ \sum_{d' \in Prec(d)} \phi(d') & \text{sinon.} \end{cases} \\ max_G|T_{max}| &= \sum_{\substack{d\in D \\ deg_{out}(d)=0}} \phi(d). \end{aligned} \tag{B.3}$$

Maintenant, la question est de pouvoir mesurer la quantité de chaînes gagnées par la compression *via* les chaînes maximales, c'est-à-dire d'obtenir une borne sur la quantité $|T| - |T_{max}|$. Dans le cas d'une trajectoire à une seule chaîne de taille 2, il y a toujours une chaîne maximale. Comme par construction la compression par chaîne maximale ne fait que supprimer des chaînes de la trajectoire, on peut affirmer que zéro la borne minimale de chaîne supprimée :

$$0 \leq |T| - |T_{max}|. \tag{B.4}$$

Pour la borne maximale, il s'agit du cas où T contient le plus grand nombre de chaînes possibles. En effet, l'existence de toute chaîne d'une taille supérieure ou égale à 3, favorise la compression par chaîne maximale, qui permet de la représenter en une seule chaîne, tandis que T la représente en plusieurs chaînes (pour une chaîne de taille trois, il faut lui adjoindre ses deux sous-chaînes de taille deux). Ainsi la borne maximale est atteinte lorsque T contient toutes les chaînes possibles sur G. On peut également considérer le plus grand support minimal G possible. Pour cela on pose un corpus qui contient uniquement des documents publiés à des dates distinctes, et qu'il y a un arc entre tous les documents qui respecte la temporalité. La quantité maximale de chaînes sur ce support est de $2^{|D|} - |D| - 1$, tandis que la quantité minimale de chaînes maximales est de 1. On peut également, sur ce même support, évaluer la quantité maximale de chaînes maximales à l'aide de l'équation B.3. En numérotant les documents du plus ancien, noté d_1, au plus récent, noté d_n on obtient :

$$\begin{aligned} max_G|T_{max}| = \phi(d_n) &= \sum_{i<n} \phi(d_i), \\ &= \phi(d_{n-1}) + \sum_{i<n-1} \phi(d_i), \\ &= \phi(d_{n-1}) + \phi(d_{n-1}), \\ &= 2\phi(d_{n-1}). \\ &= 2^{n-2}\phi(d_2). \end{aligned} \tag{B.5}$$

Cette récurrence est correcte à condition que $n > 2$. Dans les cas $n = 1$ et $n = 2$ on montre facilement que $\phi(d_n) = 1$. Ainsi, pour les corpus de plus de deux documents, on a dans le cas où le support minimal est le plus grand :

$$max_G|T_{max}| = 2^{|D|-2} = \frac{2^{|D|}}{4} = \frac{max_G|T|}{4} - O(|D|). \tag{B.6}$$

La compression par trajectoires maximales reste dans le pire cas d'une complexité mémoire inférieure mais similaire à la taille de la trajectoire seule. Utiliser T_{max} est une bonne approche pour représenter de grandes trajectoires qui vérifient l'hypothèse d'héritage des sous-chaînes. Lorsque l'hypothèse d'héritage des sous-chaînes n'est pas vérifiée, comme cela arrive en règle générale, on peut agir de plusieurs manières pour réduire la quantité de chaînes. La première est de différencier les chaînes dont toutes les sous-chaînes sont présentes dans la trajectoire et les autres. À l'aide d'un marquage on peut compresser les premières, la quantité de telles chaînes donne un indicateur du respect de l'hypothèse d'héritage des sous-chaînes et est utilisée dans l'Annexe C pour analyser cette hypothèse. La seconde est de considérer que les sous-chaînes sont de toutes manières de la redondance, et que tout traitement sur la sur-chaîne donne également des informations sur la sous-chaîne. Il s'agit de la méthode que nous utilisons, en prenant certaines précautions, dans les chapitres 4 et 5 (cf. section 4.2.1) pour travailler sur la liste des chaînes.

Cependant les ensembles de chaînes maximales ont deux défauts. Le premier est qu'il n'est pas évident d'utiliser notre méthode de calcul de la trajectoire en ne conservant qu'un ensemble de chaînes maximales. Il est possible de le faire, mais cela requiert des calculs supplémentaires de compression et de décompression des ensembles de chaînes. D'autant plus que le second défaut des ensembles de chaînes maximales et qu'ils peuvent encore présenter un certain degré de redondance : Par exemple les chaînes $d_1d_2d_3d_4$ et $d_1d_2d_3d_5$ peuvent toutes deux être maximales, mais la sous-chaîne $d_1d_2d_3$ est représentée deux fois. Nous présentons dans la section suivante une représentation mémoire de la trajectoire qui non seulement élimine ce type de rendondance, mais en plus fonctionne parfaitement avec notre approche.

B.2 Représentation condensée de la trajectoire

L'ensemble $FinitEn(d)$ correspond à l'ensemble des chaînes de la trajectoire qui finissent au document d. Il s'agit d'une construction centrale dans notre méthode de calcul de la trajectoire, qu'on peut comprendre comme un dictionnaire sur les chaînes indexés par le document où la chaîne finit.

$FinitEn(d)$ est un ensemble de chaînes, aussi on peut imaginer le représenter avec des chaînes maximales. Comme nous l'avons soulevé à la section précédente, il y a toujours une question de répétition intrinsèque dans cette approche avec le cas suivant :

— État initial, $c = c_1 \cdot c_2 \in FinitEn(d)$.

— Prolongement en d', deux cas de figure :

1. Dans le cas où on ne garde que les chaînes maximales :

$$c' = c_2 \cdot d' \in FinitEn(d').$$

Il y a potentiellement une duplication de c_2.

2. Dans le cas où on garde toutes les chaînes :

$$c' = c \cdot d' \in FinitEn(d').$$

Il y a potentiellement une duplication de c.

Le prolongement de c en d' donne naissance à la duplication d'une chaîne en mémoire. Il est possible de récupérer cet espace en représentant $FinitEn(d)$ et $FinitEn(d')$ par des tables indexées, comme illustré à la Figure B.1.

Cette figure distingue deux manières de représenter les chaînes. **La représentation naïve**, dont nous avons déjà parlé, est celle présentée pour $FinitEn(d)$ (à gauche) : chaque chaîne est mise en mémoire. Au contraire, $FinitEn(d')$ (à droite) ne met en mémoire qu'une référence de la chaîne prolongée, composée d'une origine pour la table et d'une référence pour l'index. C'est une **représentation condensée**.

	$FinitEn(d)$		$FinitEn(d')$	
Index	Chaîne	Index	Origine	Référence
1	c_1	1	d_{i_1}	r_1
2	c_2	2	d_{i_2}	r_2
3	c_3	3	d_{i_3}	r_3
...	...	...	...	...
i	c	j	d	i
...	...	...	...	...
k	c_k	m	d_{i_m}	r_m

FIGURE B.1 – Tables indexées. $FinitEn(d)$ met en mémoire les chaînes entières. $FinitEn(d')$ met seulement en mémoire un pointeur vers la chaîne précédente.

Les deux approches ont, *a priori*, leurs avantages. La première permet de représenter la chaîne avec la structure de données que l'on souhaite, tandis que la seconde prend moins de place mais impose une structure de liste chaînée. Les seules opérations pénalisantes pour la structure de liste chaînée sont l'accès à un seul élément d'index arbitraire et l'ajout en milieu ou fin de liste. Or, aucune de ces deux opérations n'est nécessaire lors de notre approche pour le calcul de la cohérence. Aussi, asymptotiquement, la représentation condensée est au moins aussi efficace que la représentation naïve. La place totale occupée par une chaîne correspond à la place nécessaire pour la construire dans chacun des $FinitEn(d_i)$ précédents :

$$Espace(d_1 d_2 \dots d_l, Total) = \sum_{2 \leq i \leq l} Espace(d_1 \dots d_i, FinitEn(d_i)). \tag{B.7}$$

Ce qui donne un ordre de magnitude d'écart :

$$\begin{aligned} Espace(d_1 d_2 \dots d_l, Total, \text{Naïve}) &= \sum_{2 \leq i \leq l} O(i) = O(l^2), \\ Espace(d_1 d_2 \dots d_l, Total, \text{Condensée}) &= \sum_{2 \leq i \leq l} O(1) = O(l). \end{aligned} \tag{B.8}$$

Prolonger une chaîne n'engendre aucune redondance avec la représentation condensée. Cependant elle ne convient pas pour représenter les chaînes maximales. Dans ce cas, pour parfaitement spécifier une chaîne $c = c_b \cdot dd'$, il faut une référence au document d, une référence à une chaîne maximale $c' = c_a \cdot c_b$ et la quantité de documents à

remonter le long de cette chaîne maximale, à savoir la taille de c_b. Cette variante de la représentation condensée est présentée en Figure B.2.

	$FinitEn(d)$		$FinitEn(d')$		
Index	Chaîne	Index	Origine	Référence	taille
1	c_1	1	d_{i_1}	r_1	l_1
2	c_2	2	d_{i_2}	r_2	l_2
3	c_3	3	d_{i_3}	r_3	l_3
...	...	...	...	...	...
i	$c = c_a \cdot c_b$	j	d	i	$taille(c_b)$
...	...	...	...	...	...
k	c_k	m	d_{i_m}	r_m	l_m

FIGURE B.2 – Variante de la figure B.1, pour le cas où on ne garde que des chaînes maximales.

La représentation condensée de la trajectoire et sa variante qui ne conserve que les chaînes maximales fournissent deux représentations sans redondances et complètes de la trajectoire, la première permet efficacement son calcul et la seconde est adaptée au stockage de celle-ci. Lorsqu'il est nécessaire de parcourir plusieurs fois les chaînes pour les analyser comme c'est le cas dans les chapitres 4 ou 5, il peut être judicieux, si l'espace ne fait pas défaut, de représenter les trajectoires à l'aide d'une structure de donnée mieux adaptée au cas d'application envisagé.

Annexe C

Étude quantitative du calcul de la Trajectoire

Pour parvenir à la méthode de construction d'ensemble de chaînes cohérentes, détaillée en section 3.3, nous nous sommes appuyés sur plusieurs intuitions dont il convient de discuter les implications et de tester le bien-fondé. Notre approche repose en particulier sur deux hypothèses fondamentales concernant la cohérence d'une chaîne.

La première de ces hypothèses concerne l'estimation de la cohérence d'une chaîne. Nous postulons ainsi que la cohérence d'une chaîne c peut s'estimer à partir de la similarité deux à deux des documents constitutifs de c. La première des conséquences pratiques de cette hypothèse est que nous nous abstrayons du corpus de document D. Notre approche requiert ainsi, seulement, la matrice de similarité des documents de D. L'hypothèse appelle aussi en pratique deux nouvelles précisions : Quelles paires de documents constitutifs de c entrent effectivement en jeu dans l'estimation de la cohérence ? De quelle manière faut-il combiner ces similarités pour estimer la cohérence ? Nous appelons respectivement la réponse à ces questions la stratégie de sélection et la fonction de combinaison. Il existe certainement des manières arbitrairement sophistiquées de choisir la stratégie de sélection et la fonction de combinaison. Nous nous contenterons de montrer que des cas simples produisent

déjà des résultats intéressants. Nous reléguons l'exploration des différentes méthodes pour approcher la cohérence, qu'elles soient construites ainsi ou non, comme un sujet futur intéressant.

La seconde hypothèse fondamentale de notre approche repose sur la relation de cohérence entre une chaîne et ses sous-chaînes. C'est en exploitant cette idée que nous pouvons construire une approche itérative efficace pour balayer l'ensemble des chaînes possibles. Nous la postulons ainsi : Si c est une chaîne cohérente, alors toute sous-chaîne $c' \subset c$ est faiblement cohérente. Cela nous permet de construire dans un premier temps l'ensemble des chaînes faiblement cohérentes, $T^*_{coh}(D)$ puis de le filtrer selon le critère de cohérence (cf. section 3.3.3). Pour ce faire, nous nous appuyons sur une hypothèse similaire sur la faible cohérence : Si c est une chaîne faiblement cohérente, alors toute sous-chaîne $c' \subset c$ est faiblement cohérente. Or, étant donnée la manière dont nous estimons la cohérence d'une part, et la manière dont nous construisons $T^*_{coh}(D)$ d'autre part, rien ne garantit cette propriété. L'ensemble $T^*_{coh}(D)$ calculé par notre approche est ainsi potentiellement bruité et non exhaustif.

Nous menons dans cette annexe une étude sur le comportement de notre approche face à une matrice de similarités aléatoires, indépendantes et de loi normale. Dans ce contexte, aucune chaîne autre que celles de taille 2 ne devrait être cohérente en raison de l'indépendance des similarités. Cela permet de mettre en évidence l'existence de chaînes fortuites, qui existent par construction. Nous dégageons les propriétés de ces chaînes, typiquement elles ne respectent pas la seconde hypothèse fondamentale. Nous montrons que de telles chaînes existent aussi pour des corpus de documents réels, qu'il est possible de les filtrer à l'aide de seuils de cohérence adéquat, et que les chaînes restantes respectent au contraire la seconde hypothèse. Nous montrons également dans ce contexte théorique que notre approche capture une part non négligeable des chaînes faiblement cohérentes.

C.1 Chaînes fortuites

L'intuition de notre travail est de considérer que la similarité entre les documents capture, en partie et dans ses nuances, la cohérence des chaînes. Cependant, dans une chaîne $c = d_1 d_2 d_3$, tous les documents peuvent être similaires deux à deux, pour des raisons indépendantes les unes des autres. Notre approche va ainsi construire c comme une chaîne cohérente, tandis qu'une observation humaine de c nous montrera qu'elle ne l'est pas. On dit que c est une **chaîne cohérente fortuite**. Pour étudier ce phénomène, on peut se placer dans un cadre théorique où notre approche ne construit que de telles chaînes. Pour le modéliser, on considère que la similarité suit une loi normale (Eq. C.1) tronquée sur $[0, 1]$ et que chaque paire de documents a une similarité indépendante de celle des autres paires.

$$\forall d_1, d_2 \in D, sim(d_1, d_2) \sim \mathcal{N}(\mu, \sigma^2). \tag{C.1}$$

Cette approche modélise le bruit de la similarité au sein d'un corpus, où tous les documents sont faiblement similaires. Dans ce contexte, on peut observer la manière dont se comporte la trajectoire calculée. Nous commençons par détailler à l'aide d'indicateurs la forme d'une trajectoire, nous étudions ensuite l'impact du seuil de faible cohérence γ^* sur cette forme pour différentes fonctions de cohérence. Nous comparons ces résultats à la forme observée dans le cadre de jeux de données réels. Nous procédons de même pour le seuil de cohérence γ et l'impact de l'approche heuristique avec q_{limit}.

C.2 Chaînes stockées, maximales et totales

La trajectoire calculée est un ensemble de chaînes, noté T. Pour la différencier d'autres ensembles de chaînes que nous allons distingués, on nomme **chaînes stockées** les chaînes de T. On peut distinguer deux autres ensembles de chaînes qui sont liés à la notion de sous-chaînes de T. La **trajectoire complétée** et la **trajectoire**

compressée.

La **trajectoire complétée** de T, notée $Compl(T)$ correspond à l'ensemble des chaînes de T augmenté de leurs sous-chaînes :

$$Compl(T) = T \cup \{c, \exists c' \in T / c \subset c'\}. \tag{C.2}$$

Une des hypothèses de notre approche est que les sous-chaînes des chaînes faiblement cohérentes sont elles-mêmes faiblement cohérentes. Cette hypothèse se traduit ainsi :

$$Compl(T^*_{coh}(D)) = T^*_{coh}(D). \tag{C.3}$$

Aussi, le rapport entre la trajectoire des chaînes faiblement cohérentes calculée et sa trajectoire complétée renseigne sur la validité de l'hypothèse C.3. On nomme **chaînes totales** les chaînes de la trajectoire complétées.

A contrario, on a vu qu'il était possible de compresser la représentation d'une trajectoire en utilisant les **chaînes maximales**, pour évacuer la redondance des sous-chaînes. L'hypothèse C.3 n'étant pas nécessairement vérifiée, on ne peut compresser, dans la trajectoire calculée, que les chaînes dont toutes les sous-chaînes sont présentes. Cela nous amène a avoir une définition légérément différente d'une chaîne maximale. Une chaîne est maximale dans une trajectoire si toutes ses sous-chaînes y sont aussi et qu'elle n'est sous-chaîne d'aucune autre chaîne maximale :

$$c \in T_{max} \iff \begin{cases} \forall c' \subset c, c' \in T \\ \forall c' \neq c \in T_{max}, c \not\subset c' \end{cases}. \tag{C.4}$$

Il est ainsi possible de marquer, pour une trajectoire T, les chaînes qui sont maximales, et d'enlever leurs sous-chaînes, ce qui donne la **trajectoire compressée** $Compr(T)$. Il s'agit de l'ensemble des chaînes maximales T_{max}, auquel on adjoint toutes les chaînes de T qui ne sont sous-chaînes d'aucune chaîne maximale. Comme l'hypothèse d'héritage n'est, *a priori*, pas validée, il est possible que des chaînes soient faiblement

cohérentes sans que leurs sous-chaînes le soient.

$$Compr(T) = T_{max} \cup \{c \in T, \nexists c' \in T_{max}/c \subset c'\} \tag{C.5}$$

L'écart entre $Compr(T^*_{coh}(D))$ et $T^*_{coh}(D)$ renseigne également sur le nombre de chaînes qui valident l'hypothèse C.3. Dans la suite, par abus de langage, on nomme **chaînes maximales** toutes les chaînes de la trajectoire compressée.

Intuitivement, on peut ainsi déterminer si une trajectoire T valide globalement l'hypothèse lorsque elle est quantitativement proche de sa trajectoirée complétée : $|Compl(T)| \simeq |T|$. Ou déterminer qu'elle ne la valide globalement pas lorsque elle est quantitativement proche de sa trajectoire compressée : $|Compr(T)| \simeq |T|$.

C.3 Influence du seuil de faible cohérence γ^*

On s'intéresse tout d'abord au comportement de la trajectoire faiblement cohérente $T^*_{coh}(D)$ lorsque le seuil de faible cohérence γ^* varie. Nous proposons plusieurs stratégies pour calculer la cohérence. Il s'agit à chaque fois d'une moyenne usuelle d'un ensemble de similarités entre les documents constitutifs de la chaîne. La loi normale étant stable par la moyenne arithmétique d'une part, et les similarités étant indépendantes d'autre part, il est possible dans ce modèle de calculer la loi de probabilité de la cohérence pour une chaîne :

$$coh_{avg}(c) \sim \mathcal{N}(\mu, \frac{\sigma^2}{|select(c)|}). \tag{C.6}$$

La fonction *select* retourne l'ensemble des paires de documents de c à comparer, le cardinal $|select(c)|$ correspond à la quantité de similarités impliquées dans la moyenne arithmétique. De là, connaissant le seuil γ^*, le cardinal de $|select(c)|$ pour une chaîne de taille l, noté $f_{select}(l)$, et la fonction de répartition de la loi normale centrée réduite ϕ, on peut connaître la proportion attendue de chaînes faiblement

cohérentes de taille l, sur le total de chaînes de taille l :

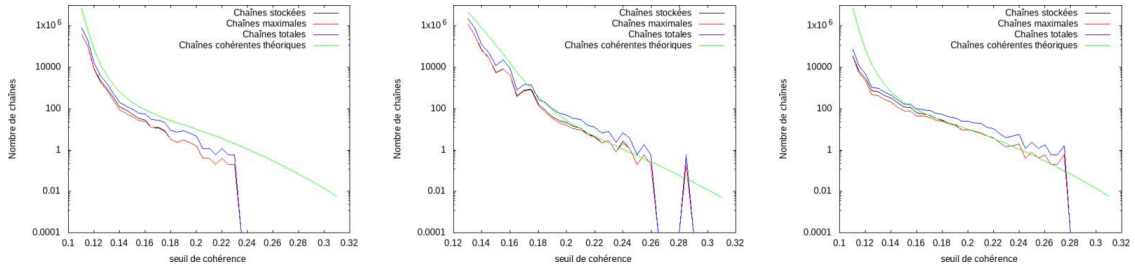
$$\frac{qte_{coh}(l)}{qte(l)} = 1 - \phi(\frac{(\gamma^* - \mu)}{\sigma}\sqrt{f_{select}(l)}). \tag{C.7}$$

On remarque d'abord que si $\gamma^* = \mu$, à savoir le bruit du corpus, on s'attend, en moyenne, à ce que la moitée des chaînes possibles soient faiblement cohérentes. Pour un corpus de $|D|$ documents cela fait environ $2^{|D|-1}$ chaînes. On s'intéresse à la quantité de chaînes faiblement cohérentes pour $\gamma^* > \mu$ et à la quantité effectivement calculées par l'approche. Cela dépend de la stratégie de sélection déployée. La Figure C.1 présente $|T^*_{coh}|$ pour trois stratégies de selection : $select_{all}$, $select_{succ}$ et $select_{fin}$ définies ainsi :

1. $select_{all}(d_1 \dots d_l)$ choisit toutes les paires de documents de la chaîne $\{(d_i, d_j), 1 \leq i < j \leq l\}$. $f_{all}(l) = \frac{l.(l-1)}{2}$.
2. $select_{succ}(d_1 \dots d_l)$ ne choisit que les maillons successifs de la chaîne $\{(d_i, d_{i+1}), 1 \leq i < l\}$. $f_{succ}(l) = l - 1$.
3. $select_{fin}(d_1 \dots d_l)$ ne choisit que les similarités relatives au dernier document de la chaîne $\{(d_i, d_l), 1 \leq i < l\}$. Cette sélection n'est pas, *a priori*, justifiée. Il n'y a pas d'intuition selon laquelle, si tous les documents d'une chaîne sont similaires au dernier, cela induise que cette chaîne soit cohérente. Cette stratégie ne construit donc pas une cohérence dans l'absolu. Elle construit cependant une cohérence conditionnelle : sachant qu'une chaîne c est cohérente, la cohérence de la chaîne prolongée $c \cdot d$ s'exprime en fonction de la similarité entre d et les documents de c. Couplée à la manière dont on construit $T^*_{coh}(D)$, la cohérence induite par $select_{fin}$ construit des chaînes $d_1 \dots d_l$ où chaque d_i est similaire aux documents d_j, pour $j < i$. Il convient donc de compter le cardinal effectif de $select_{fin}$ comme $f_{fin}(l) = \frac{l(l-1)}{2}$.

On peut émettre plusieurs remarques à la lecture des résultats de la Figure C.1 :

1. On observe d'abord que $select_{all}$ et $select_{succ}$ capturent une quantité de chaînes faiblement cohérentes du même ordre de grandeur que la quantité théorique



(a) $select_{all}$ (b) $select_{succ}$ (c) $select_{fin}$

FIGURE C.1 – Quantité de chaînes faiblement cohérentes calculées et théoriques pour différentes stratégies de sélection. Nombre de documents : 30. $\mu = 0,1$. $\sigma = 0.05$. Moyenne sur 20 tirages.

attendue. Cela signifie que l'hypothèse de construction sur laquelle se base l'approche, bâtir les chaînes $d_1 \dots d_{l-1} d_l$ à partir des chaînes faiblement cohérentes de la forme $d_1 \dots d_{l-1}$ précédemment calculées, fournit un échantillon non négligeable de l'ensemble attendu des chaînes faiblement cohérentes.

2. On peut ainsi observer que $select_{fin}$ fournit également un échantillon non négligeable de l'ensemble attendu des chaînes faiblement cohérentes. Cette stratégie est particulièrement intéressante pour son rapport qualité coût : elle permet d'obtenir des chaînes de la même qualité que $select_{all}$ pour une compléxité de calcul linéaire en la taille de la chaîne.

3. On remarque également que la quantité de chaînes stockées est généralement proche de la quantité de chaînes maximales, ce qui sous-entend que la trajectoire calculée est déjà dans une représentation essentiellement compacte vis-à-vis des chaînes maximales.

4. À l'inverse, il est rare que la quantité de chaînes de la trajectoire calculée T (*chaînes stockées*) soit proche de la quantité de chaînes de la trajectoire complétée $Compl(T)$ (*chaînes totales*). Vis-à-vis de l'hypothèse selon laquelle une chaîne cohérente doit être composée de sous-chaînes faiblement cohérentes, cela suggère que le bruit (ou un ensemble de documents indépendamment similaires les uns des autres), ne peut générer que peu de chaînes cohérentes.

5. Enfin, on remarque que pour $\gamma^* > \mu + 2\sigma$, le bruit gaussien n'engendre presque plus aucune chaîne cohérente fortuite.

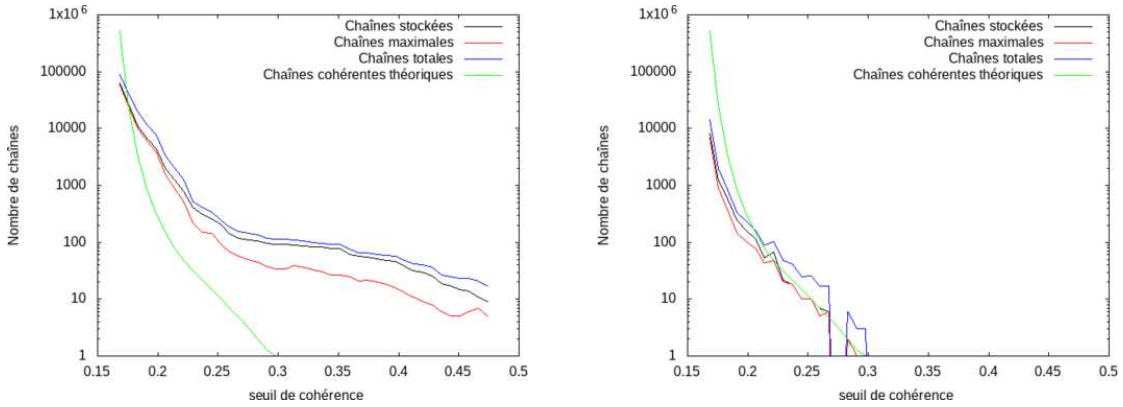
Pour évaluer ces résultats théoriques à ceux obtenus avec des données réelles, nous avons construit des matrices de similarités tirés de nos corpus de documents réels **HuffPost** et **AMINER**. Pour ce faire nous piochons au hasard une certaine quantité de documents pour chaque corpus. De là, nous construisons les matrices de similarités, pour lesquelles nous comptons la moyenne μ et la variance σ^2, qui nous serviront à nous comparer au comportement d'un jeu synthétique de bruit gaussien de même paramètres. Le tout est répété plusieurs fois pour produire les résultats donnés en Figure C.2 pour HuffPost et en Figure C.3 pour AMINER. On constate deux choses :

— Pour $\gamma^* \simeq \mu$, on retrouve la proportion de chaînes fortuites constatée dans le modèle normal. Cependant, lorsque $\gamma*$ augmente, on ne suit plus la distribution théorique pour le HuffPost, avec une quantité de chaînes de cohérences élevées. On y constate une rupture de pente sur l'échelle logarithmique aux alentours de $\gamma^* = 0,24 \simeq \mu + \sigma$, ce qui correspond à l'effondrement de la quantité de chaînes dans les données de synthèse. Cela sous-entend que, dans l'intervalle $[\mu, \mu + \sigma]$, les trajectoires calculées sur des jeux de données réels possèdent des chaînes qui ont un comportement proches des chaînes fortuites générées par un bruit gaussien.

— Au même point de rupture, on constate que la quantité de chaînes stockées $|T|$ est beaucoup plus proche de la quantité complétée des chaînes totales $|Compl(T)|$ que dans le modèle normal. Ces chaînes qui ne se comportent pas comme des chaînes fortuites semblent vérifier l'hypothèse d'héritage.

Ces deux constats indiquent qu'il existe des chaînes de bonne qualité au sens de notre hypothèse de faible cohérence, et qu'elles sont filtrables par un choix adéquat de γ^*.

Résumons ce que nous apprennent ces résultats. L'approche de calcul de la trajectoire faiblement cohérente $T^*_{coh}(D)$ présentée dans ce chapitre repose essentiellement sur deux hypothèses. Selon la première, la notion de cohérence d'une chaîne se mesure à l'aide des similarités deux à deux entre les documents constitutifs de la chaîne. Cela

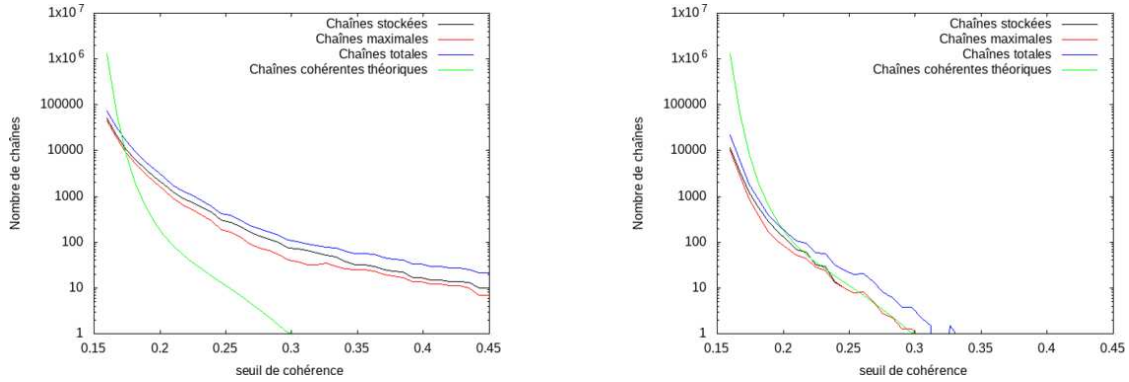


(a) HuffPost (b) Données de synthèse

FIGURE C.2 – Quantité de chaînes faiblement cohérentes calculées pour des documents tirés au hasard dans le corpus HuffPost (a). Même quantité pour des données de synthèses suivant le modèle $sim(d_i, d_j) \sim \mathcal{N}(\mu, \sigma^2)$ (b). Nombre de documents tirés : 30. Nombre de tirages : 20. Similarité moyenne : $\mu \simeq 0,1528$. Écart-type : $\sigma \simeq 0,0996$. Similarité utilisée : Cosinus de TF-IDF. Stratégie de sélection : $select_{fin}$. Combinaison : moyenne arithmétique.

implique que la donnée relative aux documents nécessaires à l'approche se réduit à la matrice de similarité entre ces documents. L'autre entrée de l'approche, est la manière dont sont choisies les similarités qui entrent dans le calcul de la cohérence et la manière dont elles se combinent. Nous nous contentons de stratégies de sélection simples et d'une combinaison par moyenne classique. Or rien n'indique, si ce n'est l'intuition, qu'un tel dispositif de calcul de la cohérence corresponde en effet à une notion de cohérence. L'analyse à partir de matrices de similarités générées par un bruit gaussien nous montre que dans ce cadre, de très nombreuses chaînes peuvent satisfaire le critère de faible cohérence γ^*, alors que dans un cadre d'indépendance des similarités, aucune chaîne ne devrait sembler cohérente. Cependant, nous montrons empiriquement que notre approche capture, même dans ce contexte, une proportion non négligeable des chaînes faiblement cohérentes. De plus, ces chaînes faiblement cohérentes fortuites sont localisées pour des valeurs de γ^* bornées autour de μ.

La cohérence est cependant soumise à la seconde hypothèse essentielle de l'approche : On s'attend à ce que toutes les sous-chaînes d'une chaîne cohérente soient faiblement cohérentes. Plus particulièrement, on s'attend à ce que toutes les sous-chaînes d'une chaîne *faiblement* cohérente soient faiblement cohérentes. Dans le



(a) AMINER (b) Données de synthèse

FIGURE C.3 – Quantité de chaînes faiblement cohérentes calculées pour des documents tirés au hasard dans le corpus AMINER (a). Même quantité pour des données de synthèses suivant le modèle $sim(d_i, d_j) \sim \mathcal{N}(\mu, \sigma^2)$ (b). Nombre de documents tirés : 30. Nombre de tirages : 20. Similarité moyenne : $\mu \simeq 0,1450$. Écart-type : $\sigma \simeq 0,1078$. Similarité utilisée : Cosinus de TF-IDF. Stratégie de sélection : $select_{fin}$. Combinaison : moyenne arithmétique.

modèle du bruit gaussien, cette dernière propriété n'est pas vérifiée pour la trajectoire que nous calculons. Si, au contraire du bruit gaussien, on prend un corpus de documents réel pour constituer la matrice de similarité, on constate d'une part la superposition de chaînes qui se comportent comme celles issues d'un bruit gaussien et d'autre de natures différentes, et qu'il existe, en dehors de la zone bruitée, un intervalle pour γ^* où la seconde hypothèse est globalement vérifiée. Ces chaînes sont donc plus proches de l'intuition que nous nous faisons de chaînes faiblement cohérentes et sont capturées par notre approche.

C.4 Influence du seuil de cohérence γ

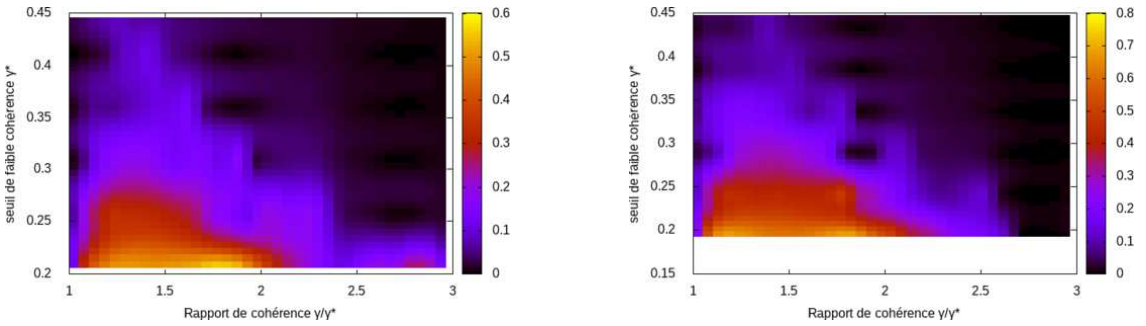
Il nous reste à vérifier quantativement que la distinction entre γ^*, le seuil de faible cohérence, et γ le seuil de cohérence permet de capturer des chaînes que nous ne capturerions pas autrement. Pour ce faire, nous comparons $T_{coh}(D, \gamma, \gamma^*)$, la trajectoire calculée lorsqu'on distingue γ et γ^* et $T_{coh}(D, \gamma, \gamma)$, la trajectoire calculée lorsqu'on ne les distingue pas, auquel cas $\gamma^* = \gamma$. Comme $\gamma^* \leq \gamma$, et étant donné la

manière dont nous construisons la trajectoire, on a la relation :

$$T_{coh}(D, \gamma, \gamma) \subset T_{coh}(D, \gamma, \gamma^*) \tag{C.8}$$

Les autres paramètres étant fixés par ailleurs, on peut mesurer l'efficacité de la distinction γ/γ^* en calculant le rapport $dif(\gamma, \gamma^*)$:

$$dif(\gamma, \gamma^*) = \frac{|T_{coh}(D, \gamma, \gamma^*)| - |T_{coh}(D, \gamma, \gamma)|}{|T_{coh}(D, \gamma, \gamma^*)|} \tag{C.9}$$



(a) AMINER (b) HuffPost

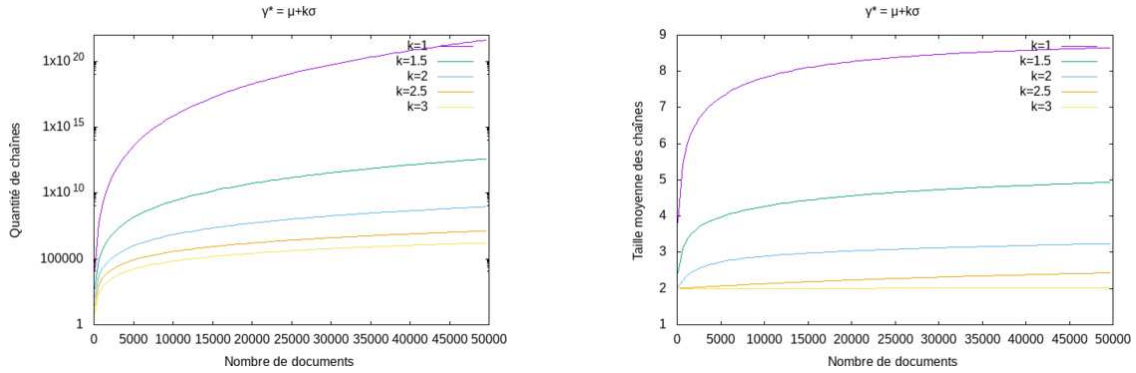
FIGURE C.4 – Évolution de $dif(\gamma, \gamma^*)$ pour AMINER (a) et HuffPost (b). Réalisé sur 20 tirages de 30 documents pour chaque jeu de données. La valeur initiale de γ^* vaut $\gamma^* = \mu + \frac{\sigma}{2}$, où μ et σ sont respectivement la similarité moyenne et l'écart-type associé pour chaque jeu de données. Les cartes de chaleurs sont extrapolées à partir d'un ensemble $20x32$ coordonnées calculées.

L'évolution de $dif(\gamma, \gamma^*)$ est donnée en Figure C.4 pour les deux jeux de données HuffPost et AMINER. On constate globalement que distinguer une faible cohérence permet de capter un plus grand nombre de chaînes cohérentes. C'est particulièrement vrai autour des valeurs basses de γ^*, ce qui pourrait conforter l'idée qu'une chaîne cohérente contient généralement des parties plus faibles que son tout. L'analyse des chaînes fortuites, du seuil de cohérence et du seuil de cohérence faible permettent jusqu'ici de justifier nos hypothèses fondamentales. Ces résultats reposent sur l'hypothèse selon laquelle nous travaillons avec une estimation convenable de la cohérence. Pour tester cette hypothèse, nous avons procédé à une campagne d'évaluation humaine de la cohérence des chaînes, présentée en section 3.4. Nous étudions maintenant

l'influence de l'heuristique de l'approche de calcul, présenté en section 3.3.5.

C.5 Passage à l'échelle et influence de q_{limit}

Jusqu'à maintenant, toutes les expérimentations ont été menées sur des jeux de données, réels ou synthétiques, de petite taille (30 documents). Lorsque le nombre de documents augmente, le nombre de chaînes potentielles augmente aussi, et en particulier, le nombre de chaînes fortuites. On observe en Figure C.5 que la quantité de chaînes faiblement cohérentes dans notre modèle de bruit gaussien croît rapidement avec le nombre de documents. Pour des valeurs de γ^* suffisamment élevées, il ne reste que les chaînes de taille 2, qui ne sont pas, sachant la confiance qu'on accorde à la fonction de similarité, des chaînes fortuites. Il semblerait donc qu'il suffise de



(a) Quantité de chaînes

(b) Longueur moyenne de chaîne

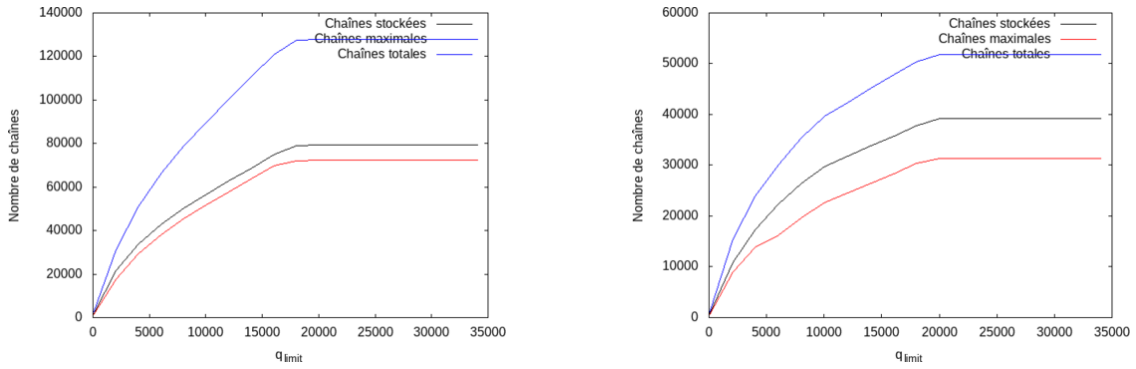
FIGURE C.5 – Évolution du modèle de bruit gaussien quand la quantité de documents augmente. k correspond à l'écart de γ^* à la moyenne μ en écart-type, par la formule $\gamma^* = \mu + k\sigma$.

s'éloigner convenablement du bruit pour éviter l'explosion combinatoire des chaînes fortuites. Il ne s'agit pas du seul phénomène qui puisse produire une explosion combinatoire lorsque le nombre de documents est élevé. Par exemple, un sous-corpus de n documents tous similaires deux à deux peut générer jusqu'à $2^n - n - 1$ chaînes. Ces sous-corpus sont courants dans les jeux de données réels, et les chaînes qu'ils engendrent sont effectivement cohérentes [1]. Rien ne garantit que l'ensemble des

1. Une question intéressante est de différencier, parmi ces chaînes cohérentes au sein d'un sous-corpus, les plus susceptibles d'être effectivement des chemins de propagation. Dans un premier temps, nous préférons agréger ces différentes chaînes, et traiter cette question par la suite.

chaînes cohérentes ait une taille raisonnable vis-à-vis du nombre de documents dans le corpus. Pour pouvoir traiter des corpus d'une taille conséquente, en toute généralité, il est donc nécessaire de se munir de garde-fous.

Pour ce faire, nous limitons la quantité de chaînes qui finit en un certain document d, limite notée q_{limit}. Ainsi, la quantité totale de chaînes calculée est bornée par $|D| * q_{limit}$. La Figure C.6 représente la quantité de chaînes calculées pour différentes valeurs de q_{limit}. On constate, comme attendu, l'existence d'un seuil lorsque q_{limit} dépasse une certaine valeur, c'est-à-dire qu'il ne bride plus aucun des documents en nombre de chaînes y finissant.

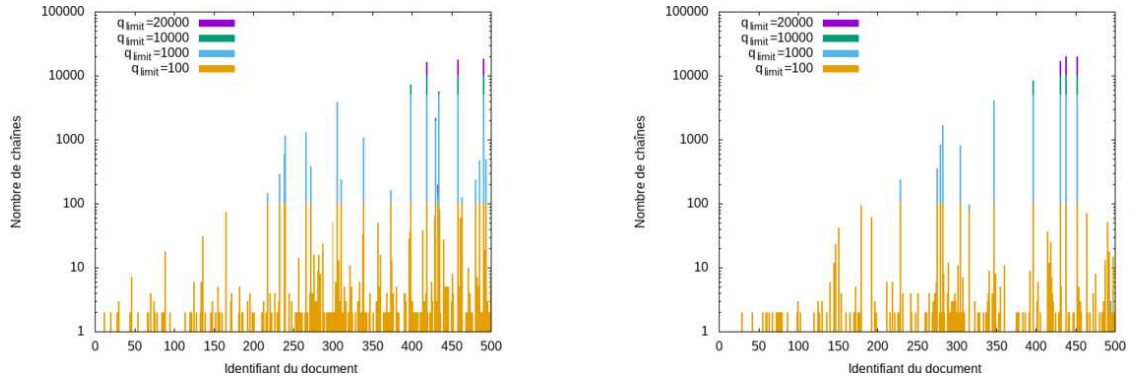


(a) AMINER (b) HuffPost

FIGURE C.6 – Nombre de chaînes faiblement cohérentes pour différentes valeurs de q_{limit}. Obtenues pour un tirage de 500 documents de AMINER (a) et un tirage de 500 documents de HuffPost (b). $\gamma^* = \mu + \sigma$ dans les deux cas.

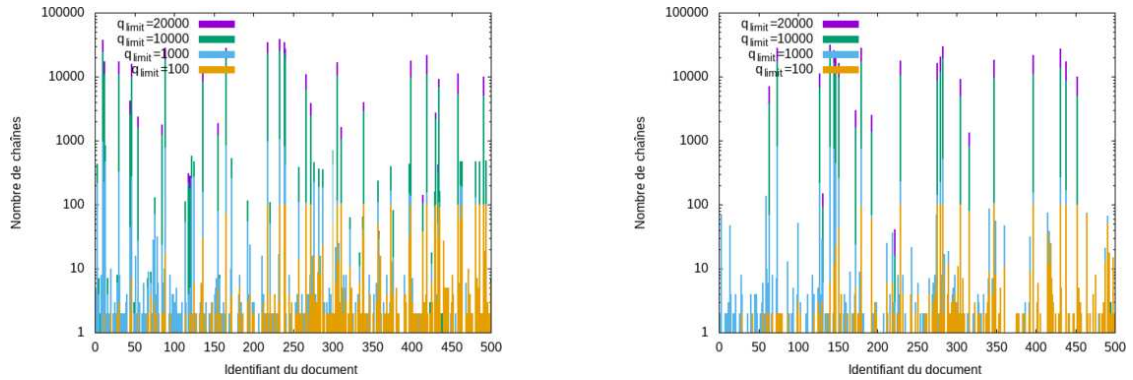
La Figure C.7 donne la quantité de chaînes finissant en chaque document selon différentes valeurs de q_{limit}. On note qu'une petite quantité de documents monopolise les grandes quantités de chaînes, et que les brider a un impact faible sur les chaînes, en dehors de ces quelques points. Il s'agit de documents capables de se greffer à la fin de nombreuses chaînes qui leurs préexistent. La Figure C.8 illustre le nombre de chaînes qui passent par un certain document. Si les chaînes induites par une hausse de q_{limit} passent par un nombre important de documents, elles ne finissent qu'en un nombre relativement faible de documents. Ces documents sont donc, structurellement, des aggrégateurs de chaînes, il s'agit d'une autre situation capable de générer une explosion du nombre de chaînes. Que ces aggrégateurs soient des anomalies de notre

méthode, ou de réels documents centraux du corpus, le fait de les brider par un q_{limit} suffisant nous laisse la capacité de les distinguer, quitte à les traiter en particulier dans un second temps, tout en contrôlant le nombre de chaînes calculées.



(a) AMINER (b) HuffPost

FIGURE C.7 – Quantité de chaînes **finissant** en chaque document pour différentes valeurs de q_{limit}. Mêmes jeux et conditions que pour la Figure C.6.



(a) AMINER (b) HuffPost

FIGURE C.8 – Quantité de chaînes **passant** en chaque document pour différentes valeurs de q_{limit}. Mêmes jeux et conditions que pour la Figure C.6.

C.6 Conclusion

Cette étude quantitative a pour but de tester différentes hypothèses que nous avons posés pour bâtir notre méthode de construction des chaînes cohérentes. Tout d'abord nous sommes parti de l'hypothèse qu'il était possible de construire la cohérence d'une chaîne à partir d'une combinaison de similarité des documents deux à deux. Nous montrons en section 3.4 que les mesures de cohérence ainsi construites

capturent bel et bien le jugement humain, à condition de considérer les bons seuils de cohérence. Il reste à savoir comment fixer ces seuils. Surtout que nous montrons dans cette section que les mesures de cohérence ainsi construites peuvent attribuer un score non nul à des chaînes de manière fortuite. Nous étudions ce comportement en construisant des matrices de similarité synthétiques, qui suivent une loi normale, à partir desquelles nous construisons des chaînes cohérentes qui sont de fait fortuites. Les chaînes cohérentes calculées pour des jeux de données réels se comportent, pour certains seuils de similarités et dans leur ensemble, de la même manière que ces chaînes fortuites. Nous estimons ainsi un critère de choix du seuil de faible cohérence γ^* qui permet d'échapper à la construction des chaînes fortuites :

$$\gamma^* \geq \mu + \sigma$$

, ou μ et σ sont la moyenne et l'écart-type de la matrice de similarité du corpus de document. Nous montrons de plus que les chaînes par-delà ce seuil respectent globalement notre hypothèse d'héritage de la faible cohérence : la sous-chaîne d'une chaîne faiblement cohérente est elle-même faiblement cohérente.

Lorsque nous avons introduit la notion de faible cohérence en section 3.3.3, nous l'avons justifié comme une méthode pour construire des chaînes cohérentes qui ne sont pas composées de sous-chaînes cohérentes. Nous montrons également que choisir $\gamma > \gamma^*$ permet effectivement d'atteindre un plus grand nombre de chaînes cohérentes qu'en fixant $\gamma = \gamma^*$. Un travail intéressant serait de faire une évaluation humaine de ces chaînes qui semblent plus cohérentes dans leur tout que dans leurs parties.

Enfin, à la section 3.3.5, nous introduisons dans notre méthode de calcul une approche heuristique pour brider les potentielles explosions combinatoires du nombre de chaînes calculées. Pour ce faire nous limitons le nombre de chaînes pouvant finir en chaque document, en fixant une constante q_{limit}. Nous montrons quantitativement que cette heuristique ne bride qu'un nombre restraint de documents même pour des valeurs relativement basses de q_{limit}, comme $q_{limit} = 100$. Cela signifie que cette heuristique nous garantie la plupart des chaînes sur la majorité des documents du

corpus, en ne discriminant que les documents avec une très forte connectivité.

Ainsi, cette étude quantitative fournit un fondement aux différents choix de notre approche et défriche les principales difficultés pour l'établissement d'une nouvelle manière de calculer l'ensemble des chaînes cohérentes.

Bibliographie

[Frobenius, 1898] Leo Frobenius. *Der Ursprung der afrikanischen Kulturen*. Gebrüder Borntraeger, 1898.

[Harris, 1954] Zellig Harris. « Distributional structure ». In : *Word* 10.23 (1954), p. 146-162.

[Firth, 1957] J. R. Firth. « A synopsis of linguistic theory 1930-55. » In : 1952-59 (1957), p. 1-32.

[Damerau, 1964] Fred J Damerau. « A technique for computer detection and correction of spelling errors ». In : *Communications of the ACM* 7.3 (1964), p. 171-176.

[Levenshtein, 1966] Vladimir I Levenshtein. « Binary codes capable of correcting deletions, insertions, and reversals ». In : *Soviet physics doklady*. T. 10. 8. 1966, p. 707-710.

[Milgram, 1967] Stanley Milgram. « The small world problem ». In : *Psychology today* 2.1 (1967), p. 60-67.

[Bass, 1969] Frank Bass. « A New Product Growth Model for Product Diffusion ». In : *Management Science* 15.5 (1969), p. 215-227.

[Jones, 1972] Karen Spärck Jones. « A statistical interpretation of term specificity and its application in

retrieval ». In : *Journal of Documentation* 28 (1972), p. 11-21.

[Granovetter, 1973] Mark S. GRANOVETTER. « The Strength of Weak Ties ». In : *American Journal of Sociology* 78.6 (1973), p. 1360-1380.

[Granovetter, 1978] Mark GRANOVETTER. « Threshold models of collective behavior ». In : *American journal of sociology* 83.6 (1978), p. 1420-1443.

[Deerwester, 1990] Scott DEERWESTER et al. « Indexing by latent semantic analysis ». In : *Journal of the American society for information science* 41.6 (1990), p. 391-407.

[Woese, 1990] Carl R WOESE, Otto KANDLER et Mark L WHEELIS. « Towards a natural system of organisms : proposal for the domains Archaea, Bacteria, and Eucarya. » In : *Proceedings of the National Academy of Sciences* 87.12 (1990), p. 4576-4579.

[Peirce, 1991] Charles Sanders PEIRCE. *Peirce on signs : Writings on semiotic*. UNC Press Books, 1991.

[Banerjee, 1992] Abhijit V BANERJEE. « A simple model of herd behavior ». In : *The quarterly journal of economics* 107.3 (1992), p. 797-817.

[Brown, 1992] Peter F BROWN et al. « Class-based n-gram models of natural language ». In : *Computational linguistics* 18.4 (1992), p. 467-479.

[Blume, 1993] Lawrence E BLUME et al. « The statistical mechanics of strategic interaction ». In : *Games and economic behavior* 5.3 (1993), p. 387-424.

[Ellison, 1993] Glenn Ellison. « Learning, local interaction, and coordination ». In : *Econometrica : Journal of the Econometric Society* (1993), p. 1047-1071.

[Shoham, 1993] Yoav Shoham. « Agent-oriented programming ». In : *Artificial intelligence* 60.1 (1993), p. 51-92.

[Tarde, 1993] Gabriel de Tarde. *Les lois de l'imitation*. Éditions Kimé, 1993.

[Allan, 1998] James Allan, Ron Papka et Victor Lavrenko. « On-line new event detection and tracking ». In : *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*. 1998, p. 37-45.

[BrinPage, 1998] Sergey Brin et Lawrence Page. « The Anatomy of a Large-Scale Hypertextual Web Search Engine ». In : *COMPUTER NETWORKS AND ISDN SYSTEMS*. 1998, p. 107-117.

[CarbonellGoldstein, 1998] Jaime Carbonell et Jade Goldstein. « The use of MMR, diversity-based reranking for reordering documents and producing summaries ». In : *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*. ACM. 1998, p. 335-336.

[BaezaYatesRibeiroNeto, 1999] Ricardo Baeza-Yates, Berthier Ribeiro-Neto et al. *Modern information retrieval*. T. 463. ACM press New York, 1999.

[Kleinberg, 1999] Jon M Kleinberg. « Authoritative sources in a hyperlinked environment ». In : *Journal of the ACM (JACM)* 46.5 (1999), p. 604-632.

[Blackmore, 2000] Susan BLACKMORE. *The meme machine*. T. 25. Oxford Paperbacks, 2000.

[Hethcote, 2000] H. HETHCOTE. « The Mathematics of Infectious Diseases ». In : *SIAM Review* 42.4 (2000), p. 599-653.

[Berger, 2001] Eli BERGER. « Dynamic Monopolies of Constant Size ». In : *Journal of Combinatorial Theory, Series B* 83.2 (2001), p. 191-200.

[Goldenberg, 2001] Jacob GOLDENBERG, Barak LIBAI et Eitan MULLER. « Talk of the Network : A Complex Systems Look at the Underlying Process of Word-of-Mouth ». In : *Marketing letters* 12.3 (2001), p. 211-223.

[McdonaldRamscar, 2001] Scott MCDONALD et Michael RAMSCAR. « Testing the distributional hypothesis : The influence of context on judgements of semantic similarity ». In : *In Proceedings of the 23rd Annual Conference of the Cognitive Science Society*. 2001, p. 611-6.

[SinghalInc, 2001] Amit SINGHAL et Google INC. « Modern information retrieval : a brief overview ». In : *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering* 24 (2001), p. 2001.

[MacyWiller, 2002] Michael W MACY et Robert WILLER. « From factors to actors : Computational sociology and agent-based modeling ». In : *Annual review of sociology* 28.1 (2002), p. 143-166.

[Watts, 2002] Duncan J WATTS. « A simple model of global cascades on random networks ». In : *Proceedings*

of the National Academy of Sciences 99.9 (2002), p. 5766-5771.

[Blei, 2003] David M Blei, Andrew Y Ng et Michael I Jordan. « Latent dirichlet allocation ». In : *Journal of machine Learning research* 3.Jan (2003), p. 993-1022.

[Kempe, 2003] David Kempe, Jon Kleinberg et Éva Tardos. « Maximizing the Spread of Influence Through a Social Network ». In : *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '03. New York, NY, USA : ACM, 2003, p. 137-146.

[Bass, 2004] Frank Bass. « Comments on "A New Product Growth for Model Consumer Durables The Bass Model" ». In : *Management Science* 50.12_supplement (2004), p. 1833-1840.

[Gruhl, 2004] Daniel Gruhl et al. « Information diffusion through blogspace ». In : *Proceedings of the 13th international conference on World Wide Web*. 2004, p. 491-501.

[Robertson, 2004] Stephen E. Robertson. « Understanding inverse document frequency : on theoretical arguments for IDF ». In : *Journal of Documentation* 60 (2004), p. 503-520.

[AdarAdamic, 2005] E. Adar et L. A. Adamic. « Tracking Information Epidemics in Blogspace ». In : *The 2005 IEEE/WIC/ACM International Conference on Web Intelligence (WI'05)*. Sept. 2005, p. 207-214.

[Al Hasan, 2006] Mohammad Al Hasan et al. « Link prediction using supervised learning ». In : *SDM06 : workshop on link analysis, counter-terrorism and security*. T. 30. 2006, p. 798-805.

[Bishop, 2006] M Christopher Bishop. « Pattern Recognition and Machine Learning ». In : *Company New York* 16.4 (2006), p. 049901.

[BleiLafferty, 2006] David M Blei et John D Lafferty. « Dynamic topic models ». In : *Proceedings of the 23rd international conference on Machine learning*. 2006, p. 113-120.

[Caporaso, 2007] J Gregory Caporaso et al. « MutationFinder : a high-performance system for extracting point mutation mentions from text ». In : *Bioinformatics* 23.14 (2007), p. 1862-1865.

[DaleyVereJones, 2007] Daryl J Daley et David Vere-Jones. *An introduction to the theory of point processes : volume II : general theory and structure*. Springer Science & Business Media, 2007.

[Leskovec, 2007] Jure Leskovec, Lada A Adamic et Bernardo A Huberman. « The dynamics of viral marketing ». In : *ACM Transactions on the Web (TWEB)* 1.1 (2007), p. 5.

[LibenNowellKleinberg, 2007] David Liben-Nowell et Jon Kleinberg. « The link-prediction problem for social networks ». In : *Journal of the American society for information science and technology* 58.7 (2007), p. 1019-1031.

[Clauset, 2008] Aaron Clauset, Cristopher Moore et M. E. J. Newman. « Hierarchical Structure and the Prediction of Missing Links in Networks ». In : *Nature* 453.7191 (2008), p. 98-101.

[Kossinets, 2008] Gueorgi Kossinets, Jon Kleinberg et Duncan Watts. « The Structure of Information Pathways in a Social Communication Network ». In : *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2008, p. 435-443.

[Manning, 2008] Christopher D Manning, Prabhakar Raghavan et Hinrich Schütze. *Introduction to information retrieval*. Cambridge university press, 2008.

[SeoCroft, 2008] Jangwon Seo et W Bruce Croft. « Local text reuse detection ». In : *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*. 2008, p. 571-578.

[Tang, 2008] Jie Tang et al. « ArnetMiner : Extraction and Mining of Academic Social Networks ». In : *KDD'08*. 2008, p. 990-998.

[Yarlett, 2008] Daniel G Yarlett. *Similarity-based generalization in language*. Stanford University, 2008.

[BalVan Boheemen, 2009] Mieke Bal et Christine Van Boheemen. *Narratology : Introduction to the theory of narrative*. University of Toronto Press, 2009.

[Dogrusoz, 2009] Ugur Dogrusoz et al. « A layout algorithm for undirected compound graphs ». In : *Information Sciences* 179.7 (2009), p. 980-994.

[GuimeràSalesPardo, 2009] Roger Guimerà et Marta Sales-Pardo. « Missing and Spurious Interactions and the Reconstruction of Complex Networks ». In : *Proceedings of the National Academy of Sciences* 106.52 (2009), p. 22073-22078.

[Leskovec, 2009] Jure Leskovec, Lars Backstrom et Jon Kleinberg. « Meme-Tracking and the Dynamics of the News Cycle ». In : *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2009, p. 497-506.

[GomezRodriguez, 2010] Manuel Gomez-Rodriguez, Jure Leskovec et Andreas Krause. « Inferring networks of diffusion and influence ». In : *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2010, p. 1019-1028.

[Jackson, 2010] Matthew O Jackson. *Social and economic networks*. Princeton university press, 2010.

[Leskovec, 2010] Jure Leskovec et al. « Kronecker Graphs : An Approach to Modeling Networks ». In : *Journal of Machine Learning Research* 11 (2010), p. 985-1042.

[MontanariSaberi, 2010] Andrea Montanari et Amin Saberi. « The Spread of Innovations in Social Networks ». In :

Proceedings of the National Academy of Sciences 107.47 (2010), p. 20196-20201.

[Newman, 2010] Mark NEWMAN. *Networks : An Introduction*. OUP Oxford, 2010.

[Rogers, 2010] Everett M. ROGERS. *Diffusion of innovations*. New York : Free Press of Glencoe, 2010.

[Saito, 2010] Kazumi SAITO et al. « Selecting Information Diffusion Models over Social Networks for Behavioral Analysis ». In : *Machine Learning and Knowledge Discovery in Databases*. Lecture Notes in Computer Science 6323. Springer, 2010, p. 180-195.

[ShahZaman, 2010] Devavrat SHAH et Tauhid ZAMAN. « Detecting sources of computer viruses in networks : theory and experiment ». In : *SIGMETRICS 2010, Proceedings of the 2010 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems*. 2010, p. 203-214.

[ShahafGuestrin, 2010] Dafna SHAHAF et Carlos GUESTRIN. « Connecting the Dots Between News Articles ». In : *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2010, p. 623-632.

[Vinh, 2010] Nguyen Xuan VINH, Julien EPPS et James BAILEY. « Information theoretic measures for clusterings comparison : Variants, properties, normalization and correction for chance ». In : *The*

Journal of Machine Learning Research 11 (2010), p. 2837-2854.

[YangCounts, 2010] Jiang YANG et Scott COUNTS. « Predicting the speed, scale, and range of information diffusion in twitter ». In : *Fourth International AAAI Conference on Weblogs and Social Media*. 2010.

[BaezaYatesRibeiro, 2011] Ricardo BAEZA-YATES, Berthier de Araújo Neto RIBEIRO et al. *Modern information retrieval*. New York : ACM Press ; Harlow, England : Addison-Wesley, 2011.

[Bakhshandeh, 2011] Reza BAKHSHANDEH et al. « Degrees of separation in social networks ». In : *Fourth Annual Symposium on Combinatorial Search*. 2011.

[Fyson, 2011] Nick FYSON, Tijl DE BIE et Nello CRISTIANINI. « Reconstruction of causal networks by set covering ». In : *International Conference on Adaptive and Natural Computing Algorithms*. Springer. 2011, p. 196-205.

[GomezRodriguez, 2011] Manuel GOMEZ-RODRIGUEZ, David BALDUZZI et Bernhard SCHÖLKOPF. « Uncovering the Temporal Dynamics of Diffusion Networks ». In : *Proceedings of the 28th International Conference on Machine Learning, ICML*. 2011, p. 561-568.

[KimLeskovec, 2011] M. KIM et J. LESKOVEC. « The Network Completion Problem : Inferring Missing Nodes and Edges in Networks ». In : *Proceedings of the 2011 SIAM International Conference on Data Mining*. 2011, p. 47-58.

[Long, 2011] Rui LONG et al. « Towards effective event detection, tracking and summarization on microblog data ». In : *International Conference on Web-Age Information Management*. Springer. 2011, p. 652-663.

[LüZhou, 2011] Linyuan LÜ et Tao ZHOU. « Link prediction in complex networks : A survey ». In : *Physica A : statistical mechanics and its applications* 390.6 (2011), p. 1150-1170.

[MenonElkan, 2011] Aditya Krishna MENON et Charles ELKAN. « Link prediction via matrix factorization ». In : *Joint european conference on machine learning and knowledge discovery in databases*. Springer. 2011, p. 437-452.

[Romero, 2011] Daniel M. ROMERO, Brendan MEEDER et Jon KLEINBERG. « Differences in the Mechanics of Information Diffusion across Topics : Idioms, Political Hashtags, and Complex Contagion on Twitter ». In : *Proceedings of the 20th International Conference on World Wide Web*. ACM, 2011, p. 695-704.

[Snowsill, 2011] Tristan Mark SNOWSILL et al. « Refining causality : who copied from whom ? » In : *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2011, p. 466-474.

[Yan, 2011] Rui YAN et al. « Evolutionary Timeline Summarization : A Balanced Optimization Framework

via Iterative Substitution ». In : *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2011, p. 745-754.

[Backstrom, 2012] Lars BACKSTROM et al. « Four degrees of separation ». In : *Proceedings of the 4th Annual ACM Web Science Conference*. 2012, p. 33-42.

[Bakshy, 2012] Eytan BAKSHY et al. « The role of social networks in information diffusion ». In : *Proceedings of the 21st international conference on World Wide Web*. 2012, p. 519-528.

[Bär, 2012] Daniel BÄR, Torsten ZESCH et Iryna GUREVYCH. « Text reuse detection using a composition of text similarity measures ». In : *Proceedings of COLING*. 2012, p. 167-184.

[BergerMilkman, 2012] Jonah BERGER et Katherine L MILKMAN. « What makes online content viral ? » In : *Journal of marketing research* 49.2 (2012), p. 192-205.

[Metzler, 2012] Donald METZLER, Congxing CAI et Eduard HOVY. « Structured event retrieval over microblog archives ». In : *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*. Association for Computational Linguistics. 2012, p. 646-655.

[MyersLeskovec, 2012] Seth A. MYERS et Jure LESKOVEC. « Clash of the Contagions : Cooperation and Competition in Information Diffusion ». In : *12th IEEE In-*

ternational Conference on Data Mining, ICDM. 2012, p. 539-548.

[Pinto, 2012] Pedro C. Pinto, Patrick Thiran et Martin Vetterli. « Locating the Source of Diffusion in Large-Scale Networks ». In : *Phys. Rev. Lett.* 109 (6 août 2012), p. 068702.

[Prakash, 2012] B. Aditya Prakash, Jilles Vreeken et Christos Faloutsos. « Spotting Culprits in Epidemics : How Many and Which Ones ? » In : *12th International Conference on Data Mining.* IEEE, 2012, p. 11-20.

[RodriguezSchölkopf, 2012] Manuel Gomez Rodriguez et Bernhard Schölkopf. « Submodular Inference of Diffusion Networks from Multiple Trees ». In : *arXiv preprint arXiv :1205.1671* (2012).

[Shahaf, 2012] Dafna Shahaf, Carlos Guestrin et Eric Horvitz. « Trains of thought : Generating information maps ». In : *Proceedings of the 21st international conference on World Wide Web.* 2012, p. 899-908.

[Wang, 2012] Liaoruo Wang, Stefano Ermon et John E. Hopcroft. « Feature-Enhanced Probabilistic Models for Diffusion Network Inference ». In : *Joint European Conference on Machine Learning and Knowledge Discovery in Databases.* Springer, 2012, p. 499-514.

[Xiong, 2012] Fei Xiong et al. « An information diffusion model based on retweeting mechanism for online

social media ». In : *Physics Letters A* 376.30-31 (2012), p. 2103-2108.

[Barbieri, 2013] Nicola Barbieri, Francesco Bonchi et Giuseppe Manco. « Topic-aware social influence propagation models ». In : *Knowledge and information systems* 37.3 (2013), p. 555-584.

[BrockmannHelbing, 2013] Dirk Brockmann et Dirk Helbing. « The hidden geometry of complex, network-driven contagion phenomena ». In : *science* 342.6164 (2013), p. 1337-1342.

[Du, 2013] Nan Du et al. « Scalable influence estimation in continuous-time diffusion networks ». In : *Advances in neural information processing systems*. 2013, p. 3147-3155.

[Gomez Rodriguez, 2013] Manuel Gomez Rodriguez, Jure Leskovec et Bernhard Schölkopf. « Structure and dynamics of information pathways in online media ». In : *Proceedings of the sixth ACM international conference on Web search and data mining*. 2013, p. 23-32.

[Hajibagheri, 2013] Alireza Hajibagheri, Ali Hamzeh et Gita Sukthankar. « Modeling information diffusion and community membership using stochastic optimization ». In : *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. 2013, p. 175-182.

[Lagnier, 2013] Cédric Lagnier et al. « Predicting information diffusion in social networks using content and

user's profiles ». In : *European conference on information retrieval*. Springer. 2013, p. 74-85.

[Luo, 2013] Wuqiong Luo, Wee Peng Tay et Mei Leng. « Identifying infection sources and regions in large networks ». In : *IEEE Transactions on Signal Processing* 61.11 (2013), p. 2850-2865.

[Mikolov, 2013] Tomas Mikolov et al. « Efficient estimation of word representations in vector space ». In : *arXiv preprint arXiv :1301.3781* (2013).

[YangZha, 2013] Shuang-Hong Yang et Hongyuan Zha. « Mixture of Mutually Exciting Processes for Viral Diffusion ». In : *Proceedings of the 30th International Conference on Machine Learning, ICML*. 2013, p. 1-9.

[Alvari, 2014] Hamidreza Alvari, Alireza Hajibagheri et Gita Sukthankar. « Community detection in dynamic social networks : A game-theoretic approach ». In : *2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2014)*. IEEE. 2014, p. 101-107.

[Alvi, 2014] Faisal Alvi, Mark Stevenson et Paul D Clough. « Hashing and Merging Heuristics for Text Reuse Detection. » In : *CLEF (working notes)*. 2014, p. 939-946.

[Bourigault, 2014] Simon Bourigault et al. « Learning Social Network Embeddings for Predicting Information Diffusion ». In : *Proceedings of the 7th ACM In-*

ternational Conference on Web Search and Data Mining. ACM, 2014, p. 393-402.

[Cheng, 2014] Justin Cheng et al. « Can cascades be predicted ? » In : *Proceedings of the 23rd international conference on World wide web*. 2014, p. 925-936.

[Daneshmand, 2014] Hadi Daneshmand et al. « Estimating diffusion network structures : Recovery conditions, sample complexity & soft-thresholding algorithm ». In : *International Conference on Machine Learning*. 2014, p. 793-801.

[DiaoJiang, 2014] Qiming Diao et Jing Jiang. « Recurrent chinese restaurant process with a duration-based discount for event identification from twitter ». In : *Proceedings of the 2014 SIAM International Conference on Data Mining*. SIAM. 2014, p. 388-397.

[Fioriti, 2014] Vincenzo Fioriti, Marta Chinnici et Jesus Palomo. « Predicting the Sources of an Outbreak with a Spectral Technique ». In : *Applied Mathematical Sciences* 8.135 (2014), p. 6775-6782.

[Guille, 2014] Adrien Guille. « Diffusion de L'information Dans Les Médias Sociaux ». Thèse de doct. Université Lumière Lyon 2, 2014.

[Jiang, 2014] Chunxiao Jiang, Yan Chen et KJ Ray Liu. « Evolutionary dynamics of information diffusion over social networks ». In : *IEEE Transactions on Signal Processing* 62.17 (2014), p. 4573-4586.

[Lauf, 2014] Aurélien Lauf. « Propagation Du Buzz Sur Internet – Identification, Analyse, Modélisation et Représentation Dans Un Contexte de Veille ». Thèse de doct. 2014.

[LeMikolov, 2014] Quoc V. Le et Tomas Mikolov. « Distributed Representations of Sentences and Documents ». In : *Proceedings of the 31th International Conference on Machine Learning, ICML*. 2014, p. 1188-1196.

[Lokhov, 2014] Andrey Y Lokhov et al. « Inferring the origin of an epidemic with a dynamic message-passing algorithm ». In : *Physical Review E* 90.1 (2014), p. 012801.

[Nematzadeh, 2014] Azadeh Nematzadeh et al. « Optimal network modularity for information diffusion ». In : *Physical review letters* 113.8 (2014), p. 088701.

[Pennington, 2014] Jeffrey Pennington, Richard Socher et Christopher D. Manning. « GloVe : Global Vectors for Word Representation ». In : *Empirical Methods in Natural Language Processing (EMNLP)*. 2014, p. 1532-1543.

[TorresMoreno, 2014] J.M. Torres-Moreno. *Automatic Text Summarization*. Cognitive science and knowledge management series. Wiley, 2014.

[VitaliRosatiSinatra, 2014] Marcello Vitali-Rosati et Michael E Sinatra. *Pratiques de l'édition numérique*. Les Presses de l'Université de Montréal, 2014.

[Zafarani, 2014] Reza Zafarani, Mohammad Ali Abbasi et Huan Liu. *Social Media Mining, An Introduction.* Cambridge University Press, 2014.

[Dermouche, 2015] Mohamed Dermouche. « Modélisation Conjointe Des Thématiques et Des Opinions – Application À L'analyse Des Données Textuelles Issues Du Web ». Thèse de doct. 2015.

[Farajtabar, 2015] Mehrdad Farajtabar et al. « Back to the Past : Source Identification in Diffusion Networks from Partially Observed Cascades ». In : *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics, AISTATS.* 2015.

[Fink, 2015] Clay Fink et al. « Complex Contagions and the Diffusion of Popular Twitter Hashtags in Nigeria ». In : *Social Network Analysis and Mining* 6.1 (déc. 2015), p. 1.

[Hu, 2015] Zhiting Hu et al. « Community Level Diffusion Extraction ». In : *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data.* ACM, 2015, p. 1555-1569.

[Kurka, 2015] David Burth Kurka, Alan Godoy et Fernando J. Von Zuben. « Online Social Network Analysis : A Survey of Research Applications in Computer Science ». In : (2015). arXiv : 1504.05655.

[Masrour, 2015] Farzan Masrour et al. « Network completion with node similarity : A matrix completion approach with provable guarantees ». In : *2015*

IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). IEEE. 2015, p. 302-307.

[Sundareisan, 2015] S. Sundareisan, J. Vreeken et B. Prakash. « Hidden Hazards : Finding Missing Nodes in Large Graph Epidemics ». In : *Proceedings of the 2015 SIAM International Conference on Data Mining*. Society for Industrial and Applied Mathematics, 2015, p. 415-423.

[Tran, 2015] Giang Tran, Mohammad Alrifai et Eelco Herder. « Timeline summarization from relevant headlines ». In : *European Conference on Information Retrieval*. Springer. 2015, p. 245-256.

[Vossen, 2015] Piek Vossen, Tommaso Caselli et Yiota Kontzopoulou. « Storylines for structuring massive streams of news ». In : *Proceedings of the First Workshop on Computing News Storylines*. 2015, p. 40-49.

[Wang, 2015a] Lu Wang, Claire Cardie et Galen Marchetti. « Socially-Informed Timeline Generation for Complex Events ». In : (2015), p. 1055-1065.

[Wang, 2015b] Peng Wang et al. « Link prediction in social networks : the state-of-the-art ». In : *Science China Information Sciences* 58.1 (2015), p. 1-38.

[Wang, 2015c] Senzhang Wang et al. « Inferring Diffusion Networks with Sparse Cascades by Structure Transfer ». In : *Database Systems for Advanced Appli-*

cations. Lecture Notes in Computer Science 9049. Springer, 2015, p. 405-421.

[WuShen, 2015] Bo Wu et Haiying Shen. « Analyzing and Predicting News Popularity on Twitter ». In : *International Journal of Information Management* 35.6 (2015), p. 702-711.

[Zhao, 2015] Qingyuan Zhao et al. « SEISMIC : A Self-Exciting Point Process Model for Predicting Tweet Popularity ». In : *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2015, p. 1513-1522.

[Zhou, 2015] Deyu Zhou, Haiyang Xu et Yulan He. « An unsupervised Bayesian modelling approach for storyline detection on news articles ». In : *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. 2015, p. 1943-1948.

[Adamic, 2016] Lada A. Adamic et al. « Information Evolution in Social Networks ». In : *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*. WSDM '16. New York, NY, USA : ACM, 2016, p. 473-482.

[Babaei, 2016] Mahmoudreza Babaei et al. « On the Efficiency of the Information Networks in Social Media ». In : *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*. ACM, 2016, p. 83-92.

[Bourigault, 2016] Simon Bourigault, Sylvain Lamprier et Patrick Gallinari. « Representation Learning for Information Diffusion Through Social Networks : An Embedded Cascade Model ». In : *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*. ACM, 2016, p. 573-582.

[Cheng, 2016] Justin Cheng et al. « Do Cascades Recur ? » In : *Proceedings of the 25th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 2016, p. 671-681.

[Dawkins, 2016] Richard Dawkins. *The selfish gene*. Oxford university press, 2016.

[Goel, 2016] Sharad Goel et al. « The structural virality of online diffusion ». In : *Management Science* 62.1 (2016), p. 180-196.

[GomezRodriguez, 2016] Manuel Gomez-Rodriguez et al. « Influence estimation and maximization in continuous-time diffusion networks ». In : *ACM Transactions on Information Systems (TOIS)* 34.2 (2016), p. 1-33.

[HassanianEsfahaniKargar, 2016] R. Hassanian-Esfahani et M. j Kargar. « A Survey on Web News Retrieval and Mining ». In : *2016 Second International Conference on Web Research (ICWR)*. 2016, p. 90-101.

[Jiang, 2016] Jiaojiao Jiang et al. « Identifying propagation sources in networks : State-of-the-art and com-

parative studies ». In : *IEEE Communications Surveys & Tutorials* 19.1 (2016), p. 465-481.

[Louzada Pinto, 2016] Julio Cesar Louzada Pinto. « Information Diffusion and Opinion Dynamics in Social Networks ». Theses. Institut National des Télécommunications, 2016.

[Nies, 2016] Tom De Nies, Erik Mannens et Rik Van de Walle. « Reconstructing Human-Generated Provenance Through Similarity-Based Clustering ». In : *Provenance and Annotation of Data and Processes*. Lecture Notes in Computer Science 9672. Springer, 2016, p. 191-194.

[Renaud, 2016] Clément Renaud, Valérie Fernandez et Gilles Puel. « Les mèmes Internet ont-ils un mode de propagation spécifique ? » In : *Réseaux* 195 (mar. 2016), p. 107-130.

[Sela, 2016] Alon Sela et al. « Comparing the Diversity of Information by Word-of-Mouth vs. Web Spread ». In : *EPL (Europhysics Letters)* 114.5 (2016), p. 58003.

[Wang, 2016a] Lu Wang. « Summarization And Sentiment Analysis For Understanding Socially-Generated Content ». Thèse de doct. Cornell University, 2016.

[Wang, 2016b] William Yang Wang et al. « A low-rank approximation approach to learning joint embeddings of news stories and images for timeline summarization ». In : *Proceedings of the 2016 Conference of the North American Chapter of the Association*

for Computational Linguistics : Human Language Technologies. 2016, p. 58-68.

[Xie, 2016] Wei Xie et al. « Topicsketch : Real-time bursty topic detection from twitter ». In : *IEEE Transactions on Knowledge and Data Engineering* 28.8 (2016), p. 2216-2229.

[Zhou, 2016] Deyu Zhou et al. « Unsupervised Storyline Extraction from News Articles. » In : *IJCAI*. 2016, p. 3014-3021.

[AllcottGentzkow, 2017] Hunt Allcott et Matthew Gentzkow. « Social media and fake news in the 2016 election ». In : *Journal of economic perspectives* 31.2 (2017), p. 211-36.

[GambhirGupta, 2017] Mahak Gambhir et Vishal Gupta. « Recent automatic text summarization techniques : a survey ». In : *Artificial Intelligence Review* 47.1 (2017), p. 1-66.

[Liu, 2017] Bang Liu et al. « Growing Story Forest Online from Massive Breaking News ». In : *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. ACM, 2017, p. 777-785.

[MeleCrestani, 2017] Ida Mele et Fabio Crestani. « Event detection for heterogeneous news streams ». In : *International Conference on Applications of Natural Language to Information Systems*. Springer. 2017, p. 110-123.

[Mozafari, 2017] Niloofar Mozafari, Ali Hamzeh et Sattar Hashemi. « Modelling information diffusion based on non-dominated friends in social networks ». In : *Journal of Information Science* 43.6 (2017), p. 801-815.

[Shen, 2017] Yanning Shen, Brian Baingana et Georgios B Giannakis. « Tensor decompositions for identifying directed graph topologies and tracking dynamic networks ». In : *IEEE Transactions on Signal Processing* 65.14 (2017), p. 3675-3687.

[Zarezade, 2017] Ali Zarezade et al. « Correlated Cascades : Compete or Cooperate ». In : *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*. 2017, p. 238-244.

[Zhang, 2017] Leihan Zhang, Ke Xu et Jichang Zhao. « Sleeping beauties in meme diffusion ». In : *Scientometrics* 112.1 (2017), p. 383-402.

[Ghalebi, 2018] Elahe Ghalebi et al. « Dynamic network model from partial observations ». In : *Advances in Neural Information Processing Systems*. 2018, p. 9862-9872.

[HuyghuesDespointes, 2018] Charles Huyghues-Despointes et al. « Extraction de chaînes cohérentes en vue de reconstuire la Trajectoire de l'information ». In : *Extraction et de Gestion des Connaissances*. Paris, France, 2018.

[Li, 2018] Yuchen Li et al. « Influence maximization on social graphs : A survey ». In : *IEEE Transactions*

on Knowledge and Data Engineering 30.10 (2018), p. 1852-1872.

[TeeTaylor, 2018] J. Tee et D. P. Taylor. « Is Information in the Brain Represented in Continuous or Discrete Form ? » In : *ArXiv e-prints* (2018). arXiv : 1805.01631.

[Zhou, 2018] Deyu Zhou, Linshen Guo et Yulan He. « Neural storyline extraction model for storyline generation from news articles ». In : Association for Computational Linguistics. 2018.

[Hou, 2019] Chenglong Hou et al. « A Survey of Deep Learning Applied to Story Generation ». In : *International Conference on Smart Computing and Communication*. Springer. 2019, p. 1-10.

[HuyghuesDespointes, 2019] Charles Huyghues-Despointes et al. « Weaving Information Propagation : Modeling The Way Information Spreads In Document Collections ». In : *Canadian Conference on Artificial Intelligence*. Springer. 2019, p. 394-399.

[Mateos, 2019] Gonzalo Mateos et al. « Connecting the dots : Identifying network structure via graph signal processing ». In : *IEEE Signal Processing Magazine* 36.3 (2019), p. 16-43.